Higher-Order RNA and DNA Hubs Shape Genome Organization in the Nucleus

Thesis by

Sofia Agustina Quinodoz

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2020
(Defended December 16, 2019)

Sofia Agustina Quinodoz
ORCID: 0000-0003-1862-5204

ACKNOWLEDGEMENTS

I am extremely thankful to have had so many people who have taken the time to support and mentor me throughout my Ph.D. both inside and outside of the lab. This has been an incredible few years and I have grown immensely as a scientist, and that would not have been possible without everyone's support during my time at Caltech.

First and foremost, I would like to thank my Ph.D. advisor, Mitch Guttman, for believing in me and in SPRITE throughout the multiple years of troubleshooting, and for giving me so many incredible opportunities to grow as a scientist. It has been so much fun to take the SPRITE method from a dream during my rotation to a reality together. I am extremely grateful for his mentorship—especially providing boundless inspiration, enthusiasm, and support to boldly go after ambitious projects in the lab, and for teaching me the importance of effectively communicating scientific results to the broader scientific community outside the lab. No matter the time of day or week, Mitch has always made time to talk and help whenever I needed advice. I will always be immensely grateful for how much time Mitch invested in training me and how much he believed in me as a scientist and my project over the years. His mentorship has played an extremely large role shaping me into the scientist that I am today, and I plan to take these skills with me for the rest of my career.

The entire Guttman Lab has been a wonderful group of scientists that I am so grateful to work with and learn from. To Alex Shishkin, who initially trained me when I started in the lab and is the original "ligation master". To Noah Ollikainen, who I have been extremely lucky to work with and to learn so much from as we explored the SPRITE data together. To Prashant Bhat, for being an incredible scientist to work with every day and for his endless help, encouragement, advice, and friendship. To Joanna Jachowicz, for being a role model for me and for always being available to talk about experiments, brainstorm fun scientific ideas, and provide advice. I am immensely grateful to have worked with an incredible "SPRITE team"—Jan Marten Schmidt, Elizabeth Detmar, and Elizabeth Soehalim—they put in so many hours working in the lab alongside of me that were invaluable to get SPRITE off the ground. To Mario Blanco and Amy Chow, for always being available to talk over experiments and answer so many questions. Additional special thanks to Ali Palla, Barbara Tabak, Charlotte Lai, Peter Chovanec, Abhik Banerjee, Tony Szempruch, and Colleen McHugh—I am so lucky for all the help, support, and advice.

I would also like to thank several faculty mentors: My committee, Alexei Aravin, Kathrin Plath, Paul Sternberg, and Barbara Wold, each of whom served as wonderful mentors throughout my Ph.D. and have made time out of their busy schedules to provide valuable advice. To Dianne Newman, for the opportunity to learn how to teach as a Bi1 TA for three years and for making time to provide career advice. To Manuel Garber, for being a wonderful collaborator to work with and learn from. To Michael Elowitz, for taking me on as a young HHMI EXROP student in 2011—which led me to Caltech for my Ph.D. To my undergraduate mentors, Bonnie Bassler, Julia van Kessell, and Eva-Maria Schoetz for training me as a young scientist and inspiring me to tackle exciting questions in my Ph.D.

Outside of the lab, thanks to all my friends, especially Alicia Rogers, Katie Schretter, Scott Saunders, Riley Galton, Nadia Herrera, Naeem Husain, Larissa Charnsangavej, Taso Dimitriadis, Constanza Jackson, Harry Nunns, Matteo Ronchi, and Samuel Ho, who have been a great support system outside of the lab and made these years such fun.

Without my family, I do not know where I would be. So many thanks to my parents and brothers, for providing encouragement and advice from afar—especially understanding and supporting my move to California to pursue my goals. I am so thankful for my parents for constantly supporting my academic studies and cultivating my passion for science throughout my life, and for all the sacrifices they have made to get me to where I am. And to my brothers, Gabriel and Rafael, for always making me laugh and reminding me to relax when I am working too hard. Extra special thanks to my extended family in Argentina for their encouragement from afar.

Finally, thanks to Aaron Lin for being the most supportive person every day during my Ph.D.—constantly reminding me to never give up as I hit challenges and always believing in me. I am so grateful for the endless encouragement, inspiration, and support over these years.

ABSTRACT

Although the entire genome is present within the nucleus of every cell, distinct genes need to be accessed and expressed in different cellular conditions. Accordingly, the nucleus of each cell is a highly organized arrangement of DNA, RNA, and protein that is dynamically assembled and regulated in different cellular states. These dynamic nuclear structures are largely arranged around functionally related roles and often occur across multiple chromosomes. These include large nuclear bodies (i.e., nucleolus, nuclear speckle), smaller nuclear bodies (i.e., Cajal bodies and histone locus bodies), and gene-gene interactions (i.e., transcription compartments and loops). Yet, what molecular components are involved in establishing this dynamic organization have been largely unknown due to a lack of methods to measure the RNA and DNA components of nuclear bodies and their spatial arrangements in the nucleus. Here, we present Split-Pool Recognition of Interactions by Tag Extension (SPRITE), which enables genome-wide detection of higher-order interactions within the nucleus. In the second chapter, we introduce SPRITE and recapitulate known structures identified by proximity ligation and identify additional interactions occurring across larger distances, including two hubs of inter-chromosomal interactions that are arranged around the nucleolus and nuclear speckles. We show that a substantial fraction of the genome exhibits preferential organization relative to these nuclear bodies. Our results generate a global model whereby nuclear bodies act as inter-chromosomal hubs that shape the overall packaging of DNA in the nucleus. In the third chapter, we provide a detailed experimental protocol for performing SPRITE and an automated computational pipeline for analyzing SPRITE data. Finally, in the fourth chapter, we present a dramatically improved implementation of the SPRITE method that enables comprehensive mapping of all classes of RNA in the nucleus, from abundant RNAs encoded from DNA repeats to low abundance RNAs such as nascent pre-mRNAs and lncRNAs. We find that RNAs localize broadly across the nucleus, with individual RNAs localizing within discrete territories ranging from nuclear bodies to individual topologically associated domains. We uncover that nascent mRNAs interact in structures corresponding to nascent mRNA chromosome territories and compartments. Together, these results uncover a central and widespread role for non-coding RNA in demarcating 3D nuclear structures within the nucleus.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. S.A. Quinodoz, P. Bhat, P. Chovanec, J.W. Jachowicz, N. Ollikainen, E. Detmar, E. Soehalim, M. Guttman. SPRITE: A genome-wide method to map higher-order 3D spatial interactions in the nucleus using combinatorial split-and-pool barcoding. *Nature Protocols*—Invited Protocol. [Submitted]

   S.A.Q. conceptualized the experimental method, developed the experimental method and generated data, and wrote the manuscript, all with input and support from M.G. and all other authors.

2. S.A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J.M. Schmidt, E. Detmar, M. Lai, A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman. Higher-order inter-chromosomal hubs shape 3-dimensional genome organization in the nucleus. *Cell*. 174(3):744–757. 2018. doi: 10.1016/j.cell.2018.05.024

   S.A.Q. conceptualized and led the project, generated experimental data, performed analyses, and wrote the paper, all with input and support from M.G. and all other authors.

3. S. Quinodoz and M. Guttman. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biology. 24:651–63. 2014. doi: 10.1016/j.tcb.2014.08.009

   S.A.Q. wrote the review input and support from M.G.

For my aunt, Dra. Maria del Carmen Cruañes, who inspired me with her endless love for her family and her incredible passion for science—having received her Ph.D. in Argentina while working full time and raising twin daughters—we miss you always.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1
Number *Page*

Chapter 2
Number *Page*

Chapter 3
Number *Page*

Chapter 4
Number *Page*

Chapter 1: Introduction

LONG NON-CODING RNAS: AN EMERGING LINK BETWEEN GENE REGULATION AND NUCLEAR ORGANIZATION

S. Quinodoz and M. Guttman

A modified version of this introduction was published as “Long noncoding RNAs: an emerging link between gene regulation and nuclear organization” in Trends Cell Biology 24:651–63. 2014.
doi: 10.1016/j.tcb.2014.08.009

1.1 ABSTRACT

Mammalian genomes encode thousands of long non-coding RNAs (lncRNAs) that play important roles in diverse biological processes. As a class, lncRNAs are generally enriched in the nucleus and specifically within the chromatin-associated fraction. Consistent with their localization, many lncRNAs have been implicated in the regulation of gene expression and in shaping three-dimensional nuclear organization. Here, we discuss the evidence that many nuclear-retained lncRNAs can interact with various chromatin regulatory proteins and recruit them to specific sites on DNA to regulate gene expression. Furthermore, we discuss the role of specific lncRNAs in shaping three-dimensional nuclear organization and explore their emerging mechanisms. Based on these examples, we propose a model that explains how lncRNAs can shape nuclear organization to regulate gene expression.

2 INTRODUCTION: RNA AND THREE-DIMENSIONAL NUCLEAR ORGANIZATION

Although the entire genome is present within the nucleus of every cell, distinct genes need to be accessed and expressed in different cellular conditions. Accordingly, the nucleus of each cell is a highly organized arrangement of DNA, RNA, and protein that is dynamically assembled and regulated in different cellular states[1-3]. These dynamic nuclear structures are largely arranged around functionally related roles and often occur across multiple chromosomes[2-4]. These include large nuclear bodies (i.e., nucleolus[5,6], nuclear speckle[7], and paraspeckle[8,9]), gene-gene interactions (i.e., transcription factories[10-12] and polycomb bodies[13-16]), and promoter-enhancer interactions[17]. Yet, what molecular components are involved in establishing this dynamic organization is still largely unknown[1-3].

It has long been suspected that RNA might play a role in organizing the structure of the nucleus. Early studies of heterogeneous nuclear RNA (hnRNA) identified a large proportion of poly(A)-modified RNA that was retained in the nucleus and was of a distinct composition from messenger RNA (mRNA) and their precursors[18-20]. Many of these RNAs were subsequently localized to precise regions of the nucleus including the nuclear speckles[21] and other chromatin-associated regions[21,22]. Subsequent studies showed that global disruption of RNA transcription, but not protein translation, led to visible rearrangements of nuclear organization[23]. These studies led to the proposal that nuclear-retained RNAs might play an important structural role in the nucleus[18,21,23]. Yet, the identities and mechanisms of these RNA components of nuclear organization were unknown.

Over the last decade many thousands of functional lncRNAs have been identified[24-27]. Recent work has highlighted that many of these lncRNAs can play important roles in diverse biological processes such as embryonic development[28-30], cardiac function[31,32], immune responses[33,34], and cancer[35-39]. As a class, these lncRNAs are generally enriched in the nucleus and specifically within the chromatin-associated fraction[27,40]. Accordingly, most work on lncRNAs have focused on their role in gene regulation and, specifically, in the recruitment of chromatin regulatory proteins to genomic DNA locations[25,41,42]. In addition to this role, several recent studies have demonstrated another important role for lncRNAs in the nucleus—that is, several lncRNAs are essential for organizing distinct nuclear structures[43-52].

While lncRNAs are likely to fall into many different classes with different mechanisms[25,41,42], in this review we focus exclusively on nuclear-retained lncRNAs that are involved in the regulation of gene expression[25,42] and in shaping three-dimensional nuclear organization[4,37,44-47,53]. Here, we discuss the evidence demonstrating that several lncRNAs can interact with various chromatin regulatory proteins, recruit them to specific sites on DNA, modify chromatin, and regulate gene expression. Furthermore, we discuss the role of specific lncRNAs in shaping three-dimensional nuclear organization and the emerging mechanisms by which they perform this role. Based on these examples, we synthesize the observed data into a model that may explain how some lncRNAs can shape nuclear organization to regulate gene expression—highlighting how these two apparently distinct roles may indeed occur through a shared mechanism.

3 MECHANISMS OF LNCRNA REGULATION OF GENE EXPRESSION THROUGH CHROMATIN REGULATION

.3.1 LNCRNAS CONTROL DIVERSE GENE EXPRESSION PROGRAMS

It is increasingly clear that many lncRNAs can act to affect various gene expression programs[25,42]. Initial evidence for the role of lncRNAs in gene regulation came from studies of mammalian X-chromosome inactivation (XCI), a process that entails silencing of an entire X chromosome in females during development[54]. This process is orchestrated by the Xist lncRNA, which is transcribed exclusively from the inactive X chromosome (Xi)[55-57] and coats the entire Xi[58]. Importantly, genetic deletion of Xist prevents XCI[59], and induction of Xist is sufficient to initiate XCI on the same chromosome from which it is transcribed[60,61]. This silencing capability is dependent on a discrete region of the lncRNA, the A-repeat domain, whose deletion prevents transcriptional silencing without affecting Xist localization across the X-chromosome[62].

There are numerous additional examples of lncRNAs that participate in the regulation of various genes. A classic example is the Air lncRNA, which is responsible for regulating the Igfr2 gene to control genetic imprinting[63-65]. Another example is HOTAIR, which affects the expression of genes in the HoxD cluster[66] and many additional genes throughout the genome[35,67]. More recently, systematic studies exploring lncRNA function have shown that a large percentage of lncRNAs in the cell have an effect on various gene expression programs[29,30,68], including those involved in embryonic development[28-30], immune responses[33,34], and cancer[35-39]. Based on these gene expression studies, various regulatory strategies have been proposed for lncRNAs, including the activation[49,69] and repression[36,54,66] of genes in *cis*[49,54] and in *trans*[35,66]. Yet, whether lncRNAs directly or indirectly regulate these target genes remains unknown.

.3.2 LNCRNAS CAN RECRUIT CHROMATIN REGULATORY PROTEINS TO GENOMIC DNA TARGETS

Insights into how lncRNAs can regulate gene expression initially came from studies of Xist. Specifically, female embryos containing a deletion of a component of the Polycomb Repressive Complex 2 (PRC2), which places repressive histone modifications on chromatin[70], failed to maintain proper XCI[71]. It was subsequently shown that the PRC2 complex was recruited to the entire Xi[38] and that the timing of PRC2 recruitment tightly coincides with the induction of Xist during development[54,72,73]. Importantly, deleting a discrete region of the Xist lncRNA, the B-F-repeat domain, causes a loss of PRC2 recruitment to the Xi without impacting the localization of Xist[74]. Surprisingly, PRC2 recruitment by Xist is not sufficient to trigger transcriptional silencing during induction of XCI as deletion of the A-repeat, the silencing domain of Xist, still recruits PRC2 without leading to transcriptional silencing[54,74]. Exactly how Xist silences transcription of genes on the Xi remains an open question[75]. While many questions about the role of Xist in PRC2 recruitment remain unanswered[75], including whether Xist physically interacts with the PRC2 complex[76] or indirectly recruits PRC2[74], it is clear that Xist is required to recruit the PRC2 complex across the X-chromosome[73,74,77].

This chromatin protein recruitment model may be more general beyond Xist and has been proposed for several other lncRNAs. For example, HOTAIR is thought to physically interact with PRC2, and loss of function of HOTAIR leads to a reduction of the PRC2-dependent H3K27me3 repressive modifications across the HoxD gene cluster[35,66], suggesting that HOTAIR recruits PRC2 to these genes and may be involved in silencing their expression. Another example is HOTTIP, which is thought to physically interact with the WDR5 protein and whose loss of function leads to a reduction of its associated H3K4me3 active histone modifications on chromatin across the HoxA gene cluster[49], suggesting that HOTTIP recruits WDR5 to these genes and may be involved in activating their expression. More generally, a large percentage of lncRNAs are thought to physically interact with various chromatin regulatory proteins including PRC2[30,40,78,79], WDR5[49,80], and other readers[30,37,79,81], writers[30,37,65,82], and erasers[30,83] of chromatin modifications **(Figure 1B).** These examples highlight how lncRNAs may both activate and repress gene expression through a common chromatin-centric recruitment mechanism **(Figure 1A).**

Recently, it has been suggested that the PRC2 complex may interact with all RNAs in the cell—including lncRNAs and mRNAs[84,85]. There is considerable debate about how many of the identified interactions between lncRNAs and chromatin proteins are specific[84-86] and whether these physical interactions occur through direct RNA-protein or indirect protein–protein contacts[75,87] (**see Box**). Yet, it is increasingly clear that at least some of these lncRNA-chromatin interactions are important for lncRNA- and chromatin-mediated gene regulation. For example, mutating the RNA binding domain of WDR5 eliminates its chromatin modification and gene regulatory activities at its target sites without impacting its catalytic activity[80]. More generally, RNAi-mediated loss of function of several lncRNAs impact the same genes as those impacted by loss of function of their associated chromatin regulatory proteins[30,40].

Together, these results suggest that many lncRNAs may recruit chromatin regulatory complexes to specific targets on genomic DNA to control gene expression **(Figure 1A)**.

.3.3 LNCRNAS MAY SCAFFOLD MULTIPLE CHROMATIN PROTEINS TO COORDINATE DISCRETE FUNCTIONS

Individual lncRNAs may interact with multiple chromatin proteins simultaneously to coordinate multiple functional roles that are required to properly regulate gene expression (**Figure 1B**). For example, HOTAIR is thought to interact with both the PRC2 histone methyltransferase and the LSD1 histone demethylase complex[83]. This interaction may be important for coordinating the removal of activating marks (LSD1) and the addition of repressive marks (PRC2) on chromatin. More generally, >30 of the lncRNAs expressed in embryonic stem cells (ESCs) are thought to interact simultaneously with multiple chromatin regulatory complexes that can read, write, and erase functionally related chromatin marks[30]. As an example, some of these ESC lncRNAs can interact with the JARID1C histone demethylase complex, the PRC2 histone methyltransferase complex, and the ESET histone methyltransferase complex[30]. This interaction may be important for coordinating the removal of activating marks (JARID1C) and addition of different repressive marks on chromatin (PRC2, ESET) (**Figure 1C**). Importantly, many of these chromatin proteins have been shown to co-localize at specific sets of genes in ESCs even though these proteins are not thought to directly interact with each other[88-90].

Furthermore, it is now clear that the Xist lncRNA coordinates at least three discrete functions to carry out its role in XCI. These functions are mediated by distinct genetic domains of the lncRNA that are required for silencing transcription (A-repeat)[62], recruitment of PRC2 (B-F-repeat)[74], and localization to chromatin (C-repeat)[91,92]—all of which are required for proper XCI. Despite these clear genetic roles, the exact molecular mechanisms by which Xist coordinates these functions is still largely unclear because the exact proteins that interact with Xist are still largely unknown.

Together, these results suggest that some lncRNAs may create unique assemblies of chromatin regulatory complexes and other protein complexes that do not normally form protein–protein interactions **(Figure 1B)**. By acting as a scaffold for regulatory proteins, lncRNAs may coordinate the regulation of gene expression by recruiting a set of proteins that are required in combination for the shared regulation of a specific set of target genes **(Figure 1C)**.

4 MECHANISMS OF LNCRNA RECRUITMENT TO GENOMIC DNA

.4.1 LNCRNAS CAN RECOGNIZE SPECIFIC GENOMIC DNA SITES THROUGH DIVERSE MECHANISMS

There are three general mechanisms that have been proposed for how lncRNAs that recruit protein complexes to genomic DNA can recognize specific target sites (**Figure 1A**). (i) RNA polymerase can tether a lncRNA to its site of transcription and, from this location, a lncRNA can act on its neighboring genes. This mechanism may explain the localization of the Neat1 lncRNA, which requires transcription to act even when large amounts of non-nascent mature RNA is present[44]. (ii) A lncRNA can interact with DNA through direct nucleic-acid hybridization. This can include traditional base-pairing interactions, which explains the specificity of the telomerase RNA component for telomeric DNA repeats[93]. Additionally, lncRNAs can interact with DNA through triplex-mediated interactions, which may explain the localization of specific ncRNAs to ribosomal DNA promoters[94]. (iii) A lncRNA can physically interact with DNA-binding proteins. A clear example of this mechanism is the localization of the Drosophila roX lncRNAs, which are dependent on their interaction with the CLAMP DNA-binding protein to recognize specific DNA-binding sites[67,95-97]. Furthermore, this mechanism may explain the localization of Xist and Firre; both lncRNAs are thought to interact with the hnRNPU/SAF-A DNA-binding protein, which is required for their localization to DNA[47,98].

However, these mechanisms—polymerase tethering, hybridization, and DNA-binding protein mediated recruitment—alone may not be sufficient to explain how a lncRNA localizes to specific sites. For example, the roX lncRNAs localize through their interaction with CLAMP, but they do not interact at all sites throughout the genome where CLAMP is localized[95,99]. Similarly, both Xist and Firre interact with hnRNPU/SAF-A[47,98], yet each localize to very different genomic DNA sites. This argues that specificity may not depend on a single factor, but may involve multiple independent factors, including those described above, that together provide localization specificity. Exactly how most lncRNAs recognize and localize to genomic DNA remains largely unknown.

.4.2 LNCRNAS CAN EXPLOIT THE THREE-DIMENSIONAL CONFORMATION OF THE NUCLEUS TO SEARCH FOR TARGETS

Recent results are pointing to a potentially general mechanism by which lncRNAs search for regulatory targets—that is, lncRNAs can exploit the three-dimensional confirmation of the nucleus to search for target sites. As an example, Xist utilizes three-dimensional nuclear organization to locate DNA target sites by first localizing to genomic sites that are in close spatial proximity to its own transcription locus[46,100]. Simply moving Xist to a different genomic location leads to its relocalization to new genomic target sites that are defined by their close spatial proximity to the new Xist integration site[46]. Other lncRNAs have also been shown to use spatial proximity to identify target sites[39,95,101]; for example, the HOTTIP lncRNA localizes across the HoxA cluster, which is in close spatial proximity to its own transcription locus[49].

This interplay between proximity-guided search and lncRNA localization is not restricted to interactions that occur on the same chromosome, but can also occur across chromosomes because regions that are present on different chromosomes can be in close spatial proximity in the nucleus[102] (**Figure 2B**). As an example, the CISTR-ACT lncRNA localizes to sites that are present on the same chromosome as well as sites on different chromosomes that are in close spatial proximity to its transcription locus[48]. This proximity-guided model may also explain the localization of the HOTAIR lncRNA, the first example of a *trans*-regulatory lncRNA. HOTAIR is transcribed from the HoxC locus and regulates the expression of the genes in the HoxD locus, which is present on a different chromosome[66]. Indeed, the Hox gene loci, despite being present on different chromosomes, often interact with each other in close spatial proximity within the nucleus[13,103,104]. Such a proximity-guided search model may explain the apparent observations of both *cis* and *trans*-mediated regulatory mechanisms of various lncRNAs and suggests that these apparently divergent mechanisms may share a common principle of proximity (**Figure 2**).

This proximity-guided search model exploits a feature that is unique for RNA, relative to proteins, which is its ability to function immediately upon transcription. In this model, the local concentration of a lncRNA depends primarily on its spatial distance from its transcription locus, such that sites that are close will have high concentration and sites that are far will have low concentration. Yet, proximity alone is not sufficient to explain interaction, because mRNAs are also present at high concentration, but do not act, in spatial proximity to their transcription locus. Similarly, the Firre lncRNA interacts with specific DNA sites that are in spatial proximity to the Firre locus, but does not interact with all sites in spatial proximity[47]. Instead, other mechanisms, such as tethering, hybridization, or DNA-binding interactions, are likely to be required for proper localization of a lncRNA to specific sites. Indeed, the roX lncRNAs interact with specific DNA sites, defined by the presence of CLAMP DNA elements, only when these sites are present in close spatial proximity[95,101]. These two components—proximity and sequence specificity—may explain the localization of many lncRNAs. Specifically, a lncRNA will have a high probability of interacting with a target site within a region of high concentration, but it will have a low probability of interacting with a target site within a region of low concentration—even if it has a high affinity for that site **(Figure 2C)**. Importantly, such a strategy would explain how lncRNAs, which are generally of lower abundance, could reliably identify their target genes by searching in close spatial proximity to their transcription locus rather than searching across the entire nucleus.

5 LNCRNAS ARE ESSENTIAL FOR THE ESTABLISHMENT AND MAINTENANCE OF NUCLEAR DOMAINS

Recent studies have highlighted another link between lncRNAs and nuclear organization—that is, several lncRNAs have been shown to play a critical role in bringing together DNA, RNA, and proteins to actively shape three-dimensional nuclear organization. We discuss examples of lncRNAs that establish nuclear domains across various levels of organization from nuclear bodies to enhancer-promoter interactions below (**Figure 3**).

.5.1 NUCLEAR BODIES: NEAT1 ESTABLISHES THE PARASPECKLE

A clear example of a lncRNA that is essential for organizing nuclear structure is Neat1[45,105], which localizes within the paraspeckle nuclear domain[45,106,107]. The paraspeckle consists of various RNAs and proteins that are spatially co-localized and is thought to be the site of nuclear retention of adenosine-to-inosine edited mRNAs[8,9]. Importantly, loss of function of the Neat1 lncRNA leads to loss of the paraspeckle domain[45,106]. Conversely, induction of the Neat1 lncRNA is sufficient to establish the paraspeckle domain[44]. Furthermore, recruitment of Neat1 to a transgenic site is sufficient to create paraspeckles at that location[44,108]. Indeed, synthetically tethering Neat1 to a genomic DNA region is sufficient to form paraspeckles, but tethering the paraspeckle-associated proteins, such as PSP1 to DNA is not sufficient to assemble paraspeckles[108]. Neat1 transcription is required for establishing and maintaining paraspeckles by recruiting paraspeckle-associated proteins to the *Neat1* genomic locus[44]. Accordingly, disruption of Neat1 transcription, even without a reduction in overall Neat1 levels, leads to the loss of paraspeckles[44].

Together, these studies demonstrate that Neat1 plays an architectural role in the establishment and maintenance of the paraspeckle nuclear domains by seeding at its transcription locus and recruiting associated proteins to create an RNA-protein nuclear compartment.

.5.2 INTRA-CHROMOSOMAL REGULATORY DOMAINS: XIST COMPACTS THE X CHROMOSOME.

One of the most well-studied examples of a lncRNA that modulates chromosomal organization is Xist. Xist is responsible for restructuring the three-dimensional genome context of the inactive X chromosome, leading to its compaction and relocation to the periphery of the nucleus[54,109]. Indeed, simply integrating the Xist lncRNA into transgenic locations, including on autosomes, is sufficient to silence, compact, and reposition the chromosome on which Xist is integrated[61,110]. Xist spreads from its transcription locus to initial sites that are in close spatial proximity[46]. From these sites, Xist then spreads across the entire X chromosome. This spreading process is known to involve significant changes to chromosome architecture across the X chromosome[54,109]. These structural changes depend on the A-repeat domain of Xist, the same domain required for silencing transcription, because deletion of the A-repeat leads to the exclusion of actively transcribed regions from the silenced X-chromosome territory[46,53].

Together, these studies demonstrate that Xist is essential for restructuring genomic DNA regions to establish an RNA-mediated silenced nuclear compartment by spreading across the X-chromosome and repositioning genes into the silenced Xist compartment[46].

.5.3 INTER-CHROMOSOMAL REGULATORY DOMAINS: FIRRE FORMS A TRANS-CHROMSOMAL COMPARTMENT.

Another example is the lncRNA Firre, a recently identified lncRNA that is required for adipogenesis[68] and is localized in a single nuclear domain[47,68]. This nuclear domain includes its own transcription locus on the X-chromosome as well as at least 5 genes that are located on different chromosomes including chromosomes 2, 9, 15, and 17[47]. Importantly, deletion of the *Firre* locus results in reduced co-localization of the *trans*-chromosomal contacts within this nuclear domain[47]. These interacting genes were previously implicated in energy metabolism, consistent with the observed phenotypic role of Firre in regulating adipogenesis[68]. Random integration of Firre into different chromosomal regions leads to the emergence of new nuclear foci, suggesting that Firre may be sufficient to create a nuclear compartment at its integration sites. Taken together, these results suggest that Firre is required to maintain and may even be required to establish the formation of a trans-chromosomal nuclear compartment containing target genes of shared function.

.5.4 ENHANCER-PROMOTER INTERACTIONS: ERNAS CAN PROMOTE CHROMOSOMAL LOOPING.

Another example of lncRNA-mediated formation of nuclear structures are lncRNAs that are transcribed from active enhancer regions (eRNAs)[39,50,111,112]. Recent studies have shown that several eRNAs can play a role in mediating chromosomal interactions between enhancer regions and their associated promoters[37,112,113]. For example, estrogen-induced[113] and androgen-induced eRNAs[37] have been shown to maintain DNA looping between enhancer and promoter regions and, through this interaction, promote gene activation of estrogen-responsive genes and androgen-receptor-activated genes, respectively. Together, these results suggest that some eRNAs are required to maintain the three-dimensional chromosomal looping between an enhancer and its associated promoter.

While much is still unknown about how these eRNAs work, initial insights are emerging from two specific eRNAs that are highly expressed in prostate cancer[37]. The PRNCR1 eRNA binds to the enhancer regions of androgen-receptor regulated genes and is thought to recruit the DOT1L histone methyltransferase to the enhancer. This chromatin protein recruitment in turn recruits a second lncRNA, PCGEM1, to the same region. The PCGEM1 lncRNA is thought to interact with Pygo2, an H3K4me3 reader, which can recognize the methylation groups on active promoter regions[37]. Through the recruitment of these proteins, these eRNAs appear to facilitate looping between the enhancer and promoter regions, leading to the subsequent activation of the target gene[37]. These results suggest that eRNAs can recruit chromatin regulatory proteins to create high affinity interactions between different regions of DNA and, through this, act to reposition enhancer and promoter regions into close spatial proximity.

Collectively, these results and others[43] demonstrate that several lncRNAs play an important role in establishing and maintaining higher-order nuclear structures across various levels of nuclear organization from nuclear bodies to enhancer-promoter interactions.

6 A PROPOSED MODEL FOR LNCRNA-MEDIATED ORGANIZATION OF NUCLEAR STRUCTURE

While there is some evidence that lncRNAs can recruit chromatin regulators, modify chromatin structure, regulate gene expression, search in spatial proximity, and reposition genes into a nuclear domain, how these mechanisms work together to create a dynamic nuclear compartment remains unclear. Important insights can be derived from studies of nuclear body formation, which depends on many molecules—including RNA, DNA, and protein—coming together into a single nuclear region[4,114,115]. This process requires the localization of an initial nucleating factor, which seeds organization and recruits other factors to this location[4,115]. For example, simply tethering individual RNAs or proteins that are present in the Cajal bodies to a random location in the genome is sufficient to seed the formation of a new Cajal body at that site[108,116]. In the context of well-studied nuclear bodies, such as the Cajal body and the nucleolus, the proteins involved have domains that allow them to self-interact, thereby creating preferential interactions between molecules of the same identity[114,115,117,118]. This self-organization creates a high local concentration of a defined set of molecules within a spatially confined region around the initial nucleating factor[115].

This process may also explain the assembly of other functional nuclear structures; DNA containing common chromatin modification patterns[16] or DNA that is bound by shared proteins, such as PRC2[13-15] or various transcription factors[10-12], can cluster together in three-dimensional proximity. While the exact mechanism that leads to the formation of these particular long-range interactions is still largely unclear[119], it appears that a similar self-organization property may be involved because molecules of shared identity preferentially interact in three-dimensional proximity[120,121].

Based on the studies discussed above, we propose a model for how lncRNAs may organize nuclear architecture (**Figure 4**). This model is an extension of those originally proposed for Neat1[44] and Xist[46]. In this model, lncRNAs can **seed** organization by creating domains of high local lncRNA concentration near their site of transcription. This would allow the lncRNAs to **scaffold** various protein complexes and thereby **nucleate** lncRNA-protein complex assembly, increasing the effective concentration of proteins within this domain[4,44]. Then, lncRNAs can interact with high affinity target sites to achieve specificity and **recruit** lncRNA-protein complexes to specific target sites. At these targets, lncRNA-protein complexes may **modify** chromatin state and, through this, may act to **reposition** DNA sites into new nuclear domains of shared chromatin modification or protein occupancy[46]. Importantly, whether chromatin modifications or other mechanisms, such as self-organization based on the recruitment of shared protein complexes, are what drive repositioning remains to be tested. This proposed model may not be restricted to the formation of DNA compartments, but may also explain the spatial assembly of RNA and protein domains in the nucleus through a similar lncRNA-centric mechanism[44,108,118].

This process may involve **iterative** steps by which the lncRNA, while actively transcribed, can continue to seed, nucleate, modify, and reposition genes into an expanding nuclear domain[46]. For example, Xist spreads to new sites on the X chromosome by interacting in spatial proximity with the genes that have not yet been silenced and then repositioning these genes into the growing silenced nuclear compartment[46]. Once established, the lncRNA may **maintain** this domain through continued transcription from a location in close spatial proximity to the newly formed compartment. For example, Neat1 is required to maintain paraspeckles through an ongoing process of transcription[44].

7 POSSIBLE IMPLICATIONS OF LNCRNA-MEDIATED NUCLEAR ORGANIZATION IN GENE REGULATION

It is increasingly clear that there are functional nuclear domains that contain shared chromatin modification patterns[16] or protein occupancy[2,10,12,14]. However, not all DNA that is modified or bound by a specific protein in the nucleus is spatially localized within a single nuclear domain[13,16]. We hypothesize that different lncRNAs may establish these specific nuclear domains by scaffolding and recruiting distinct combinations of proteins. (**Figure 5A**). For example, the nucleus contains multiple discrete functional domains that are enriched for polycomb protein occupancy (polycomb bodies)[14,15]; one such domain is the inactive X-chromosome[77], which is established by a specific lncRNA and is spatially distinct from other polycomb-enriched domains in the nucleus.

While nuclear organization is known to be highly dynamic between cell states, how this organization is dynamically established during various processes, such as cellular differentiation, remains unclear[11,103,122,123]. We hypothesize that some lncRNAs might act as “organizational centers” to establish cell-type specific nuclear domains that organize genes of similar function in close three-dimensional proximity. Such a role is consistent with the observation that lncRNAs exhibit extraordinary cell-type specificity[26,27,124], in contrast to proteins, which are often reused in multiple cellular contexts[125]. In this model, nuclear compartments can be dynamically organized simply through the activation or repression of a single lncRNA gene **(Figure 5B).**

This hypothesized role of lncRNAs as organizational centers might represent an ideal strategy for how nuclear-localized lncRNAs could act to regulate gene expression. Because lncRNAs are generally expressed at low abundance, the probability of coordinately finding multiple target genes that are distributed throughout the nucleus would be low, potentially leading to heterogeneous expression of these genes. There are two theoretical solutions: increase lncRNA abundance or cluster target genes in spatial proximity. While both approaches solve the challenge of finding distributed genes, increasing the levels of a lncRNA may not be an optimal solution because this may lead to sub-saturation of a lncRNA scaffold with its associated regulatory proteins **(Figure 1B)**. Therefore, lncRNA regulation of multiple distributed genes requires a tradeoff between the optimality of finding all genes (high lncRNA expression) with the optimality of interacting with all required regulatory proteins (low lncRNA expression). Spatial clustering would provide an ideal solution because it would enable a lncRNA to easily find all of its targets based on spatial proximity, where the lncRNA is in high local concentration, while ensuring saturation of the lncRNA regulatory complexes to coordinately regulate all of its target genes.

8 BOX 1: EXPERIMENTAL METHODS TO DEFINE LNCRNA-PROTEIN INTERACTIONS

There are several common methods for purifying lncRNA-protein complexes including protein-based and RNA-based purification methods. For a more complete discussion of these methods and their strengths and limitations, see McHugh et al. 2014[87].

Briefly, most lncRNA-chromatin interactions[36,40,78], including the Xist-PRC2 interaction[76,86], have been identified using ‘native purification’ methods, which purify RNA-protein complexes under physiological conditions. The advantage of these methods is that they preserve the native complexes present in the cell. Yet, these methods also have several limitations including the potential for the identification of RNA-protein interactions that form in solution, which do not reflect interactions occurring in the cell[126,127]. Because of these issues, there has been some debate about the biological significance of interactions detected by these methods[75,85,86], including the Xist-PRC2 interaction, with some arguing that they are non-specific[75].

One way to distinguish *in vivo* interactions from interactions that form subsequently in solution is by crosslinking RNA-protein complexes in the cell and purifying the complex under denaturing conditions[127]. Methods such as CLIP utilize UV crosslinking, which crosslinks directly interacting RNA and protein molecules, to purify complexes using high-stringency wash conditions followed by separation on a denaturing SDS-PAGE gel[128,129]. A limitation of this method is that UV will not capture interactions that occur through a complex of multiple proteins[130]. This has restricted its adoption for mapping many chromatin regulatory proteins because the precise protein within most chromatin regulatory complexes that might directly interact with a lncRNA is unknown.

Other crosslinking methods such as formaldehyde, which crosslinks nucleic acid–protein as well as protein–protein interactions, can eliminate the need to know the exact interacting protein while enabling purification in high-stringency conditions[30,131]. Indeed, several studies have used this approach to map numerous chromatin regulatory proteins, including PRC2 and WDR5, and have identified a more specific set of interactions than previously identified by native purifications[30,80]. Yet, adapting this formaldehyde approach to a denaturing strategy is challenging since a denaturing SDS-PAGE gel will no longer resolve the purified complex.

Furthermore, because formaldehyde crosslinks across a larger physical distances than UV, many of the interactions identified by this method might not reflect physical interactions between a lncRNA and chromatin complex[75]. For example, this approach will also identify chromatin proteins and lncRNAs that are in close proximity within a DNA locus; such proximity will likely occur for nascent transcripts and the many activating chromatin complexes bound near their transcription locus.

In the absence of the ability to define a lncRNA-protein interaction using direct crosslinking and denaturing conditions, it is unclear how to confidently define *in vivo* physical interactions using biochemical methods. In such cases, complimentary genetic methods are essential to demonstrate the functional importance of an identified lncRNA-protein interaction.

9 REFERENCES

1. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-82 (2013).

2. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nat Struct Mol Biol* **20**, 290-9 (2013).

3. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800 (2007).

4. Mao, Y.S., Zhang, B. & Spector, D.L. Biogenesis and function of nuclear bodies. *Trends Genet* **27**, 295-306 (2011).

5. McStay, B. & Grummt, I. The epigenetics of rRNA genes: from molecular to chromosome biology. *Annu Rev Cell Dev Biol* **24**, 131-57 (2008).

6. Melese, T. & Xue, Z. The nucleolus: an organelle formed by the act of building a ribosome. *Curr Opin Cell Biol* **7**, 319-24 (1995).

7. Lamond, A.I. & Spector, D.L. Nuclear speckles: a model for nuclear organelles. *Nat Rev Mol Cell Biol* **4**, 605-12 (2003).

8. Fox, A.H. & Lamond, A.I. Paraspeckles. *Cold Spring Harb Perspect Biol* **2**, a000687 (2010).

9. Fox, A.H., et al., et al. Paraspeckles: a novel nuclear domain. *Curr Biol* **12**, 13-25 (2002).

10. Schoenfelder, S., et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53-61 (2010).

11. Wei, Z., et al. Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell Stem Cell* **13**, 36-47 (2013).

12. Osborne, C.S., et al. Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol* **5**, e192 (2007).

13. Bantignies, F., et al. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* **144**, 214-26 (2011).

14. Cheutin, T. & Cavalli, G. Polycomb silencing: from linear chromatin domains to 3D chromosome folding. *Curr Opin Genet Dev* **25C**, 30-37 (2014).

15. Delest, A., Sexton, T. & Cavalli, G. Polycomb: a paradigm for genome organization from one to three dimensions. *Curr Opin Cell Biol* **24**, 405-14 (2012).

16. Sexton, T., et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-72 (2012).

17. Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Curr Opin Cell Biol* **25**, 387-94 (2013).

18. Herman, R.C., Williams, J.G. & Penman, S. Message and non-message sequences adjacent to poly(A) in steady state heterogeneous nuclear RNA of HeLa cells. *Cell* **7**, 429-37 (1976).

19. Perry, R.P., Kelley, D.E. & LaTorre, J. Synthesis and turnover of nuclear and cytoplasmic polyadenylic acid in mouse L cells. *J Mol Biol* **82**, 315-31 (1974).

20. Brawerman, G. & Diez, J. Metabolism of the polyadenylate sequence of nuclear RNA and messenger RNA in mammalian cells. *Cell* **5**, 271-80 (1975).

21. Huang, S., Deerinck, T.J., Ellisman, M.H. & Spector, D.L. In vivo analysis of the stability and transport of nuclear poly(A)+ RNA. *J Cell Biol* **126**, 877-99 (1994).

22. He, D.C., Nickerson, J.A. & Penman, S. Core filaments of the nuclear matrix. *J Cell Biol* **110**, 569-80 (1990).

23. Nickerson, J.A., Krochmalnic, G., Wan, K.M. & Penman, S. Chromatin architecture and nuclear RNA. *Proc Natl Acad Sci U S A* **86**, 177-81 (1989).

24. Guttman, M., et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-7 (2009).

25. Guttman, M. & Rinn, J.L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339-46 (2012).

26. Cabili, M.N., et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).

27. Derrien, T., et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).

28. Sauvageau, M., et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**, e01749 (2013).

29. Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. & Bartel, D.P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-50 (2011).

30. Guttman, M., et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).

31. Klattenhoff, C.A., et al. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570-83 (2013).

32. Han, P., et al. A long noncoding RNA protects the heart from pathological hypertrophy. *Nature* (2014).

33. Carpenter, S., et al. A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789-92 (2013).

34. Wang, P., et al. The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science* **344**, 310-3 (2014).

35. Gupta, R.A., et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-6 (2010).

36. Huarte, M., et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-19 (2010).

37. Yang, L., et al. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* **500**, 598-602 (2013).

38. Gutschner, T., et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* **73**, 1180-9 (2013).

39. Trimarchi, T., et al. Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia. *Cell* **158**, 593-606 (2014).

40. Khalil, A.M., et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-72 (2009).

41. Wang, K.C. & Chang, H.Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**, 904-14 (2011).

42. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**, 145-66 (2012).

43. Bergmann, J.H. & Spector, D.L. Long non-coding RNAs: modulators of nuclear structure and function. *Curr Opin Cell Biol* **26**, 10-8 (2014).

44. 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).

45. Clemson, C.M., et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**, 717-26 (2009).

46. Engreitz, J.M., et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013).

47. Hacisuleyman, E., et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**, 198-206 (2014).

48. Maass, P.G., et al. A misplaced lncRNA causes brachydactyly in humans. *J Clin Invest* **122**, 3990-4002 (2012).

49. Wang, K.C., et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-4 (2011).

50. Lam, M.T., Li, W., Rosenfeld, M.G. & Glass, C.K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39**, 170-82 (2014).

51. Shichino, Y., Yamashita, A. & Yamamoto, M. Meiotic long non-coding meiRNA accumulates as a dot at its genetic locus facilitated by Mmi1 and plays as a decoy to lure Mmi1. *Open Biol* **4**, 140022 (2014).

52. Shimada, T., Yamashita, A. & Yamamoto, M. The fission yeast meiotic regulator Mei2p forms a dot structure in the horse-tail nucleus in association with the sme2 locus on chromosome II. *Mol Biol Cell* **14**, 2461-9 (2003).

53. 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 & development* **20**, 2223-37 (2006).

54. Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annual review of genetics* **36**, 233-78 (2002).

55. Brown, C.J., et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38-44 (1991).

56. Brockdorff, N., et al. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329-31 (1991).

57. Brockdorff, N., et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515-26 (1992).

58. Clemson, C.M., McNeil, J.A., Willard, H.F. & Lawrence, J.B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *The Journal of cell biology* **132**, 259-75 (1996).

59. Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131-7 (1996).

60. Rasmussen, T.P., Wutz, A.P., Pehrson, J.R. & Jaenisch, R.R. Expression of Xist RNA is sufficient to initiate macrochromatin body formation. *Chromosoma* **110**, 411-20 (2001).

61. Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Molecular cell* **5**, 695-705 (2000).

62. Wutz, A., Rasmussen, T.P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature genetics* **30**, 167-74 (2002).

63. Kulinski, T.M., Barlow, D.P. & Hudson, Q.J. Imprinted silencing is extended over broad chromosomal domains in mouse extra-embryonic lineages. *Curr Opin Cell Biol* **25**, 297-304 (2013).

64. Yamasaki, Y., et al. Neuron-specific relaxation of Igf2r imprinting is associated with neuron-specific histone modifications and lack of its antisense transcript Air. *Hum Mol Genet* **14**, 2511-20 (2005).

65. Nagano, T., et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-20 (2008).

66. Rinn, J.L., et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-23 (2007).

67. Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol Cell* (2011).

68. Sun, L., et al. Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110**, 3387-92 (2013).

69. Orom, U.A., et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

70. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-9 (2011).

71. Wang, J., et al. Imprinted X inactivation maintained by a mouse Polycomb group gene. *Nat Genet* **28**, 371-5 (2001).

72. Mak, W., et al. Mitotically stable association of polycomb group proteins eed and enx1 with the inactive x chromosome in trophoblast stem cells. *Current biology : CB* **12**, 1016-20 (2002).

73. Silva, J., et al. Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Developmental cell* **4**, 481-95 (2003).

74. da Rocha, S.T., et al. Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol Cell* **53**, 301-16 (2014).

75. Brockdorff, N. Noncoding RNA and Polycomb recruitment. *RNA* **19**, 429-42 (2013).

76. 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* **322**, 750-6 (2008).

77. Plath, K., et al. Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**, 131-5 (2003).

78. Zhao, J., et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939-53 (2010).

79. Kaneko, S., et al. Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Mol Cell* **53**, 290-300 (2014).

80. Yang, Y.W., et al. Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *Elife* **3**, e02046 (2014).

81. Yap, K.L., et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell* **38**, 662-74 (2010).

82. Yang, L., et al. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* **147**, 773-88 (2011).

83. Tsai, M.C., et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-93 (2010).

84. Kaneko, S., Son, J., Shen, S.S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20**, 1258-64 (2013).

85. Davidovich, C., Zheng, L., Goodrich, K.J. & Cech, T.R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20**, 1250-7 (2013).

86. Cifuentes-Rojas, C., Hernandez, A.J., Sarma, K. & Lee, J.T. Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Mol Cell* **55**, 171-85 (2014).

87. McHugh, C.A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol* **15**, 203 (2014).

88. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-9 (2009).

89. Tahiliani, M., et al. The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. *Nature* **447**, 601-5 (2007).

90. Cloos, P.A., Christensen, J., Agger, K. & Helin, K. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes Dev* **22**, 1115-40 (2008).

91. Beletskii, A., Hong, Y.K., Pehrson, J., Egholm, M. & Strauss, W.M. PNA interference mapping demonstrates functional domains in the noncoding RNA Xist. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 9215-20 (2001).

92. Sarma, K., Levasseur, P., Aristarkhov, A. & Lee, J.T. Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. *Proc Natl Acad Sci U S A* **107**, 22196-201 (2010).

93. Harrington, L.A. & Greider, C.W. Telomerase primer specificity and chromosome healing. *Nature* **353**, 451-4 (1991).

94. Schmitz, K.M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**, 2264-9 (2010).

95. Soruco, M.M., et al. The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes Dev* **27**, 1551-6 (2013).

96. Simon, M.D., et al. The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A* **108**, 20497-502 (2011).

97. Wang, C.I., et al. Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. *Nat Struct Mol Biol* **20**, 202-9 (2013).

98. Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K. & Nakagawa, S. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell* **19**, 469-76 (2010).

99. McElroy, K.A., Kang, H. & Kuroda, M.I. Are we there yet? Initial targeting of the Male-Specific Lethal and Polycomb group chromatin complexes in Drosophila. *Open Biol* **4**, 140006 (2014).

100. Simon, M.D., et al. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504**, 465-9 (2013).

101. Quinn, J.J., et al. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat Biotechnol* (2014).

102. Williams, A., Spilianakis, C.G. & Flavell, R.A. Interchromosomal association and gene regulation in trans. *Trends Genet* **26**, 188-97 (2010).

103. Denholtz, M., et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**, 602-16 (2013).

104. Noordermeer, D., et al. Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *Elife* **3**, e02557 (2014).

105. Hutchinson, J.N., et al. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).

106. Sunwoo, H., et al. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**, 347-59 (2009).

107. Chen, L.L. & Carmichael, G.G. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol Cell* **35**, 467-78 (2009).

108. Shevtsov, S.P. & Dundr, M. Nucleation of nuclear bodies by RNA. *Nat Cell Biol* **13**, 167-73 (2011).

109. Splinter, E., et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development* **25**, 1371-83 (2011).

110. Lee, J.T. & Jaenisch, R. Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. *Nature* **386**, 275-9 (1997).

111. Orom, U.A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154**, 1190-3 (2013).

112. Lai, F., et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501 (2013).

113. Li, W., et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-20 (2013).

114. Dundr, M. & Misteli, T. Biogenesis of nuclear bodies. *Cold Spring Harb Perspect Biol* **2**, a000711 (2010).

115. Misteli, T. The concept of self-organization in cellular architecture. *J Cell Biol* **155**, 181-5 (2001).

116. Kaiser, T.E., Intine, R.V. & Dundr, M. De novo formation of a subnuclear body. *Science* **322**, 1713-7 (2008).

117. Misteli, T. Concepts in nuclear architecture. *Bioessays* **27**, 477-87 (2005).

118. Lewis, J.D. & Tollervey, D. Like attracts like: getting RNA processing together in the nucleus. *Science* **288**, 1385-9 (2000).

119. Gonzalez, I., Mateos-Langerak, J., Thomas, A., Cheutin, T. & Cavalli, G. Identification of Regulators of the Three-Dimensional Polycomb Organization by a Microscopy-Based Genome-wide RNAi Screen. *Mol Cell* **54**, 485-99 (2014).

120. Schaaf, C.A., et al. Cohesin and polycomb proteins functionally interact to control transcription at silenced and active genes. *PLoS Genet* **9**, e1003560 (2013).

121. Atchison, M.L. Function of YY1 in Long-Distance DNA Interactions. *Front Immunol* **5**, 45 (2014).

122. Zhang, H., et al. Intrachromosomal looping is required for activation of endogenous pluripotency genes during reprogramming. *Cell Stem Cell* **13**, 30-5 (2013).

123. Phillips-Cremins, J.E., et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-95 (2013).

124. Guttman, M., et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-10 (2010).

125. Ponten, F., et al. A global view of protein expression in human cells, tissues, and organs. *Mol Syst Biol* **5**, 337 (2009).

126. Mili, S. & Steitz, J.A. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* **10**, 1692-4 (2004).

127. Darnell, R.B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1**, 266-86 (2010).

128. Ule, J., Jensen, K., Mele, A. & Darnell, R.B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376-86 (2005).

129. Wang, Z., Tollervey, J., Briese, M., Turner, D. & Ule, J. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods* **48**, 287-93 (2009).

130. Brimacombe, R., Stiege, W., Kyriatsoulis, A. & Maly, P. Intra-RNA and RNA-protein cross-linking techniques in Escherichia coli ribosomes. *Methods Enzymol* **164**, 287-309 (1988).

131. Singh, G., Ricci, E.P. & Moore, M.J. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods* **65**, 320-32 (2014).

10 FIGURES

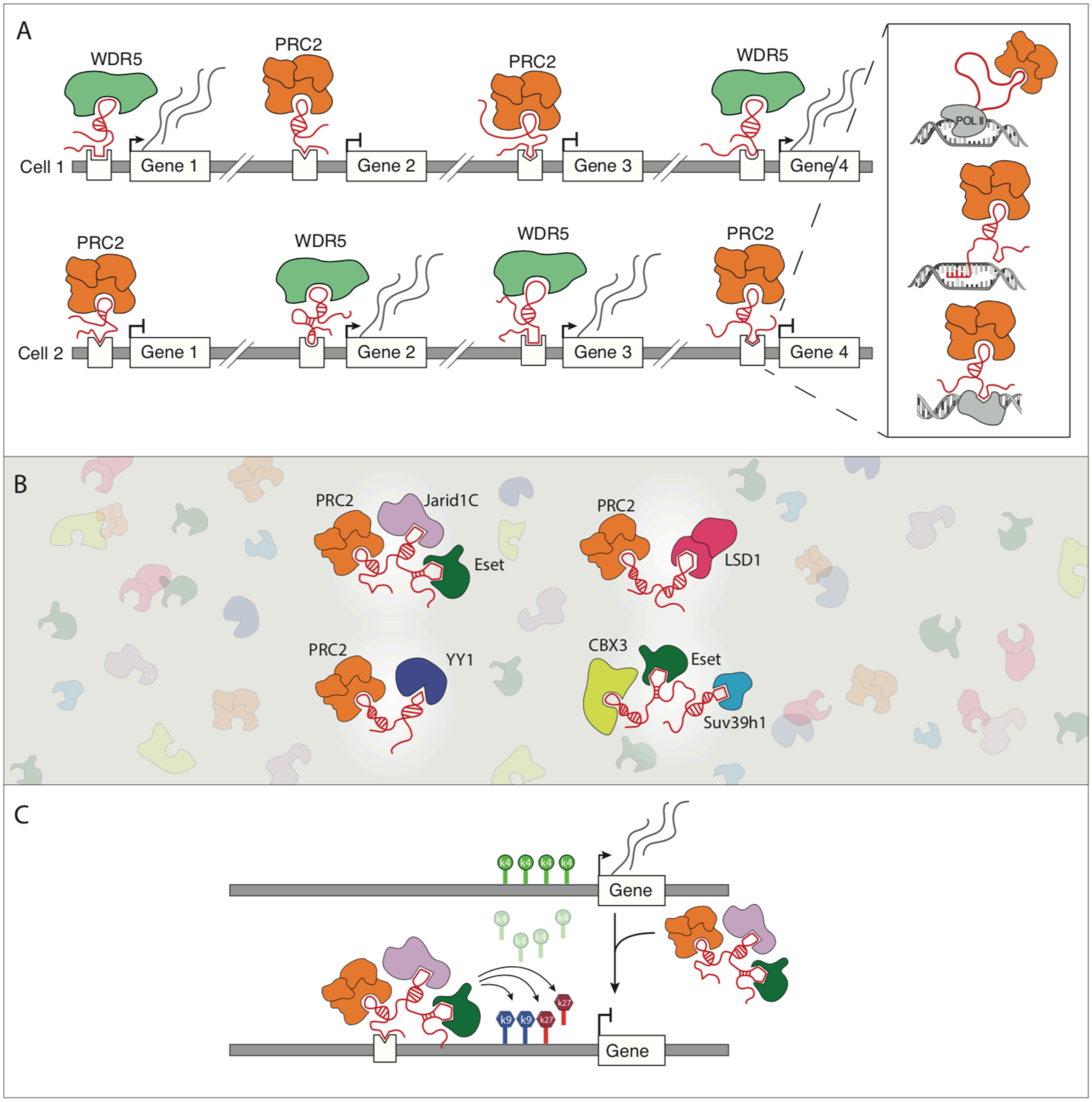


Figure 1. lncRNA-mediated regulation of gene expression through the recruitment of chromatin regulatory proteins.

(a) Different cell types express distinct lncRNAs that can differentially recruit these same chromatin regulatory proteins, including the repressive PRC2 complex and the activating WDR5 chromatin modifying protein, to specific genes. Inset: lncRNAs can recruit these complexes by binding to target sites through three mechanisms: tethering to its nascent transcription locus (top panel), directly hybridizing to genomic targets (middle panel), or interacting with a DNA-binding protein (bottom panel). **(b)** Different lncRNAs can scaffold unique assemblies of chromatin regulatory complexes. lncRNAs are generally expressed at lower levels relative their associated chromatin proteins (background). **(c)** lncRNAs may act to coordinate the regulation of gene expression at specific target locations. In this illustration, a lncRNA that can scaffold PRC2, JARID1C, and ESET may act to remove H3K4me3 and place H3K27me3 and H3K9me2, thereby coordinating the repression of transcription.

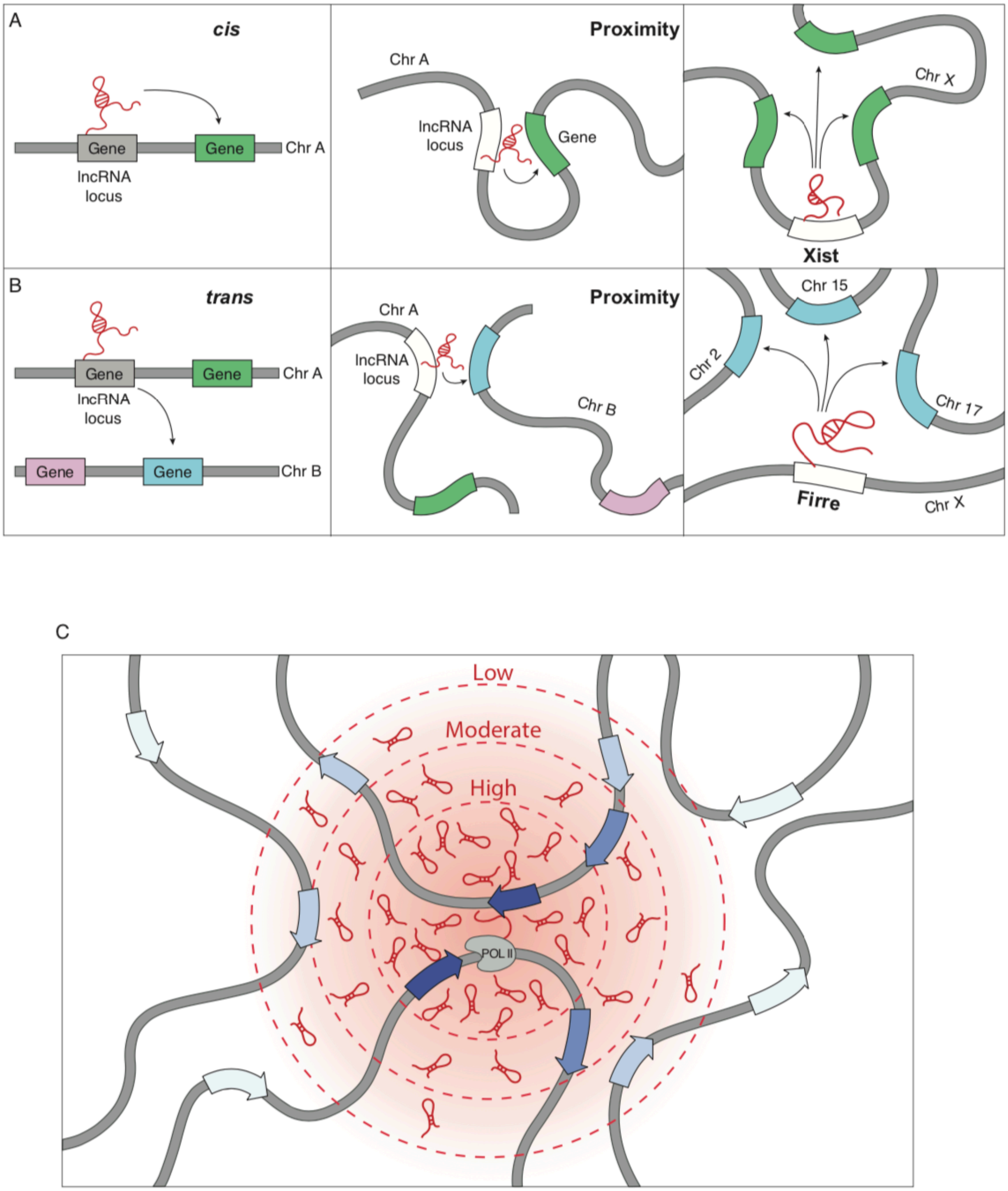


Figure 2. lncRNAs can utilize a proximity-guided search to localize to target genes.

(a) lncRNAs can regulate genes (green box) on its own chromosome (left panel). In the nucleus, this regulation can occur if the lncRNA locus is in close physical proximity to its target sites (middle panel). For instance, Xist localizes to genes across the X chromosome (right panel). **(b)**

lncRNAs can also regulate expression of genes on different chromosomes (blue box, left panel). In the nucleus, this can also occur when the lncRNA locus and its targets are in close proximity (middle panel). An example is Firre, which localizes to targets that are present across several chromosomes (right panel). **(c)** The concentration of a lncRNA will be highest (dark red—inner circle) near its site of transcription and will decrease (light red—outer circles) the further the distance is from its site of transcription, creating a concentration gradient of lncRNA abundance (red cloud, intensity indicates average lncRNA concentration). This spatial gradient establishes a nuclear domain with a high lncRNA concentration, where they can interact with site-specific targets (dark blue arrows). Conversely, lncRNAs outside of the nuclear domain will have a lower probability of interacting with site-specific targets (light blue arrows) due to decreased lncRNA concentration.

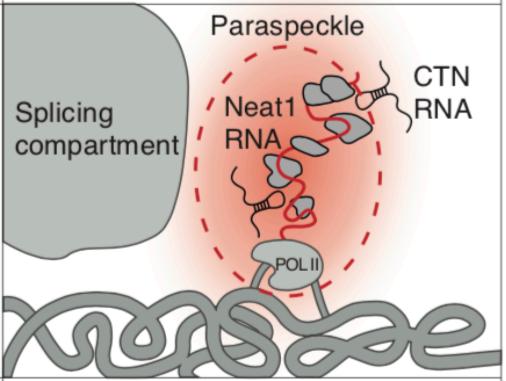
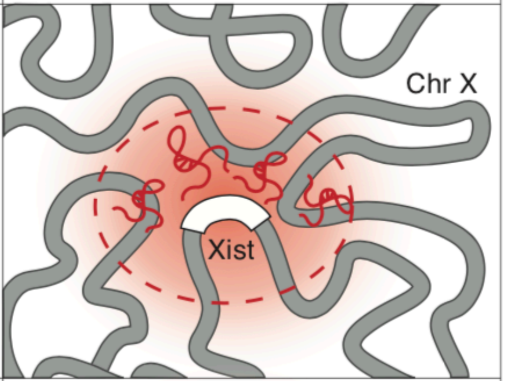
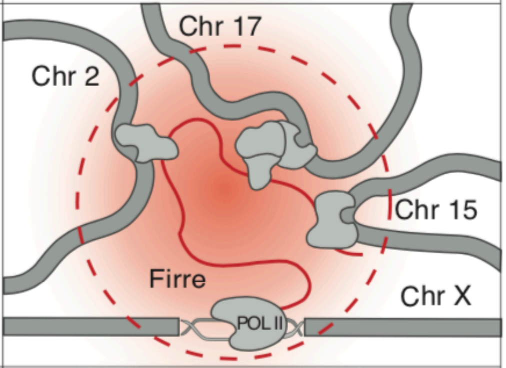
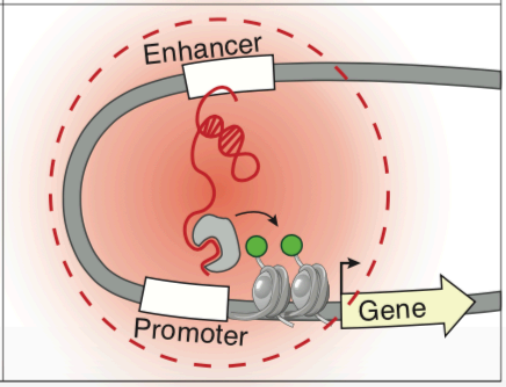
lncRNA	Level of nuclear organization	
Neat1	Nuclear bodies	 Paraspeckle Splicing compartment Neat1 RNA CTN RNA POL II
Xist	Intrachromosomal regulatory domain	
Firre	Interchromosomal regulatory domain	
eRNA	Enhancer-promoter interactions	

Figure 3. lncRNAs can shape three-dimensional nuclear architecture across various levels of organization.

(a) Actively transcribed Neat1 (red line) is required to establish the formation of the paraspeckle nuclear body (red cloud), which is an RNA-protein (gray) nuclear domain that is the site of nuclear retention of RNAs such as the CTN RNA (black). **(b)** Xist (red line) establishes an intra-chromosomal nuclear domain (red cloud) by nucleating near its transcription site (white box) and spreading to DNA sites in spatial proximity to its locus. **(c)** Firre establishes an inter-chromosomal nuclear domain and brings together targets on chromosomes 2, 15, and 17 into close physical proximity to its transcriptional locus on the X chromosome. **(d)** Enhancer RNAs (eRNAs) maintain the interaction between enhancer and promoter regions and may do this by interacting with proteins that can modify chromatin.

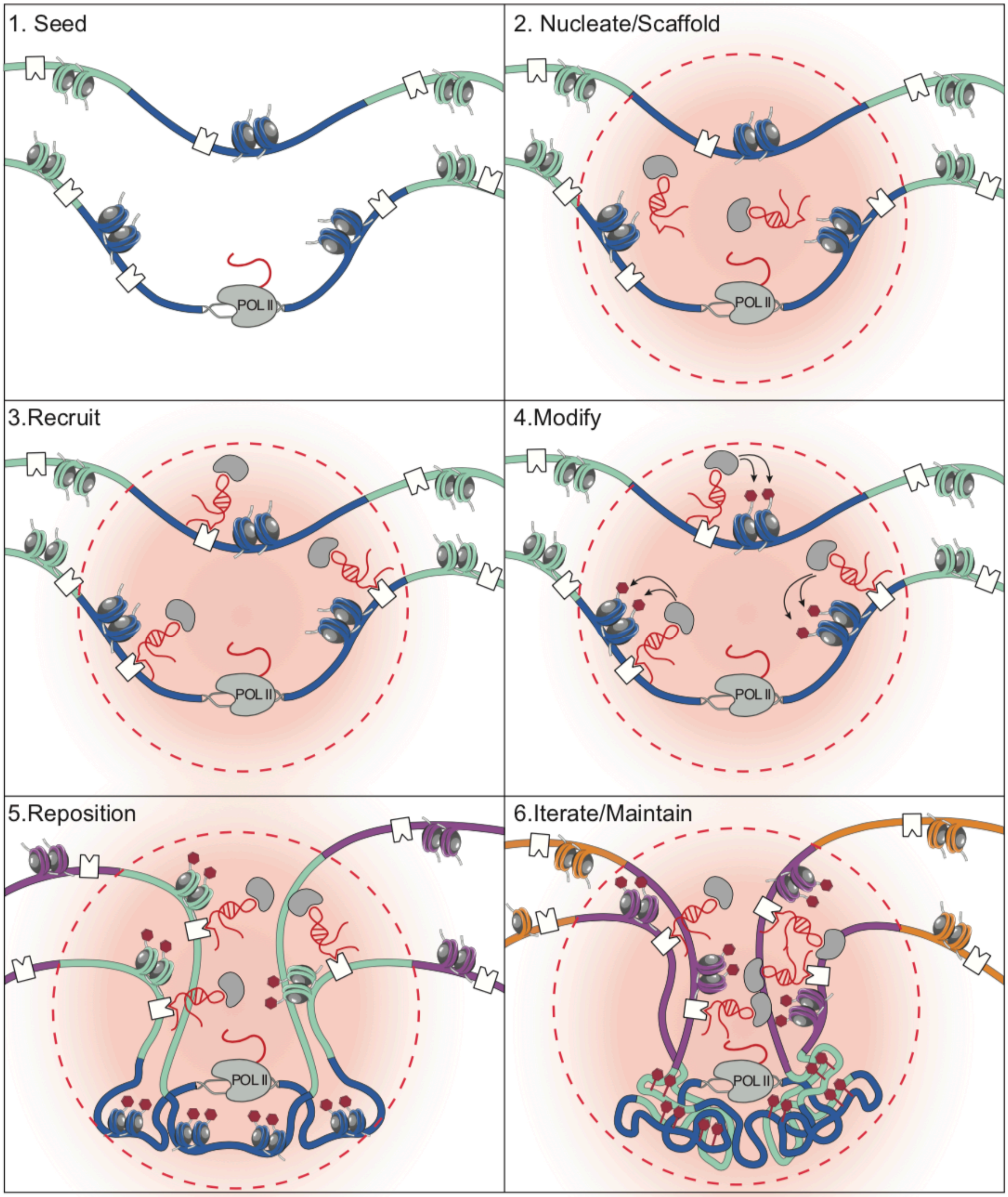


Figure 4. A model for how lncRNAs can dynamically shape nuclear organization.

The proposed steps involved in lncRNA-mediated assembly of nuclear organization roughly based on the proposed models for the Neat1[44] and Xist[46] lncRNAs. (i) Transcription of a lncRNA can **seed** the formation of a lncRNA nuclear domain. (ii) lncRNAs can bind to proteins in the nucleus (gray circles) to **scaffold** protein complexes. Formation of these complexes will **nucleate** the formation of a spatial compartment (red cloud, dashed lines) near the transcriptional locus of the lncRNA. (iii) lncRNAs can bind to specific DNA sites (white squares) to **recruit** lncRNA-protein complexes to target sites. (iv) By recruiting these complexes to DNA, lncRNAs can guide chromatin modifications (blue histones), such as repressive histone modification (red marks). (v) Modified chromatin may be compressed and **repositioned** into a new nuclear region. (vi) As the lncRNA continues to be transcribed from its transcriptional locus, it may **iteratively** bind to DNA sites (green regions), modify target sites, and reposition DNA into the lncRNA nuclear domain. This continuous process may act to **maintain** the nuclear domain established by a lncRNA.

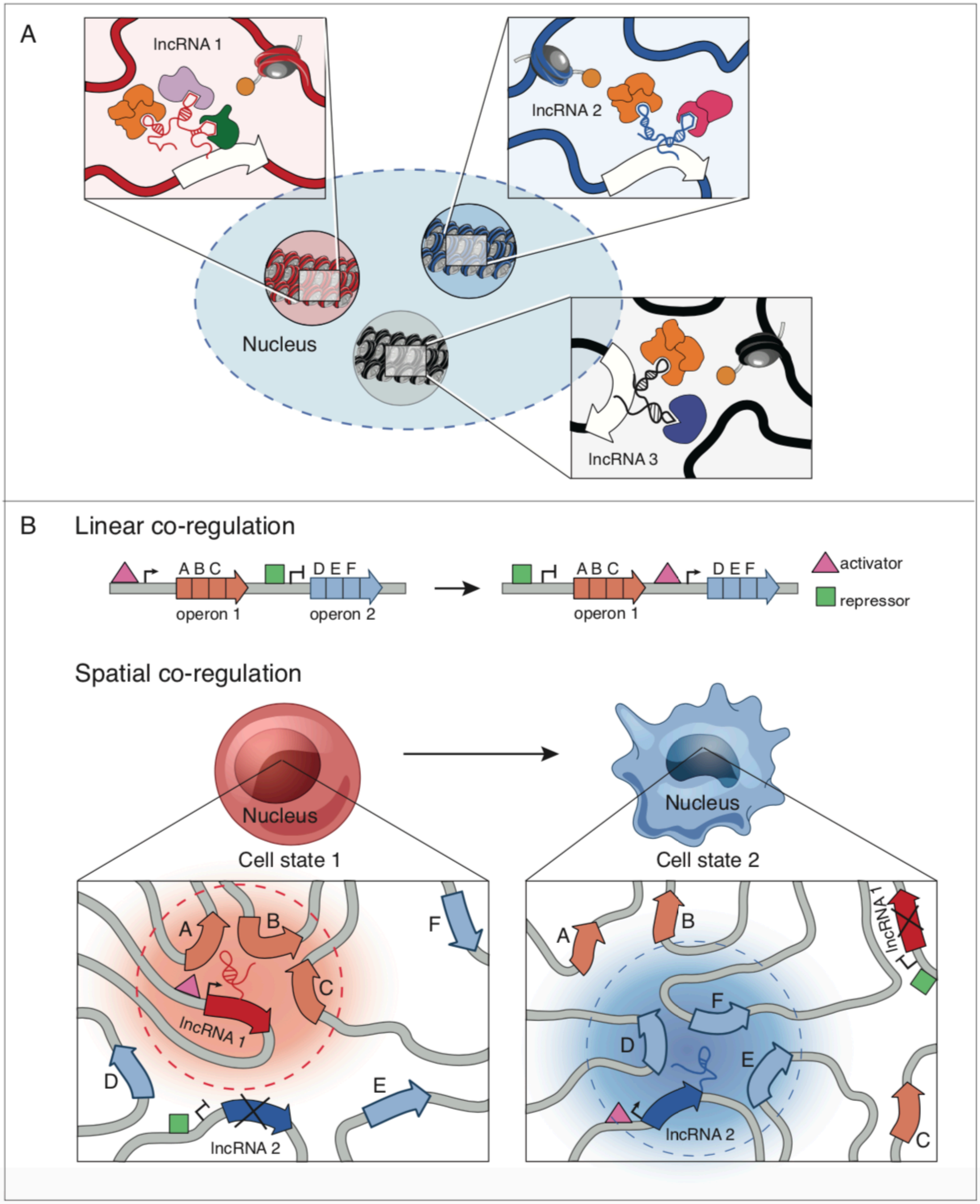


Figure 5. A hypothesis for how lncRNAs may act to assemble dynamic and specific nuclear domains.

(a) Nuclear domains that share the same proteins can interact in different regions of the nucleus. Zoom-in panels: We hypothesize that different lncRNAs may act to distinguish between these domains by scaffolding and assembling distinct domains. **(b)** Through linear co-regulation, operons can simultaneously regulate sets of genes (A, B, C and D, E, F) with shared regulatory functions. Activators (pink triangles) and repressor (green boxes) control operon expression under a particular cell state. We hypothesize that through spatial co-regulation, lncRNAs may nucleate the formation of nuclear domains to co-localize target genes upon induction of lncRNA expression. For instance, upon induction of lncRNA1, genes A, B, and C are co-regulated in a nuclear domain (red cloud, dashed lines). Under a different cell state, lncRNA1 expression is repressed, leading to the breakdown of the lncRNA1 nuclear domain and expression of lncRNA2 leads to formation of another nuclear domain (blue cloud, dashed lines) containing genes D, E, and F.

Chapter 2

HIGHER-ORDER INTER-CHROMOSOMAL HUBS SHAPE 3D GENOME ORGANIZATION IN THE NUCLEUS

Sofia A. Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, Alexander A. Shishkin, Prashant Bhat, Yodai Takei, Vickie Trinh, Erik Aznauryan, Pamela Russell, Christine Cheng, Marko Jovanovic, Amy Chow, Long Cai, Patrick McDonel, Manuel Garber, and Mitchell Guttman

A modified version of this chapter was published as “Higher-order inter-chromosomal hubs shape 3-dimensional genome organization in the nucleus” in *Cell*. 174(3):744–757. 2018.
doi: 10.1016/j.cell.2018.05.024

2.1 ABSTRACT

Eukaryotic genomes are packaged into a three-dimensional structure in the nucleus. Current methods for studying genome-wide structure are based on proximity ligation. However, this approach can fail to detect known structures, such as interactions with nuclear bodies, because these DNA regions can be too far to directly ligate. Accordingly, our overall understanding of genome organization remains incomplete. Here, we develop Split-Pool Recognition of Interactions by Tag Extension (SPRITE), which enables genome-wide detection of higher-order interactions within the nucleus. Using SPRITE, we recapitulate known structures identified by proximity ligation and identify additional interactions occurring across larger distances, including two hubs of inter-chromosomal interactions that are arranged around the nucleolus and nuclear speckles. We show that a substantial fraction of the genome exhibits preferential organization relative to these nuclear bodies. Our results generate a global model whereby nuclear bodies act as inter-chromosomal hubs that shape the overall packaging of DNA in the nucleus.

2.2 INTRODUCTION

Although the same genomic DNA is packaged in the nucleus of each cell, different sets of genes are expressed in different cell states. Despite significant progress over the past decade, there are still many unanswered questions about how the genome is organized within the nucleus and how these structures change across different cell states.

Current methods for genome-wide mapping of 3-dimensional (3D) genome structure rely on proximity ligation (e.g., Hi-C), which works by ligating the ends of DNA regions that are in close spatial proximity in the nucleus followed by sequencing to map pairwise interactions. These methods have shown that the genome is largely organized around chromosome territories, such that most DNA interactions occur within an individual chromosome (Gibcus and Dekker, 2013). These interactions include chromatin loops that connect genomic DNA regions such as enhancers and promoters, local interacting neighborhoods of DNA called topologically associated domains

(TADs), and compartments where DNA regions interact based on transcriptional activity (Pombo and Dillon, 2015).

Another widely used method for studying nuclear structure is microscopy, which involves in situ imaging of DNA, RNA, and protein in the nucleus. These methods have shown that specific regions of the genome, including specific inter-chromosomal interactions, can organize around nuclear bodies (Hu et al., 2010). For example, RNA Polymerase I transcribed ribosomal DNA (rDNA) genes, which are encoded on several distinct chromosomes, localize within the nucleolus (Pederson, 2011). In addition, specific examples of RNA polymerase II (PolII) transcribed genes have been shown to localize near the periphery of nuclear speckles (Khanna et al., 2014), a nuclear body that contains various mRNA processing and splicing factors (Spector and Lamond, 2011). These observations, and others (Branco and Pombo, 2006; Lomvardas et al., 2006), demonstrate that genome interactions can occur beyond chromosome territories and organize around nuclear bodies.

Yet, despite the power of each of these methods for mapping nuclear structure, there is a growing appreciation that microscopy and proximity ligation measure different aspects of genome organization (Giorgetti and Heard, 2016; Williamson et al., 2014). Specifically, microscopy measures the 3D spatial distances between DNA sites within single cells, whereas proximity ligation measures the frequency with which two DNA sites are close enough in the nucleus to directly ligate (Dekker, 2016). This difference is particularly significant when considering DNA regions that organize around nuclear bodies, which can range in size from 0.5–2μm (Pederson, 2011), and therefore may be too far apart to directly ligate. This may explain why proximity-ligation methods do not identify known interactions between chromosomes that organize around specific nuclear bodies.

These differences between proximity ligation and microscopy highlight a challenge for generating comprehensive maps of genome structure. Specifically, it remains unclear whether the specific inter-chromosomal interactions identified by microscopy represent special cases or broader principles of global genome organization. Additionally, both methods are limited to measuring simultaneous contacts between a small number (~2–3) of genomic regions and therefore cannot measure how multiple DNA sites simultaneously organize within the nucleus (O’Sullivan et al., 2013). Accordingly, current methods cannot generate a global picture of genome organization, which is critical for addressing key questions, such as why some specific PolII-transcribed regions associate with nuclear speckles while others do not (Shopland et al., 2003), which DNA regions organize simultaneously around the same nuclear body, and how genomic DNA is organized in the nucleus relative to multiple nuclear bodies, chromosome territories, and other features.

To address these technical challenges, we develop a method called SPRITE that moves away from proximity ligation and enables genome-wide detection of multiple DNA interactions that occur simultaneously within the nucleus. Using SPRITE, we recapitulate known genome structures identified by Hi-C, including chromosome territories, compartments, TADs, and loop structures, and identify that many of these occur within higher-order structures in the nucleus. Because SPRITE does not rely on proximity ligation, it identifies interactions that occur across larger spatial distances than can be observed by Hi-C. These long-range interactions include two major hubs of inter-chromosomal interactions. By extending SPRITE to simultaneously measure RNA and DNA interactions, we find that these two inter-chromosomal hubs correspond to DNA organization around the nucleolus and nuclear speckles, respectively. Moreover, DNA regions within these hubs can simultaneously organize around the same nuclear body within individual cells. We show that gene-dense and highly transcribed PolII regions organize around nuclear speckles and gene poor, and therefore transcriptionally inactive, regions that are centromere proximal organize around the nucleolus. In addition to the regions that directly associate with these nuclear bodies, we find that a substantial fraction of the genome exhibits preferential spatial positioning relative to each of these nuclear bodies. Importantly, these preferential spatial distances quantitatively correspond to functional and structural properties, including the density of active PolII within a genomic region. Together, our results provide a global model of genome organization whereby nuclear bodies act as inter-chromosomal hubs that shape the overall 3D packaging of DNA in the nucleus.

2.3 RESULTS

2.3.1 SPRITE: A GENOME-WIDE METHOD TO IDENTIFY HIGHER-ORDER DNA INTERACTIONS IN THE NUCLEUS

We sought to develop a genome-wide method that enables mapping of higher-order interactions that occur simultaneously between multiple DNA sites within the same nucleus. To do this, we developed SPRITE, a method that does not rely on proximity ligation. SPRITE works as follows: DNA, RNA, and protein are crosslinked in cells, nuclei are isolated, chromatin is fragmented, interacting molecules within an individual complex are barcoded using a split-pool strategy, and interactions are identified by sequencing and matching all reads that contain identical barcodes (**Figure 1A, Methods**).

Specifically, we uniquely barcode each molecule in a crosslinked complex by repeatedly splitting all complexes across a 96-well plate ("split"), ligating a specific tag sequence onto all DNA molecules within each well ("tag"), and then pooling these complexes into a single well ("pool"). After several rounds of split-pool tagging, each molecule in an interacting complex contains a unique series of ligated tags, which we refer to as a barcode (**Figure 1A**). Because all molecules in a crosslinked complex are covalently linked, they will sort together in the same wells throughout each round of the split-pool tagging process and will contain the same barcode, whereas the molecules in separate complexes will sort independently and therefore will obtain distinct barcodes. Therefore, the probability that molecules in two independent complexes will receive the same barcode decreases exponentially with each additional round of split-pool tagging. For example, after 6 rounds of split-pool tagging, there are ~10^{12} possible unique barcode sequences, which exceeds the number of unique DNA molecules present in the initial sample (~10^{9}). After split-pool tagging, we sequence all tagged DNA molecules and match all reads with shared barcodes (**Methods**). We refer to all unique DNA reads that contain the same barcode as a SPRITE cluster (**Figure 1A**).

To confirm that DNA reads within a SPRITE cluster represent interactions that occur in the same nucleus and are not formed by spurious association or aggregation in solution, we mixed crosslinked lysates from human and mouse cells prior to performing SPRITE and found that

~99.8% of all SPRITE clusters contained only human or mouse reads, but not both (**Methods**, **Figure S1A**).

SPRITE differs from previous methods in several ways. In contrast to Hi-C, SPRITE can measure multiple DNA molecules that simultaneously interact within an individual nucleus, provides information about interactions that are heterogeneous from cell-to-cell, and is not restricted to measuring DNA interactions that are close enough in the nucleus to directly ligate. In contrast to Genome Architecture Mapping (GAM) (Beagrie et al., 2017), another proximity-ligation-independent method, SPRITE can be performed without requiring specialized equipment or training, is faster to perform, and does not require extensive whole genome amplification. Furthermore, because SPRITE does not rely on proximity ligation or whole genome amplification, it can be extended beyond DNA to directly incorporate RNA simultaneously. We describe these specific features of SPRITE in the following sections.

2.3.2 SPRITE ACCURATELY MAPS KNOWN GENOME STRUCTURES ACROSS VARIOUS RESOLUTIONS

To test whether SPRITE can accurately map genome structure, we compared the results obtained by SPRITE to those measured by Hi-C. Specifically, we generated SPRITE maps in two mammalian cell types that have been previously mapped by Hi-C—mESCs (mES) (Dixon et al., 2012) and human lymphoblastoid cells (GM12878) (Rao et al., 2014). We generated ~1.5 billion sequencing reads from each sample and matched reads containing the same barcode to obtain ~50 million SPRITE clusters from each sample (**Methods**). These SPRITE clusters range in size from 2 reads to >1000 reads per cluster (**Figure S1C**). To directly compare SPRITE and Hi-C, we converted SPRITE clusters into pairwise contact frequencies by enumerating all pairwise contacts observed within a single cluster and down-weighting each pairwise contact by the total number of molecules contained within the cluster (**Figure 1A**). This normalization prevents large SPRITE clusters from disproportionately impacting the pairwise frequency maps (**Methods**).

Overall, the pairwise contact maps generated with SPRITE are highly comparable to Hi-C maps, with similar structural features observed across all levels of genomic resolution. At a genome-wide level, we observe a clear preference for interactions that occur within the same chromosome (**Figure 1B**). At a chromosome-wide level, we observe similar A and B compartments as identified by Hi-C in both the mouse and human data (Spearman ρ= 0.85, 0.93, respectively) (**Figure S1D-F**). These correspond to locations of active and inactive transcription (Gibcus and Dekker, 2013). At 40kb resolution, we observe TADs, where adjacent DNA sites organize into highly self-interacting domains separated by boundaries that preclude interactions with other neighboring regions (**Figure 1D**). The location and strength of TAD boundaries, measured by insulation scores across the genome, are highly correlated in Hi-C and SPRITE for both mouse and human (Spearman ρ= 0.90, 0.94) (**Figure S1G-I**). Finally, at 25kb resolution, we observe specific "looping" interactions that connect local regions that contain the expected convergent CTCF motif orientation previously described for loop structures (**Figure 1E, Figure S1J**) (Sanborn et al., 2015). More generally, we find that the loops previously identified by Hi-C are strongly enriched within our SPRITE data at 10kb resolution (**Figure S1K-L, Methods**).

2.3.3 SPRITE IDENTIFIES HIGHER-ORDER INTERACTIONS THAT OCCUR SIMULTANEOUSLY

In addition to confirming pairwise genome structures identified by Hi-C, SPRITE can also directly measure multiple DNA regions that interact simultaneously within an individual cell, which we refer to as higher-order interactions. Although microscopy and specific proximity-ligation methods can also map higher-order interactions (Darrow et al., 2016; Olivares-Chauvet et al., 2016), these are largely restricted to mapping 3-way contacts. In contrast, SPRITE is not restricted in the number of simultaneous DNA contacts.

To explore the higher-order structures identified by SPRITE, we enumerated all interactions that occur simultaneously between 3 or more independent genomic regions ($k \geq 3$), which we refer to as a *k*-mer. Similar to pairwise contact frequencies, one of the largest determinants of *k*-mer frequency in our data is the linear genomic distance separating each region in the *k*-mer. To account for this, we computed an enrichment score by normalizing the frequency of the observed *k*-mer by the average frequency observed across random *k*-mers that retain the same genomic distance (**Figure S2A, Methods**).

Overall, we identified >310,000 *k*-mers (1Mb resolution, *k*=3–14 regions) that were observed in at least 5 independent SPRITE clusters, occurred at a frequency that exceeded 90% of the random permutations, and occurred >4-fold more frequently than the average of the permuted regions (**Figure S2B, Table S1A, Methods**). Importantly, the frequency of observing even a single higher-order SPRITE cluster that does not represent a set of interactions that occur within the same individual cell is extremely low (<0.2% of all clusters, **Figure S1A**).

These enriched *k*-mers include various higher-order genomic DNA structures, including active compartments, gene clusters, and consecutive loop structures.

(i) Active Compartments. We observed highly enriched *k*-mers that connect multiple A compartment (transcriptionally active) regions that are non-contiguous and span large distances of the same chromosome. Specifically, we observed tens of thousands of individual SPRITE clusters that contain reads from at least three different A compartment regions that span at least 100 megabases within an individual chromosome (**Table S1B-C**, **Figure 2A, S2C**). This suggests that active DNA regions may interact within higher-order compartments (**Figure 2B)**.

(ii) Gene Clusters. We identified >75 SPRITE clusters that connect three non-contiguous genomic regions within the human *HIST1* cluster that encode 55 human histone genes (**Figure 2C**). Notably, these SPRITE clusters skip the intervening transcriptionally inactive regions. The frequency of SPRITE clusters that connect the three *HIST1* gene clusters was never observed in any of the 100 randomly permuted *k*-mers containing the same genomic distance. This result suggests that multiple histone gene clusters simultaneously interact, consistent with observations that histone genes localize within specific nuclear bodies referred to as histone locus bodies (Nizami et al., 2010) (**Figure 2D**).

(iii) Consecutive Loops. Previous Hi-C studies suggested that consecutive loops may form higher-order interactions that bring together three distinct regions of the genome (Sanborn et al., 2015). Consistent with this, we observe several examples of highly enriched *k*-mers that correspond to consecutive loop structures (**Figure S2D**). For example, we observe >19 SPRITE clusters that contain reads corresponding to three loop anchor points on human chromosome 8 (**Figure 2E**). This suggests that multiple consecutive loops may occur simultaneously within the same cell (**Figure 2F**).

In addition to these examples, we also observe >11,000 enriched SPRITE *k*-mers corresponding to simultaneous 3-way interactions of TAD regions containing multiple enhancers and highly transcribed regions previously reported by GAM (**Table S1D-E, Figure S2E**). Taken together, these results suggest that SPRITE can detect multiple DNA interactions that occur simultaneously.

2.3.4 SPRITE IDENTIFIES INTERACTIONS THAT OCCUR ACROSS LARGE GENOMIC DISTANCES

Because SPRITE does not rely on proximity ligation, which requires two DNA sites to be close enough to form a ligation junction (Dekker, 2016; Giorgetti and Heard, 2016), we reasoned that SPRITE might identify additional interactions that occur at further nuclear distances than those identified by Hi-C (**Figure 3A**). Indeed, we noticed that the number of pairwise contacts observed between two genomic regions as a function of their linear genomic distance ("distance decay") occurs at different rates when comparing the Hi-C and SPRITE data (**Figure S3A**). Specifically, SPRITE identifies a significantly larger number of pairwise contacts that are present at larger linear genomic distances. Importantly, these additional long-range interactions correspond to an increased number of contacts between specific genomic regions expected to interact (**Table S2A**). For example, we observe a significant increase in interactions occurring between non-local active compartments separated by >100Mb (**Figure S3B**).

Because SPRITE uses fragmentation to generate clusters of crosslinked interactions within the nucleus, we reasoned that small SPRITE clusters may represent interactions that are close in 3D space, whereas larger SPRITE clusters may represent interactions crosslinked across farther distances (**Figure 3A**). To test this, we stratified the SPRITE clusters based on their number of reads and generated pairwise contact maps. For small SPRITE clusters (2–10 reads), we observe a distance decay rate comparable to the rate observed by Hi-C, with most contacts occurring in close linear distances. Indeed, SPRITE clusters containing 2–10 reads also show pairwise contact maps that are similar to Hi-C (**Figure S3C**). In contrast, for the larger SPRITE clusters (11–1000+ reads), a larger number of contacts occur at longer genomic distances and the number of interactions at these longer distances increases with cluster size (**Figure 3B**). These different cluster sizes represent interactions that correspond to different structural features (**Table S2A**). For example, the small clusters preferentially identify interactions within TADs and within local compartment regions, while larger clusters correspond to increased interactions between distinct TADs as well as distal compartment regions (**Figure 3C, S3C-D**).

These results demonstrate that SPRITE captures longer-range interactions than are observed by Hi-C and that the distances that interactions occur can be measured using SPRITE clusters of different sizes (**Figure 3A**).

2.3.5 INTER-CHROMOSOMAL INTERACTIONS ARE PARTITIONED INTO TWO DISTINCT HUBS

We also observed many inter-chromosomal interactions that are identified within the larger SPRITE clusters (11–1000+ reads), but are not observed in the smaller clusters or by Hi-C (**Figure 3D, S3E-G**).

To explore these inter-chromosomal interactions, we built a graph connecting all 1Mb regions in the mouse genome containing a significant pairwise interaction (p-value < 10^{-10}) (**Figure 3E**). These interactions segregate into two discrete “hubs”, such that a large number of contacts occur within each hub, but no interactions occur between the two hubs. These hubs contain different functional properties: the first hub corresponds to gene-poor and therefore transcriptionally inactive regions, whereas the second hub corresponds to gene-dense regions that are highly transcribed by RNA polymerase II, enriched for active chromatin modifications, and contain other features of active transcription (**Figure 3F, S3J**, see **Methods**). Based on these properties, we refer to these hubs as the “inactive hub” and “active hub”, respectively (**Table S2B-C**).

Importantly, we observed two similar inter-chromosomal hubs in the human genome that displayed comparable functional properties (**Figure S3K-N**, **Table S2D-E**). Given the similar properties of the mouse and human hubs, we focused on mouse embryonic stem (mES) cells for our subsequent characterization of these hubs.

2.3.6 RNA-DNA SPRITE REVEALS THAT THE INACTIVE INTER-CHROMOSOMAL HUB IS ORGANIZED AROUND THE NUCLEOLUS

To understand where in the nucleus these inter-chromosomal interactions occur, we first explored the inactive hub and noticed that several of the DNA regions in this hub are linearly close to genomic DNA regions that encode ribosomal RNAs (rDNA, see **Methods**). Because rDNA regions are known to be organized and transcribed within the nucleolus (Pederson, 2011), we hypothesized that the inactive hub regions may organize around the nucleolus.

To test this, we explored whether the DNA regions in this hub are associated with ribosomal RNA localization, which demarcates the nucleolus (Pederson, 2011). Specifically, we adapted the SPRITE protocol to enable simultaneous mapping of interactions between RNA and DNA molecules by ligating an RNA-specific adaptor that enable simultaneous tagging of both DNA and RNA during each round of the split-pool procedure (see **Methods**, **Figure S4A-B**). Using this approach, we mapped the interactions of ribosomal RNA on genomic DNA and found it was specifically enriched over the genomic DNA regions contained within the inactive hub (**Figure 4A**). In fact, ribosomal RNA enrichment across the genome is correlated with how frequently a region contacts the inactive hub (**Figure 4A**, **S4C**).

To confirm that the inactive hub represents DNA sites located near the nucleolus in situ, we performed DNA FISH combined with immunofluorescence for nucleolin, a protein marker of the nucleolus. Specifically, we selected DNA FISH probes for 4 genomic regions in the inactive hub and 3 control regions on the same chromosomes. We selected an additional control region on a chromosome lacking any inactive hub regions (**Figure 4B**). We calculated the 3D distance between each allele and the nearest nucleolus and found that inactive hub regions are dramatically closer to the nucleolus than negative control regions (average ~750nm closer, **Figure S4D**). In the majority of cells analyzed, at least one allele of the inactive hub DNA regions directly contacts the periphery of the nucleolus (~61% of cells, **Figure 4C-D**, **Figure S4D-E)**. Therefore, we refer to this hub as the nucleolar hub. Our results confirm previous observations that the nucleolus can act as an anchor for inactive chromatin regions (Padeken and Heun, 2014).

Because many genomic regions are contained within the nucleolar hub, we hypothesized that multiple DNA sites simultaneously interact around a single nucleolus. Consistent with this, we observed >1,200 SPRITE clusters that contain simultaneous interactions between at least three distinct genomic regions on different chromosomes in the nucleolar hub (**Figure 4E, Table S3**). To confirm these inter-chromosomal contacts occur through co-localization at the same nucleolus, we performed 2-color DNA FISH combined with immunofluorescence for nucleolin and measured the frequency of co-association at the same nucleolus (**Movie S1**). We observed that two regions in the nucleolar hub were ~7-times more likely to co-occur around the same nucleolus compared to a nucleolar hub region and control region (**Figure S4F**). Importantly, the frequency of co-occurrence of a pair of DNA sites at the same nucleolus measured by microscopy is highly correlated with the frequency at which these genomic DNA regions co-occur within the same SPRITE clusters (Pearson r=0.99, **Figure 4F**). This demonstrates that SPRITE quantitatively measures the frequency at which DNA sites co-occur at a nuclear body within single cells.

Together, these results provide a genome-wide map of DNA regions that spatially organize around the nucleolus. While other studies have previously mapped individual regions that contact the nucleolus across a population of cells (Németh et al., 2010), our results provide a genome-wide 3D picture of how multiple DNA sites arrange simultaneously around the nucleolus.

2.3.7 THE ACTIVE INTER-CHROMOSOMAL HUB IS ORGANIZED AROUND NUCLEAR SPECKLES

We noticed that the genomic DNA regions within the active hub are strongly enriched for U1 spliceosomal RNA and Malat1 lncRNA localization (Engreitz et al., 2014) (**Figure S5A**) and the level of their localization is highly correlated with how frequently a DNA region interacts with the active hub (Spearman ρ =0.80 and 0.74, **Figure 5A**, **S5B**). Because the U1 and Malat1 RNAs are known to localize at nuclear speckles (Hutchinson et al., 2007), a nuclear body that contains proteins involved in mRNA splicing and processing (Spector and Lamond, 2011), we hypothesized that inter-chromosomal interactions occurring between regions in the active hub may be spatially organized around nuclear speckles.

To test this, we performed DNA FISH combined with immunofluorescence for SC35, a well-known protein marker of nuclear speckles. We selected FISH probes targeting 3 DNA regions contained in the active hub and 2 control regions on the same chromosome not in the active hub. We also selected another control region within the inactive hub (**Figure 5B**). We calculated the 3D distance between each region and the closest nuclear speckle and found that all 3 active hub regions are consistently closer to nuclear speckles compared with the 3 control regions (**Figure 5C-E**, **S5C-G**). Indeed, for active hub regions, we observe a dramatic increase in the number of cells where at least one allele directly touches a nuclear speckle relative to control regions (~13-fold, **Figure S5G**). Despite this preferential organization near the nuclear speckle, the number of cells in which an active region directly contacts the nuclear speckle is relatively low (~10% of cells), which is consistent with previous observations by live-cell microscopy that individual genomic DNA interactions with a nuclear speckle are transient (Khanna et al., 2014). Based on these observations, we refer to the DNA regions in this hub as the nuclear speckle hub.

We hypothesized that nuclear speckle hub regions may simultaneously organize around the same nuclear speckle. Indeed, we identified >690 SPRITE clusters containing at least three distinct active hub regions that were present on different chromosomes (**Table S4**, **Figure 5F**). Consistent with this, we observe two DNA sites in the speckle hub are >8-times as likely to be within 1μm of each other by microscopy compared to an active and control region (**Figure S5H**). Indeed, we observe two pairs of genomic DNA regions in the active hub preferentially organized near the same nuclear speckle in 2 out of 15 cells that were measured (**Figure 5G**). In contrast, we did not observe even a single example of an active hub and control region that were organized near the same speckle in 21 cells that were measured. However, because there are many nuclear speckles in a given cell (~20/nucleus) and DNA interactions with an individual nuclear speckle can be transient (~10% of regions directly contact any speckle), it is challenging to robustly quantify the frequency at which multiple DNA regions simultaneously associate around the same nuclear speckle using microscopy (see **Methods**).

Our results confirm previous observations that specific actively transcribed regions can interact with nuclear speckles (Brown et al., 2008; Khanna et al., 2014; Shopland et al., 2003) and extend this observation genome-wide by providing a map of DNA interactions around nuclear speckles. Moreover, our results suggest that multiple actively transcribed DNA regions can arrange simultaneously around nuclear speckles to form higher-order inter-chromosomal interactions.

2.3.8 NUCLEAR BODIES CONSTRAIN THE OVERALL 3D ORGANIZATION OF GENOMIC DNA IN THE NUCLEUS

We considered that nuclear bodies might play an important role in defining the overall arrangement of genomic DNA in the nucleus because they organize large hubs of inter-chromosomal interactions. To address this, we focused on how genomic regions that are not within these hubs are spatially positioned relative to each nuclear body. We considered 3 possibilities: these regions show (i) random spatial positioning with respect to either nuclear body (random preference), (ii) spatial positioning that linearly decays as a function of genomic distance from a hub-associated region (linear preference), or (iii) specific non-linear spatial preferences to either nuclear body (non-linear preference).

To test these possibilities, we calculated the average number of SPRITE contacts for each 1Mb region in the genome relative to regions in the nucleolar or nuclear speckle hubs (**Figure 6A**). Interestingly, we find that a large fraction of genomic regions exhibit preferential contacts with either hub (**Figure 6A)**, such that regions that frequently contact the nucleolar hub are depleted relative to the nuclear speckle hub, and vice versa (**Figure 6B**). Importantly, these preferential contacts do not occur exclusively at regions in close linear distance to hub regions, as would be expected if this organization occurred through a linear "dragging" effect of the chromatin polymer. For example, several non-contiguous regions on mouse chromosome 11 have high speckle hub contact frequencies despite being linearly far from speckle hub regions (**Figure 6B**). Moreover, several non-contiguous genomic regions preferentially contact the nucleolar hub even though chromosome 11 does not contain any nucleolar hub regions (**Figure 6B**). These results suggest that a large fraction of genomic DNA regions show preferential non-linear spatial arrangement to either the nucleolus or nuclear speckle.

To confirm that these spatial preferences accurately represent 3D distances of DNA sites to these nuclear bodies in situ, we performed DNA FISH combined with immunofluorescence for nucleolin or SC35. Specifically, we selected 9 DNA regions across a range of SPRITE contact frequencies relative to the nucleolus, including nucleolar hub regions, a speckle hub region, and 4 regions with different intermediate spatial preferences (**Figure S6B-C**). In all cases, the 3D distances between each DNA region and the nucleolus is strongly correlated with its SPRITE contact frequency to the nucleolar hub (Pearson r = 0.98, **Figure 6C, S6D**). Similarly, we selected 9 DNA regions across a range of SPRITE contact frequencies relative to the speckle hub, including speckle hub regions, a nucleolar hub region, and 3 regions with different intermediate spatial preferences (**Figure S6B-C**). The 3D distance between each DNA region and a nuclear speckle is strongly correlated with its SPRITE contact frequency to the nuclear speckle hub (Pearson r = 0.98, **Figure 6D, S6B-D**). These results demonstrate that SPRITE provides accurate quantitative measurements of 3D spatial distances across the nucleus.

2.3.9 FUNCTIONAL AND STRUCTURAL PROPERTIES DEFINE PREFERENTIAL ORGANIZATION TO NUCLEAR BODIES

To understand the basis of these spatial preferences, we examined the structural and functional properties of the DNA regions positioned close to each nuclear body.

(i) Nucleolar preference. We found that regions that are linearly close to the centromere are closer to the nucleolus (**Figure 6E,** Spearman ρ = 0.76). Notably, these results are consistent with previous observations that centromeres often co-localize on the periphery of the nucleolus (Pollock and Huang, 2009; Tjong et al., 2016) **Figure S7A-B**). However, not all genomic regions close to centromeres are close to the nucleolus because actively transcribed regions are excluded from the nucleolar compartment even when they reside in linear proximity to a centromere (**Figure S6E**). Because actively transcribed regions are preferentially positioned away from the nucleolus, the genomic DNA regions that are closer to the nucleolus tend to correspond to inactive chromatin (**Figure S6A,E**).

Importantly, not all inactive regions are positioned close to the nucleolus, they can also arrange close to the nuclear lamina (Peric-Hupkes et al., 2010). Given that the nuclear lamina is another nuclear structure known to organize inactive chromatin, we explored whether lamina-associated DNA regions form preferential interactions. Indeed, we observe an increased number of DNA contacts between genomic regions associated with the nuclear lamina; however, these lamina-associated interactions generally occur between regions that are linearly close to each other rather than between chromosomes (**Figure S7C-D**). Although both compartments are associated with inactive chromatin, we do not observe a global relationship between genomic regions that are close to the nucleolar hub and regions that are associated with the nuclear lamina (**Figure S7E**, spearman $\rho = 0.01$). In contrast, regions that are closer to the nuclear speckle are highly depleted for nuclear lamina association (Spearman $\rho = -0.71$, **Figure S7F**).

(ii) Nuclear Speckle preference. Regions that are closer to nuclear speckles are strongly associated with high levels of active PolII transcription (Spearman $\rho = 0.88$, **Figure 6E, S6F, Methods**). Yet, we find that the transcriptional activity of an individual gene alone does not explain its distance to nuclear speckles because genomic DNA regions that are not transcribed, but are contained within highly transcribed gene-dense regions, tend to be closer to nuclear speckles (**Figure 6F**). Conversely, highly transcribed genes within otherwise inactive genomic regions tend to be farther from nuclear speckles (**Figure 6F**). These results demonstrate that the density of PolII transcription within a genomic neighborhood, rather than transcriptional activity of individual genes, defines proximity to nuclear speckles. This explains why only some of the specific actively transcribed genes previously studied organize close to nuclear speckles (Shopland et al., 2003) and may explain previous observations that actively transcribed gene-dense regions can “loop out” from the core chromosome territory (Branco and Pombo, 2006; Mahy et al., 2002).

Together, our results provide a global picture of the structural and transcriptional properties that define spatial positioning relative to nuclear bodies within the nucleus (**Figure 7**).

2.4 DISCUSSION

2.4.1 AN INTEGRATED MODEL OF HOW THE GENOME IS PACKAGED IN THE NUCLEUS

We described SPRITE, a method that enables genome-wide mapping of higher-order DNA interactions that occur simultaneously within the nucleus. SPRITE fills a critical gap among current methods by bridging the information derived from microscopy with the ability to generate high-resolution genome-wide maps. In doing so, SPRITE provided new biological insights into how DNA is packaged in the nucleus at multiple levels.

Molecular Insights. SPRITE provides a genome-wide molecular picture of all DNA regions that contact specific nuclear bodies. These results confirm previous observations made by microscopy for a limited number of DNA regions (Brown et al., 2008) and extends this molecular picture genome-wide and at higher resolution. This explains why some, but not other, active PolII-transcribed regions are organized around nuclear speckles and why some, but not other, inactive regions are organized around the nucleolus.

Spatial Insights. SPRITE provides a genome-wide spatial picture of how multiple DNA regions organize around the same nuclear body. This extends previous observations of a small number of specific DNA regions that can organize simultaneously around the same nuclear bodies (Brown et al., 2008; Strongin et al., 2014) to a global spatial picture where DNA organization around nuclear bodies form large spatial hubs of higher-order inter-chromosomal contacts.

Quantitative Global Insights. SPRITE provides a quantitative map of where DNA regions are organized relative to nuclear bodies, structural features, and other genomic regions. Specifically, we uncover quantitative preferences that spatially relate all genomic DNA regions to each nuclear body. Our results indicate that organization around nuclear bodies act as a dominant feature of global genome organization where: (i) a significant proportion of the genome preferentially organizes closer to one of these nuclear bodies and that (ii) organization around these bodies can lead to closer spatial organization of regions on different chromosomes. Because these spatial preferences correspond to PolII-transcriptional status, they may be dynamic between cell states.

Together, these results suggest an integrated and global picture of genome organization where individual genomic regions across chromosomes organize around nuclear bodies to shape the overall packaging of genomic DNA in a highly regulated and dynamic manner (**Figure 7**).

Although it remains unclear whether spatial organization around nuclear bodies directly impacts transcription or whether it is a consequence of PolII occupancy within a genomic region, spatial segregation may provide regulatory advantages by segregating factors into regions of high local concentration within the nucleus. For example, organization of DNA near nuclear speckles could increase the efficiency of post-transcriptional mRNA processing by concentrating splicing and processing factors, which are enriched in the nuclear speckles, near actively transcribed genes. Future work will be needed to determine how this spatial organization is established, its functional role, and its dynamics across cell states.

2.4.2 SPRITE PROVIDES A POWERFUL METHOD FOR STUDYING 3-DIMENSIONAL SPATIAL ORGANIZATION

This global model represents just one example of how SPRITE can be used to uncover new aspects of 3D genome organization in the nucleus. SPRITE provides several features that make it a powerful tool that can be applied to explore many open questions regarding organization and function in the nucleus.

Higher-order spatial interactions. SPRITE provides a genome-wide map of higher-order interactions that occur simultaneously in 3D spatial proximity within an individual nucleus. In contrast to proximity-ligation and microscopy methods, SPRITE is not limited in the number of simultaneous interactions that can be measured. This will enable exploration of additional higher-order interactions, such as spatial clusters of individual genes (e.g., olfactory receptor genes (Lomvardas et al., 2006)) and multiple enhancers that simultaneously interact with a promoter.

Global spatial maps. SPRITE accurately measures 3D spatial distances across a wide-range of nuclear distances. Because SPRITE does not rely on proximity ligation, it is not restricted to identifying interactions between molecules that are close enough to directly ligate. This ability to measure longer-range distances and the ability to measure crosslinked complexes of different sizes, enables quantitative and global reconstruction of 3D spatial distances across the nucleus.

Simultaneous RNA and DNA maps. SPRITE is not restricted to measuring DNA molecules, but can also simultaneously map RNA within crosslinked complexes. Because RNA demarcates various nuclear bodies, including the nucleolus and nuclear speckles, this allowed us to define specific DNA hubs as organizing around these bodies. SPRITE can be extended to include direct measurements of additional RNAs to enable direct mapping of genome structure relative to other RNA-demarcated structures (Rinn and Guttman, 2014) as well as for exploring enhancer-promoter interactions and their corresponding nascent transcription levels.

More generally, SPRITE represents a powerful new framework for spatial mapping because it provides genome-wide data that is highly analogous to microscopy and can be used to explore large numbers of high-resolution interactions that occur simultaneously in 3D space. Beyond its current applications, SPRITE can be extended in several ways. For example, SPRITE can be used to measure other spatial interactions beyond the nucleus, such as preferential associations of RNA in the cytoplasm (e.g., RNA phase separated bodies (Decker and Parker, 2012)). Furthermore, SPRITE can be extended to incorporate protein localization using pools of barcoded antibodies (Frei et al., 2016) to generate combinatorial and spatial maps of DNA, RNA, and/or protein. In addition, SPRITE can be extended to generate global single-cell maps by split-pool tagging of all molecules within individual cells (Ramani et al., 2017). These applications will enable exploration of previously inaccessible questions regarding the relationship between 3D genome structure and gene regulation within the nucleus and their dynamics across time.

2.5 REFERENCES

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. *4*, 780–788.

Brown, J.M., Green, J., Neves, R.P. Das, Wallace, H.A.C., Smith, A.J.H., Hughes, J., Gray, N., Taylor, S., Wood, W.G., Higgs, D.R., et al. (2008). Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. J. Cell Biol. *182*, 1083–1097.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., Shamim, M., et al. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proc. Natl. Acad. Sci. U. S. A. *113*, E4504-12.

Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harb. Perspect. Biol. *4*.

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

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

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science (80-.). *341*, 1237973–1237973.

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

Frei, A.P., Bava, F.-A., Zunder, E.R., Hsieh, E.W.Y., Chen, S.-Y., Nolan, G.P., and Gherardini, P.F. (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat. Methods *13*, 269–275.

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

Giorgetti, L., and Heard, E. (2016). Closing the loop: 3C versus DNA FISH. Genome Biol. *17*, 215.

Hu, Y., Plutz, M., and Belmont, A.S. (2010). Hsp70 gene association with nuclear speckles is Hsp70 promoter specific. J. Cell Biol. *191*, 711–719.

Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics *8*, 39.

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

Khanna, N., Hu, Y., and Belmont, A.S. (2014). HSP70 transgene directed motion to nuclear speckles facilitates heat shock activation. Curr. Biol. *24*, 1138–1144.

Li, W., Gong, K., Li, Q., Alber, F., and Zhou, X.J. (2015). Hi-Corrector: A fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. Bioinformatics *31*, 960–962.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal Interactions and Olfactory Receptor Choice. Cell *126*, 403–413.

Mahy, N.L., Perry, P.E., and Bickmore, W.A. (2002). Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. J. Cell Biol. *159*, 753–763.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and Van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Németh, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Péterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Längst, G. (2010). Initial genomics of the human nucleolus. PLoS Genet. *6*.

Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal body and histone locus body. Cold Spring Harb. Perspect. Biol. *2*.

O'Sullivan, J.M., Hendy, M.D., Pichugina, T., Wake, G.C., and Langowski, J. (2013). The statistical-mechanics of chromosome conformation capture. Nucl. (United States) *4*.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296–300.

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

Pederson, T. (2011). The nucleolus. Cold Spring Harb. Perspect. Biol. *3*, 1–15.

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

Pollock, C., and Huang, S. (2009). The perinucleolar compartment. J. Cell. Biochem. *107*, 189–193.

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

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat. Methods *14*, 263–266.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rinn, J.L., and Guttman, M. (2014). RNA and dynamic nuclear organization. Science (80-.). *345*, 1240–1241.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. Nat. Methods *12*, 323–325.

Shopland, L.S., Johnson, C. V., Byron, M., McNeil, J., and Lawrence, J.B. (2003). 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.

Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*, 959–965.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb. Perspect. Biol. *3*, 1–12.

Strongin, D.E., Groudine, M., and Politz, J.C.R. (2014). Nucleolar tethering mediates pairing between the *IgH* and *Myc* loci. Nucleus *5*, 474–481.

Suzuki, H., Kurihara, Y., Kanehisa, T., and Moriwaki, K. (1990). Variation in the distribution of silver-stained nucleolar organizing regions on the chromosomes of the wild mouse Mus musculus. Mol. Biol. Evol. *7*, 271–282.

Takei, Y., Shah, S., Harvey, S., Qi, L.S., and Cai, L. (2017). Multiplexed Dynamic Imaging of Genomic Loci by Combined CRISPR Imaging and DNA Sequential FISH. Biophys. J. *112*, 1773–1776.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X.J., Le Gros, M.A., et al. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc. Natl. Acad. Sci. *113*, E1663–E1672.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

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

2.6 MAIN FIGURE TITLES AND LEGENDS

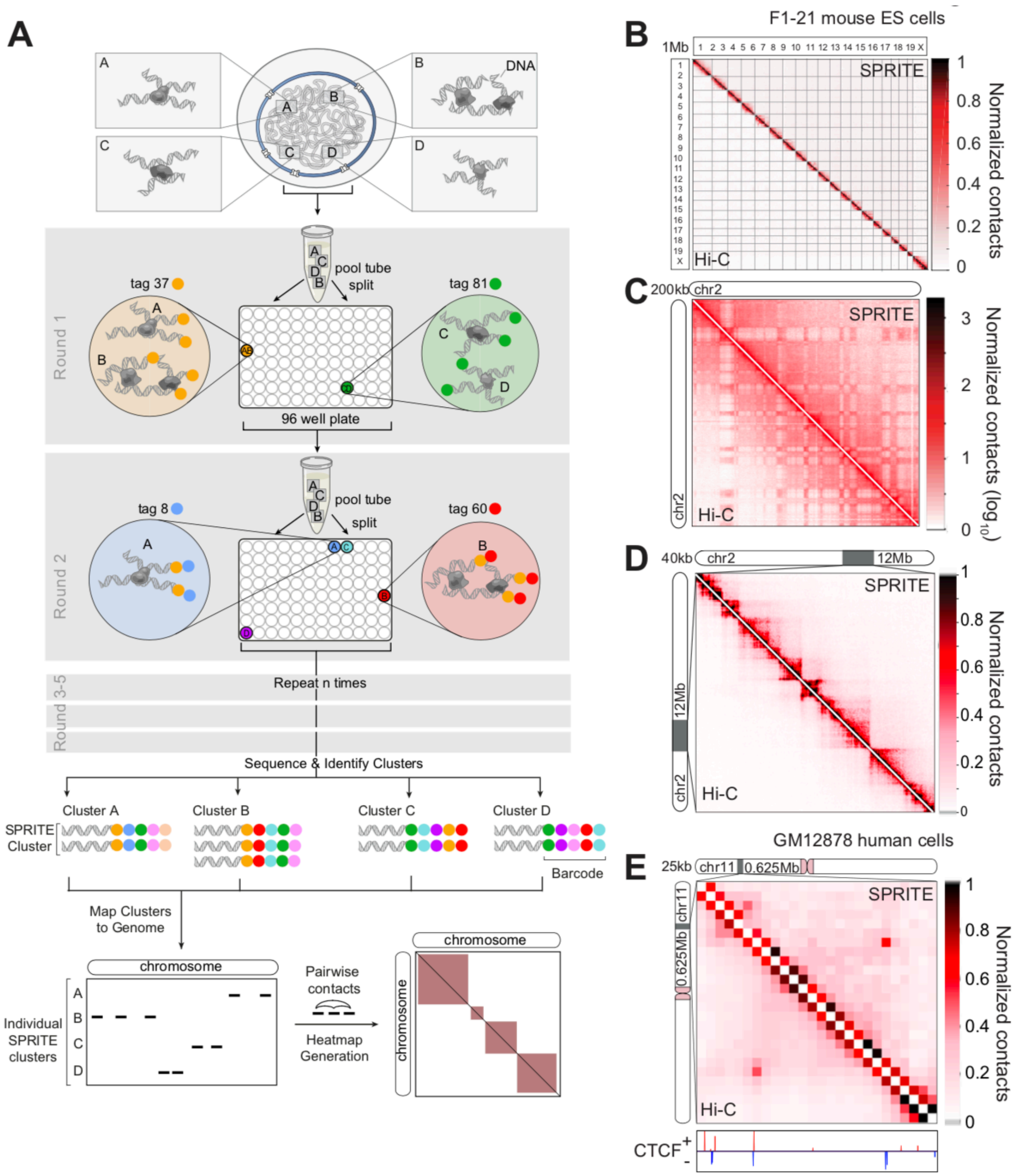


Figure 1. SPRITE accurately maps known genome structures across various resolutions. (A) Schematic of the SPRITE protocol. Crosslinked DNA is split into a 96-well plate and tagged with a unique sequence (colored circle) and then pooled into one tube. This split-and-pool process is

repeated with tags sequentially added. DNA is sequenced, and tags are matched to generate SPRITE clusters. (B-D) Comparison of SPRITE (upper diagonal) and Hi-C (Dixon et al., 2012) (lower diagonal) in mESCs (mESCs) (B) across all chromosomes at 1 Mb resolution, (C) on chromosome 2 at 200kb resolution (shown in log scale), and (D) within a 12 Mb region at 40 kb resolution. (E) Comparison of SPRITE and Hi-C (Rao et al., 2014) in human GM12878 cells within a 625 kb region at 25 kb resolution. CTCF binding (ENCODE) is colored based on motif orientation.

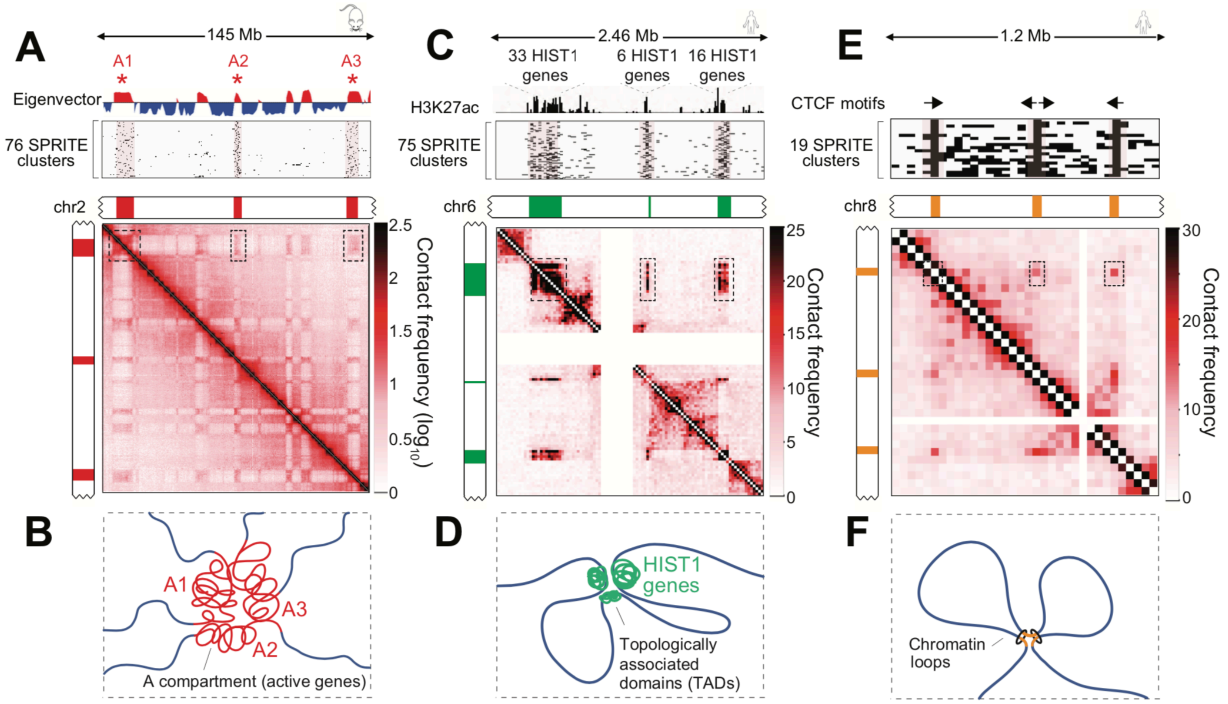


Figure 2. SPRITE identifies higher-order interactions that occur simultaneously. (A) Compartment eigenvector showing A (red) and B (blue) compartments on mouse chromosome 2 (Top). Individual SPRITE clusters (rows) containing reads mapping to at least 3 distinct regions (*) (Middle). Pairwise contact map at 200kb resolution (Bottom). (B) Schematic of multiple A compartment interactions. (C) H3K27ac ChIP-seq signal across a 2.46 Mb region on human chromosome 6 corresponding to 3 TADs containing 55 histone genes (Top). SPRITE clusters containing reads in all 3 TADs (Middle). Pairwise contact map at 25kb resolution (Bottom). (D) Schematic of higher-order interactions of *HIST1* genes (green). (E) CTCF motif orientations at 3 loop anchors on human chromosome 8 (Top). SPRITE clusters overlapping all 3 loop anchors (Middle). Pairwise contact map at 25kb resolution (Bottom). (F) Schematic of higher-order interactions between consecutive loop anchors.

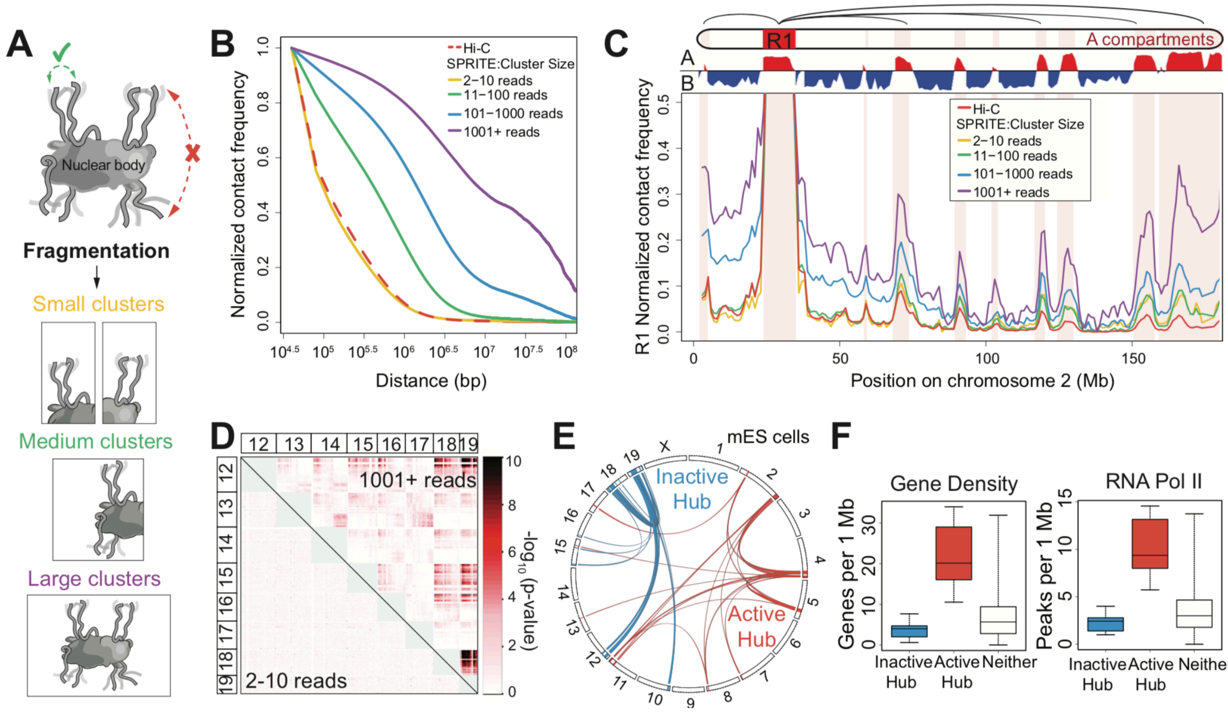


Figure 3. SPRITE identifies interactions across large genomic distances and across chromosomes. (A) Proximity-ligation methods identify interactions that are close enough to directly ligate (green check), but miss those that are too far apart to ligate (red x). SPRITE identifies all crosslinked interactions within a complex and measures different DNA cluster sizes generated by fragmentation of the nucleus. (B) Relationship between contact frequency observed by Hi-C and different SPRITE cluster sizes relative to linear genomic distance in mESCs. (C) Contact frequency between a specific region (R1: 25–34 Mb) and other regions on mouse chromosome 2 for different SPRITE cluster sizes and Hi-C. Red shaded areas represent A compartment. (D) Interaction *p*-values are shown for SPRITE clusters of size 2–10 reads (lower diagonal) and 1001+ reads (upper diagonal) between mouse chromosomes 12 through 19. (E) Circos diagram of two sets of significant inter-chromosomal interactions are shown in blue (inactive hub) and red (active hub). (F) Box plots of gene density (left) and RNA polymerase II occupancy (right) for regions in the inactive hub (blue), active hub (red), or neither hub (gray).

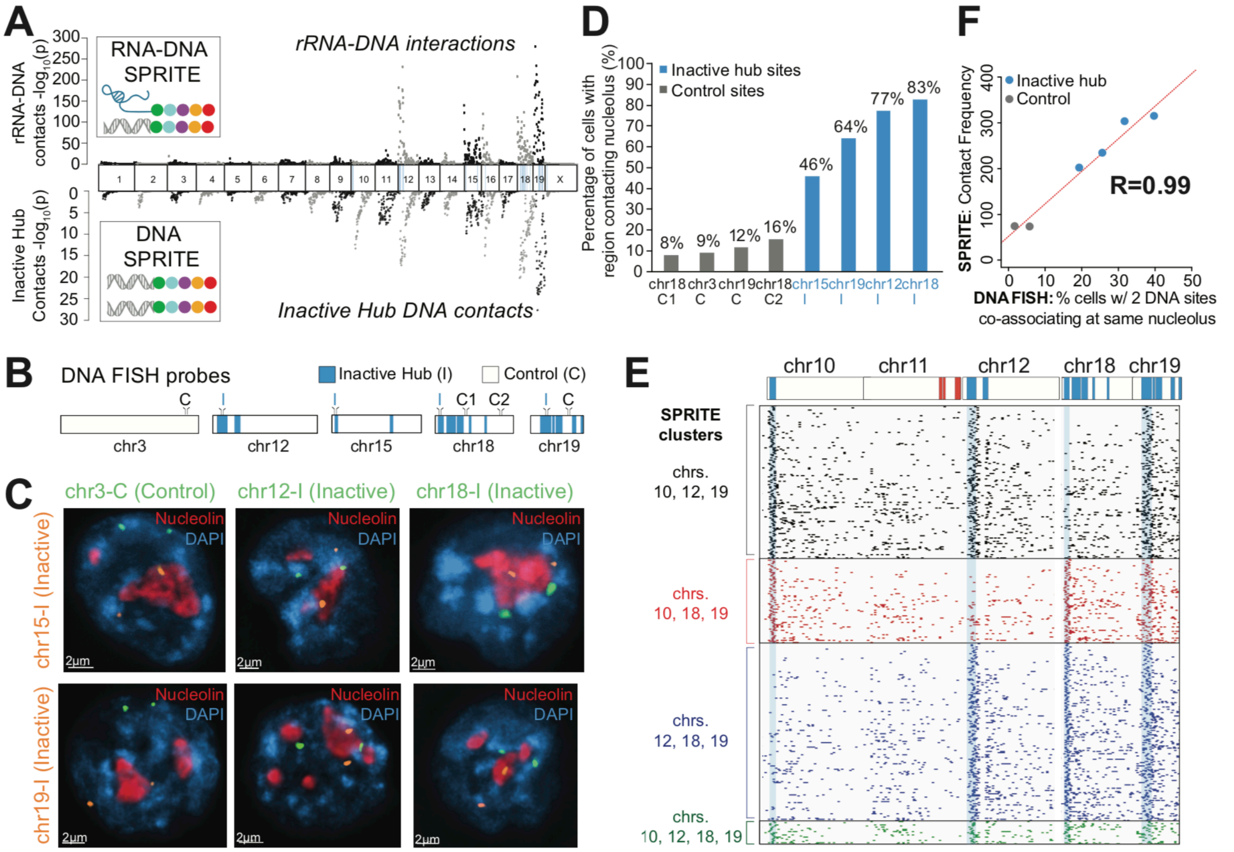


Figure 4. Genomic DNA in the inactive hub is organized around the nucleolus. (A) Ribosomal RNA (rRNA) localization across the mouse genome identified using RNA-DNA SPRITE (top) compared to the DNA SPRITE contact frequency with regions in the inactive hub (bottom). (B) Locations of probe regions used for DNA FISH experiments. (C) Example images from immunofluorescence for nucleolin (red) combined with DNA FISH for six different pairs of DNA FISH probes (orange and green) and DAPI (blue). (D) Percent of cells that overlap nucleolin (distance = 0 µm) for 8 different probe regions, 4 control regions (gray) and 4 inactive hub regions (blue) measured in 50–155 cells/region. (E) Example of individual SPRITE clusters (rows) containing reads from different combinations of inactive hub regions (blue) on chromosomes 10, 12, 18, and 19 binned at 1Mb resolution. (F) Comparison of co-association of two DNA sites on different chromosomes around the same nucleolus measured by microscopy (x-axis) and SPRITE co-association frequency (y-axis) for six pairs of regions (see details in Figure S4F, n = 50–64 cells).

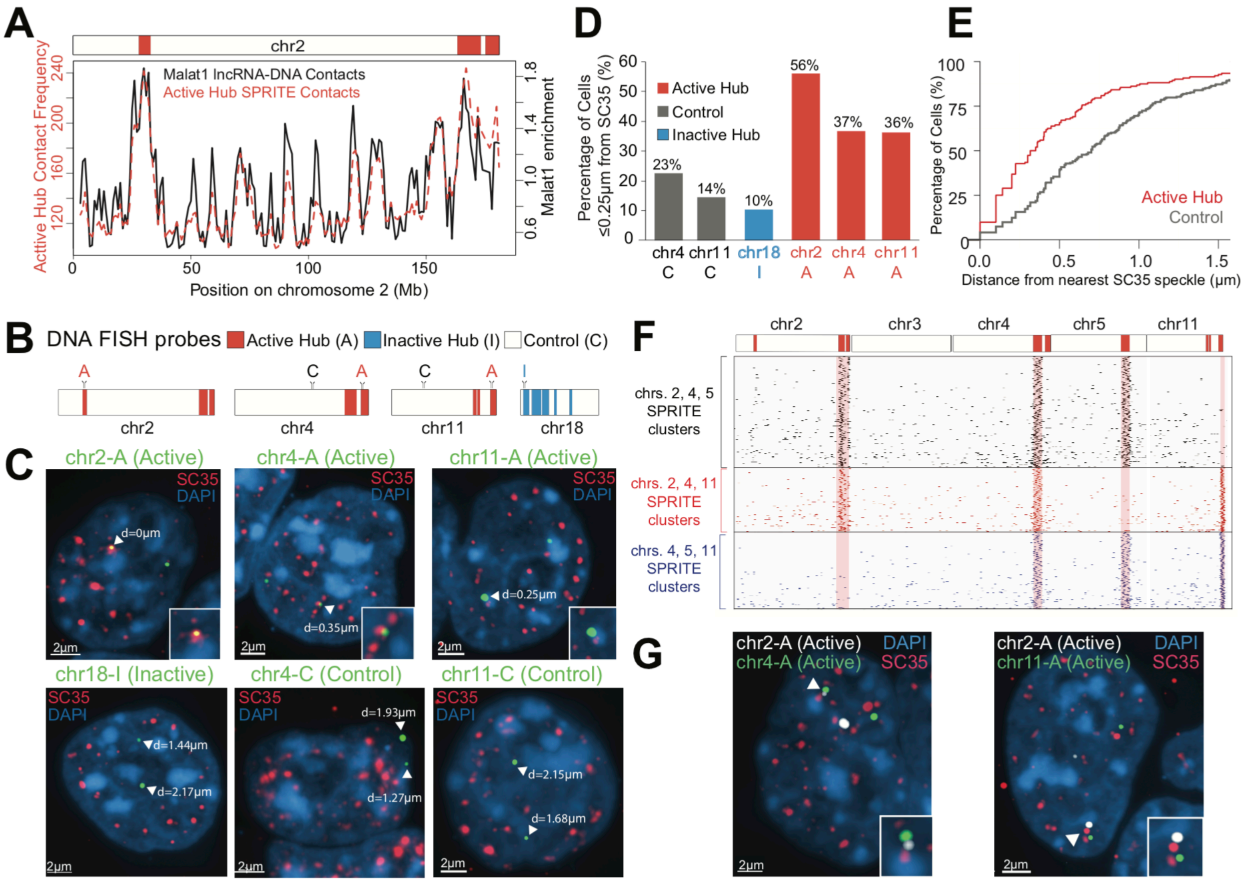


Figure 5: Genomic DNA in the active hub is organized around nuclear speckles. (A) Malat1 lncRNA localization (black) (Engreitz et al., 2014) compared to SPRITE contact frequency with regions in the active hub (red). (B) Locations of probe regions used for DNA FISH experiments. (C) Example images from immunofluorescence for SC35 (red) combined with DNA FISH for six DNA regions (green) and DAPI (blue) performed in formaldehyde fixed cells. Arrowhead: 3D distance to SC35 is noted. (D) Percentage of cells with at least 1 allele within 0.25 μm of SC35 (n = 41–90 cells). See Figure S5E for further quantitation. (E) Cumulative frequency of minimum 3D distance to SC35 for active hub (red) and control (gray) regions. (F) Example individual SPRITE clusters (rows) containing reads from different combinations of 3 active hub regions on chromosomes 2, 4, 5, and 11. (G) Images of 2 active hub regions on different chromosomes that are close to the same nuclear speckle.

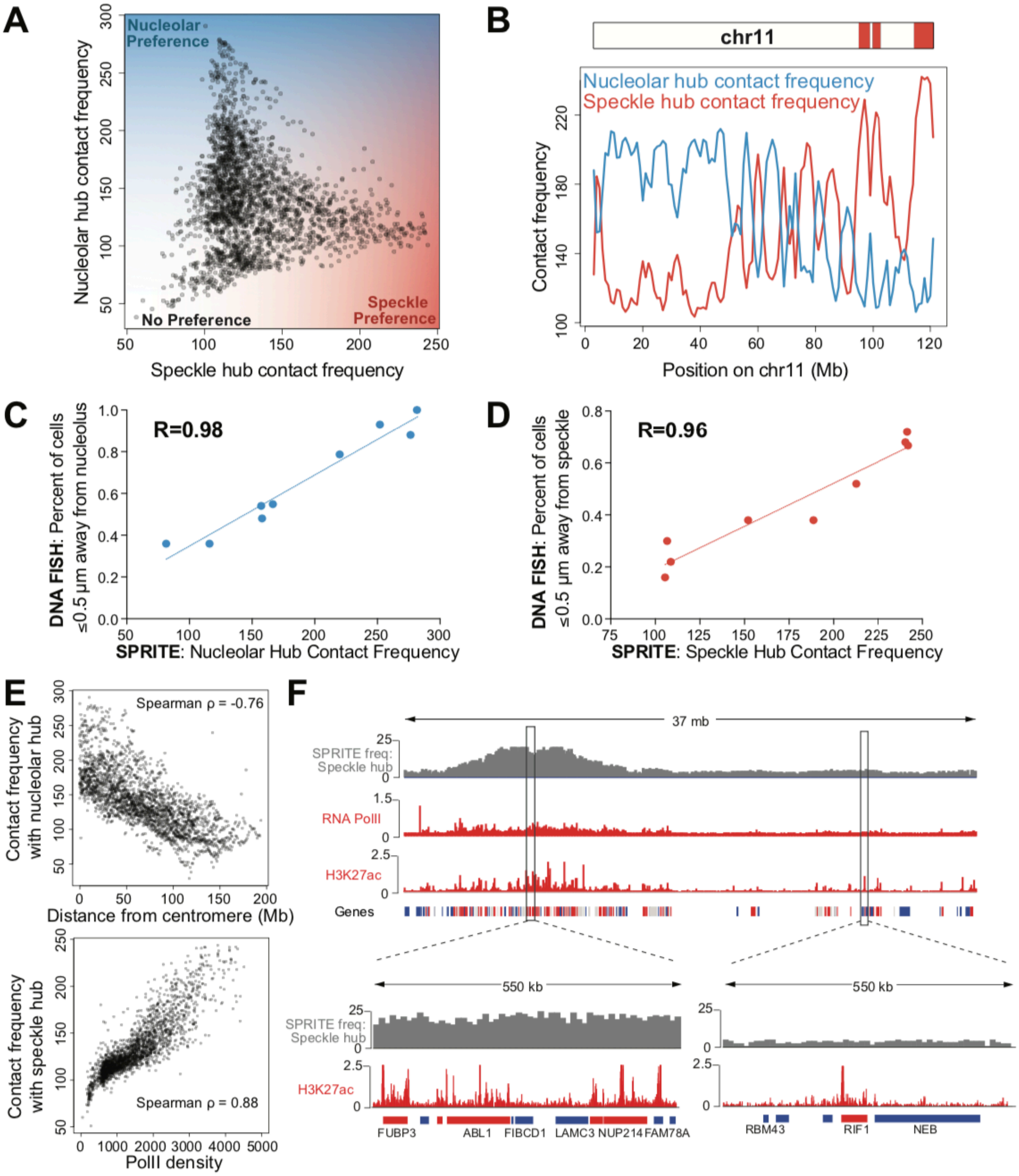


Figure 6. Preferential DNA distance to the nucleolus and nuclear speckles constrain overall genome organization. (A) SPRITE contact frequency to the nucleolar hub (y-axis) or speckle hub (x-axis) for each 1Mb genomic bin in mES cells. (B) SPRITE contact frequency to the nucleolar hub (blue) or speckle hub (red) across mouse chromosome 11. Red boxes represent active hub regions. (C) SPRITE contact frequency to the nucleolar hub (x-axis) compared to DNA FISH

contact frequency to the nucleolus as measured by microscopy across 50–155 cells/region (y-axis). (D) SPRITE contact frequency to the speckle hub (x-axis) compared to DNA FISH contact frequency to nuclear speckles as measured by microscopy (y-axis) across 50–51 cells/region. See Figure S6B-D for further details. (E) SPRITE contact frequencies with the nucleolar hub (y-axis) and centromere distance (x-axis) (top) and SPRITE contacts with the speckle hub (y-axis) and PolII density (x-axis, ENCODE) (bottom). (F) SPRITE contact frequency with the speckle hub compared to RNA PolII and H3K27ac signal (ENCODE) across a region on chromosome 2. Highly expressed (red, FPKM>10), moderately expressed (gray FPKM=2–10), or inactive (blue, FPKM=0–2) genes are indicated. Zoom-in: chr2:31.4–30.0 Mb (left) and chr2:51.7–52.3 Mb (right).

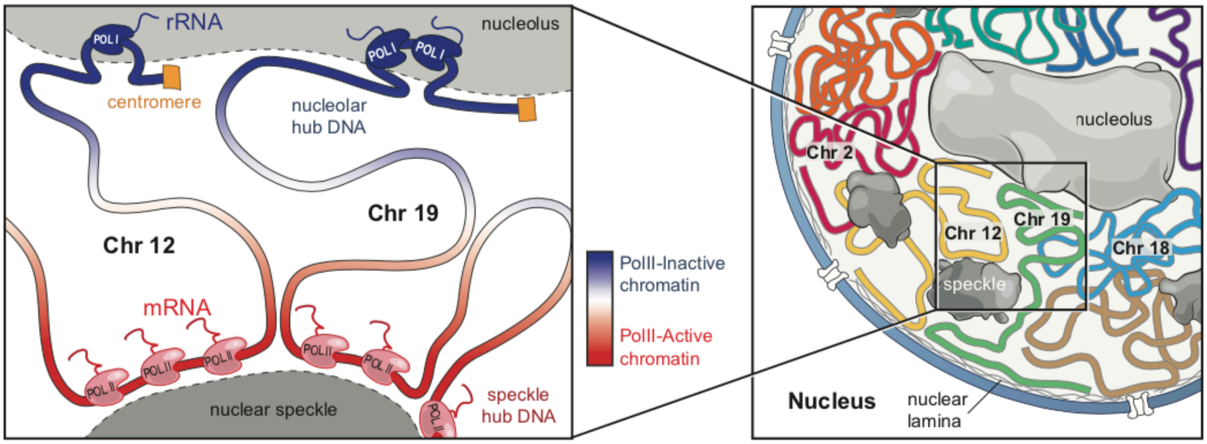
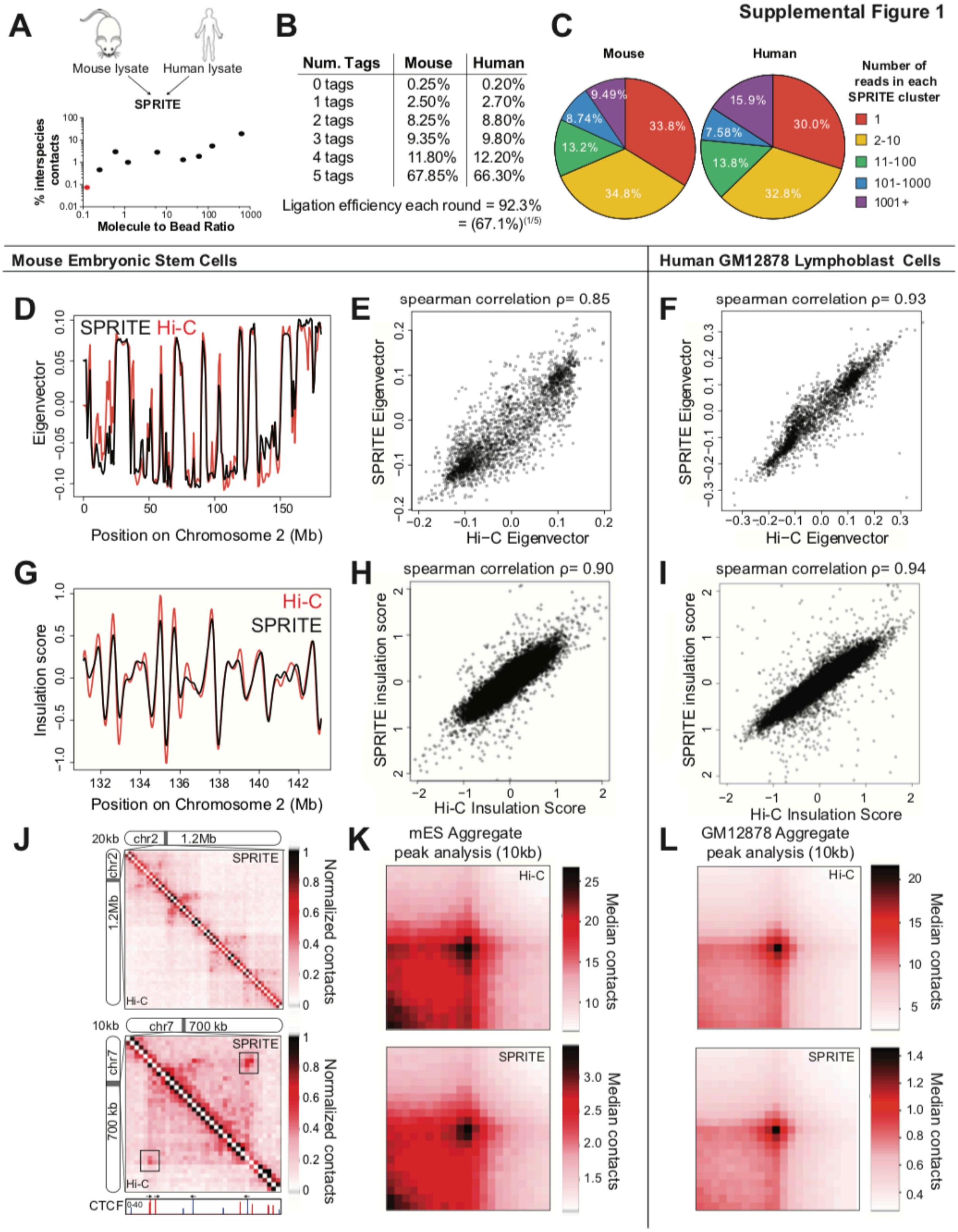


Figure 7. A global model for how nuclear bodies shape overall three-dimensional genome organization in the nucleus. Left Panel: DNA regions containing a high density of PolII associate with the nuclear speckle, while genomic regions linearly close to ribosomal DNA or centromeric regions associate with the nucleolus. This leads to co-association of multiple DNA regions around the same nuclear body to create spatial hubs of inter-chromosomal contacts. In addition to the genomic regions directly associating around these nuclear bodies, other DNA regions exhibit preferential organization, such that regions with higher levels of PolII density are closer to the nuclear speckle (red gradient) and regions with lower levels of PolII density are closer to the nucleolus (blue gradient). Right panel: These overall constraints act to shape the global layout of genomic DNA in the nucleus. DNA regions on the same chromosome tend to be closer to each other (colored lines). Yet, regions on different chromosomes containing similar properties organize around a nuclear body and can be closer to each other than to other regions contained on the same chromosome.

2.7 SUPPLEMENTAL FIGURE TITLES AND LEGENDS



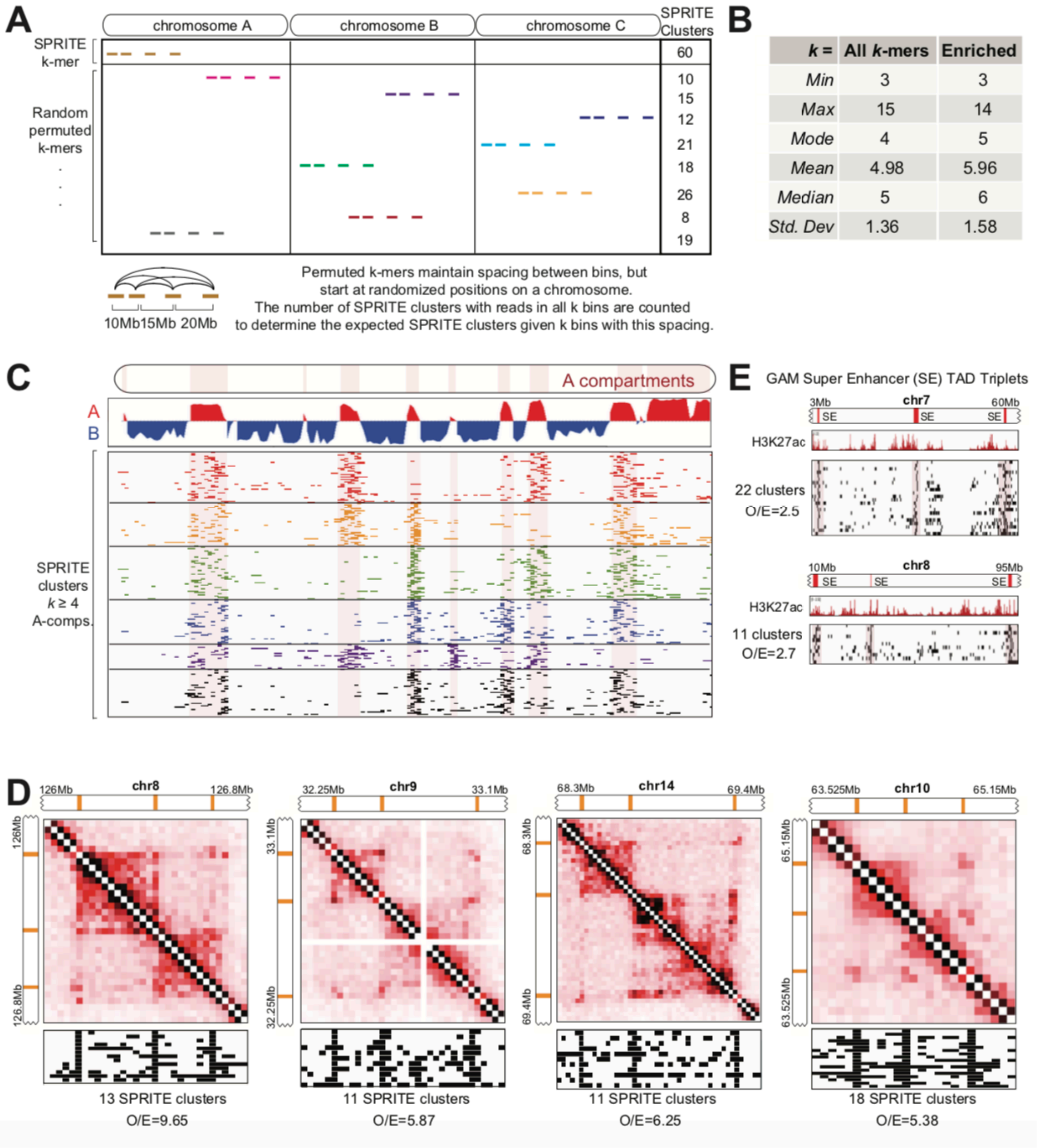
Num. Tags	Mouse	Human
0 tags	0.25%	0.20%
1 tags	2.50%	2.70%
2 tags	8.25%	8.80%
3 tags	9.35%	9.80%
4 tags	11.80%	12.20%
5 tags	67.85%	66.30%

Ligation efficiency each round = 92.3% = $(67.1\%)^{(1/5)}$

Figure S1. SPRITE accurately recapitulates genome structure measured by Hi-C, Related to Figure 1. (A) Estimate of the frequency of spurious contacts that are identified within individual SPRITE clusters. Noise was estimated by coupling crosslinked lysate from mouse ES cells and human HEK293T cells to NHS beads. The ratio of material per bead was coupled at different amounts (x-axis) and the percentage of clusters and pairwise contacts containing human and mouse reads were computed (inter-species contacts, y-axis). The red point indicates the molecule-to-bead ratio used for the SPRITE experiments in this paper. (B) Distribution of reads containing 0, 1, 2, 3, 4, 5 tags for mouse (left) and human (right) experiments. The estimate for ligation efficiency each round is determined by taking the 5th root of the fraction of reads with all 5 tags. (C) Distribution of SPRITE cluster sizes for mouse (left) and human (right) experiments. The percentage of reads was calculated for different SPRITE cluster sizes (1, 2–10, 11–100, 101–100 and over 1001) and reported as the percentage of total reads. Cluster size is defined as the number of reads with the same barcode. (D-L) SPRITE in mouse ES cells and human GM12878 lymphoblast cells was compared to Hi-C data generated in Dixon et al.(Dixon et al., 2012) and Rao et al.(Rao et al., 2014), respectively. (D) Compartment eigenvector for mouse chromosome 2 calculated using SPRITE (black) and Hi-C (red) contact maps from mouse ES cells binned at 1Mb resolution. Positive and negative values correspond to the A and B compartments, respectively. (E) Genome-wide correlation between compartment eigenvectors calculated using SPRITE (y-axis) and Hi-C (x-axis) contact maps from mouse ES cells binned at 1Mb resolution. (F) Genome-wide correlation between compartment eigenvectors calculated using SPRITE (y-axis) and Hi-C (x-axis) contact maps from human GM12878 cells binned at 1Mb resolution. (G) Insulation score profile for a region on mouse chromosome 2 (shown in Figure 1D) calculated using SPRITE (black) and Hi-C (red) contact maps from mouse ES cells binned at 40kb resolution. Local minima correspond to boundary regions. (H) Genome-wide correlation between insulation scores calculated using SPRITE (y-axis) and Hi-C (x-axis) contact maps from mouse ES cells binned at 40kb. (I) Genome-wide correlation between insulation scores calculated using SPRITE (y-axis) and Hi-C (x-axis) contact maps from human GM12878 cells binned at 40kb. (J) Examples of SPRITE and Hi-C contact maps binned at 20kb resolution (top) and 10kb resolution (bottom) showing chromatin loop interactions. CTCF ChIP-seq peaks are shown according to their positive (red) or negative (blue) motif orientation. (K) Aggregate peak analysis heatmaps for Hi-C (top) and SPRITE (bottom) in mouse ES cells binned at 10kb resolution. 1493 loops obtained from Rao

et al.(Rao et al., 2014) were used in this analysis. Heatmaps show the median contact map values for each pair of 10kb bins in regions +/- 200kb of the loops. (L) Aggregate peak analysis heatmaps for Hi-C (top) and SPRITE (bottom) in human GM12878 cells binned at 10kb resolution. 5789 loops obtained from Rao et al.(Rao et al., 2014) were used in this analysis. Heatmaps show the median contact map values for each pair of 10kb bins in regions +/- 200kb of the loops.

Supplemental Figure 2



k =	All k-mers	Enriched
Min	3	3
Max	15	14
Mode	4	5
Mean	4.98	5.96
Median	5	6
Std. Dev	1.36	1.58

Figure S2. SPRITE measures higher-order interactions across known genome structures, Related to Figure 2. (A) Cartoon representation of the method for identifying enriched higher-order *k*-mer frequencies. Given an observed *k*-mer, we randomly permute *k*-mers by randomizing the location of the genomic positions within the *k*-mer while preserving the spacing between its

reads. The observed frequency is then normalized by the expected frequency derived from the randomly permuted *k*-mers. See Methods for further details. (B) Statistics of the size distribution observed in all enumerated (left) and enriched (right) *k*-mers in mES cells at 1Mb resolution. Enriched *k*-kmers are defined as those that are observed in at least 5 independent SPRITE clusters, occur >4-fold more frequently than the average of the permuted *k*-mers, and are observed at a frequency that is more than 90% of the permuted structures. (C) Example SPRITE clusters spanning four A compartment regions on mouse chromosome 2. The compartment eigenvector showing A (red) and B (blue) compartments is shown on top. Rows correspond to individual SPRITE clusters and colored lines denote 1Mb bins with at least 1 read. Each colored group represents a different combination of four A compartment regions interacting across several SPRITE clusters (max cluster size n=20 1Mb bins). Red shaded area demarcates the A compartment regions. Enrichments compared to randomly permuted *k*-mers spanning 4 or more A compartment regions in mES cells are listed in **Table S1C**. (D) Four examples of 3-way interactions between 3 loop anchors in human GM12878 cells. SPRITE contact maps are shown at 25kb resolution above the individual SPRITE clusters that contain all three loop anchors (max cluster size n=100 25kb bins) (E) Examples of enriched SPRITE clusters corresponding to super-enhancer TAD triplets previously identified using GAM in mES cells(Beagrie et al., 2017) (max cluster size n=100 1Mb bins).

Supplemental Figure 3

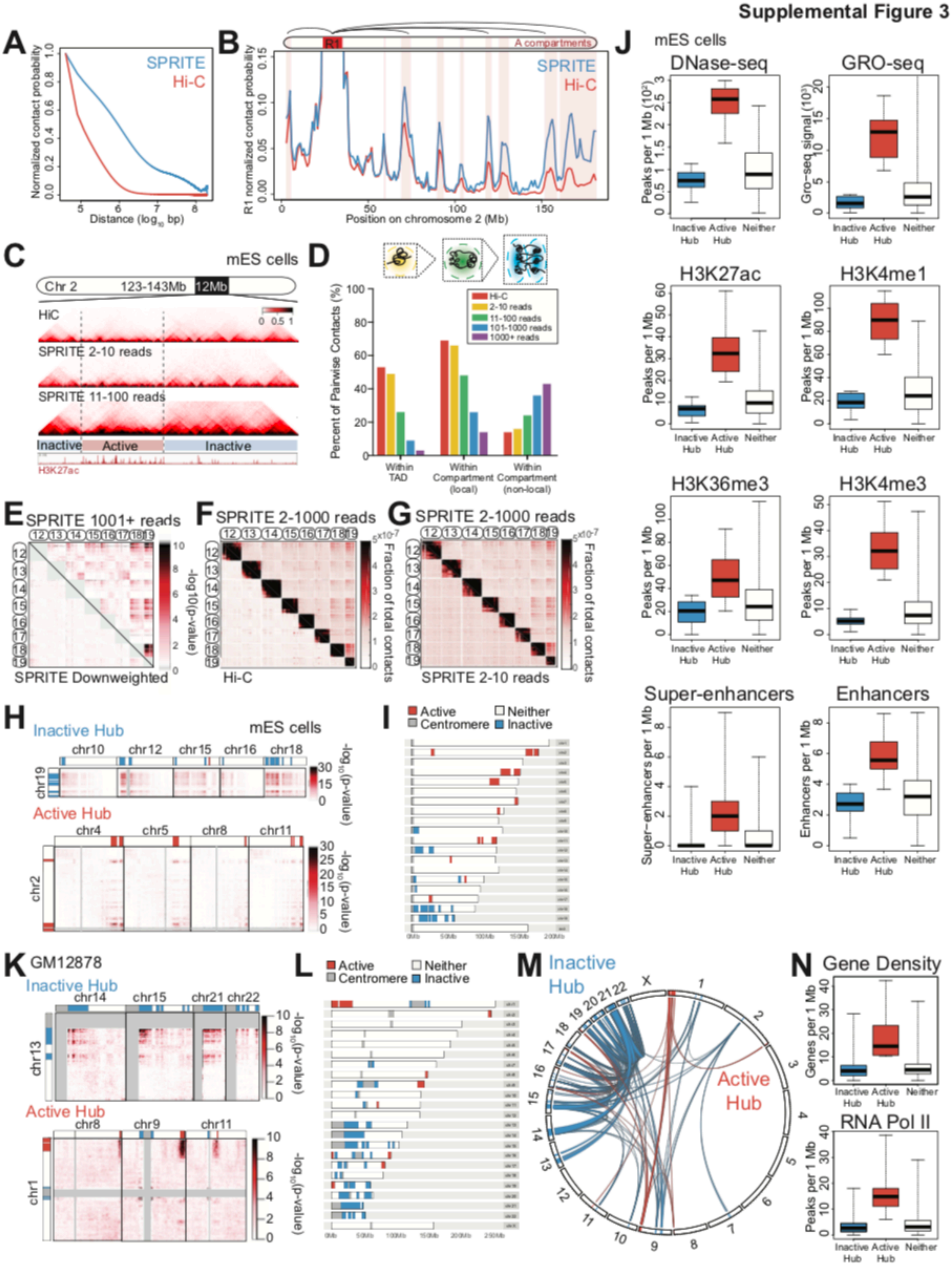


Figure S3. SPRITE identifies long-range intra-chromosomal interactions and hubs of inter-chromosomal interactions, Related to Figure 3. (A) Relationship between contact frequency as a function of linear genomic distance observed in SPRITE (blue) and Hi-C(Dixon et al., 2012) (red) for mouse ES cells. (B) Contact frequency between a specific A (red) compartment region (R1: 25–34 Mb) and all other regions on mouse chromosome 2. Pairwise contact frequencies are shown for SPRITE (blue) and Hi-C (red). Red shaded areas represent A compartment regions. (C) Comparison of Hi-C contact map (top) with SPRITE contact maps based on SPRITE clusters with 2 to 10 reads (middle) and 11 to 100 reads (bottom) in a 12Mb region on mouse chromosome 2 binned at 40kb resolution. Contact map values are shown as the fraction of total contacts. H3K27ac ChIP-seq signal (ENCODE) from mouse ES cells is shown below. (D) Different SPRITE cluster sizes capture different scales of interactions. The percentage of pairwise contacts are quantified within structures of different sizes in the nucleus—TADs, local compartment regions, and non-local compartment regions for SPRITE clusters containing 2–10 reads (yellow), 11–100 reads (green), 101–1000 reads (blue) and 1001 or more reads (purple) compared against Hi-C (red). (E) Inter-chromosomal contacts between chromosomes 12 through 19 in mES cells binned at 1Mb resolution. Interaction p-values are shown for SPRITE clusters with 1001+ reads (upper diagonal) and for SPRITE contacts that have been down-weighted by cluster size (lower diagonal). *p*-values were calculated based on inter-chromosomal interaction frequencies and are shown in units of -$\log_{10}$ (*p*-value). (F) Comparison of SPRITE (upper diagonal) and Hi-C (lower diagonal) inter-chromosomal contacts between chromosomes 12 through 19 in mES cells binned at 1Mb resolution. SPRITE contacts are based on clusters with 2 to 1000 reads. Values shown are the fraction of total contacts. (G) Comparison of inter-chromosomal contacts between chromosomes 12 through 19 based on SPRITE clusters with 2 to 1000 reads (upper diagonal) or 2 to 10 reads (lower diagonal) in mES cells binned at 1Mb resolution. Values shown are the fraction of total contacts. (H, K) Examples of inter-chromosomal interactions that comprise the Inactive Hub (top, blue) and Active Hub (bottom, red) in mES and GM12878 cells, respectively. Inactive and active hub regions are colored on the chromosome ideogram as blue and red, respectively. Interaction p-values were calculated using unweighted contact frequencies from SPRITE clusters with 2 to 1000 reads and are shown in units of –$\log_{10}$ (*p*-value). (I, L) Ideogram showing inactive hub (blue) and active hub (regions) on each mouse and human chromosome, respectively. Centromere regions are demarcated in gray. (J) Box plots showing the distribution of different features associated with

high RNA polymerase II transcription (DNase-seq from ENCODE, GRO-seq (Jonkers et al., 2014)), active chromatin marks (H3K27ac, H3K4m1, H3K36me3, H3K4me3 from ENCODE), and regulatory features (super-enhancers, enhancers(Whyte et al., 2013)) in mES cells. (M) Circos diagram of significant inter-chromosomal interactions in GM12878 cells. Interactions are identified from unweighted inter-chromosomal contact maps in SPRITE clusters containing 2 to 1000 reads. Interactions with *p*-values less than 10^{-8} are shown. Two distinct networks of non-interacting connections are shown in blue (inactive hub) and red (active hub). (N) Box plots showing the distribution of gene density (top) and RNA polymerase II (Pol II) occupancy (bottom) for regions in the inactive hub, nucleolar hub, or neither hub in GM12878 cells. Gene density is calculated as genes per 1Mb and Pol II occupancy is calculated as the number of ChIP-seq peaks per 1Mb (from ENCODE).

Supplemental Figure 4

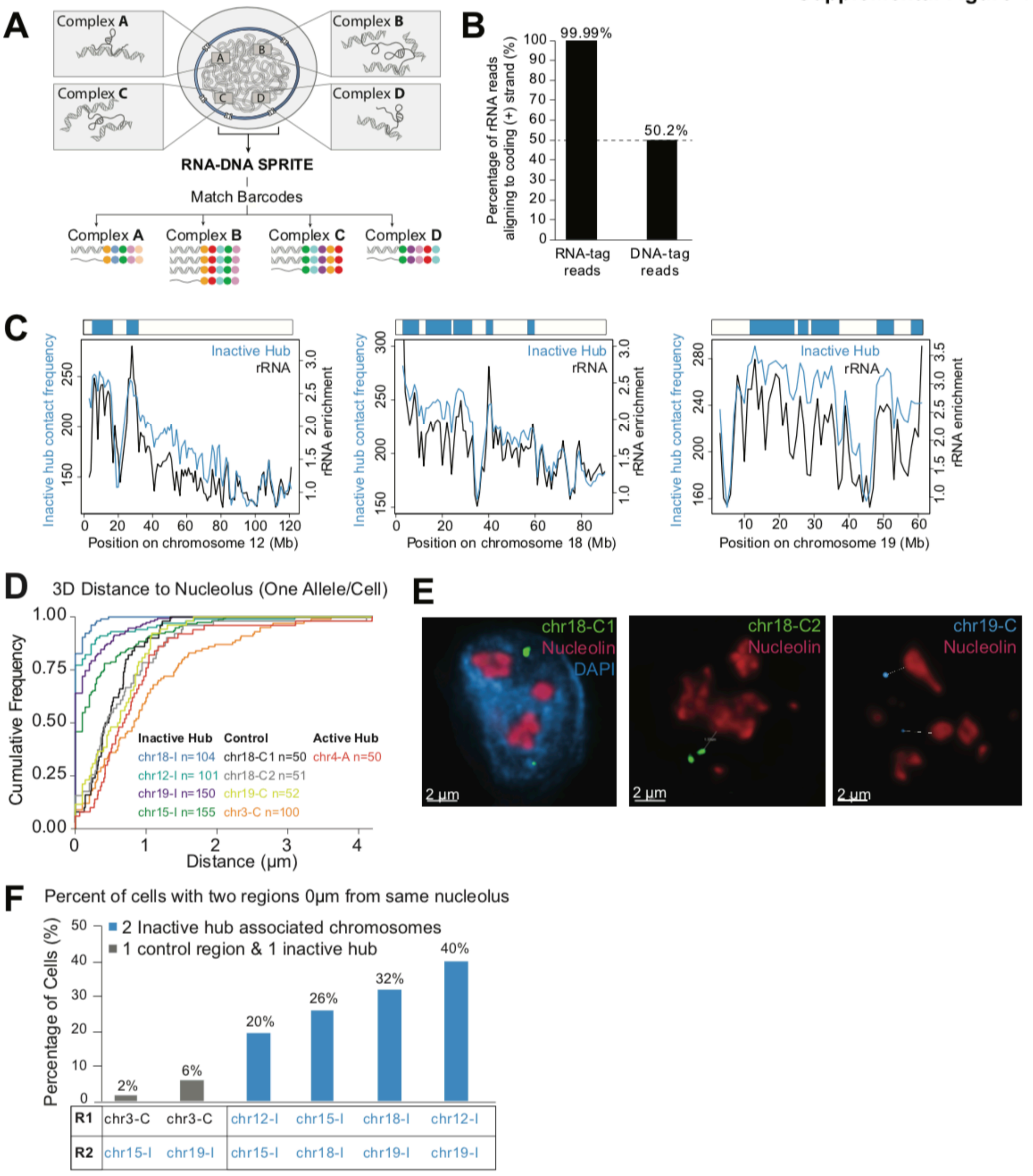


Figure S4. Genomic DNA in the inactive hub is organized around the nucleolus, Related to Figure 4. (A) A schematic of the RNA-DNA SPRITE procedure. SPRITE was adapted to simultaneously tagged RNA and DNA molecules in crosslinked complexes during each round of split-pool tagging. (B) Most reads in the RNA-DNA SPRITE maps containing an RNA-specific

tag aligned to the transcribed (+) strand of ribosomal RNA, in contrast to reads containing a DNA-specific tag which uniformly aligned with either of the two strands. (C) Comparison of ribosomal RNA localization from RNA-DNA SPRITE maps (black) and average inter-chromosomal contact frequency with regions in the inactive hub (blue) binned at 1Mb resolution for mouse chromosomes 12, 18, and 19. Inactive hub regions are shown in blue at the top of each plot. (D) Cumulative distributions for the 3D distances of 8 probe regions and the most proximal nucleolus (only the distance of the nearest allele per cell is shown). Four probe regions in the inactive hub are shown in blue, green, teal, and purple. Four control probe regions not in the inactive hub are shown in black, gray, orange, and yellow. One active hub probe region is shown in red. Number of cells measured (n) is listed for each probe. (E) Example images from immunofluorescence for nucleolin (red) combined with DNA FISH for three different DNA FISH control region probes (chr18-C1, chr18-C2, and chr19-C). (F) Comparison of the frequency of co-association of two DNA regions on different chromosomes at the same nucleolus measured by DNA FISH. Percentage of cells with two regions at the same nucleolus is reported if at least one allele for six different pairs of DNA FISH probes are 0 μm from the same nucleolus (n = 50 to 64 cells). Pairs of probes with one control region and one inactive hub region are shown in gray, and pairs of two inactive hub region probes are shown in blue.

Supplemental Figure 5

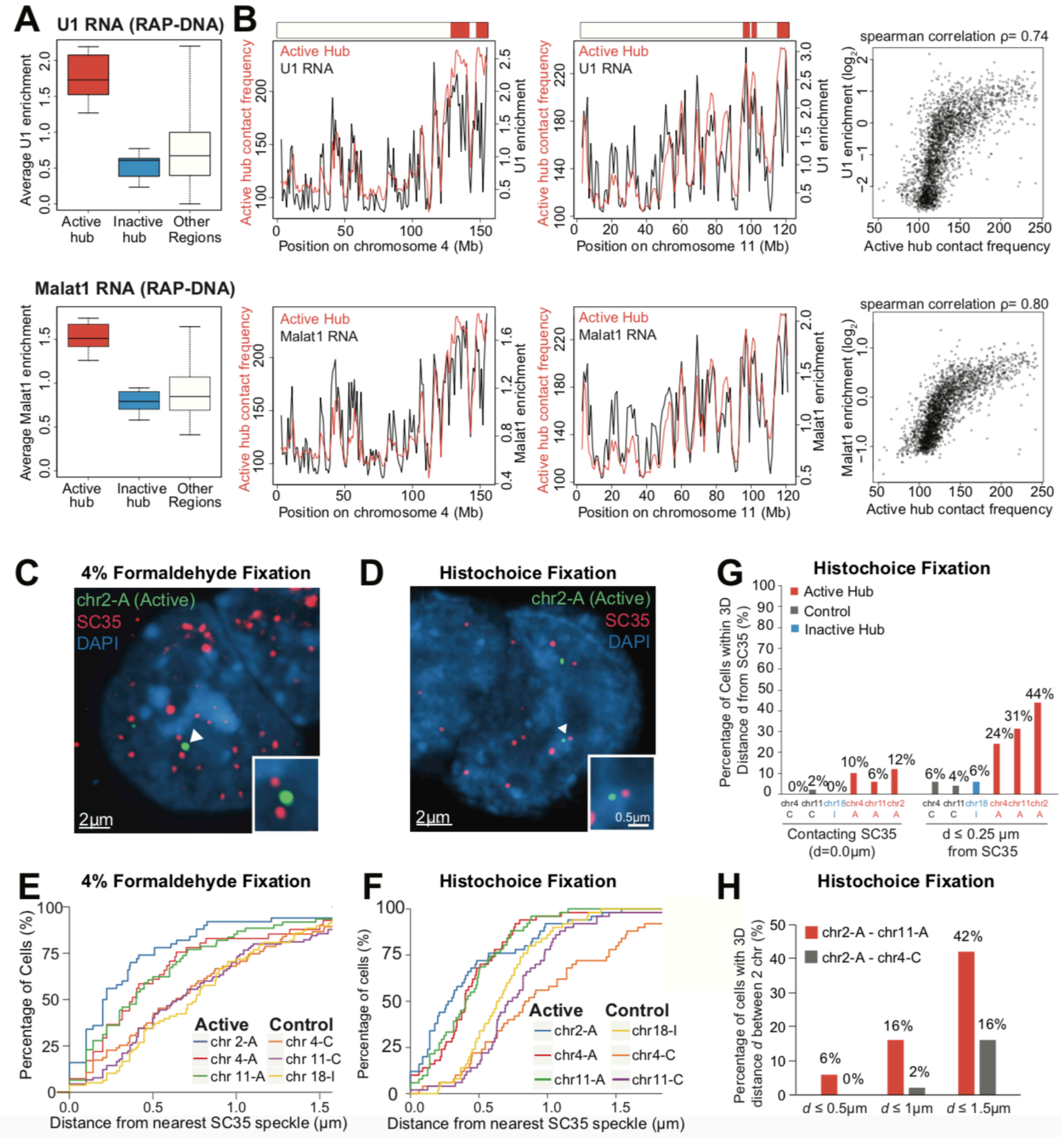
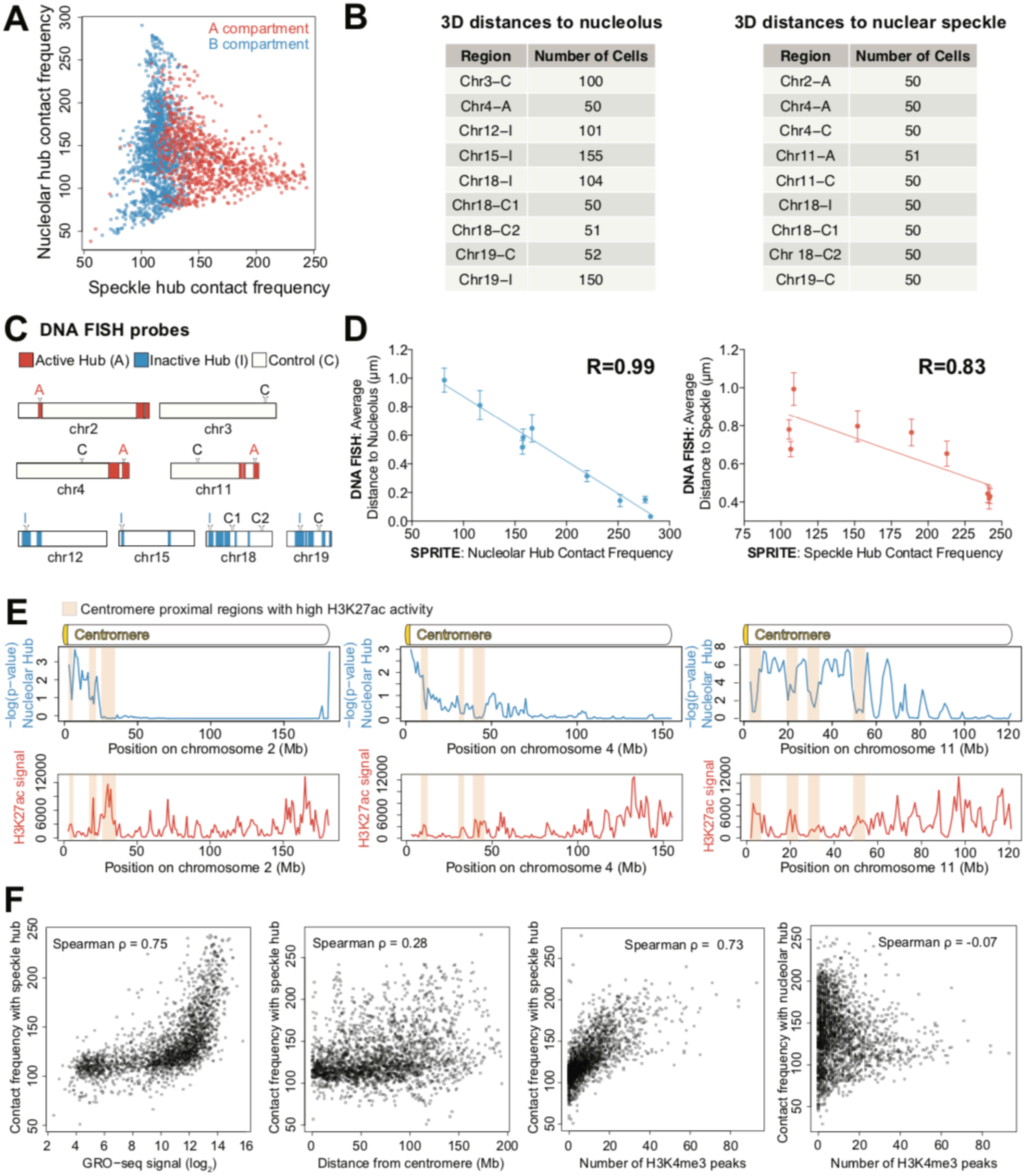


Figure S5. Genomic DNA regions in the active hub are organized around the nuclear speckles, Related to Figure 5. (A) Box plots showing the distribution of U1 RNA enrichment to genomic DNA (top) and Malat1 RNA enrichment to genomic DNA (bottom) from Engreitz et al.(Engreitz et al., 2014) for regions in the active hub, inactive hub, or neither. (B) Comparison of U1 or Malat1 RNA-DNA (black) enrichment with average inter-chromosomal contact frequency with regions in the active hub binned at 1Mb resolution for mouse chromosomes 4 (left) and 11

(middle). Genome-wide correlation plots are shown on the right. (C, D) Example images from immunofluorescence for SC35 (red) combined with DNA FISH (green) for an active hub probe region on chromosome 2 (chr2-A) in cells fixed using 4% formaldehyde and histochoice prior to DNA FISH/IF, respectively. Inset images show the region around the arrowhead. DAPI (blue) is used to demarcate the nucleus. (E, F) Cumulative distributions for the 3D distances of 6 probe regions and the most proximal speckle (only the distance of the nearest allele per cell is shown) in cells fixed using 4% formaldehyde (n = 50 to 155 cells) and histochoice (n = 41 to 90 cells) prior to DNA FISH/IF, respectively. Three probe regions in the active hub are shown in blue, red, and green. Three control probe regions not in the active hub are shown in yellow, orange, and purple. (G) Comparison of the 3D distance to the most proximal nuclear speckle for 6 different probe regions, including 3 control regions (gray/blue) and 3 active hub regions (red). Proximity to nuclear speckles is calculated for each probe region as the percent of cells with at least 1 allele within 0 μm (left) and 0.25 μm (right) of an SC35 speckle for cells crosslinked using histochoice prior to DNA FISH/IF. (H) Comparison of the percentage of cells where two probes are within a given distance (0.5 μm, 1.0 μm and 1.5 μm) for two active hub probes (red) and an active/control pair (gray) measured in cells fixed using histochoice prior to DNA FISH/IF. The large distance between the two DNA regions likely reflects the fact that the individual DNA sites are organized around nuclear speckles rather than directly contacting each other.

Supplemental Figure 6



3D distances to nucleolus

Region	Number of Cells
Chr3–C	100
Chr4–A	50
Chr12–I	101
Chr15–I	155
Chr18–I	104
Chr18–C1	50
Chr18–C2	51
Chr19–C	52
Chr19–I	150

3D distances to nuclear speckle

Region	Number of Cells
Chr2–A	50
Chr4–A	50
Chr4–C	50
Chr11–A	51
Chr11–C	50
Chr18–I	50
Chr18–C1	50
Chr 18–C2	50
Chr19–C	50

Figure S6. Preferential organization relative to the nucleolus and nuclear speckles shapes the overall organization of genomic DNA in the nucleus, Related to Figure 6. (A) Genome-wide comparison of average SPRITE contact frequency to regions in the nucleolar hub (y-axis) or

speckle hub (x-axis). Each dot corresponds to a 1Mb bin. Regions within the A and B compartment are colored red and blue, respectively. (B) Number of cells imaged and measured for 3D distances to the nucleolus and nuclear speckle reported in Figures 6C-D and S6D. (C) A schematic representation showing the location of the various DNA FISH probes used in Figures 6C-D and S6D. **(**D) Comparison of DNA FISH average 3D distances from nucleoli (left) or nuclear speckles (right) (y-axis) with average SPRITE inter-chromosomal contact frequencies with the nucleolar and nuclear speckle hub regions (x-axis), respectively. Average 3D distances are measured by DNA FISH and IF experiments using nucleolin and SC35 as markers for nucleoli and nuclear speckles, respectively, from cells fixed with histochoice prior to IF/DNA FISH. The closest allele in each cell is used to measure the 3D distance to each nuclear body. Data points and error bars represent mean ± SEM. See Figure S6B-C for cell numbers and probes used for quantitation of IF/DNA FISH. (E) Examples of decreased nucleolar hub interactions in centromere-proximal regions with high levels of transcriptional activity (light red box). Although there is a general trend of centromere-proximal regions to interact with the nucleolar hub, highly active regions near centromeres are depleted in nucleolar hub interactions (blue). H3K27ac ChIP-seq (ENCODE) signal (red) is shown below. (F) Comparison of average inter-chromosomal contact frequencies with the speckle hub and nascent transcription levels (GRO-seq signal (Jonkers et al., 2014)), linear distance from centromeres, and number of H3K4me3 peaks (ENCODE). Comparison of average SPRITE inter-chromosomal contact frequencies with the nucleolar hub and H3K4me3 peaks.

Supplemental Figure 7

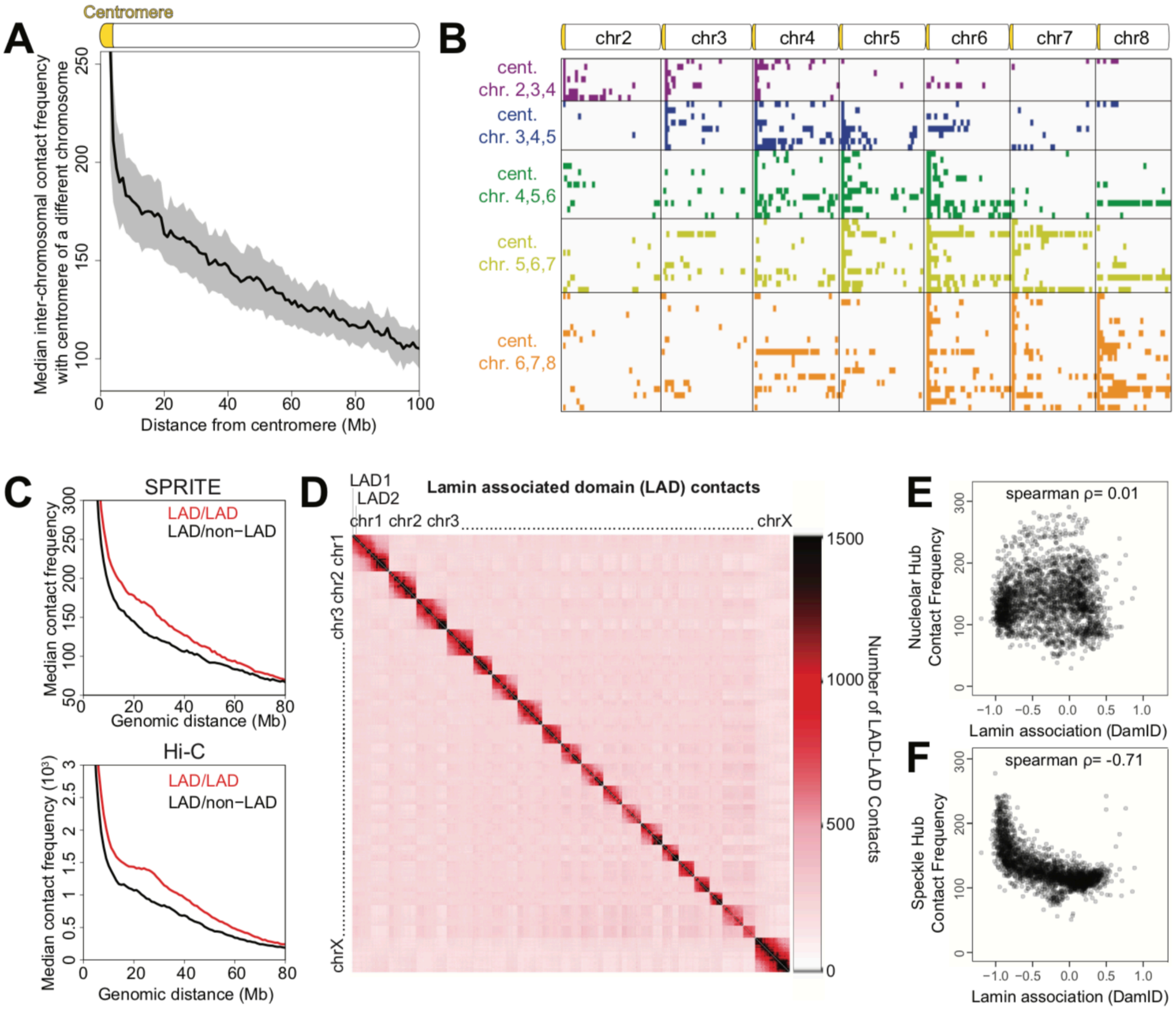


Figure S7. Model for how nuclear bodies shape global three-dimensional genome organization in the nucleus, Related to Figure 7. (A) Median inter-chromosomal contact frequency of the most proximal region mapped to the centromere in the reference genome of each chromosome with all genomic sites on different chromosomes at different genomic distances in mES cells. (B) Example SPRITE clusters containing multi-way (k≥3) interactions between centromere-proximal regions on different chromosomes in mouse ES cells. Rows show individual SPRITE clusters and each line denotes a 1Mb bin with at least 1 read within these clusters (max cluster size n=500 1Mb bins). (C) Median contact frequency between two Lamin Associated Domains (LADs) regions (or one LAD and one non-LAD region) as a function of linear genomic distance on the same chromosome measured by SPRITE (top) and Hi-C (bottom). Lamina

association data comes from DamID signal generated by Meuleman et al.(Meuleman et al., 2013). (D) ICE-normalized inter-chromosomal contact map of all LADs on all chromosomes in mES cells. Each bin represents a LAD (LAD1, LAD2, etc.). (E) Comparison of lamina association signal (DamID) and nucleolar hub contact frequencies in mES cells. (F) Comparison between lamina association signal and speckle hub contact frequencies in mES cells.

2.8 METHODS

2.8.1 EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Culture and Lines Used in Analysis

Mouse ES cell lines were cultured in serum-free 2i/LIF medium and maintained at an exponential growth phase as previously described (Engreitz et al., 2014). SPRITE DNA-DNA maps were generated in female ES cells (F1 2-1 line, provided by K. Plath), a F1 hybrid wild-type mouse ES cell line derived from a 129 × *castaneous* cross. SPRITE RNA-DNA maps were generated in the pSM33 ES cell line (provided by K. Plath), a male ES cell line derived from the V6.5 ES cell line, which expresses Xist from the endogenous locus under the transcriptional control of a Tet-inducible promoter and the Tet transactivator (M2rtTA) from the Rosa26 locus. We induced Xist expression in these cells using doxycycline (Sigma, D9891) at a final concentration of 2 μg/mL for 6–24 hours.

Human GM12878 cells, a female lymphoblastoid cell line obtained from Coriell Cell Repositories, were cultured in RPMI 1640 (Gibco, Life Technologies), 2 mM L-glutamine, 15% fetal bovine serum (FBS; Seradigm), and 1X penicillin-streptomycin and maintained at 37°C under 5% CO_2. Cells were seeded every 3–4 days at 200,000 cell/mL in T25 flasks, maintained at an exponential growth phase, and passaged or harvested before reaching 1,000,000 cell/mL.

HEK293T, a female human embryonic kidney cell line transformed with the SV40 large T antigen was obtained from ATCC and 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.

2.8.2 METHOD DETAILS

Split-Pool Recognition of Interactions by Tag Extension (SPRITE)

Crosslinking. Cells were crosslinked in a single-cell suspension to ensure that we obtain individual crosslinked nuclei rather than crosslinked colonies of cells. GM12878 lymphoblast cells, which are grown in suspension, were pelleted and media was removed prior to crosslinking. Mouse ES cells, which are adherent, were trypsinized to remove from plates prior to crosslinking in suspension. Specifically, 5 mL of Trypsin Versene Phosphate (TVP) (1 mM EDTA, 0.025% Trypsin, 1% Sigma Chicken Serum; pre-warmed at 37°C) was added to each 15 cm plate, then rocked gently for 3–4 minutes until cells start to detach from the plate. Afterwards, 25 mL of wash solution (DMEM/F-12 supplemented with 0.03% Gibco BSA Fraction V, pre-warmed at 37°C) was added to each plate to inactivate the trypsin. Cells were lifted into a 15 mL or 50 mL conical tube, pelleted at 330 g for 3 minutes, and then washed in 4 mL of 1X Phosphate Buffered Saline (PBS) per 10 million cells. During all crosslinking steps and washes, volumes were maintained at 4 mL of buffer or crosslinking solution per 10 million cells. After pelleting, cells were pipetted to disrupt clumps of cells and crosslinked in suspension with 4 mL of 2mM disuccinimidyl glutarate (DSG, Pierce) dissolved in 1X PBS for 45 minutes at room temperature. DSG was removed, and cells were pelleted (as above) and washed with 1X PBS. A solution of 3% formaldehyde (FA Ampules, Pierce) in 1X PBS was added to cells for 10 minutes at room temperature. Formaldehyde was immediately quenched with addition of 200 µl of 2.5 M glycine per 1 mL of 3% FA solution. Cells were pelleted, formaldehyde was removed, and cells were washed three times with 0.5% BSA in 1X PBS that was kept at 4°C. Aliquots of 10 million cells were allocated into 1.7 mL tubes and pelleted. Supernatant was removed and cells were flash frozen in liquid nitrogen and stored in -80°C until lysis.

Chromatin Isolation. Crosslinked cell pellets (10 million cells) were lysed using the nuclear isolation procedure previously described in the HT-ChIP protocol. Specifically, cells were incubated in 1 mL of Nuclear Isolation Buffer 1 (50 mM Hepes pH 7.4, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 140 mM NaCl, 0.25% Triton X, 0.5% NP-40, 10% Glycerol, 1X PIC) for 10 minutes on ice. Cells were pelleted at 850 *g* for 10 minutes at 4°C. Supernatant was removed, 1 ml of Lysis Buffer 2 (50 mM Hepes pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X PIC) was added and incubated for 10 minutes on ice. Nuclei were obtained after pelleting and supernatant was removed (as above), and 550 µL of Lysis Buffer 3 (50 mM Hepes pH 7.4, 1.5

mM EDTA, 1.5 mM EGTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% NLS, 1X PIC) was added and incubated for 10 minutes on ice prior to sonication.

Chromatin Digestion. After nuclear isolation, chromatin was digested via sonication of the nuclear pellet using a Branson needle-tip sonicator (3 mm diameter (1/8” Doublestep), Branson Ultrasonics 101-148-063) at 4°C for a total of 1 minute at 4–5 W (pulses of 0.7 seconds on, followed by 3.3 seconds off). DNA was further digested using 2–6 µL of TurboDNAse (Ambion) per 10 µL of sonicated lysate (equivalent to ~200,000 cells), in 1x DNase Buffer (Diluted from 10x DNase Buffer: 200 mM Hepes pH 7.4, 1 M NaCl, 0.5% NP-40, 5 mM $CaCl_2$, 25 mM $MnCl_2$) at 37°C for 20 minutes. Concentrations of DNase were optimized to obtain DNA fragments of approximately 150–1000 bp in length, which is needed for sequencing. DNase activity was quenched by adding 10 mM EDTA and 5 mM EGTA.

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 overnight. The DNA was purified using Zymo DNA Clean and Concentrate columns per the manufacturer’s specifications with minor adaptations, such as binding to the column with 7X Binding Buffer to improve yield. Molarity of the DNA was calculated by measuring the DNA concentration using the Qubit Fluorometer (High-Sensitivity (HS) dsDNA kit) and the average DNA sizes were estimated using the Agilent Bioanalyzer (HS DNA kit).

NHS-bead coupling. We used these numbers to calculate the total number of DNA molecules per microliter of lysate. We coupled the lysate to NHS-activated magnetic beads (Pierce) overnight at 4°C in 1 mL of 0.1% SDS in 1X PBS rotating on a HulaMixer Sample Mixer (Thermo). Specifically, we coupled $1x10^{10}$ DNA molecules to 1.75 mL of beads (mouse) and $5x10^{10}$ DNA molecules to 2 mL of beads (human). We obtain roughly 50% coupling efficiency of molecules to the beads, which effectively halves the ratio of molecules coupled per bead. This coupling ratio was selected to ensure that most beads contained less than 0.125 to 0.25 complexes 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. At this loading concentration of 0.125 complexes per bead, we find that <0.2% of SPRITE clusters contain any inter-species contacts and <0.1% of pairwise contacts contain any spurious pairing of human and mouse fragments that arise due to bead coupling (**Figure S1A**).

After coupling lysate to NHS beads overnight, we quench the beads with 1 mL of 0.5 M Tris pH 8.0 for 1 hour at 4°C rotating on a HulaMixer. We then wash the beads four times at 4°C in 1mL of modified RLT buffer (1X Buffer RLT supplied by Qiagen with added 10 mM Tris pH 7.5, 1 mM EDTA, 1 mM EGTA, 0.2% NLS, 0.1% Triton X, 0.1% NP-40) for 3 minutes each. Next, beads are washed in 1 mL of SPRITE wash buffer (1X PBS, 5 mM EDTA, 5 mM EGTA, 5 mM DTT*,* 0.2% Triton X, 0.2% NP-40, 0.2% sodium deoxycholate) twice at 50°C and once at room temperature for 5 minutes each. These washes remove any material that is not covalently attached to the beads. Prior to performing all enzymatic steps, buffer is exchanged on the beads through two rinses using 1 mL of SPRITE Detergent Buffer (20 mM Tris pH 7.5, 50 mM NaCl, 0.2% Triton X, 0.2% NP-40,0.2% sodium deoxycholate). These detergents are used throughout the protocol to prevent bead aggregation, which could result in spurious interactions. 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 rpm (Eppendorf Thermomixer) to avoid bead settling and aggregation, and all enzymatic steps were inactivated by adding 0.5–1 mL modified RLT buffer to the NHS beads.

DNA Repair. 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, T4 Polynucleotide Kinase (NEB) treatment is performed at 37°C for 1 hour, the enzyme is quenched using 1 mL Modified RLT buffer, and then buffer is exchanged with two washes of 1 mL SPRITE Detergent Buffer to beads at room temperature. Next, the NEBNext End Repair Enzyme cocktail (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) at 37°C for 1 hour, and inactivated and buffer exchanged as specified above.

Split-pool ligation. The beads were then repeatedly split-and-pool ligated over five rounds with a set of "DNA Phosphate Modified" (DPM), "Odd", "Even" and "Terminal" tags (see **SPRITE Tag Design** below for details). The DPM tag is ligated by an "Odd" tag. The "Odd" and "Even" tags were designed so that they can be ligated to each other over multiple rounds, such that after Odd is ligated, then Even ligates the Odd tags, and then Odd can ligate the Even tags. This can be repeated such that the same two plates of tags can be used over multiple rounds of split-pool tagging without self-ligation of the adaptors to each other. Finally, a set of barcoded Terminal tags are ligated at the end to attach an Illumina sequence for final library amplification. In this study, we performed five rounds total of split-and-pool ligation in the following order: DPM, Odd, Even, and Terminal tag. Over each round, the samples are split across a 96-well plate in 4.4 µL of SPRITE Detergent Buffer per well to prevent aggregation of beads, which would result in spurious interactions. Each plate contained 2.4 µL of 96 different tags at a concentration of 45 µM. 10 µL of 2X Instant Sticky-End Ligation Master Mix (NEB) and 3.2 µL of Ultra Pure H_2O (Invitrogen) was added to each well of the 96-well plate, for a final concentration of 1X Instant Sticky-End Ligation Master Mix per well. All ligations were performed at 20°C for 1 hour with shaking at 1600 rpm for 30 seconds every 5 minutes. Following every round of split-pool ligation, we inactivated the ligase via addition of 60 µL of modified RLT buffer to every well, which prevents spurious ligation of tags in the pooled tube. The sample was then pooled into a single 1.7 mL tube. After removing modified RLT buffer from the beads, remaining free tags were removed by washing the beads in 1 mL SPRITE wash buffer three times at 45°C for 3 minutes each. We then performed buffer exchange into SPRITE Detergent Buffer by adding 1 mL of Buffer and exchanging three times. We ensured that the majority of DNA molecules within a crosslinked complex are barcoded by optimizing the ligation efficiency such that >90% of DNA molecules are ligated during each round of split-pool tagging (**Figure S1B**).

Estimating sequencing depth. SPRITE interactions are defined based on the sequences that share the same tags. Accordingly, it is essential to sequence as many of the barcoded molecules in a complex as possible in order to identify interactions. Therefore, the number of unique molecules that are sequenced dramatically affects the likelihood of identifying interacting molecules. To address this, we optimized the loading density of our sequencing sample based on the number of unique molecules contained in the sample. Our goal is to load approximately equimolar unique molecules as the number of sequencing reads available. Specifically, based on our simulations, we have found that sequencing with ~1–3X coverage of reads per the number of unique molecules will ensure that most molecules are sampled. This follows Poisson sampling where 1-1/e^c of molecules are sampled at a given *c* coverage. For example, 3X, 2X, and 1X coverage samples approximately 95%, 86%, and 63% of interactions, respectively. In this study, most libraries were sampled with approximately 1.5–2X coverage.

To estimate the number of unique molecules in our sample, we measure the amount of material present on beads prior to reverse crosslinking all interactions. To do this, we take an aliquot of the sample and reverse crosslink to elute (as above) the DNA, which is then cleaned and amplified for 9–12 cycles. We then measure the molarity using the Qubit and Bioanalyzer (as above). The number of unique molecules in the aliquot prior to amplification is back calculated from a standard curve and adjusted to account for loss during the cleanup. This is used to estimate the number of unique molecules in the remaining crosslinked sample. In addition to optimizing molarity, because this dilution results in approximately 1% aliquots of the total sample being separately eluted and amplified, this effectively serves as another round of split-pool barcoding as each library is tagged with a unique barcoded Illumina primer. This further reduces the probability that molecules in different clusters obtain the same barcodes.

Sequencing library generation. We ensured that the number of unique DNA molecules to be sequenced (prior to amplification) does not exceed the number of molecules that can be sequenced (~150–300 million reads). Thus, aliquots were selected to contain approximately 50–150 million unique molecules. Each aliquot was digested with proteinase K (NEB) for 1 hour at 50°C in proteinase K Buffer (20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton X, 0.2% SDS), and their crosslinks were reversed overnight at 65°C. DNA was isolated using the Zymo DNA Clean and Concentrator columns using 7x Binding Buffer to increase yield. Libraries were amplified using Q5 Hot Start Mastermix (NEB) with primers that add the full Illumina adaptor sequences. After amplification, the libraries are cleaned up using 0.7X SPRI (AMPure XP) twice to remove excess primers and adaptors.

Mapping RNA and DNA simultaneously using SPRITE. To map RNA and DNA interactions simultaneously, the SPRITE protocol was performed with the following modifications: (i) Upon coupling of lysate to NHS beads, RNA overhangs caused by fragmentation are repaired by a combination of treatment with FastAP (Thermo) and T4 Polynucleoide Kinase (NEB) with no ATP at 37°C for 15 minutes and 1 hour, respectively. RNA was subsequently ligated with a "RNA Phosphate Modified" (RPM) tag using T4 RNA Ligase 1 (ssRNA Ligase), High Concentration (NEB) at 20°C for 1 hour (Shishkin et al., 2015). The RPM tag is designed with a 5' ssRNA overhang and 3' dsDNA sticky end for sequential ligation of DNA tags to the RNA (see **SPRITE Tag Design**). (ii) RNA was converted to cDNA using Superscript III (Thermo) using a manganese reverse transcriptase protocol (Siegfried et al., 2014) to promote reverse transcription through formaldehyde crosslinks on RNA. After cDNA synthesis, cDNA was selectively eluted from NHS beads using RNaseH (NEB) and RNase Cocktail (Ambion). cDNA was ligated with a unique cDNA tag as previously described (Shishkin et al., 2015), which serves as a RNA-specific identifier during sequencing.

SPRITE Tag Design

All sequence tags were designed to contain at least four mismatches from all other tags to prevent incorrect assignments due to sequencing errors. The 5' end of each sequence tag was designed with a modified phosphorylated base (IDT) to enable ligation. To obtain dsDNA tags, the ssDNA top and bottom strands of the tags were annealed in 1X Annealing Buffer (100 mM Tris-HCl pH 7.5, 2 M LiCl, 1.5 mM EDTA, 0.5 mM EGTA) by heating at 90°C for 2 minutes and slowly cooled to room temperature by reducing 1°C every 10 seconds in a thermocycler.

Framework of barcoding scheme. In order to enable an arbitrarily large number of tags to be added to DNA, we designed a scheme that enabled reuse of the same sets of tags. In this scheme, an initial tag is ligated to all DNA ends (DPM) or RNA ends (RPM). These RNA and DNA universal tags contain the same sticky-end overhang that is complimentary to the 5' end of a set of tags referred to as "Odd" tags. These Odd tags contain a unique 3' sticky end that is recognized exclusively by a set of "Even" tags, which contain a 3' sticky end that is complementary to the Odd tags. In this scheme, the number of tags can be increased to as many rounds as needed, but eliminates chimera formation within a single round of split-pool tagging. We explain each tag's design in greater detail below. Sequences of all tags are in **Table S5**.

DPM tag. The 5’ end of the top and bottom strands of the DPM tag have a modified phosphate group that allows for ligation to dA-tailed genomic DNA and subsequent ligation of the Odd tag. DPM contains a sequence of nine nucleotides that is unique to each of the 96 DPM tags (purple region). Each DPM tag contains a sticky-end overhang that ligates to the Odd set of adaptors (green region). The DPM tag also contains a partial sequence that is complementary to the universal Read1 Illumina primer, which is used for library amplification (gray region).

```
5′ Phos AAACACCCAAGATCGGAAGAGCGTCGTGTA   3′ Spcr
        ||||||||||||||||||||
3'     TTTTGTGGGTTCTAGCCTTCTGTACTGTTCAGT 5′ Phos
```

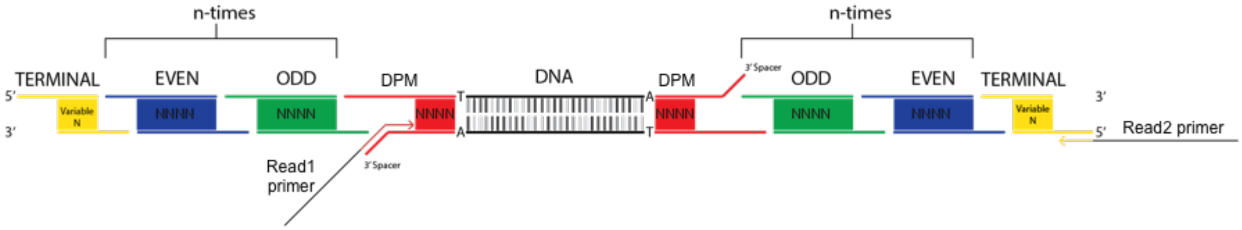
Because the DPM tag will ligate to both ends of the double-stranded DNA molecule, we designed the DPM tag to ensure that we would only read the barcode sequence from one sequencing read (Read2), rather than both. To achieve this, we included a 3’ spacer on the top strand. This prevents the top strand of the Odd tag from ligating to genomic DNA. This modification is also critical for successful amplification of the barcoded DNA by preventing hairpin formation of the single-stranded DNA during the initial PCR denaturation because otherwise both sides of the tagged DNA molecule would have complementary barcode sequences.

“Odd” and “Even” Tags. We designed two sets of tags called the “Odd” and “Even” set. Both the Odd tags and Even tags have modified 5’ phosphate groups to allow for ligation. The Even tags are designed to have a sticky end that anneals to the Odd tags, and the Odd tags are designed to contain a sticky end that anneals to the Even tags. The Odd tags are ligated in the 1st, 3rd, 5th, … rounds of the SPRITE process and the Even tags are ligated 2nd, 4th, 6th, … rounds of SPRITE. Each of the Even and Odd tags contain a unique sequence of seventeen nucleotides.

Terminal Tag. The Terminal tags contain a sticky end that ligates to the Odd tags (green), though a Terminal tag can also be designed to ligate to an Even tag. The Terminal tag only contains a modified 5’ phosphate on the top strand. The bottom strand contains a region (gray) that contains part of the Illumina read 2 sequence, which allows for priming and incorporation of the full-length barcoded Read2 Illumina adaptor. The Terminal tag contains a unique sequence of nine nucleotides (bold).

```
5' Phos      AGTTGTCACCATAATAAGATCGGAAGA              3'
                    ||||||||||||||||||||
3'                  TGGTATTATTCTAGCCTTCTCGTGTGCAGAC  5'
```

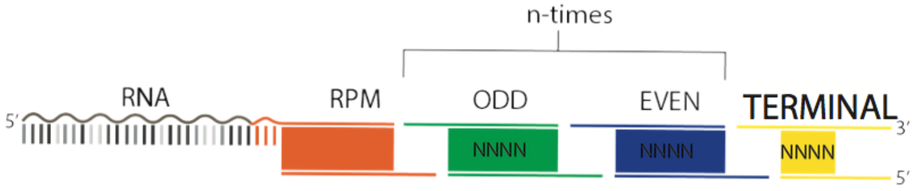
Final DNA structure. After SPRITE, the genomic DNA contains a DPM tag ligated on both ends as well as the Odd, Even, and Terminal tags. We call the full tag sequence a barcode. The product is represented below:



RNA Barcoding. For RNA tagging, we use the same approach as above, except the first ligation to the RNA is an RPM tag. The RPM tag is designed with a ssRNA overhang to specifically ligate RNA molecules using a single-stranded RNA ligase. This RNA-specific ligation tags RNA molecules in order to distinguish a molecule as RNA, rather than DNA, on the sequencer. The RPM tag contains a distinct sequence relative to DPM (pink region) and serves as a RNA-specific tag to mark each read as RNA. However, the sticky end is identical to that contained on the DPM tag (green sequence) to enable barcoding of both DNA and RNA simultaneously. The bottom strand of the RPM tag (TGACTTGCTGACGCTAAGTCCATCCTATCTACATCCG) is phosphorylated after ligation of the RPM tag to RNA to ensure that the RPM tags do not form chimeras and ligate to each other during the ssRNA ligation of the RPM tag. The 3’ spacer on the top strand (/5Phos/rArUrCrArGrCrArCrCrCrGrGATGTAGATAGGATGGACTTAGCGTCAG/3SpC3/) of the RPM tag prevents ligation of single-stranded RPM molecules and from forming chimeras during ligation.

```
5' rArUrCrArGrCrArCrCrCrGrGATGTAGATAGGATGGACTTAGCGTCAG          3'spcr
                        | |  ||||||||||||||||||||||||||||
3'                      G C  CTACATCTATCCTACCTGAATCGCAGTCGTTCAGT  5'
```

Final RNA structure. The final RNA product after SPRITE contains the RPM, Odd, Even, and Terminal tags. We call the full tag sequence a barcode. The product is represented below:



cDNA tag. In order to amplify cDNA molecules, we ligate a cDNA tag to the 3'end of all cDNA molecules. The cDNA tag contains a five-nucleotide sequence that identifies the tagged molecule as RNA in read 1 during sequencing (blue). The cDNA tag also contains a sequence that is part of the Illumina Read 1 primer (green). It is 5'phosphate modified to ligate to the 3'end of cDNA, and contains a 3'spacer to prevent chimeras of tags.

```
5′ /5Phos/actgaAGATCGGAAGAGCgtcgtgtaggg/3SpC3/ 3′
```

Final Library Amplification Primers. DNA and RNA libraries are amplified using common primers that incorporate the full Illumia sequencing adaptors. These are the Read 1 primer (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) and the Read 2 primer (CAAGCAGAAGACGGCATACGAGATGCCTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT). The Read 1 primer amplifies the top strand of the DPM tag on DNA and cDNA tag on RNA, and adds the Illumina Read 1 sequence to each molecule. The Read 2 primer amplifies the Terminal tag on both DNA and cDNA, and adds the Illumina Read 2 sequence to the molecule.

SPRITE Data Processing, Cluster Generation, and Heatmap Generation

All SPRITE data was generated using Illumina paired-end sequencing on the HiSeq 2500 or NextSeq 500. Read pairs were generated with at least 115 x 100 bps. The reads have the following structure: Read 1 contains genomic DNA positional information and the DPM or cDNA tag, and Read 2 has the remaining tags. See **Table S6A** for number of human and mouse reads during all steps of filtering prior to cluster generation.

Barcode identification. SPRITE barcodes were identified by parsing the first DNA tag sequence from the beginning of Read 1 and the remainder of the tags were parsed from Read 2. We identified these tag sequences using a hashtable populated with the known sets of DPM, Odd, Even, and Terminal tags that were added to these samples. We allowed for up to two mismatches to each internal tag of the Odd and Even tags to account for possible sequencing errors. Because the tags were designed to contain at least four mismatches to any other tag sequence, this enables robust error correction. For DPM and Terminal tag alignments, we did not tolerate any mismatches due to their shorter unique barcode sequences. We excluded any reads that did not contain a full set of all ligated tags (DPM, Odd, Even, Odd, Terminal) in the order expected from the experimental procedure. We also excluded all read pairs where we could not unambiguously map the barcode. We stored the barcode string to the name of each read in the FASTQ file.

Alignment to genome. We aligned each read to the appropriate reference genome (mm9 for mouse and hg19 for human) using Bowtie2 (v2.3.1) with the default parameters and with the following deviations. We trimmed the 11 base pair tag sequence (DPM) from Read 1 using the --trim5 11 parameter. To account for a short genomic fragment that might lead to additional tag sequences being included on the 3'end, we used a local alignment search (--local). The corresponding SAM file was sorted and converted to a BAM file using SAMtools v1.4. The barcode string is stored in the name of each record of the BAM file.

Repeat Masking and Filtering Low-Complexity Sequences. We filtered the resulting BAM file for low quality reads, multimappers, and repetitive sequences. First, we removed all alignments with a MAPQ score less than 10 or 30 (heatmaps were generated for both). Second, we removed all reads that had >2 mismatches to the reference genome. Third, we removed all alignments that overlapped a region that was masked by Repeatmasker (UCSC, milliDiv < 140) using bedtools (v2.26.0). Fourth, we removed any read that aligned to a non-unique region of the genome by excluding alignments mapping to regions generated by the ComputeGenomeMask program in the GATK package (readLength=35nt mask). In the human maps, all reads that overlap with an annotated *HIST* gene were removed for analysis of the histone locus in **Figure 2**.

Identifying SPRITE clusters. To define SPRITE clusters, all reads that have the same barcode sequence were grouped into a single cluster. To remove possible PCR duplicates, all reads starting at the same genomic position with identical barcodes were removed. We generated a SPRITE cluster file for all subsequent analyses where each cluster occupies one line of the resulting text file containing the barcode name and genomic alignments.

Visualizing SPRITE clusters. Multi-way interactions were identified by counting how often multiple genomic regions simultaneously interact in individual SPRITE clusters. In Figures 2, S2, 4, and 5, rows show individual SPRITE clusters there multiple regions simultaneously interact and black lines denote genomic bins (typically 1Mb or 25kb) with at least one read within these clusters.

RNA-DNA SPRITE analysis

RNA and DNA sequences were separated by the presence of a cDNA tag or DPM tag in the first 9 nucleotides of the Read 1 sequence, respectively. RNA-tagged reads, identified by the cDNA tag, were aligned to ribosomal RNA sequences (28S, 18S, 5S. 5.8S, 4.5S, 45S) as well as other RNAs of interest such as snoRNAs, snRNAs including spliceosomal RNAs, and Malat1. Any RNA sequences that did not align to this set of RNA genes of interest were aligned to the mm9 genome using bowtie2. DNA-tagged sequences, containing the DPM tag, were aligned to mm9. All RNA and DNA reads were subsequently filtered by edit distance and MAPQ score as described above; DNA reads were additionally filtered with a DNA mask file. See **Table S6B** for number reads during all steps of filtering prior to cluster generation.

Generating pairwise contacts and heatmaps from SPRITE data. To compute the pairwise contact frequency between genomic bins i and j, we counted the number of SPRITE clusters that contained reads overlapping both bins. Specifically, we counted the number of unique SPRITE clusters overlapping the two bins and not the number of reads contained within them. In this way, if a SPRITE cluster contains multiple reads that mapped to the same genomic bin, we only count the SPRITE cluster once to eliminate possible PCR duplicates. Because the number of pairwise contacts scales quadratically based on the number of reads (n) contained within a SPRITE cluster, larger clusters will contribute a disproportionally large number of the contacts observed between any two bins. To account for this, we reasoned that a minimally connected graph containing n

reads would contain *n*-1 contacts. Therefore, we down-weighted each of the *n*(*n*-1)/2 pairwise contacts in a SPRITE cluster such that each pairwise contact has a weight of 2/*n*. In this way, the total contribution of pairwise contacts from a cluster is proportional to the minimally connected edges in the graph. This also ensures that the number of pairwise contacts contributed by a cluster is linearly proportional to the number of reads within a cluster.

Pairwise contacts were computed using multiple different bin sizes (10kb, 20kb, 25kb, 40kb, 50kb, 200kb, 250kb, 1Mb) to generate contact maps at different genomic resolutions. SPRITE contact maps were normalized by read coverage using Hi-Corrector (Li et al., 2015). Contacts occurring within the same bin (i.e., along the diagonal, i=j) were not considered to avoid any chance of possible PCR duplicates generating false-positive interactions. All low coverage bins are masked in heatmaps.

SPRITE processing details and scripts for performing this processing are available at: https://github.com/GuttmanLab/sprite-pipeline/wiki

Human-mouse Mixing Experiment

Human HEK293T cells and mouse pSM33 cells were crosslinked, lysed, and DNase digested separately. The two lysates were then combined at equimolar concentration and coupled to NHS beads at a ratio of 620, 125, 60, 25, 6, 1.2, 0.5, 0.25, 0.125 molecules per NHS bead. Reads were aligned to both hg19 and mm9. All reads aligning to both species were removed, and species-specific reads were used to determine the amount of inter-species contacts and normalized by the expected number of contacts. For the experiments in the paper, we selected a coupling efficiency of 0.125 to 0.25 molecules/bead because it provided a small number of spurious contacts while minimizing the number of beads used in the experiment.

Comparison of SPRITE and Hi-C Data

Hi-C contact maps for mESCs (Dixon et al., 2012) and human GM12878 cells (Rao et al., 2014) were normalized by read coverage using Hi-Corrector (Li et al., 2015).

Compartments. We identified A and B compartments and insulation scores for SPRITE and Hi-C using *cworld* (Dekker lab, https://github.com/dekkerlab/cworld-dekker). To calculate compartment eigenvectors, we used the *cworld* script "matrix2compartment.pl" with default parameters with contact maps binned at 1Mb resolution as input. For human chromosomes, compartment eigenvectors were calculated separately for each chromosome arm. Bias in coverage of reads towards the A compartment was calculated as the observed percentage of total reads that aligned to regions in the A compartment divided by the expected percentage of total A compartment reads assuming a uniform distribution of read coverage. We note that differences in distance decay profiles between different cluster sizes, such as those observed in **Figure 3B-C**, occur despite comparable genomic coverage in both the SPRITE and Hi-C data (**Table S2A**).

TADs. To calculate insulation scores, we used the *cworld* script "matrix2insulation.pl". For human insulation scores, we used the parameters "--ss 100000 --im iqrMean --is 600000 --ids 400000" with contact maps binned at 50kb resolution. For mouse insulation scores, we used the parameters "--ss 80000 --im iqrMean --is 480000 --ids 320000" with contact maps binned at 40kb resolution.

Loops. We performed aggregate peak analysis (APA) on mouse and human contact maps in both HiC and SPRITE data binned at 10kb resolution as previously described (Rao et al., 2014). Positions for loops with end points separated by at least 200kb were obtained from mouse CH12-LX cells (1493 loops) and human GM12878 cells (5789 loops) (Rao et al., 2014). Aggregate contact maps were computed for the median contact frequency in regions +/- 200kb of the loops.

Analysis of Higher-order *K*-mer Interactions

We enumerated all higher-order *k*-mers represented in the SPRITE data at 1 megabase (Mb) resolution. We retained *k*-mers that were observed in at least 5 independent SPRITE clusters. We found that the greatest determinant of *k*-mer frequency, similar to pairwise frequency, was linear genomic distance. Accordingly, to assess the significance of a given *k*-mer, we compared the observed frequency of a given *k*-mer to the expected frequency of other *k*-mers containing the same genomic distance at different positions across the genome (**Figure S2A**). See "Determining significant higher-order *k*-mer interactions" below for further details on how significance was computed for each *k*-mer.

SPRITE Cluster-Size: Comparison of SPRITE Contacts in Different Cluster Sizes

An individual SPRITE cluster is defined as a set of reads that all contain the same barcode sequence. Accordingly, the size of a SPRITE cluster is defined as the number of reads that have the same barcode sequence. We separated SPRITE clusters into four groups: clusters with 2 to 10 reads, 11 to 100 reads, 101 to 1000 reads, and clusters with 1001 or more reads. Contact maps were generated separately for each group of clusters as described above but without down-weighting for cluster size. We analyzed the relationship between genomic distance and contact frequency by computing the average contact frequency between bins separated by 40kb, 80kb, 120kb and so forth up to 100Mb. To compare heatmaps across different cluster sizes, we normalized contact frequencies by the maximum observed value such that overall contact frequency ranges from 0 to 1 for each cluster size. To examine A compartment interactions in different cluster sizes, we calculated the average contact frequency between all 1Mb bins on mouse chromosome 2 and the bins within a 9Mb A compartment region (25 to 34Mb) and normalized these values to range from 0 to 1 for each cluster size. These contact frequencies were down-weighted by cluster size (as described above).

Additionally, we compared each SPRITE cluster-size group according to the percent of contacts that occurred between two regions in the same TAD, regions in the same compartment (A or B), or regions on the same chromosome. Two regions were considered to be in the same "local" or "contiguous" compartment only if no compartment switches occurred in the linear genomic span between them. Two regions were considered to be in the same "non-local" or "discontiguous" compartment if compartment switches exist between them, but they are nevertheless both in A compartment regions or both in B compartment regions.

Defining the "Active" and "Inactive" Inter-chromosomal DNA Hubs

Hub definitions. The two "active" and "inactive" inter-chromosomal hubs were identified solely based on DNA contacts. Specifically, we identified significant inter-chromosomal interactions occurring between all 1Mb genomic regions (see below). These significant contacts cluster into 2 "sets" of DNA regions, such that there is a large degree of interconnectivity among genomic regions within set 1 and among regions within set 2. There is no connectivity between regions in set 1 and regions in set 2. Because of these two properties, we defined these sets of interactions as two distinct "hubs" of inter-chromosomal contacts. However, once these hubs were identified based solely on DNA contact frequencies, we investigated what features are associated with the DNA regions contained within each hub (see below). Specifically, we noticed that genomic regions in one of the hubs is primarily gene poor and depleted of RNA PolII signal. In contrast, genomic regions in the other hub were highly gene dense and contained high levels of RNA PolII occupancy. Based on these properties, we simply referred to these hubs as the "inactive hub" and "active hub" respectively.

Analysis of inter-chromosomal interactions. To identify significant inter-chromosomal interactions, we removed all intra-chromosomal contacts from the ICE-normalized inter-chromosomal heatmap. We then calculated an interaction *p*-value using a one-tailed binomial test where the expected frequency assumes a uniform distribution of inter-chromosomal contacts. We used contact maps binned at 1Mb resolution based on SPRITE clusters containing 2 to 1000 reads without down-weighting for cluster size. We built a graph where nodes represent a 1Mb bin and edges represent connections between 2 bins. We filtered edges to reduce potentially spurious contacts that may be caused by outlier bins by looking for consistency of contacts across at least 3 consecutive bins. Specifically, we only included an edge between two bins (i and j) when the edge connecting i and j was significant and all interacting pairs i±1 and j±1 were also significant. This approach produced two networks of inter-chromosomal interactions that were defined as the inactive hub (nucleolar hub, p-values $\leq 10^{-10}$) and active hub (active hub, p-values $\leq 10^{-8}$).

Identifying features associated with each hub. To identify features that distinguish these hubs and the rest of the genome, we calculated various properties, such as average gene density and average number of Pol II ChIP-seq peaks for each genomic region in these hubs and compared these to a set of control regions not contained in either hub, but with the same distribution of lengths. In **Figure 3F**, gene density is calculated as the number of genes per 1Mb region and Pol II occupancy is calculated as number of ChIP-seq peaks per 1Mb region (ENCODE). The properties we analyzed included the following: gene density, number of enhancers (Whyte et al., 2013), number of super-enhancers (Whyte et al., 2013), histone modifications (H3K27ac, H3K4me1, H3K36me3, H3K4me3), GRO-seq signal (Jonkers et al., 2014), number of Pol II ChIP-seq peaks, number of DNase-seq peaks. Histone modification, Pol II ChIP-seq and DNase-seq data were obtained from ENCODE. Because one of the two hubs was highly enriched in active chromatin marks, RNA PolII, and GRO-seq signal we defined it as the "active hub", while the other was depleted for these features of high transcriptional activity and was therefore defined as the "inactive hub".

SPRITE Contact Frequency With the Nucleolar and Active Hubs

We defined contact frequency with the nucleolar and active hubs for each 1Mb bin based on the average inter-chromosomal contact frequency with all regions in the nucleolar and active hub, respectively. *p*-values were calculated for these contact frequencies using a one-tailed binomial test where the expected frequency assumed a uniform distribution of inter-chromosomal contacts and are shown in units of $-\log_{10}(p\text{-value})$ in the corresponding figures. Contact frequencies with these hubs were compared to RNA localization (described above) and data from ChIP-seq (e.g., Pol II, H3K4me3) and GRO-seq experiments obtained from ENCODE and from (Jonkers et al., 2014), respectively.

Analysis of Inter-chromosomal Centromere Interactions

Mouse inter-chromosomal centromere interactions were defined as interactions between the closest bin to the centromere that can be mapped ("peri-centromeric bin", usually 3–4Mb) and any given region on a different chromosome. To analyze these interactions, we calculated the median contact frequency between the peri-centromeric bin of each mouse chromosome and bins on different chromosomes at specific distances ranging from 1Mb to 100Mb. We then computed the mean contact frequency for each centromere distance across all mouse chromosomes. We excluded chromosomes 14 and X from this analysis because they do not have reads that uniquely map within 4Mb of the centromere.

Analysis of Lamina-Associated Domain Interactions

Lamina-associated domains (LADs) were obtained from Peric-Hupkes et al., 2010 (Peric-Hupkes et al., 2010). We calculated a genome-wide LAD interaction heatmap by summing the number of contacts between all pairs of LADs and normalizing by read coverage using Hi-Corrector (Li et al., 2015). To compare the frequency of LAD/LAD contacts with LAD/non-LAD in SPRITE with Hi-C, we classified each 1Mb bin as "LAD" or "non-LAD" based on whether or it had a Lamin B1-DamID score above the median. We then computed the median contact frequency between bins separated by specific distances ranging from 1Mb to 80Mb for pairs of bins classified as LAD/LAD and LAD/non-LAD.

Defining Genomic Regions Near Ribosomal DNA Clusters

The precise location of ribosomal DNA genes in both mouse and human are unknown because they are not mapped in the reference genomes. However, approximate locations have been reported based on non-sequencing methods. In mouse, rDNA genes are encoded from the centromere-proximal regions of chromosomes 12, 15, 16, 18, and 19 (Suzuki et al., 1990). In human, rDNA genes are encoded on chromosomes 13, 14, 15, 21 and 22 (Németh et al., 2010; Pederson, 2011). Importantly, the locations of rDNA genes can be strain-dependent in mice (Strongin et al., 2014; Suzuki et al., 1990). For instance, on chromosome 15, the *129* allele of the F1-21 hybrid mouse line does not include rDNA genes, while it does on the *Cast* allele (Strongin et al., 2014). This hybrid line was used for all DNA-DNA mapping methods and nucleolar hub identification. Instead, we performed our rRNA-DNA maps in another mouse cell line (pSM33), which is derived from a *C57BL/7* x *129SV* mouse, which are reported to contain rDNA genes on both alleles of chromosome 15. This difference in the rDNA locations between strains may explain why we observe a stronger enrichment of rRNA on chromosome 15 in the rRNA-DNA maps than in the DNA nucleolar hub contact maps (**Figure 4A**).

Ribosomal RNA (rRNA) Localization Quantification

To quantify the localization of ribosomal RNA (rRNA) across the genome, we split SPRITE clusters into two groups, one with clusters that contained at least 1 rRNA read (rRNA positive clusters) and the other with clusters lacking any rRNA reads (rRNA negative clusters). We then calculated the ratio of rRNA positive clusters to rRNA negative clusters for each 1Mb bin, normalizing for total number of clusters in each group, and defined this ratio as the rRNA enrichment for each bin. We calculated rRNA enrichment *p*-values for each bin using a one-tailed binomial test where the expected frequency was based on the rRNA negative clusters and are shown in units of $-\log_{10}(p\text{-value})$ in **Figure 4A**. To quantify Malat1 and U1 localization, we obtained Malat1 and U1 RAP-DNA alignments from Engreitz et al.(Engreitz et al., 2014) and calculated Malat1 and U1 enrichment for each 1Mb bin by normalizing to input RAP-DNA alignments.

Measurement of Distance Between DNA Loci and Nuclear Bodies Using Microscopy (Histochoice Fixation)

DNA FISH Combined with Immunofluorescence. DNA fluorescence in situ hybridization (DNA FISH) was performed with Agilent SureFISH DNA FISH probes following the manufacturer's protocol with adaptations noted below. Probe sets used for FISH were designed by Agilent technologies using their standard procedures against genomic regions defined in **Table S6C**. Female F1-21 cells were cultured on Poly-D Lysine (Sigma-Aldrich) and gelatin (Sigma-Aldrich) coated coverslips. The coverslips were fixed using 300uL of Histochoice for 10 minutes at room temperature, then dehydrated through incubation in a series of graded ethanol concentrations up to 100% ethanol, and air dried. The coverslips were then turned cell-side down onto a 5uL mixture of a custom probe set targeting a selected DNA locus (Agilent) and SureFISH Hybridization Buffer (Agilent, G9400A). The coverslips and probe mixture were denatured for 8 minutes at 83°C, then incubated at 37°C overnight in a humidified chamber. The following morning, coverslips were washed with FISH Wash Buffer 1 (Agilent, G9401A) at 73°C for 2 min on a shaking incubator at 300 rpm, and FISH Wash Buffer 2 (Agilent, G9402A) at room temperature for 1 minute. Coverslips were then rehydrated and suspended in 1X PBS in preparation for immunofluorescence staining.

Following DNA FISH probe hybridization, immunofluorescence was performed. Coverslips were permeabilized with 0.1% Triton X in PBS at room temperature for 10 minutes, then blocked with 1X blocking buffer (Abcam ab126587) in PBS at room temperature for one hour. The coverslips were then incubated with primary antibodies in a humidified chamber at room temperature for one hour for anti-Nucleolin or overnight at 4°C for anti-SC35. The coverslips were washed with 0.1% Triton X in PBS at room temperature, then incubated with secondary antibodies in a humidified chamber at room temperature for one hour. Coverslips were washed with PBS and H_2O and mounted on slides in ProLong® Gold antifade reagent with DAPI (Life Technologies, P36931). The primary antibodies used for IF were rabbit polyclonal anti-Nucleolin (Abcam; ab22758; 1:1000) and mouse monoclonal anti-SC35 (Abcam; ab11826; 1:200). The secondary antibodies used for IF were Alexa Fluor® 647 goat anti-rabbit IgG (H+L) (Thermo Fisher Scientific; A21244; 1:300) and DyLight® 650 goat anti-mouse IgG (H+L) (Bethyl; A90-116D5; 1:300).

Microscopic imaging. DNA FISH/IF samples were imaged using a Leica DMI 6000 Deconvolution Microscope, with a z-stack collected for each channel (4 μm; step size, 0.2 μm). The objectives used were the Leica HC PL APO 63x/1.30 GLYC CORR CS2 objective and the Leica HCX PL APO 100X/1.40- 0.70na OIL objective. Samples were also imaged with a ZEISS Laser Scanning Microscope (LSM) 800 with the ZEISS i Plan-Apochromat 63x/1.4 OIL DIC M27 objective, with a z-stack collected for each channel (step size, 0.37 μm). Deconvolution was performed using Huygens Professional version 17.04 (Scientific Volume Imaging, The Netherlands, software available at http://svi.nl) using the built-in theoretical point spread function, the classic maximum likelihood estimation (CMLE) algorithm, a signal to noise ratio of 20, and 50 iterations.

Calculating distance between DNA loci. The nuclei of individual cells were identified by DAPI staining, and cells containing two spots per DNA FISH channel were identified manually. Images were cropped to only contain the identified cell. Analysis of cells in three dimensions was performed using Imaris version 8.4.1 (Bitplane Inc, software available at http://bitplane.com) with the ImarisXT module. Both alleles for each DNA locus were defined by applying the Imaris “Spot” function (diameter = 0.5µm, background subtraction) on the corresponding fluorescent channel. The distance between DNA loci was calculated by running the XTension “Distances Spots to Surfaces” function and manually recording the smallest distance between alleles of differing loci.

Calculating distance between DNA loci and nuclear bodies. The nucleolus and nuclear speckle, identified by immunofluorescence of nucleolin and SC35, respectively, were defined in Imaris by performing the Imaris “Surface” function (detail = 0.126 µm, absolute intensity). Custom Imaris XTensions were used to calculate a distance transform approximating Euclidean distance for the region outside of the generated surface (“Batch Process Function” by Pierre Pouchin and “Distance Transformation Outside Object For Batch” by Matthew Gastinger, obtained on open.bitplane.com). The edges of the surface served as boundary voxels; regions inside the surface were assigned a distance transform value of 0. The distance of an allele to the nucleolus or nuclear speckle was defined as the minimum distance transform value of the corresponding spot, from the edge of the surface to the nearest edge of the DNA FISH sphere.

Measurement of Distance Between DNA Loci and Nuclear Speckles Using Microscopy (Formaldehyde Fixation)

Immunofluorescence. Immunofluorescence was performed followed by non-barcoded DNA sequential FISH (DNA seqFISH) (Takei et al., 2017) with modified steps. Female F1-21 cells were cultured and fixed on Poly-D Lysine (Sigma-Aldrich) and human laminin (BioLamina LN511) coated coverslips (Thermo Scientific Gold Seal 3421) using 4% Formaldehyde (Pierce) for 10 minutes at room temperature, washed with 1X PBS, and stored in 70% ethanol for more than overnight at -20°C. The coverslips were air dried, incubated with 0.2 µm blue fluorescent (365/415) beads (Thermo Scientific F8805) with 2000-fold dilution in 2X SSC at room temperature for 5 minutes for the alignment of images. The coverslips were then washed twice with 2X SSC, and blocked with blocking buffer (5% BSA GEMINI 700-106P, 1X PBS and 0.3% Triton X) at room temperature for 30 minutes. The coverslips were washed with 0.3% Triton X in

1X PBS at room temperature, and incubated with a primary antibody of mouse monoclonal anti-SC35 (Abcam; ab11826; 1:200) in Antibody Dilution Buffer (1% BSA, 1X PBS, and 0.3% Triton X) at 4°C overnight. The coverslips were then washed with 0.3% Triton X in 1X PBS at room temperature, and incubated with secondary antibody DyLight® 488 goat anti-mouse IgG (H+L) (Bethyl; A90-116D2; 1:300) in Antibody Dilution Buffer at room temperature for 1 hour. Coverslips were washed with 1xPBS, incubated with DAPI and imaged in Anti-Bleaching Buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 2X SSC, 3 mM Trolox [Sigma 238813], 0.8% D-glucose, 100-fold diluted Catalase [Sigma C3155] and 0.5 mg/mL Glucose oxidase [Sigma G2133]). Imaging conditions are described below under “Microscopic Imaging”.

DNA seqFISH. Following the immunofluorescence, non-barcoded DNA seqFISH experiments were performed on the same coverslips. Single-stranded DNA FISH probes were designed against specific mouse regions defined in **Table S6D**, purchased as an oligoarray pool (Twist Bioscience), and generated with limited cycle PCR, in vitro transcription, reverse transcription as described previously (Takei et al., 2017). After immunofluorescence imaging, the coverslips were washed with 1X PBS twice at room temperature, and incubated with 0.1 mg/mL RNaseA (Thermo Scientific) at 37°C for 1 hour. The coverslips were washed and dried with 1x PBS, 70% ethanol and 100% ethanol. The coverslips were then heated at 90°C for 10 minutes with 50% formamide in 2X SSC. The coverslips were then hybridized with the probe pool at 37°C overnight in 50% formamide, 2X SSC, and 0.1 g/mL dextran sulfate (Sigma-Aldrich D8906). After incubation with the primary probe pool, the coverslips were washed with 55% Wash Buffer (55% formamide, 0.1% Triton X, 2X SSC) at room temperature for 30 minutes, and hybridized with readout probes that were 15-nucleotides in length (IDT), which can bind to the readout sequences on the primary probes, coupled to Alexa Fluor 647 (Lifetech) or Cy3B (GE Healthcare) at 50 nM final concentration at room temperature for 20 minutes in 10% EC Buffer (10% ethylene carbonate [Sigma-Aldrich E26258], 2X SSC, and 0.1 g/mL dextran sulfate [Sigma-Aldrich D4911]). The coverslips were then washed with 10% Wash Buffer (10% formamide, 0.1% Triton X and 2X SSC) at room temperature for 5 minutes, stained with DAPI and imaged in Anti-Bleaching Buffer. Imaging conditions are described below under “Microscopic Imaging”. Following the imaging, the coverslips were washed with 2X SSC once at room temperature, incubated in 55% Wash Buffer at room temperature for 5 minutes for readout probe displacement, and then washed three times with 2X SSC. To check the readout probe displacement, the coverslips were imaged with all imaging channels in Anti-Bleaching Buffer. The coverslips were then re-hybridized with another set of readout probes at 50 nM final concentration at room temperature for 20 minutes in 10% EC Buffer, stained with DAPI and imaged in Anti-Bleaching Buffer. In total, three rounds of hybridizations with two colors were carried out for DNA seqFISH to image six regions.

Microscopic Imaging. Samples were imaged with a microscope (Leica DMi8 automated) equipped with a confocal scanner unit (Yokogawa CSU-W1), a sCMOS camera (Andor Zyla 4.2 PLUS), 63X oil objective lens (Leica NA 1.40), and a motorized stage (ASI MS2000). Lasers from CNI and filter sets from Semrock were used. Images were acquired with 0.35 µm z steps.

Calculating distance between DNA loci and nuclear bodies. Image processing was carried out using ImageJ and MATLAB. Images between different imaging rounds were registered and aligned using 405 nm channel images, which contain DAPI and 0.2 µm blue fluorescent beads signals, with MATLAB's imregtform and imwarp functions. To remove the effects of chromatic aberration, 0.1 µm TetraSpeck beads (Thermo Scientific T7279) were used to create geometric transforms to align different fluorescence channels. Cells with two DNA FISH spots for a given DNA loci were visually identified and confirmed. Nuclei of individual cells were identified by DAPI staining, and the precise location of DNA FISH spots were defined as described previously (Takei et al., 2017). The nuclear speckles identified by immunofluorescence of SC35 were defined using ImageJ's rolling ball background subtraction algorithm with a radius of 3 pixels, followed by ImageJ's auto threshold algorithm. The distance of an allele to the nuclear speckle was defined as the minimum distance of a DNA FISH spot to the boundary of any nuclear speckle in the corresponding cell.

Technical challenges associated with quantitatively measuring multi-way interactions using seqFISH. There are significant technical challenges that precludes us from quantitatively measuring the co-occurrence of multiple DNA sites around a nuclear body across many individual cells by microscopy. Specifically, following computational and manual image processing, approximately 10–15% of all cells imaged have DNA FISH signal identified for a single DNA region of interest. Accordingly, the percentage of cells that contain FISH signals for two DNA regions in the same cell is approximately 1–2%. This is further compounded when increasing to 3-way interactions in the same cell (0.1–0.3% of cells). Specifically, to identify 2 examples where two distinct DNA regions organize around the same nuclear speckle required imaging and analysis of ~800 individual cells. 15 (of ~800) individual cells contained FISH signals for two active hub DNA regions in the *same* cell. 2 (of 15) cells showed simultaneous co-association around the same speckle. In contrast, for an active hub and control DNA region, we observe 0 (of 21) cells that showed simultaneous co-association around the same speckle.

Comparison of SPRITE and DNA FISH

To compare SPRITE and DNA FISH measurements, we used SPRITE contact frequencies from contact maps binned at 1 Mb resolution based on SPRITE clusters containing 2 to 1000 reads without down-weighting for cluster size. We note that similarly high correlations between SPRITE and DNA FISH measurements were observed when large SPRITE clusters (>100 reads) are included in the analyses (consistent with the observation of inter-chromosomal interactions in larger clusters in **Figure 3D**), both with and without weighting for cluster size (**Table S6E**). SPRITE contact frequencies were obtained for 1 Mb bins that overlapped with each DNA FISH probe region and compared to DNA FISH distance measurements with the corresponding probe region.

2.8.3 QUANTIFICATION AND STATISTICAL ANALYSIS

Determining Significant Higher-order *K*-mer Interactions

We found that the greatest determinant of *k*-mer frequency, similar to pairwise frequency, was linear genomic distance. Accordingly, to assess the significance of a given *k*-mer, we compared the observed frequency of a given *k*-mer to the expected frequency of *k*-mers containing the same genomic distance. To do this, for a given *k*-mer, we computed the genomic distance separating each region in the *k*-mer and randomly sampled regions across the genome containing the same linear genomic distances and computed the number of SPRITE clusters containing these *k*-mers (**Figure S2A**). SPRITE cluster counts were normalized by cluster size to define a weighted score and prevent large SPRITE clusters from dominating the number of *k*-mer observations. For *k*-mers of interest, enrichment was defined as the observed weighted SPRITE counts divided by the average across 100 random permutations. Genome-wide analysis was performed across 10 random permutations to identify an initial subset of enriched *k*-mers. We also retained the number of permutations that had an observed frequency larger than observed for the *k*-mer of interest and we report this percentile to rank each higher-order *k*-mer.

Defining Significant Inter-Chromosomal Contacts

To identify significant inter-chromosomal interactions, we removed all intra-chromosomal contacts from the ICE-normalized inter-chromosomal heatmap. We then calculated an interaction *p*-value using a one-tailed binomial test where the expected frequency assumes a uniform distribution of inter-chromosomal contacts. We used contact maps binned at 1Mb resolution based on SPRITE clusters containing 2 to 1000 reads without down-weighting for cluster size. We built a graph where nodes represent a 1Mb bin and edges represent connections between 2 bins. We filtered edges to reduce potentially spurious contacts that may be caused by outlier bins by looking for consistency of contacts across at least 3 consecutive bins. Specifically, we only included an edge between two bins (i and j) when the edge connecting i and j was significant and all interacting pairs i±1 and j±1 were also significant.

Plots and Statistical Analysis

Plots and statistics were generated in GraphPad Prism version 7.0c, MATLAB R2016a (MathWorks), R version 3.3.1 (Pearson and Spearman correlation coefficients), and Microsoft Excel v16.10. For all microscopy measurements, the exact value of the number of cells used (*n*) and precision measurements used (mean ± SEM) is reported in the corresponding figure legends and also in Figures S4D, S6B.

2.8.4 DATA AND SOFTWARE AVAILABILITY

Detailed SPRITE protocols are available at http://guttmanlab.caltech.edu/protocols.php. SPRITE software is available at https://github.com/GuttmanLab/sprite-pipeline/wiki. All datasets reported in this paper are available at the Gene Expression Omnibus with accession number GEO: GSE114242.

2.9 SUPPLEMENTAL ITEMS

Table S1. All enriched SPRITE intra-chromosomal *k-mers* and SPRITE Enrichments of top GAM Triplets in mES cells, Related to Figure 2. We report the following statistics for each *k*-mer: (i) The number of individual SPRITE clusters containing the observed *k*-mer (observed count) (ii) number of SPRITE clusters containing the *k*-mer normalized by the cluster size (normalized observed score), (iii) the frequency the *k*-mer is observed relative to the average permutation computed on the cluster-size normalized scores (observed/expected), (iv) the largest permutation value observed (max permuted score), (v) percent of random permutations that the observed value exceeds (percent greater), (vi) total linear distance spanned by a *k*-mer (range), (vii) the number of reads contained in any of the *k* bins (cumulative coverage). **(A)** All *k*-mers at 1 megabase resolution that were observed in at least 5 independent SPRITE clusters, at an observed frequency that exceeded 90% of the random permutations, and occurred at least 4-times more frequently than the average of the permuted regions. Each *k*-mer was randomly permuted 10 times in a manner that preserves genomic distance between all regions within the *k*-mer. **(B)** All *k*-mers observed in SPRITE clusters that contain reads from three different A compartment regions that span at least 100Mb within an individual chromosome. These *k*-mers were randomly permuted 100 times in a manner preserving genomic distance between these regions. **(C)** All *k*-mers observed in SPRITE clusters that contain reads four different A compartment regions that span at least 100Mb within an individual chromosome. These *k*-mers were randomly permuted 100 times in a manner preserving genomic distance between these regions. **(D)** SPRITE enrichments for all top TAD triplets and super-enhancer TAD triplets previously identified by Beagrie et al. 2017 (Beagrie et al., 2017). Each *k*-mer was randomly permuted 10 times in a manner that preserves genomic

distance between all regions within the *k*-mer. **(E)** Enriched SPRITE *k*-mers of the top TAD triplets and super-enhancer TAD triplets previously identified by Beagrie et al. 2017 (Beagrie et al., 2017). Each *k*-mer was randomly permuted 10 times in a manner that preserves genomic distance between all regions within the *k*-mer. Enrichments are reported for top GAM *k*-mers that occurred at least 2-times more frequently than the average of the permuted regions in SPRITE clusters containing up to 100 reads.

Table S2. Contacts captured and read coverage by SPRITE clusters of different sizes compared to Hi-C and annotated Active and Inactive Hub regions in mES cells, Related to Figure 3. (A) Fraction of contacts identified within specific structures in the nucleus such as within TADs, local compartments regions, and non-local compartment regions and read coverage in A and B compartments in different SPRITE cluster sizes (2–10, 11–100, 101–1000, 1000+ read clusters) compared to Hi-C in mES cells (Dixon et al., 2012). **(B-E)** Annotated genomic DNA regions defined as inactive/nucleolar hubs (B,D) or active/speckle hubs (C,E) in mES cells (mm9) (B-C) and GM12878 cells (hg19) (D-E) at 1Mb resolution.

Table S3. *k*-mers spanning three or more nucleolar hub regions on different chromosomes, Related to Figure 4. Enrichments of *k*-mers containing at least three distinct nucleolar hub regions that were present on different chromosomes compared to inter-chromosomal *k*-mers that preserve the size of the hub regions and were randomly permuted 100 times while preserving inter-chromosomal spacing. Column definitions are the same as in Table S1.

Table S4. *k*-mers spanning three or more active hub regions on different chromosomes, Related to Figure 5. Enrichments of *k*-mers containing at least three distinct active hub regions that were present on different chromosomes compared to inter-chromosomal *k*-mers that preserve the size of the hub regions and were randomly permuted 100 times while preserving inter-chromosomal spacing. Column definitions are the same as in Table S1.

Table S5. Sequences of SPRITE barcoded adaptors used, Related to Figure 1. Barcoded DPM, Odd, Even, Terminal adaptors that were used in this study and their sequences are listed in this table. The “top” and “bottom” strands for each annealed dsDNA oligo are reported. The indexed primer barcodes used for Illumina library prep are also provided.

Table S6. Supplemental Information Related to STAR Methods. (A) SPRITE reads during all steps prior to cluster generation. **(B)** SPRITE reads in all steps prior to all steps of cluster generation for RNA-DNA tagged libraries. RNA and DNA reads, identified by a read containing an RNA-specific or DNA-specific tag are each processed separately. **(C)** DNA FISH probes used in this paper - Histochoice fixation. **(D)** Table M4: DNA seqFISH probes used in this paper - 4% Formaldehyde fixation. **(E)** Correlation between SPRITE contact frequencies with the nucleolar hub and 3D distances to the nucleolus measured by microscopy for different cluster sizes. SPRITE is highly correlated with DNA FISH for most cluster sizes, regardless of weighting or no weighting by cluster size. R-values >0.85 in bold. * represents the values currently shown in the analyses for Figure 4F.

Movie S1. Nucleolar hub regions co-associate around the same nucleolus in an individual cell, Related to Figure 4. Nucleolar hub regions on chromosome 15 (orange) and 18 (green) co-associate around the same nucleolus (red), marked by Nucleolin using immunofluorescence. DAPI is used to stain DNA in the nucleus.

Chapter 3

SPRITE: A genome-wide method to map higher-order 3D spatial interactions in the nucleus using combinatorial split-and-pool barcoding

Sofia A. Quinodoz, Prashant Bhat, Peter Chovanec, Joanna W. Jachowicz, Noah Ollikainen,
Elizabeth Detmar, Elizabeth Soehalim, Mitchell Guttman

A modified version of this chapter was submitted as "SPRITE: A genome-wide method to map higher-order 3D spatial interactions in the nucleus using combinatorial split-and-pool barcoding" to *Nature Protocols*.

3.1 ABSTRACT

A fundamental question in gene regulation is how DNA is packaged within the nucleus of each cell to influence cell-type specific gene expression. We recently developed Split-Pool Recognition of Interactions by Tag Extension (SPRITE) that enables mapping of higher-order interactions that occur within the nucleus. SPRITE works by crosslinking interacting DNA molecules and mapping their spatial arrangements through an iterative split-and-pool barcoding method. Specifically, all molecules within a crosslinked complex are barcoded by repeatedly splitting all complexes across a 96-well plate, ligating DNA molecules with a unique tag sequence, and pooling all complexes into a single well before repeating the tagging again. Because all molecules in a crosslinked complex are covalently attached, they will sort together throughout each round of split-and-pool and will obtain the same series of SPRITE tags, which we refer to as a barcode. The DNA and their associated barcodes are sequenced, and all reads sharing identical barcodes are matched to reconstruct interactions. SPRITE accurately maps pairwise DNA interactions within the nucleus and measures higher-order spatial contacts up to thousands of simultaneously interacting molecules. Here, we provide a detailed protocol describing how to perform SPRITE, which includes cell crosslinking, chromatin fragmentation, end repair, split-pool barcoding, PCR amplification of SPRITE libraries, and high-throughput sequencing. We demonstrate the experimental steps of SPRITE with a detailed video protocol. Furthermore, we provide an automated computational SPRITE pipeline available on GitHub that allows experimenters to seamlessly generate SPRITE interaction matrices starting with raw fastq files. The protocol takes approximately 4–5 days from cell crosslinking to high-throughput sequencing for the experimental steps and 1 day for data processing.

3.2 INTRODUCTION

Although all cells in an organism share the same genomic DNA sequence packaged within the nucleus of each cell, different genes are accessed and expressed across these different cell types. It has become increasingly clear that the three-dimensional organization of the genome plays a crucial gene regulatory role in various biological processes, including during embryonic development and cellular differentiation[1–4]. Current methods for genome-wide mapping of 3-dimensional (3D) genome structure rely on proximity ligation[5–7] (e.g., Hi-C), which works by ligating the ends of DNA regions that are in close spatial proximity in the nucleus followed by sequencing to map pairwise interactions. These techniques have revealed that the genome is organized into structures ranging from larger chromosome territories down to smaller features such as compartments[2,5], TADs)[8], and loops[9].

Although proximity-ligation methods have led to important progress in understanding the multiple layers of 3D genome organization, these approaches are limited to identifying DNA interactions that are close enough to directly ligate[10]. Accordingly, these techniques often fail to detect known larger-scale structures identified by microscopy, such as interactions that occur around nuclear bodies. For example, DNA interactions occurring between multiple DNA regions simultaneously co-associating around nuclear bodies, which can range in size from 0.5 to 2 μm, may therefore be too far apart to directly ligate[11]. As a result, these larger-scale interactions may be missed by proximity-ligation methods. Additionally, proximity-ligation-based methods are primarily limited to measuring pairwise interactions between primarily 2 genomic loci, and therefore cannot measure how many DNA sites simultaneously organize within structures in the nucleus. Accordingly, current methods are limited in their ability to generate comprehensive global models of how multiple genomic loci are arranged in 3D space. These limitations make it difficult to answer long-standing questions such as which DNA regions simultaneously co-localize around the same nuclear body, how DNA is organized in various features of the genome (i.e., compartments, TADs, loops), and why some genes exhibit preferential association with distinct nuclear bodies, but other genes do not.

To address these limitations, we developed SPRITE, a proximity-ligation-independent method that enables genome-wide detection of multiple DNA interactions that co-occur within the nucleus[12]. Using SPRITE, we identify genome structures previously discovered by Hi-C, including chromosome territories, compartments, TADs, and loop structures. Furthermore, because SPRITE is not constrained to mapping pairwise interactions, we were able to identify many DNA interactions that occur within higher-order structures in the nucleus. Finally, because SPRITE is not constrained to mapping interactions that are close enough to directly ligate, we can map the landscape of long-range intra- as well as inter-chromosomal interactions that organize around nuclear bodies, including the nucleolus and nuclear speckles.

3.3 OVERVIEW OF SPRITE

SPRITE is a method used for unbiasedly mapping genome-wide higher-order interactions between DNA molecules. Briefly, SPRITE works as follows (**Figure 1**):

Day 1:

(1) **Crosslinking.** Cells are crosslinked using a combination of formaldehyde and a protein–protein crosslinker called disuccinimidyl glutarate (DSG).

(2) **Lysis.** Cells are lysed, and the DNA is fragmented using sonication and DNasing.

(3) **NHS coupling.** Crosslinked lysate is covalently coupled to NHS-ester beads overnight.

Day 2:

(4) **End repair and dA-tailing.** DNA is blunt-ended and dA-tailed through treatment with End Repair and dA-tailing enzymes.

(5) **DPM ligation.** DPM adaptor is ligated to the ends of end-repaired DNA prior to four additional rounds of split-pool barcoding.

(6) **Split-pool barcoding.** Complexes coupled to beads are split into a plate of 96 unique tag adapters. Ligase adds each of the tagged adapters to the DPM adapter. The sample is then pooled into a single reservoir, and split again into a 96-well plate for another round of tagging. 5 total

rounds of split-pool barcoding are performed to achieve over 8 billion unique combinations of sequences. The entire split-pool barcoding procedure can be completed in a single day.

Day 3:

(7) **Reverse crosslinking.** Proteinase K is added to digest proteins and samples are incubated overnight to reverse crosslinks.

(8) **Library preparation.** PCR is performed to add sequencing adaptors and amplify the libraries prior to sequencing.

Day 4: Sequencing and analysis

(9) **Sequencing.** High-throughput sequencing identifies DNA molecules that were present in the same crosslinked complex.

(10) **Cluster Generation.** Reads sharing the same barcode are matched and aligned to the genome.

(11) **Analysis.** Using the SPRITE clusters, pairwise and higher-order maps of DNA interactions in the nucleus are constructed at various resolutions.

3.4 ADVANTAGES OF SPRITE

Higher-order spatial interactions:

Because SPRITE is not constrained to mapping pairwise interactions, one of the advantages of SPRITE is that it can measure interactions between multiple molecules interacting in the same three-dimensional space. This enables construction of higher-order spatial interactions between DNA molecules such as those including multiple genes involved in a shared biological process (e.g., histone gene clusters). This is in contrast to similar genome-wide mapping methods, such as HiC, which are restricted to mapping pairwise interactions.

Global Spatial Maps:

Because SPRITE utilizes a proximity-ligation-independent method to detect interactions, it is not limited to identifying only those molecules that are close enough to directly ligate. Instead, larger SPRITE clusters can detect long-range interactions including those occurring between regions on separate chromosomes. For example, SPRITE clusters containing over 1000+ DNA reads in a crosslinked complex identify longer-range interactions occurring between multiple DNA loci that are simultaneously associating around the nucleolus, as well as other SPRITE clusters containing 10–100 reads can identify inter-chromosomal interactions between genomic sites associating with nuclear speckles. This ability of SPRITE to measure interactions across longer-range distances enables measurement of crosslinked complexes of different sizes in the nucleus to reconstruct various close-range and longer-range interactions in the nucleus.

3.5 APPLICATIONS OF SPRITE

We and others have already applied SPRITE to many different cell types commonly used, including fibroblasts, ESCs, immune cells, and neuronal cells. SPRITE can be applied to model systems including *A. thaliana, C. elegans, Drosophila, Xenopus*, zebrafish, and so forth. SPRITE can be applied in these and other model organisms to investigate a variety of biological contexts, including, but not limited to, enhancer-promoter interactions, multi-way contacts between chromosomal regions, interactions with nuclear bodies, and inter-chromosomal hubs of gene activity within the nucleus. These studies will establish principles of 3D spatial interactions across evolutionary time.

Broadly speaking, SPRITE represents a robust framework to map spatial interactions of other biomolecules (e.g., RNA and protein) and how they co-occur in 3D space. For example, SPRITE can be applied to explore spatial interactions beyond the nucleus, such as the RNA composition of biomolecular condensates like stress granules in the cytoplasm[13]. Other spatial organization of different molecules can also be explored. For example, SPRITE can be extended to map protein localization using a pool of barcoded antibodies[14,15] to generate combinatorial and spatial maps of DNA, RNA, and/or protein. Such applications of SPRITE will extend our ability to research 3D spatial mapping of these biomolecules that have long remained a challenge.

Genome assembly: Constructing an accurate reference genome for a model organism is critical for those doing sequence analysis. Recently, Hi-C has been used to construct genomes[16,17] given the fact that most interactions between DNA molecules occur within, as opposed to between chromosomes. SPRITE similarly can be applied to genome assembly generation but has the added benefit of multi-way information.

3.6 EXPERIMENTAL DESIGN

Cell culture:

Cells can be cultured using standard guidelines for the cell type best suited for the experimenter’s needs. We performed SPRITE on F1-21 mESCs and human GM12878 lymphoblast cells. Details about cell culturing conditions can be found in Quinodoz et al. Cell 2018[12].

Adaptor and Barcode Design:

The SPRITE adaptors and tag ligation scheme that is central to the SPRITE process are described in **Box 1**.

Cell number:

While we typically crosslink 5–10M cells for a SPRITE experiment, the final amount of material inputted into the SPRITE split-and-pool tagging steps corresponds to DNA recovered from approximately 3000 cells. However, we typically work with more cells during the crosslinking steps as (i) loss can occur during the crosslinking procedure and (ii) DNase optimization steps require enough DNA to test multiple dilutions of DNase and visualization in a gel. We recommend ensuring that at least >50–100ng of DNA is used for DNase testing to quality control (QC) DNA sizes post-DNasing by gel electrophoresis with at least 4 DNase concentrations tested. We have successfully generated SPRITE heatmaps and libraries from approximately 200,000 cells (recovered from flow cytometry) and can typically crosslink 1–2M cells with enough DNA to do several tests for DNA sizes post-sonication and DNase chromatin fragmentation.

Crosslinking and cell lysis:

Cells are dual crosslinked with 2mM disuccinimidyl glutarate (DSG) and 3% formaldehyde which favors fixing larger chromatin complexes and longer-range interactions. We have tested reducing crosslinking to 1% formaldehyde combined with DSG and also obtain successful SPRITE libraries with known interactions such as TADs, compartments, and chromosome territories. We note that the current fragmentation conditions have been optimized for cells crosslinked with 3% formaldehyde and DSG. Crosslinking with 1% or 3% formaldehyde without DSG will work, but applying the same amount of sonication (1 minute, 4–5 W on a Branson Sonicator) will result in SPRITE libraries almost completely devoid of interactions. We recommend optimizing fragmentation if crosslinking conditions are varied to ensure interactions remained during sonication. After crosslinking, cells are lysed with a nuclear isolation protocol modified from the Amit Lab HT-ChIP protocol[18] prior to chromatin fragmentation.

To prevent cells from clumping in these strong crosslinking conditions, we recommend resuspending the crosslinked pellet uniformly in a smaller volume (1 mL) of DSG crosslinking solution using a P-1000 micropipette before adding the full volume of crosslinking solution. If the full volume of DSG crosslinking solution is added without first resuspending the pellet, it will be almost impossible to completely break up the pellet and will result in cell clumps being crosslinked together. We recommend checking nuclei integrity under a microscope before starting SPRITE on a new cell type.

Chromatin fragmentation:

SPRITE fragmentation is performed using a light sonication followed by DNasing to obtain DNA sizes for DNA amplification. Optimization of fragmentation conditions (amount of sonication, amount/timing of DNase) is a critical step in establishing the protocol for the first time. The length of sonication might vary from 30 seconds to several minutes depending on the sonicator used. We have found that sonication time is critical to break open nuclei into smaller crosslinked structures containing many DNA molecules crosslinked together. Over-sonication results in libraries devoid of any multi-way DNA interactions (singlet SPRITE clusters), whereas under-sonication can result in libraries where entire cells are tagged rather than smaller-scale structures in the nucleus. See **Box 2** for details on how to assay cell fragmentation using microscopy prior to SPRITE and the relationship between sonication and SPRITE cluster sizes. We also note that certain ChIP protocols and other chromatin fragmentation protocols perform high-speed spins following sonication, this will result in loss of SPRITE clusters and cause loss of multi-way DNA interactions in SPRITE.

DNase treatment might vary in Turbo DNase concentration, depending on cell number, ploidy, crosslinking strength, and the desired DNA fragment size. To optimize DNase timing and conditions, DNase samples with varying enzyme concentration for 20 minutes, immediately quench with EDTA and EGTA on ice, and assay DNA sizes for concentration as described in the protocol. We recommend optimizing DNA sizes primarily ranging between 100bp-1kb (**Box 2c**). If an appropriate combination of solubilization and DNA fragment sizes cannot be obtained by varying the amount of sonication or DNase, then the amount of lysate going into an experiment can be reduced and the strength of the crosslinking may also be reduced.

NHS coupling:

Crosslinked lysate is coupled to NHS-activated magnetic beads overnight. It is important to measure the concentration of DNA in the lysate prior to NHS-bead coupling to ensure that multiple crosslinked complexes do not become covalently coupled to the same bead, which would result in spurious interactions (**Box 3**). The NHS beads allow us to 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 are performed with shaking at either 1200 rpm or 1600 rpm using an Eppendorf Thermomixer to avoid bead settling and aggregation. After each step in the protocol, all enzymatic steps are inactivated by adding 1 mL modified RLT buffer and then buffer exchanged with 3 washes of 1mL SPRITE wash buffer onto the NHS beads. Washing steps shown described in detail, including recommendations on how to avoid bead loss during each step, in the video protocol (**Video 1**).

DPM adapter ligation:

After chromatin fragmentation, DNA is blunt-ended and dA-tailed using DNA end repair and dA-tailing enzymes. This single dATP is necessary to ligate the DPM adapter in the next step. We highly recommend performing a DPM ligation QC (see **Box 4**) prior to starting split-pool barcoding to ensure DPM ligation was successful and that there are enough DNA molecules ligated with DPM for the subsequent SPRITE ligations.

Split-pool barcoding and final library amplification

The SPRITE method works by splitting crosslinked lysate across a 96-well plate (**Figure 2**). Each well of the 96-well plate contains a unique tag (DPM) to which the DNA molecules are ligated.

The ligation reactions are stopped, pooled, and split again into a new 96-well plate containing different, unique tags than the first (Odd). This is repeated over multiple rounds, where the Odd and Even tags are designed to be alternated over multiple rounds of ligation and are named as such as they ligate the 1st, 3rd, 5th,... and 2nd, 4th, and 6th,... rounds, respectively. If n rounds of tag ligation are performed, 96^n unique barcodes are generated. We typically ligate 5 tags, creating over 8 billion unique barcodes. To reduce the cost of barcoded oligos purchased for SPRITE, a single Odd and Even plate of tags may be purchased and ligated over multiple rounds, rather than purchasing an additional set of barcoded DPM and Terminal barcode.

A final step of barcoding involves splitting the beads after ligation of the Terminal tag into smaller aliquot sizes, reverse crosslinking overnight, column purification of the DNA, and PCR amplification of the barcoded DNA using distinct Illumina sequencing primers for each SPRITE aliquot. This effectively serves as another round of split-pool barcoding as each library is tagged with a unique barcoded Illumina primer. This further reduces the probability that molecules in different clusters obtain the same barcodes.

SPRITE Data Processing, Cluster Generation, and Heatmap Generation

SPRITE libraries (**Figure 3a**) can be sequenced on any sequencing platform. The key parameters are ensuring that read length is long enough to read all barcodes and its associated genomic DNA regions. We sequenced using Illumina paired-end sequencing on the HiSeq 2500 or NextSeq 500 with at least 95 x 100 bps. The reads have the following structure: Read 1 contains genomic DNA positional information and the DPM tag, and Read 2 has the remaining Odd, Even, and Terminal tags.

The SPRITE computational pipeline is as follows (**Figure 4a**):

Adaptor trimming: Illumina adaptors are removed using Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). To remove DPM sequences and account for short genomic fragments that could lead to readthrough and the presence of tag sequences on the 3' end, we perform a second round of trimming using Cutadapt[19].

Barcode identification: This step identifies the barcodes contained on each of the sequenced reads. SPRITE barcodes are identified by parsing the first DNA tag sequence from the beginning of Read 1 and the remainder of the tags are parsed from Read 2. We identify these tag sequences by searching each read for the sets of DPM, Odd, Even, and Terminal tags that were ligated to these samples. We allowed for up to two mismatches for each Odd and Even tags to account for possible sequencing errors. Because the tags were designed to contain at least four mismatches to any other tag sequence, this enables robust barcode identification. For DPM and Terminal tag alignments, we do not tolerate any mismatches due to their shorter unique barcode sequences. After barcode identification, the barcode string is appended to each read name in the FASTQ file. We exclude any reads that did not contain a full set of all ligated tags (DPM, Odd, Even, Odd, Terminal) in the order expected from the experimental procedure.

Cutadapt DPM trimming: To ensure reads are properly aligned to genomic DNA, DPM sequences are trimmed from the beginning of Read 1.

Bowtie 2 Genome alignment: We align each trimmed read to the mouse or human reference genomes (mm10 for mouse and hg38 for human) using Bowtie2. The corresponding SAM file is sorted and converted to a BAM file using SAMtools.

Filter repeats, blacklisted regions: We filter the resulting BAM file for low quality reads, multimappers, and repetitive sequences. First, we remove all alignments with a MAPQ score less than 20. Then, we remove all alignments overlapping a blacklisted[20] or a genomic region masked by Repeatmasker (UCSC, milliDiv < 140) using bedtools[21].

Generate SPRITE clusters: To define SPRITE clusters, all reads that share the same barcode sequence are matched and grouped into a single cluster. To remove possible PCR duplicates, all reads within a cluster containing the same genomic position are removed. We generate a SPRITE cluster file for each SPRITE aliquot separately, as interactions can only occur between DNA molecules derived from a single SPRITE aliquot rather than between different SPRITE aliquots. All subsequent analyses are performed on the SPRITE cluster file, where each cluster occupies one line of the resulting text file containing the barcode name and genomic alignments. SPRITE clusters capture various multi-way contacts between multiple reads that all share the same barcode. These clusters range in size from 2–10 reads in a single SPRITE cluster up to >1000 reads in a single SPRITE cluster (**Figure 3c**).

Make heatmap matrix: The cluster file is used to generate a pairwise contact matrix. This can be visualized as a heatmap. Heatmaps are generated by enumerating all pairwise contacts within each SPRITE cluster and either with or without down-weighting of individual contacts. Because the number of pairwise contacts enumerated from a SPRITE cluster scales quadratically with the number of reads, down-weighting each contact by the cluster size ensures that larger clusters do not dominate the heatmap (**Figure 4b**). We recommend using "n-over-two" weighting for initial heatmap generation to visualize known genome structures (e.g., TADs, compartments, chromosome territories).

Visualizing SPRITE clusters: We identified multi-way interactions by counting how often multiple genomic regions simultaneously interact in individual SPRITE clusters. In **Figure 5**, rows show individual SPRITE clusters where multiple regions simultaneously interact. Specifically, black boxes correspond to genomic bins (typically 1Mb or 25kb in size) that contain at least one read in a given SPRITE cluster. We observe multi-way contacts corresponding to interactions across a range of genomic resolutions, including A compartments, TADs, and chromatin loops.

Expertise Needed to Implement the Protocol

The SPRITE procedure can be performed by any molecular biology lab with access to the reagents and equipment described in detail in the reagents and equipment sections. Access to a sequencing facility is needed to sequence final SPRITE libraries. Furthermore, using the SPRITE computational pipeline requires intermediate level experience installing and running data analysis software.

Limitations

One of the limitations of SPRITE is the amount of sequencing depth needed to identify DNA structural patterns like chromosome loops. For many genomic architecture mapping methods, it remains a limitation of high-resolution features the need to sequence deeply. For example, we need approximately 5 million reads to visualize chromosome territories, 25 million reads to visualize compartments, 200 million reads for TADs, and 1 billion reads for loop interactions. We imagine that enrichment procedures could be implemented, similar to those implemented for ChIA-PET, Hi-ChIP, or PLAC-seq[22–24], to select for substructures of interest for specific needs of the experiment (i.e., enhancers and promoters).

A second limitation of SPRITE is the narrow window of chromatin fragmentation optimization to get meaningful contacts. Over-fragmentation results in sparse clusters versus under-fragmentation results in clusters too large to generate high confidence interactions (see **Box 2a-b**). We highly recommend optimizing and QCing the fragmentation conditions when working with a new cell line, sonicator machine, or other variables that may affect DNA size.

There is also a high upfront cost associated with SPRITE, in particular the 96-well plate of DPM adapters which must be ordered with a 3' Spcr modification. To reduce the cost, one may modify the protocol to use a single DPM adapter sequence, performing DPM in a single tube before commencing with split-pool barcoding. Importantly, in this scenario the experimenter will need to utilize an additional round of barcoding to generate enough combinatorial complexity to distinguish unique complexes as well as increasing the sequencing length on read 2 to sequence an additional barcode. Terminal adaptor ligation may also be performed in a single well in exchange for an extra round of Odd or Even ligation; however, for the Terminal ligation we recommend purchasing 4 different Terminal adaptors of variable lengths from the set of 96 tags provided in **Supplemental Table S2** to introduce a stagger and prevent any issues during sequencing from low-complexity sticky-end sequences (see **Box 1**). However, wherever possible, we recommend performing SPRITE as described with the first round of barcoding being the 96-well plate of DPM adapters.

3.7 REAGENTS

- SPRITE tags (Supplemental Table S2; IDT, custom order)
    - CRITICAL: Order all plates of SPRITE tags (DPM, Odd, Even, and Terminal) resuspended to 200 µM in nuclease-free water. Use extreme care to avoid cross-contamination between wells at all times when working with stock plates. Always centrifuge stock plates at 1000 xg for 1 minute prior to opening.
- Indexing SPRITE Library Amplification primers (Supplemental Table S2; IDT, custom order)
- Buffer RLT (Qiagen, cat. no. 79216)
    - CAUTION: Modified RLT Buffer contains guanidine thiocyanate which when mixed with bleach produces hydrogen cyanide gas and hydrogen chloride gas. Be careful to ensure that all liquid modified RLT buffer waste is disposed of in its own waste container. Solids that have touched modified RLT buffer such as tips and reservoirs should also be discarded in a separate solid modified RLT buffer container.
- Calcium chloride solution (Sigma-Aldrich, cat. no. 21115)
- Manganese chloride solution (Sigma-Aldrich, cat. no. M1787)
    - CRITICAL: It is imperative that manganese chloride is used for DNase digestion and NOT Turbo DNase buffer that comes with the kit. Mn(II) generates the appropriate DNA ends necessary for end repair and further downstream library prep steps.
- EDTA (0.5 M, pH 8.0; ThermoFisher Scientific, cat. no. 15575020)
- EGTA (0.5 M, pH 8.0; Fisher Scientific, cat. no. 50255957)
- Glycerol (Sigma-Aldrich, cat. no. G5516)
- Sodium chloride
- Distilled water (DNase/RNase-free; Thermo Fisher Scientific, cat. no. 10977015)
- Triton X-100 Detergent (Sigma-Aldrich, cat. no. T8787)
- NP-40 Surfact Amps Detergent (Thermo Fisher Scientific, cat. no. 28324)
- N-Lauroylsarcosine sodium salt solution (Sigma-Aldrich, cat. no. L7414)
- Tris-HCl buffer (1 M, pH 7.5; Thermo Fisher Scientific, cat. no. 15567027)
- Tris-HCl buffer (1 M, pH 8.0; Thermo Fisher Scientific, cat. no. 15568025)

- Sodium deoxycholate (Sigma-Aldrich, cat. no. D6750)
- Lithium chloride solution (8M; Sigma-Aldrich, cat. no. L7026)
- SDS (20% solution; Thermo Fisher Scientific, AM9820)
- HEPES Buffer pH 7.4 (Teknova, cat. no. H1030)
- RNase-Free BSA (American Bio, cat no. AB01243-00050)
- PBS 7.4 (1x), No Calcium, No Magnesium, Liquid (Thermo Fisher Scientific, cat. no. 10010049)
- Disuccinimidyl glutarate (DSG) (50 mg, Thermo Fisher Scientific, cat. no. 20593)
- Dimethyl sulfoxide (DMSO) (Sigma-Aldrich, cat. no. D2650)
- Formaldehyde Ampules (16%, methanol-free; Thermo Scientific Pierce, cat. no. PI28908)
- Glycine, >99% (Sigma-Aldrich, cat. no. G7403-250G)
- Protease Cocktail Inhibitor tablets (Sigma-Aldrich, cat. no. 04693159001)
- Turbo DNase (2 U/μL; ThermoFisher Scientific, cat. no. AM2238)
- Proteinase K (800 U/mL, New England Biolabs, cat. no. P8107S)
- PierceTM NHS-Activated Magnetic Beads (Thermo Fisher Scientific, cat. no. 88827)
    - CRITICAL: We strongly recommend sealing opened bottles with parafilm and storing the bottle with desiccant in a centrifuge tube. Magnetic beads are moisture-sensitive.
- NEBNext End Repair Module (contains End Repair Enzyme Mix and Reaction Buffer; New England Biolabs, cat. no. E6050)
- Klenow Fragment (3' to 5' exo-) (New England Biolabs; cat. no. M0212)
- NEBNext dA-Tailing Module (New England Biolabs; cat. no.E6053)
- dA-Tailing Reaction Buffer (New England Biolabs; cat. no. B6059S)
- Instant Sticky-end Ligase Master Mix (New England Biolabs; cat. no. M0370)
- NEBNext Quick Ligation Reaction Buffer (New England Biolabs; cat. no. B6058S)
- 1,2-Propanediol (25 mL; Sigma-Aldrich, cat. no. 398039)
    - CRITICAL: Protect 1,2-Propanediol from light by placing inside it inside a drawer.
- Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs, cat. no. M0494)
- Agencourt AMPure XP Magnetic Beads (Beckman Coulter, cat. no. A63880)
- Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. no. Q32854)
- (Optional) Poly-D-lysine hydrobromide (Sigma, cat. no. P6407-5MG)

- (Optional) ProLong Gold Antifade Mountant with or without DAPI (Thermo Fisher Scientific; cat. no. P36935)
- (Optional) Micro Slides (VWR® Superfrost® Plus Micro Slide, Premium; cat. no. 48311-703)
- (Optional) Hoechst 33342 Solution (Thermo Fisher Scientific; cat. no. 62249)
- (Optional) SYBR™ Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific; cat. no. S11494)
- (Optional) YOYO™-3 Iodide (Thermo Fisher Scientific; cat. no. Y3606)

Mouse embryonic stem cell (mES) Cell Culture Specific Reagents:

- Trypsin-EDTA (0.25%), phenol red (Thermo Fisher Scientific; cat. no. 25200-056)
- Chicken Serum, USA origin, sterile-filtered, cell culture tested (Sigma; cat. no. C5405-100ML)
- DMEM/F-12 (1:1) (Thermo Fisher Scientific; cat. no. 11330-057)
- Gibco BSA Fraction V (7.5%) (Thermo Fisher Scientific; cat. no. 15260037)

3.8 EQUIPMENT

- Branson needle-tip sonicator (1/8’’ Doublestep tip, 3 mm diameter; Branson Ultrasonics, cat. no. 101-148-063)
  - NOTE: Covaris M220 Focused-ultrasonicator (Covaris, cat. No. 500295) or other sonicators may also be used. Exact conditions will need to be optimized when using a new sonicator.
- Low-retention pipette tips (20 µL, 200 µL, and 1000 µL; Rainin cat. Nos. 30389226, 30389240, and 30389213)
- Eppendorf™ 96-Well twin.tec™ PCR Plates, 96 well, semi-skirted (Eppendorf, cat. no. 10049-108)
- Pipetting Reservoirs (VWR cat. no. 10015-232)
- PCR tubes (0.2 ml; USA Scientific, cat. no. 1402-4700)
- LoBind microcentrifuge tubes (1.5 ml; Eppendorf, cat. no. 022-43-108-1)

- Centrifuge tubes (15 ml and 50 ml; Genesee Scientific Corporation, cat. Nos. 28-103 and 28-108)
- RNA Clean and Concentrator-5 Kit with Capped Columns (Zymo Research, cat. no. R1018)
- Gel Electrophoresis System
- Gel Electrophoresis Equipment
- Benchtop microcentrifuge
- Plate Centrifuge
- High Sensitivity DNA Kit for the Agilent Bioanalyzer (Agilent Technologies, cat. no. 5067-4626)
- High Sensitivity D1000 Screentape and Reagents (Sample Buffer and Ladder) for the Agilent 2200 TapeStation (Agilent Technologies, cat. Nos. 5067-5584 and 5067-5585)
- Qubit Fluorometer 2.0 (Thermofisher Scientific, cat. no. Q32866)
- Thermal mixer C (Thermomixer C; Eppendorf, cat. no. Z605271)
- Eppendorf SmartBlock 1.5mL and PCR 96 thermoblocks (Eppendorf, part nos. 5360000038 and 5306000006)
- Magnetic racks for 1.5mL, 15 mL tubes (DynaMag-2 and DynaMag-15 Magnets; Thermofisher Scientific, cat nos. 12321D and 12301D)
- Magnetic rack for 96 well plate (DynaMag-96 Side Magnet; Thermo Fisher Scientific, cat. no. 12331D)
- Thermal cycler (Eppendorf Mastercycler pro; Eppendorf, cat. No 95043-592)
- Agilent Bioanalyzer (Agilent Technologies, model no. G2939A)
- Agilent 2200 Tapestation (Agilent Technologies, manual part no. G2964-90003 Rev. C)
- HulaMixer Sample Mixer (ThermoFisher Scientific; 15920D)
- Vortex mixer (Analog vortex mixer; VWR, model no. 58816-123)
- Sequencer (Illumina, model no. HiSeq 2500)
- Single channel Pipettes (P2, P20, P200, P1000)
- 12-well Multi-channel pipettes (P20, P200)
- Low-retention tips

3.9 COMPUTER, PROGRAMS, AND SOURCE CODE

- The SPRITE pipeline has been tested on a high performance computing cluster running CentOS 7 and a local environment with 30GB of RAM, an i7-8750H CPU running Ubuntu 18.04.3 LTS. Local runtime for fastq files with around 45 million reads was approximately 7 hrs.
- Snakemake pipeline software (https://snakemake.readthedocs.io/en/stable/)
- Conda package (https://docs.conda.io/projects/conda/en/latest/) or miniconda package (https://docs.conda.io/en/latest/miniconda.html)
- Python 3.7.3 (https://www.python.org/)
- Java 8 (https://www.java.com/en/download/)
- R software v3.6.1 (https://www.r-project.org/)
- Hi-Corrector v1.2 software (https://github.com/jasminezhoulab/Hi-Corrector)
- Bowtie2 v2.3.5
- Bedtools v2.29.0
- Multiqc v1.6
- Samtools v1.9
- Trim galore! V0.6.2
- Cutadapt v2.5
- Pigz v2.3.4
- Fastqc v0.11.8
- Python packages:
    - Pysam v0.15.0.1
    - Numpy v1.17.2
- R packages
    - Ggplot2 v3.1.1
    - Gplots v3.0.1.1
    - Readr v1.3.1
    - Optparse v1.6.2

3.10 REAGENT SET UP

10% Triton X-100 (vol/vol)

Prepare 50 mL of 10% (vol/vol) Triton X-100 by diluting 5 mL of Triton X-100 with 45 mL of nuclease-free water. Store at room temperature (22ºC) for up to one year.

10% DOC (wt/vol)

Prepare 10 mL of 10% (wt/vol) DOC by dissolving 1 gram of DOC in 10 mL of nuclease-free water. Protect from light and store at room temperature for up to one month.

CRITICAL: DOC is light sensitive. Cover dissolved DOC with aluminum foil and place inside a drawer to protect DOC from light.

0.5 M DSG

Prepare 0.5 M DSG by adding 306 μL DMSO to one bottle containing 50 mg DSG. Store remaining DSG -20ºC for up to one year.

Trypsin Versene Phosphate Buffer (TVP)

1 mM EDTA, 0.025% Trypsin, 1% Sigma Chicken Serum

Post-TVP Wash Solution

DMEM/F-12 supplemented with 0.03% Gibco BSA Fraction V.

2 mM DSG Crosslinking Solution

Make an appropriate amount of 2 mM DSG crosslinking solution by adding 16 μL of 0.5M DSG per 4 mL of room temperature 1xPBS.

3% Formaldehyde Crosslinking Solution

Prepare a fresh 3% FA solution by combining 750µL of 16% FA for every 3.25 mL room temperature 1xPBS.

Scraping Buffer

1 x PBS pH 7.5 with 0.5% BSA, *Store at 4*°C

Cell Lysis Buffer A

Prepare 10 mL of Cell Lysis Buffer A by mixing 500 µL of 1M Hepes pH 7.4, 200 µL of 0.5 M EDTA, 200 µL of 0.5 M EGTA, 280 µL of 5 M NaCl, 250 µL of 10% Triton X-100, 500 µL of 10% NP-40, 1 mL of 100% glycerol, and 7.07 mL of nuclease-free water.

Cell Lysis Buffer B

Prepare 10 mL of Cell Lysis Buffer B by mixing 500 µL of 1M Hepes pH 7.4, 30 µL of 0.5 M EDTA, 30 µL of 0.5 M EGTA, 400 µL of 5 M NaCl, and 9.04 mL of nuclease-free water.

Cell Lysis Buffer C

Prepare 10 mL of Cell Lysis Buffer C by mixing 500 µL of 1M Hepes pH 7.4, 30 µL of 0.5 M EDTA, 30 µL of 0.5 M EGTA, 200 µL of 5 M NaCl, 100 µL of 10% DOC, 250 µL of 20% NLS, and 8.89 mL of nuclease-free water.

✓ Note: EDTA and EGTA may be removed from Cell Lysis Buffer C depending on the application. DNA fragmentation has been optimized in this protocol with EDTA and EGTA.

100x Mn/Ca Mix

Make a 100x solution of MnCl2 and CaCl2 by mixing 250 µL of 1 M MnCl2, 50 µL of 1 M CaCl2 and 700 µL of nuclease-free water.

10x SPRITE DNase Buffer

Prepare 1 mL of 10x SPRITE DNase Buffer by mixing 200 µL of 1 M Hepes pH 7.4, 200 µL of 5 M NaCl, 50 µL of 10% NP-40, and 10 µL of 100x Mn/Ca mix, and 540 µL of nuclease-free water.

25x DNase Stop Solution

Prepare 1 mL of 25x DNase Stop Solution by mixing 500 µL of 0.5 M EDTA, 250 µL of 0.5 M EGTA, and 250 µL of nuclease-free water.

Hoechst Dye Stock Solution

Prepare the Hoechst dye stock solution by dissolving the contents of one vial (100 mg) in 10 mL of deionized water (diH2O) to create a 10 mg/mL (16.23 mM) solution.

✓ Note: Hoechst dye has poor solubility in water, so sonicate as necessary to dissolve. The 10 mg/mL Hoechst stock solution may be stored at 2–6°C for up to 6 months or at ≤–20°C for longer periods.

Hoechst Staining Solution

Prepare the Hoechst staining solution by diluting the Hoechst® stock solution 1:2,000 in 1xPBS.

SPRITE ProK Buffer

Prepare 10 mL of SPRITE ProK buffer by mixing 200 µL of 1M Tris-HCl pH 7.5, 200 µL of 5 M NaCl, 500 µL of 10% Triton X-100, 100 µL of 20% SDS, 200 µL of 0.5 M EDTA, 200 µL of 0.5 M EGTA, and 8.6 mL of nuclease-free water.

Coupling Buffer

Prepare 1 mL of coupling buffer by adding 5 ul of 20% SDS to 995 ul 1 x PBS.

Critical: Do not place this on ice as it will crash out.

Modified RLT Buffer

Prepare 50 mL of modified RLT buffer by mixing 47.8 mL of Qiagen Buffer RLT with 100 µL of 0.5 M EDTA, 100 µL of 0.5 M EGTA, 500 µL Tris-HCl pH 7.5, 500 µL of 10% Triton X-100, 500 ul of 10% NP-40, 500 µL of 20% NLS.

SPRITE Wash Buffer

Prepare 50 mL of SPRITE wash buffer by mixing 1 mL of 1 M Tris-HCl pH 7.5, 500 µL of 5 M NaCl, 1 mL of 10% DOC, 1 mL of 10% Triton X-100, 1 mL of 10% NP-40, and 45.5 mL of nuclease-free water.

Homemade SPRITE Ligation Mastermix (3x)

Prior to pipetting the viscous NEB Quick Ligation Reaction and Instant Sticky-end Ligation Mastermix, thaw all reagents on a room temperature rotator until they are completely in solution. Combine 1600 µL of 5x NEBNext Quick Ligation Reaction Buffer, 600 µL of 1,2-Propanediol, and 1000 µL 2x Instant Sticky-end Ligation Master Mix.

10X Annealing Buffer

Prepare 10x annealing buffer by mixing 250 µL of 8 M LiCl, 100 µL of 1 M Tris-HCl pH 7.5, and 650 µL of nuclease-free water.

Poly-D-Lysine Hydrobromide

Dilute to 0.1mg/mL following the Sigma recommendations.

2.5 M Glycine Stop Solution

Dissolve 9.38g of Glycine in nuclease-free water to 50mL. Store up to 1 month at room temperature.

90 µM Annealed SPRITE Tags (DPM, Odd, Even, and Terminal stock plates)

Generate stock plates of all annealed SPRITE adapters prior to starting split-and-pool steps. To generate a stock 96-well plate of SPRTIE tags, anneal each well of the “Top” plate with its corresponding “Bottom” plate (e.g., well A1 “DPM_Top” with well A1 “DPM_Bot”). Specifically combine 18 µL of each 200 µM “top” strand with 18 µL of the corresponding 200 µM “bottom” strand in a new semi-skirted LoBind 96-well plate. Add 4 µL of 10X annealing buffer to each well of the plate. Carefully seal the plate with foil and vortex gently for 30 seconds to mix. Centrifuge the plate at 1000 xg for 1 min. Anneal by heating the plate to 95^{o}C for 5 minutes, and slowly cooling to 25^{o}C in a thermal cycler with a heated lid (ramp -1^{o}C/s). Store at -20^{o}C.

4.5 µM Annealed SPRITE Tags (Diluted DPM, Odd, Even, And Terminal Stock Plates)

Generate 4.5 µM working stock plates by diluting 5 µL of each 90 µM annealed stock plate with 95 µL of 1X annealing buffer into a new semi-skirted LoBind 96-well plate. Carefully seal the plate with foil and vortex gently for 30 seconds to mix. Centrifuge the plate at 1000 xg for 1 min. Store at -20^{o}C.

4.5 µM SPRITE Plates (Working Stocks)

Centrifuge the 4.5 µM DPM, Odd, Even, and Terminal stock plates at 1000 xg for 1 min. Carefully remove the foil seal. Generate multiple working stock plates by aliquoting 2.4 µL of each well of the 4.5 µM SPRITE stock plates into its corresponding well of a new semi-skirted LoBind 96-well plate. We recommend making multiple aliquots that can be stored for multiple experiments. Carefully seal the plate with foil. Centrifuge the plate at 1000 xg for 1 min. Store at -20^{o}C.

3.11 PROCEDURE

3.11.1 CELL CULTURE AND CROSSLINKING (TIMING 3 HOURS)

1. Seed and culture adherent cells on 10–15 cm plates under recommended conditions. This protocol details crosslinking multiple plates of cells in one suspension, but it is important to maintain consistency in lysate batches. We typically freeze 5–10 million cells per pellet in a 1.7 mL microcentrifuge tube.
2. For mESCs, the cells are lifted with TVP trypsin solution. An hour before starting, warm TVP and TVP wash solution at 37°C.
3. Pre-chill one bottle (≥100mL) of 1 x PBS (without Magnesium and without Calcium) at 4°C, keep one bottle (≥100mL) of 1 x PBS at room temperature. Store scraping buffer at 4°C.
4. Aspirate media from plates.
5. Wash cells gently with 10 mL of 1x PBS.
6. Remove PBS. For mESCs, cells are lifted by adding 5 mL TVP to each 15 cm plate and rock gently for 3–5 minutes at 37°C until cells begin to detach from the plate **?Troubleshooting**
7. Add 25 mL wash solution to each plate. Vigorously resuspend cells in the wash solution and transfer from the plate to a 50 mL centrifuge tube. Rinse the plate with extra wash solution and add to the 50 mL centrifuge tube. The same cells from different plates can be pooled into the same tube.
8. Centrifuge the cells for 3 minutes at 330xg at room temperature. Wash cells by resuspending in 4 mL room temperature 1 x PBS per 10 million cells and transfer to a 15 mL conical.
9. Centrifuge the cells for 3 minutes at 330xg.
10. Remove supernatant. Resuspend cells in DSG crosslinking solution (4 mL per 10 million cells) ✖. Rotate gently at room temperature for 45 minutes.

    ✖CRITICAL: It is vital that at the beginning of the crosslinking process the pellet is uniformly in suspension. To achieve this, completely resuspend the pellet in 1 mL of DSG crosslinking solution using a P-1000 micropipette. After the pellet is completely dissolved, add the remaining volume of DSG crosslinking solution. If you add the full volume of DSG crosslinking solution without first resuspending the pellet, it will be almost impossible to completely break up the pellet and will result in cell clumps being crosslinked together.

11. Centrifuge cells for 4 minutes at 1000xg at room temperature. Discard supernatant.
12. Wash cells with 4 mL room temperature 1 x PBS per 10 million cells ✖.
13. Centrifuge cells for 4 minutes at 1000xg at room temperature. Discard supernatant.

    ✖CRITICAL: As before, each time you resuspend the cell pellet, whether to wash or to put in formaldehyde or scraping buffer, ensure that the pellet is completely resuspended in the solution. Achieve this by first resuspending the pellet in 1 mL of the appropriate solution using a P-1000 micropipette, then adding the remaining volume.
14. Resuspend cell pellet in freshly prepared 3% formaldehyde solution (4 mL per 10 million cells). Rock gently at room temperature for 10 minutes.

    ✖CRITICAL: We strongly recommend making working solution of 3% formaldehyde FRESH every time by opening a new ampule to minimize methanol conversion. Do not use an ampule opened more than ~30 min before crosslinking the cells.
15. Add 200 µL of fresh 2.5 M glycine stop solution ✖ per 1 mL of cell suspension.

    ✖ CRITICAL: Ensure that the 2.5 M glycine stop solution was made within a month from its use.
16. Rock gently at room temperature for 5 minutes.
17. Centrifuge cells at 4°C for 4 minutes at 1000xg.
18. Discard formaldehyde supernatant in an appropriate waste container. From here, keep cells at 4°C.
19. Resuspend cell pellet in 4°C scraping buffer and gently rock for 1–2 minutes.
20. Pellet cells at 4°C for 4 minutes at 1000xg. Discard supernatant in formaldehyde waste container.
21. Resuspend cell pellet in cold scraping buffer again and gently rock for 1–2 minutes. Pellet as before and discard supernatant.
22. Resuspend pellet in 1 mL of scraping buffer per 10 million cells.
23. Aliquot 10 million cells each into 1.7 mL microcentrifuge tubes and pellet at 4°C for 5 minutes at 2000xg. Remove supernatant.

    Pause Point: Flash freeze pellets in liquid nitrogen and store pellets at -80°C.

3.11.2 CELL LYSIS (TIMING 1.5 HOURS)

1. Chill Lysis Buffers A, B, and C on ice.
2. If using an electronic chiller for the sonication chamber, pre-chill to 4°C.
3. Thaw a cell pellet (10 million cells ✓) on ice for two minutes.

   ✓ Note: Fewer cells may be used per SPRITE experiment. As low as 200,000 cells (fragmented with less DNase) have yielded successful SPRITE data sets.
4. Add 700 µL of Lysis Buffer A supplemented with 1 x Proteinase Cocktail Inhibitor (PIC) to each pellet and resuspend fully. **?Troubleshooting**
5. Incubate mixtures on ice for 10 minutes.
6. Pellet cells at 4°C for 8 minutes at 850xg.
7. Discard the supernatant, taking care not to disturb the pellet.
8. Add 700 µL of Lysis Buffer B supplemented with 1 x PIC to each cell pellet and resuspend fully.
9. Incubate mixtures on ice for 10 minutes.
10. Pellet cells at 4°C for 8 minutes at 850xg.
11. Discard the supernatant, taking care not to disturb the pellet.
12. Add 550µL of Lysis Buffer C supplemented with 1 x PIC to each 10 million nuclei pellet and resuspend.
13. Incubate mixture on ice for 8 minutes.
14. Sonicate each sample at 4–5 watts for 1 minute: 1 pulse for 0.7 seconds ON, 3.3 seconds OFF. During and after sonication, keep lysate at 4°C. A Branson needle-tip sonicator kept at 4°C was used for this protocol.
15. If sonicating more than one pellet of the same cell and conditions and you do not wish to keep them as biological replicates, pool all lysates together and split again into 10 million cell aliquots. This ensures that all samples in each tube are equally processed and DNased in the subsequent steps.
16. (Optional) QC the sonicated lysate by visualizing crosslinked complexes by microscopy (**Box 2a**).
17. Move directly into DNA fragmentation.✓

   Pause Point: Flash freeze lysates in liquid nitrogen and store pellets at -80C.

   ✓ CRITICAL: At times, we may have to go back to lysate stock to optimize the DNA fragmentation. We recommend storing smaller aliquots of the lysate stock in several 10 ul aliquots to minimize freeze-thaw cycles. Multiple freeze-thaw cycles of the lysate have shown a decrease the quality of coupling observed in SPRITE datasets.

3.11.3 DNA FRAGMENTATION (TIMING 3 HOURS)

1. If starting from frozen lysate, thaw one tube of lysate on ice.
2. Fragment the DNA with DNase. To obtain a desired DNA size distribution, perform several reactions with varying DNase concentrations ✓. **?Troubleshooting**

   ✓ Note: We typically perform SPRITE on DNA that has a size distribution from 50–1000 base pairs with an average size between 200–300 base pairs. In general, a 20% DNase reaction achieves this. (**Box 2c**)

Stock Solution	**Volume**
10X SPRITE DNase Buffer✓	2 µL
Lysate	10 µL
Turbo DNase from ThermoFisher	0.8 / 0.9 / 1 / 2 / 3 µL
H_2O	7.2 / 7.1 / 7 / 6 / 5 µL
Total	20 µL

   ✓ CRITICAL: Do NOT use 10x Turbo DNase Buffer provided with the enzyme. Instead, use the 10x SPRITE DNase Buffer components listed above. Using the incorrect 10x Turbo DNase Buffer so will result in nicked DNA that cannot be amplified.
3. Incubate at 37° C for 20 minutes.
4. Add 1 µL of 25X DNase Stop Solution to each sample to terminate the reaction.
5. Reverse the crosslinks for half of each sample. Flash freeze the remaining half of the DNased sample and store at -80° C ✓.

   ✓ Note: Reversing the crosslinks on only half the DNased sample ensures that there is remaining crosslinked sample ready for SPRITE. This negates the need to DNase the entire tube of lysate (10 million cells) which could yield a different size distribution than intended.

Stock Solution	**Volume**
DNased Lysate	10 µL
SPRITE ProK Buffer	82 µL
Proteinase K	8 µL
Total	100 µL

6. Incubate for at 65° C for 2 hours at minimum ✓.

   ✓ Alternatively, you may reverse crosslink for 55° C for 1 hour, then increasing to 65° C for overnight incubation (>12 hrs).

7. Clean up samples by following the protocol provided in the Zymo RNA Clean and Concentrator-5 Kit (>17 nt), binding in 2 volumes of RNA Binding Buffer and 1 total volume of 100% ethanol. Elute in 10 µL of H_2O.
8. Determine concentration of DNA in each sample by following the directions provided with the Qubit dsDNA HS Assay Kit. **?Troubleshooting**
9. Determine the size distribution of DNA in each sample by following the directions provided with either the High Sensitivity DNA Kit for the Agilent Bioanalyzer or the D1000 DNA HS Screentape for the Agilent 2200 TapeStation. Number of unique molecules can be calculated as shown in **Box 3c** and **Supplemental Table S1**.
10. If none of these concentrations of Turbo DNase led to ideal fragmentation, adjust concentrations and repeat the DNasing until optimal fragmentation is achieved. DNA for SPRITE generally has a size distribution from 50–1000 base pairs with an average size between 200–300 base pairs ✓.

    ✓ Note: Increasing to above 5 µL of Turbo DNase (25% of the reaction volume) may lower the efficiency of the enzyme. Instead, we recommend diluting the lysate 1:1 by combining 5 µL of lysate with 5 µL H_2O to make up a 10µL total volume. Lysate may be diluted further down if DNasing is not sufficient with high concentration of enzyme.

11. (Optional) DNase the entire batch of crosslinked lysate at the identified optimal DNase concentration ✓.

    ✓ Note: DNasing the batch of crosslinked lysate is not necessary for SPRITE if half of the DNased material was saved from Step 45.

Stock Solution	**Volume**
10X DNase Buffer	110µL
Lysate	550µL
Turbo DNase from ThermoFisher	X µL
H_2O	X µL to reach final volume
Total	1100µL

12. Incubate at 37° C for 20 minutes.
13. Add 44 µL of 25 x DNase Stop Solution to each sample to terminate the reaction.
14. Flash freeze DNased lysate and store at -80° C.

3.11.4 COUPLING (TIMING 2 HOURS OR OVERNIGHT COUPLING)

1. Prior to coupling, first calculate the number of molecules to couple to NHS beads using **Supplemental Table S1** and as described in **Box 3**.
2. Approximately 15–30 minutes prior to starting coupling, bring the bottle of Pierce NHS-activated beads to room temperature. After opening, immediately seal the bottle with parafilm and store beads in desiccant to prevent condensation.
3. Gently invert the bottle containing the Pierce NHS-activated beads in N,N-dimethylacetamide (DMAC) until there is a uniform suspension. Being careful not to

introduce water into the bottle, transfer 2 mL of NHS beads into a clean 1.7 mL LoBind tube. Place the tube on a magnetic rack to capture the beads.

CRITICAL: It is crucial to have an optimal bead to molecule ratio for the library preparation and SPRITE processes to avoid spurious interactions as a result of multiple crosslinked complexes coupling to the same NHS bead. We aim to bind at a 1:4 to 1:1 ratio of DNA molecules:beads. **Box 3** describes how to calculate the number of molecules to couple to NHS beads for a given SPRITE lysate in detail. To determine the microliter amount of lysate to couple we calculate the DNA molarity from the concentration and average size measurements obtained in the DNA fragmentation step of the protocol. Molarity is multiplied by the lysate volume (10 µL of crosslinked, DNased lysate remains in the tube) to obtain DNA molecule number.

4. Remove the DMAC and wash beads with 1 mL ice-cold 1mM HCl.
5. Wash beads with 1 mL ice-cold 1 x PBS.
6. Add 500 µL Coupling Buffer and 10 µL of 0.5 M EDTA to the beads. Vortex until resuspended.
7. After calculating the number of molecules to couple to NHS beads, dilute X µL of DNased lysate into a total volume of 500 µL Coupling buffer.
8. Combine the DNased lysate solution to the Coupling Buffer solution.
9. Incubate the lysate and beads overnight at 4° C on a mixer (or 2 hours at room temperature).
10. Place beads on a magnet and remove 500 µL of flowthrough.
11. (Optional) This flowthrough aliquot can be saved to determine coupling efficiency if the DPM QC fails.
12. Add 500µL 1M Tris pH 7.5 (3M ethanolamine pH 9.0 can also be used) to the beads and incubate on a mixer at 4° C for at least 45 minutes. This ensures that all NHS beads will be quenched with protein from bound lysate or Tris, and will not bind enzymes in the following steps.
13. Wash beads twice in 1 mL of modified RLT buffer.
14. Wash beads three times with 1 mL of SPRITE wash buffer.

15. Spin the beads down quickly in a microcentrifuge and place back on the magnet to remove any remaining liquid.

3.11.5 PHOSPHORYLATION AND END REPAIR (TIMING 1 HOUR)

1. Blunt the 5’ and 3’ ends of the DNA molecules to prevent unwanted ligation by adding the following mixture to the beads:

Stock Solution	**Volume**
H2O	212.5µL
End Repair Reaction Buffer (10X)	25µL
End Repair Enzyme Mix	12.5µL
Total	250µL

2. Incubate on a thermomixer for 15 minutes at 24° C, 1200 RPM.
3. Wash once with 1 mL of modified RLT buffer.
4. Wash three times with 1 mL of SPRITE wash buffer.
5. Spin the beads down quickly in a microcentrifuge and place back on the magnet to remove any remaining liquid.
6. Add dATP to the 3’ ends of each DNA molecule to allow for ligation of the DPM adaptor by adding the following mixture to the beads:

Stock Solution	**Volume**
H2O	215µL
dA-Tailing Reaction Buffer (10X)	25µL
Klenow Fragment (exo-)	10µL
Total	250µL

7. Incubate on a thermomixer for 15 minutes at 37° C, 1200 rpm. If ligating the DPM adaptor barcode on the same day, set up the reaction during this incubation.
8. Wash once with 1 mL of modified RLT buffer.
9. Wash three times with 1 mL of SPRITE wash buffer.
10. Spin the beads down briefly in a microcentrifuge and place back on the magnet to remove any remaining liquid.

3.11.6 DPM, ODD, EVEN, ODD, AND TERMINAL Y ADAPTOR LIGATIONS (TIMING 10 HOURS)

There are 96 adaptors that are designed to ligate onto the DNA molecules. These DPM adaptors are kept in a 96-well stock plate at 4.5 µM. The ligation reaction between the adaptors and the DNA occurs in a 96-well plate. The following steps that detail set up are designed for optimum efficiency during the process. All ligation steps include SPRITE wash buffer, which contains detergents to prevent beads from aggregating, sticking to the plastic tips and tubes, and for even distribution of the beads across a 96-well plate. We have verified that these detergents do not significantly inhibit ligation efficiency.

1. Make Ligation Master Mix for five rounds of SPRITE (DPM + four extra tags). Split the master mix evenly into each well of a 12-well strip tube by pipetting 260 µL into each well. Keep on ice ✓.

Stock Solution	**Volume**
NEBNext Quick Ligation Reaction Buffer (5X)	1600 µL
Instant Sticky-end Ligation Master Mix (2X)	1000 µL
1,2-Propanediol	600 µL
Total	3200 µL

✓ Note: Homemade Ligation Master Mix can be stored for over 1 week at -20°C.

2. Create a dilute SPRITE wash buffer by mixing 1100 µL of SPRITE wash buffer with 792 µL of H2O .
3. Accounting for bead volume, add the dilute SPRITE wash buffer to the beads to achieve a final volume of 1.075 mL. Ensure that the beads are equally resuspended in the buffer. Distribute the beads equally into a 12-well strip tube by aliquoting 89.6 µL of beads into each well.
4. Centrifuge the DPM adaptor stock plate at 1000 xg for 1 minute before removing the foil seal. Aliquot 2.4 µL from the stock plate of DPM adaptors to a new low-bind 96-well plate ✓. Be careful to ensure that there is no mixing between wells at any point of the process to avoid cross-contamination of barcodes. Use a new pipette tip for each well. After transfer is complete, seal both plates with a new foil seal.

   ✓ Note: This step can be done in advance, in bulk, so that these plates are ready-to-use.
5. Centrifuge the 96-well plate containing the aliquoted adaptors, and then remove the foil seal.
6. Aliquot 11.2 µL of beads into each well of the 96-well plate that contains 2.4 µL of the DPM adaptors. Be careful to ensure that there is no mixing between wells at any point of the process. Use a new pipette tip for each well. Also be careful to ensure that there are no beads remaining in the pipette tip✓.

   ✓ Note: Pipetting the beads out slowly and against the well tips yields clean tips.
7. Carefully add any remaining beads to individual wells on the plate in 1 µL aliquots.
8. Aliquot 6.4 µL of Ligation Master Mix into each well, mixing by pipetting up and down 10 times ✓.

   CRITICIAL 1: Be careful to ensure that there is no mixing between wells at any point of the process. Use a new pipette tip for each well.

   CRITICAL 2: After mixing, pipetting the beads out slowly and against the well tips yields clean tips and minimizes bead loss.
9. The final reaction components and volumes for each well should be as follows:

Stock Solution	**Volume**
Beads + SPRITE Wash Buffer + H2O Mix	11.2 µL

DPM Adaptor (4.5 uM)	2.4 µL
Ligation Master Mix	6.4 µL
Total	20 µL

10. Seal the plate with a foil seal and incubate on a thermomixer for 60 minutes at 20° C, shaking for 30 seconds at 1600 rpm every five minutes to prevent beads from settling to the bottom of the plate ✓.

    CRITICAL: Ligation time is critical for high efficiency of ligation each round. We have tested ligation at 5, 15, 30, 45, and 60 minute reaction times and 60 minutes ligation time has significantly higher yields over the other times, but not significant more than a 15 minute incubation.
11. After incubation, centrifuge the plate at 1000 xg for 1 minute before removing the foil seal.
12. Pour modified RLT buffer into a sterile plastic reservoir, and transfer 60 µL of modified RLT buffer into each well of the 96-well plate to stop the ligation reactions. It is not necessary to use new tips for each well.
13. Pool all 96 stopped ligation reactions into a second sterile plastic reservoir.
14. Place a 15 mL conical tube on an appropriately sized magnetic rack and transfer the ligation pool into the conical. Capture all beads on the magnet, disposing all modified RLT buffer in an appropriate waste receptacle.
15. Remove the 15 mL conical containing the beads from the magnet and resuspend beads in 1 mL SPRITE wash buffer. Transfer the bead solution to a microcentrifuge tube.
16. Wash three times with 1 mL of SPRITE wash buffer.
17. Spin the beads down briefly in a microcentrifuge and place back on the magnet to remove any remaining liquid.
18. (Optional) QC2 (**Box 4**): Resuspend the beads in SPRITE ProK buffer so that the final beads + buffer volume is 1 mL. Remove a 5% aliquot (50 µL) into a separate 1.7 mL microcentrifuge tube.

19. Place the remaining 95% of beads back on the magnetic rack, remove the SPRITE ProK buffer, and store beads in 1 mL of modified RLT buffer. Keep beads at 4° C overnight.
20. Remove modified RLT buffer from the remaining 95% aliquot and wash three times with 1 mL of SPRITE wash buffer.
21. Spin the beads down briefly in a microcentrifuge and place back on the magnet to remove any remaining liquid.
22. Repeat Steps 81 through 96 for the remaining four SPRITE rounds (Odd, Even, Odd, Terminal Y). After pooling each round, wash beads once in 1mL modified RLT buffer and three times in 1mL SPRITE wash buffer.

3.11.7 FINAL LIBRARY PREPARATION AND SEQUENCING (TIMING 3 HOURS)

1. Resuspend the beads in SPRITE ProK buffer so that the combined volume of beads and buffer is 1 mL total.
2. Remove eight 5% aliquots into clean 1.7 mL microcentrifuge tubes and elute the barcoded DNA from the beads. This serves as an additional round of barcoding so that the total number of unique barcode combinations exceeds the number of DNA molecules in the sample. The remaining lysate on beads can either be stored in modified RLT buffer at 4° C (for up to 1 week) or also eluted in 5% aliquots.

Stock Solution	**Volume**
Beads in SPRITE ProK Buffer	50 µL
SPRITE ProK Buffer	42 µL
Proteinase K	8 µL
Total	100 µL

3. Incubate at 65° C overnight.

4. Place the microcentrifuge tubes on a magnet and capture the beads. Transfer supernatant from each aliquot into eight separate microcentrifuge tubes.
5. To maximize recovery of barcoded material, rinse the beads with 25 µL of water.
6. Vortex, and re-capture the beads. Transfer supernatant from each aliquot to its respective tube for a combined volume of 125 µL per tube. Discard the beads.
7. Follow the protocol provided in the Zymo RNA Clean and Concentrator Kit (>17 nt), binding with 2 volumes of RNA Binding Buffer and 1 total volume of 100% ethanol. Elute in 20 µL of H2O.
8. Amplify the barcoded DNA.

Stock Solution	**Volume**
Barcoded DNA (cleaned)	20 µL
First Primer (100uM)	1 µL
Second Primer (100uM)	1 µL
H_2O	3 µL
2x Q5 Hot Start Master Mix	25 µL
Total	50 µL

a. PCR Program:

Initial denaturation: 98° C - 180 seconds

4 cycles:

98° C -10 seconds

68° C - 30 seconds

72° C - 40 seconds

8 cycles:

98° C -10 seconds

70° C - 30 seconds

72° C - 40 seconds

Final extension: 72° C - 180 seconds

Hold 4° C

9. Clean the PCR reaction and size select for your target libraries. The total length of our barcode on one amplified product is around 160 base pairs and each target DNA molecule is no fewer than 100 base pairs. Agencourt AMPure XP beads are able to size select while cleaning the PCR reaction of unwanted products.
10. Add 0.7 x volume (35 µL) AMPure XP beads to the sample for a total volume of 85 µL and mix thoroughly.
11. Incubate for 10 minutes at room temperature.
12. Place the beads on an appropriately sized magnet to capture the beads and the bound DNA. Wait a few minutes until all the beads are captured.
13. Remove the supernatant and discard.
14. Wash beads twice with 80% ethanol by pipetting ethanol into the tube while beads are captured, moving the tube to the opposite side of the magnet so that beads pass through the ethanol, and then removing the ethanol solution.
15. Quickly spin down the beads in a microcentrifuge, re-capture on magnet, and remove any remaining ethanol.
16. Air-dry beads while the tube is on the magnet for 5–10 minutes, or until the beads appear dry.
17. Elute the amplified DNA from the beads by resuspending the beads in 50 µL of H2O.
18. Repeat the size-select clean up with 0.7 x AMPure XP beads (add directly to the eluted DNA/bead mix), eluting finally in 12 µL H_2O ✓.

✓ Note: To ensure all library material is eluted from beads, elute twice with 6 µL H_2O. Most of the material will be removed in the first elution, and any remaining material will be removed in the second.

19. Determine the concentration of the barcoded libraries by following the directions provided with the Qubit dsDNA HS Assay Kit. Final libraries are generally between 0.3 ng/µL and 1.5 ng/µL. **?Troubleshooting**
20. Determine the size distribution and average size of the barcoded libraries by following the directions provided with the HS DNA Kit for the Agilent Bioanalyzer. Average sizes are generally around 350–450 base pairs (**Figure 3a**).
21. Input values from Steps 120 and 121 into **Supplemental Table S1** to estimate the number of unique DNA molecules present in the sample prior to library amplification. Use this sheet to estimate how many reads are necessary to sequence each SPRITE library to saturation (see **Box 5**).
22. Pool together SPRITE libraries such that the total sum of reads necessary to sequence each library to saturation is less than or equal to the number of reads available on your sequencing platform. To visualize chromosome territories, we have sequenced 5–10 million reads per sample as a QC step. We recommend sequencing approximately 20–25 million reads to visualize compartments, 200–300 million reads for TADs, and 1 billion reads for loop interactions.
23. Sequence the pooled SPRITE libraries using an Illumina sequencing kit. For example, we use a TruSeq Rapid SBS v1 Kit–HS (200 cycle) kit on an Illumina HiSeq v2500 platform. For a standard SPRITE experiment (five rounds of barcoding: DPM, Odd, Even, Odd, and Terminal ligations), sequence with paired-end reads with 50–90 bp for Read1 (DPM and the genomic DNA sequence), 8 bp index, and 95–125 bp for Read 2 (barcodes). See **Box 1** for important considerations for read lengths to ensure that all SPRITE tags and the corresponding genomic DNA sequenced.

3.11.8 COMPUTATIONAL PIPELINE AND DATA ANALYSIS (TIMING ~5–8 HOURS)

The following computational steps are detailed in the SPRITE pipeline instructions in the up to date SPRITE wiki: https://github.com/GuttmanLab/sprite-pipeline/wiki. See **Box 6** for a detailed step-by-step explanation of the commands used in the SPRITE pipeline.

1. Clone the SPRITE pipeline repository from GitHub:

```
git clone https://github.com/GuttmanLab/sprite-pipeline.git
```

2. Download the required dependencies for successfully running the protocol.

   a. Conda
   b. Java 8
   c. Python 3
   d. Snakemake
   e. An installation of Hi-Corrector (included in the scripts folder)

3. Run fastq2json.py[25] to compile the paths of all SPRITE fastq files into a .json file using the following command:

   *Note: This script will automatically identify all files ending in '_R1.fastq.gz' or '_R2.fastq.gz' within the specified fastq directory.

```
python fastq2json.py –fastq_dir /path/to/fastq/directory
```

4. Assign values and paths to each parameter referenced in the config.yaml file. Users are expected to generate or download their own Bowtie2 genome index. Example bed files for masking can be found on the SPRITE GitHub. Users can download a pre-made Bowtie2 mouse genome index here: https://ftp.ensembl.org//pub/release-95/fasta/mus_musculus/dna//Mus_musculus.GRCm38.dna_sm.toplevel.fa.gz and human genome index here: ftp://ftp.ensembl.org/pub/release-97/gtf/homo_sapiens/Homo_sapiens.GRCh38.97.gtf.gz

5. Run the SPRITE pipeline using the following command.

```
sh run_pipeline.sh
```

6. Check ligation efficiency by opening the [sample_name].ligation_efficiency.txt in any common text editor. We routinely achieve greater than 75% of reads containing all 5 barcodes (**Figure 3b**). **?Troubleshooting**

7. Check cluster-size distribution by opening the cluster_sizes.png file. A successful SPRITE library achieves a wide distribution of cluster sizes (**Figure 3c)**. Unsuccessful

experiments result in libraries that are primarily singlets and are therefore devoid of interactions. **?Troubleshooting**

8. Check the alignment statistics, trimming logs, and all other steps along the pipeline by opening the multiqc file.
9. Open the DNA interaction matrices to visualize the heatmaps. The pipeline outputs a single heatmap, but users can easily modify the script to generate heatmaps at various resolutions. An example of DNA interaction heatmaps can be seen in **Figure 5**.

3.12 TROUBLESHOOTING

Step	**Problem**	**Possible Reason**	**Solution**
Step 6	5 minutes have passed and cells have not started to lift from the plate.	Forgot to wash with PBS; residual serum will inhibit trypsinization	Quench the trypsin immediately with wash solution to prevent over-trypsinization and premature lysis of cells. Break cells off the plate by washing vigorously with wash solution.
Step 27	Pellet is not resuspending and remains clumpy even after vigorous pipetting even after vigorous pipetting.	Overcrosslinking	Either thaw a new pellet and repeat lysis resuspending pellet fully by pipetting repeatedly and lightly vortexing, or repeat crosslinking steps.
Step 27	Cell loss during crosslinking.	Cells crosslinked in suspension are more prone to loss on the plastic surfaces	Adherent cells may also be crosslinked on a plate rather than trypsinizing and crosslinking in solution to reduce cell loss.
Step 42	Fragment sizes are far less than 200 bp or far greater than 1000 bp.	Short fragments: over-DNased Long fragments: under DNased, reduced activity of enzyme	Repeat DNasing step with more titration conditions to achieve optimal size range (50–1000 bp) and average size (200–300 bp). If under DNasing is occurring, reduce the amount of lysate input into the DNasing reaction.
Step 48	DNA yield is lower than expected.	Incomplete reverse crosslinking	We typically reverse crosslink for 55° C for 1 hour, then increasing to 65° C for overnight incubation (>12 hrs).

Step 120	Library yield is lower than expected, despite a successful DPM. library obtained during DPM QC.	Unsuccessful SPRITE ligations.	Make sure the correct SPRITE barcodes have been ligated in the correct order. You can QC for efficient ligations by PCR from DPM to 2Pbarcoded.
Step 130	Ligation efficiency is lower than 50%.	(1) Residual RLT buffer not completely washed away prior to subsequent round of ligation. (2) Barcoded tags not annealed.	(1) Ensure at least three washes are performed to completely buffer exchange into SPRITE wash buffer prior to proceeding with the next round of split-and-pool. (2) Confirm tags are annealed in an agarose gel by running an a few annealed tags alongside the corresponding unannealed DNA tags to confirm a visible size shift upon annealing.
Step 131	SPRITE library is unpaired.	(1) Not sampling enough (2) Over-fragmentation (3) Spun out pellet containing multi-way contacts	(1) Resequence library at higher sequencing depth (2) Redo cell lysis and reduce DNA fragmentation (QC by microscopy, see Box 2) (3) Avoid high-speed centrifugation after sonication or DNasing. (Do not pellet and take supernatant)

3.13 TIMING

Cell Culture and Crosslinking: 3 h

Cell Lysis: 1.5 h

DNA Fragmentation: 3 h

Coupling: 2 h or overnight

Phosphorylation and End Repair: 1 h

DPM, Odd, Even, Odd, and Terminal Y Adaptor Ligations: ~ 10 h

Final Library Preparation: 3 h

Deep Sequencing: ~ 1 d

Computational Pipeline and Data Analysis: ~5–8 h

Box 2: Quality Control to Visualize Crosslinked Complexes: 4 h

Box 5: Check to Determine Ligation Efficiency of the DPM Adaptor: 3.5 h

3.14 ANTICIPATED RESULTS

SPRITE libraries typically range in size from 300 bp - 1.3kb, corresponding to the sum of all ligated SPRITE barcodes (128bp), the Illumina library amplification primer sequences (124bp), and the genomic DNA insert (50bp-1kb) (**Figure 3a**). We typically achieve >75% total reads tagged with all 5 barcodes identified, which corresponds to a ligation efficiency of approximately 95% each round ($0.95^{5\ \text{rounds}} = 0.75$) (**Figure 3b**). To QC the ligation efficiency, a "ligation efficiency" file is calculated during the SPRITE pipeline. For SPRITE libraries that have generated heatmaps recapitulating known structures such as TADs, compartments, and chromosome territories, we observe approximately 65% or more clusters with at least two or more interactions, and that the SPRITE clusters capture a range of structures containing ~30% of reads correspond to clusters with 2–10 reads, and a distribution of interactions corresponding to larger cluster sizes containing 11–100 reads, 101–1000 reads, and more than 1000 reads per cluster (**Figure 3c**).

One of the critical points for sequencing SPRITE data relies on the fact that each molecule in an interacting SPRITE cluster must be sequenced at least once. For these reasons, when sequencing data is done processing, we estimate the molecule complexity of our sequencing library using Preseq software from Smith lab[26]. Although many labs typically avoid PCR duplicates in sequencing, for SPRITE it is critical to sequence to saturation; we typically aim to achieve 1.5x coverage of each molecule (**Box 5**). Without sequencing to saturation, we have found that libraries will have very few paired SPRITE clusters and very few interactions will be observed as all molecules that were interacting in a crosslinked complex may not have been sequenced. To sequence libraries to saturation, we measure the library molarity post-PCR, and provide a calculator to estimate the number of reads to sequence a given SPRITE library (**Supplemental Sheet S1**). As a first-time user of the protocol, we recommend eluting, reverse crosslinking, and PCR amplifying SPRITE aliquots of different sizes (e.g., 0.5%, 1%, 2%, 5%, 10%) to identify a library with 4–10 million unique molecules pre-PCR, and sequencing the library to saturation (e.g., iSeq or MiSeq). We have observed chromosome territories and a broad distribution of SPRITE clusters for libraries that are sequenced to saturation with as few as 4 million reads.

We have applied SPRITE to several cell lines including mESCs, human lymphoblasts, human H1 ES cells and HFF cells, and others. In all of these cases, smaller SPRITE clusters containing 2–10 or 11–100 reads per cluster capture primarily close-range interactions occurring on the same chromosome corresponding to TADs and compartments (**Figure 3d**). Larger-range structures such as chromosome territories and interactions between chromosomes occurring around nuclear bodies are enriched in larger SPRITE clusters containing 101–1000 and more than 1001 reads per cluster (**Figure 3e**). Notably, one of the most common inter-chromosomal contacts are between gene-poor or transcriptionally active regions on chromosomes containing nucleolar organizing regions (NOR) (chromosomes 12, 15, 18, and 19 in mouse ES cells) that make contacts around the “inactive nucleolar hub”. Another common set of inter-chromosomal contacts is between highly active, gene-dense regions on chromatin (as defined by RNA Pol II Density by ChIP-seq) that tend to co-localize in the same 3D space that we describe as the “active speckle hub”.

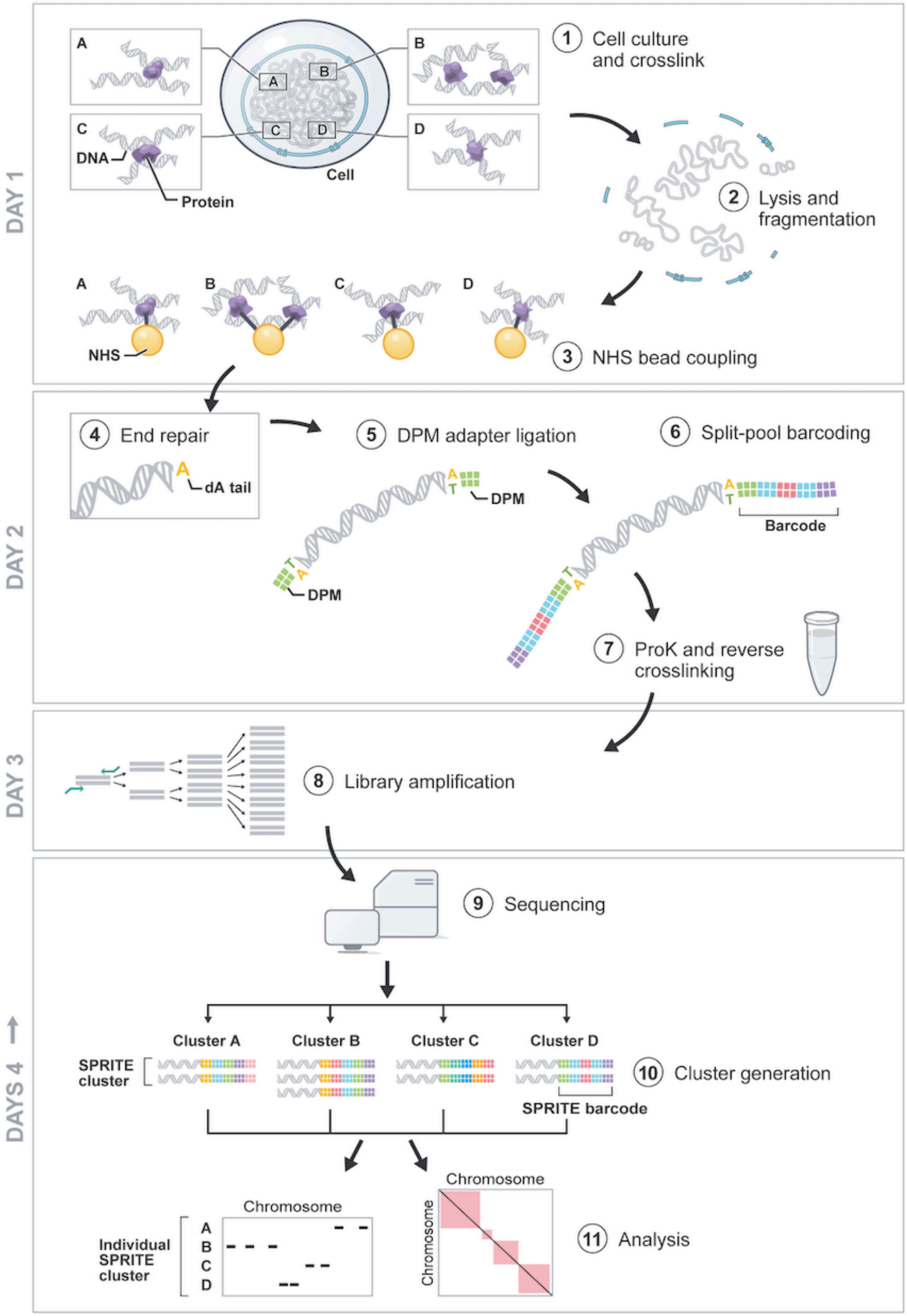
3.15 DATA AND CODE AVAILABILITY

Example DNA SPRITE datasets have been deposited on the 4DN Data Portal under accession numbers 4DNFI8ZROQ87 and 4DNFIY9HL35V.

The DNA SPRITE software is available for download on the Guttman Lab GitHub page at https://github.com/GuttmanLab/sprite-pipeline/. Version v0.2 is explained in detail within this paper.

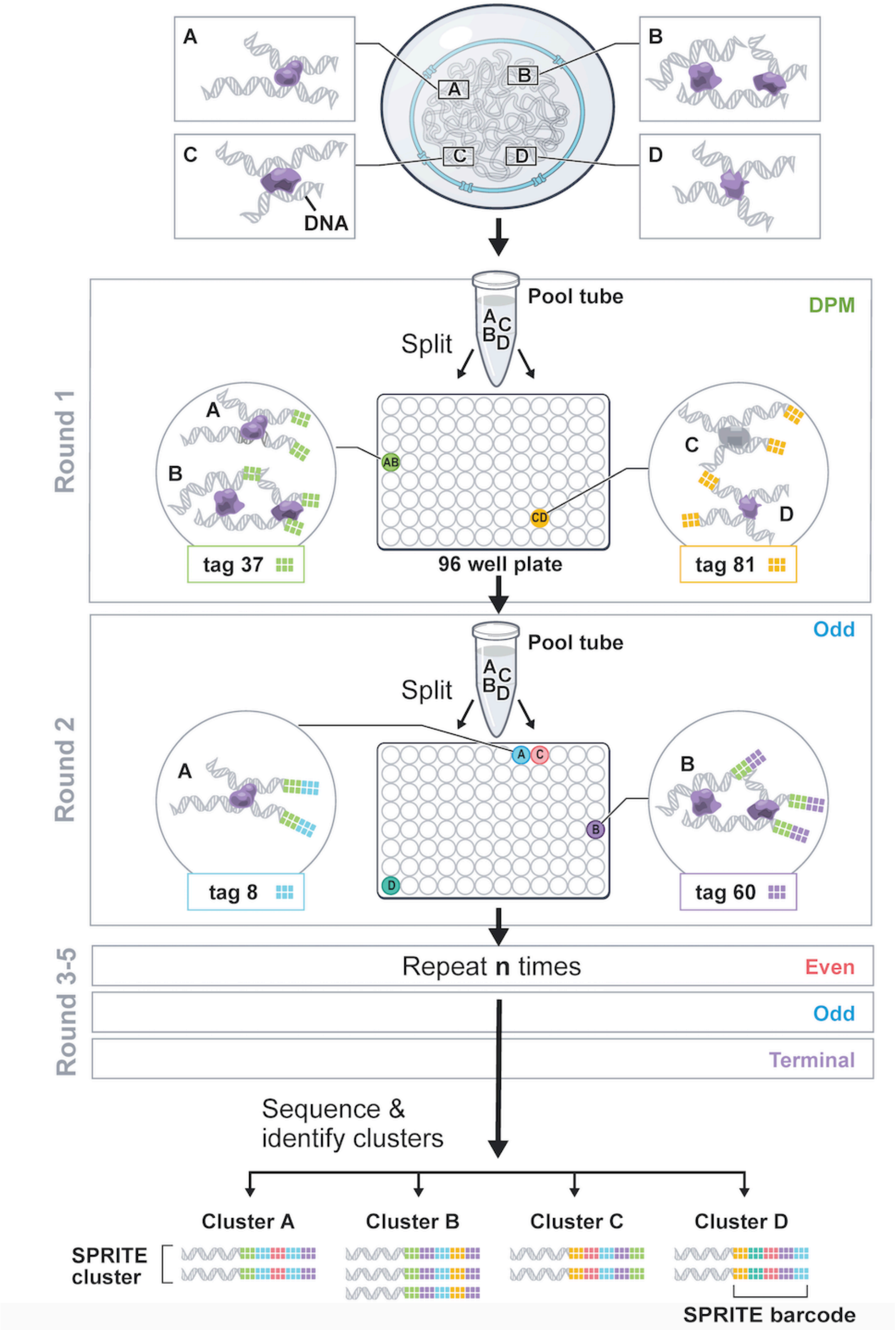
3.16 MAIN FIGURE LEGENDS

Figure 1: Overview of SPRITE procedure.



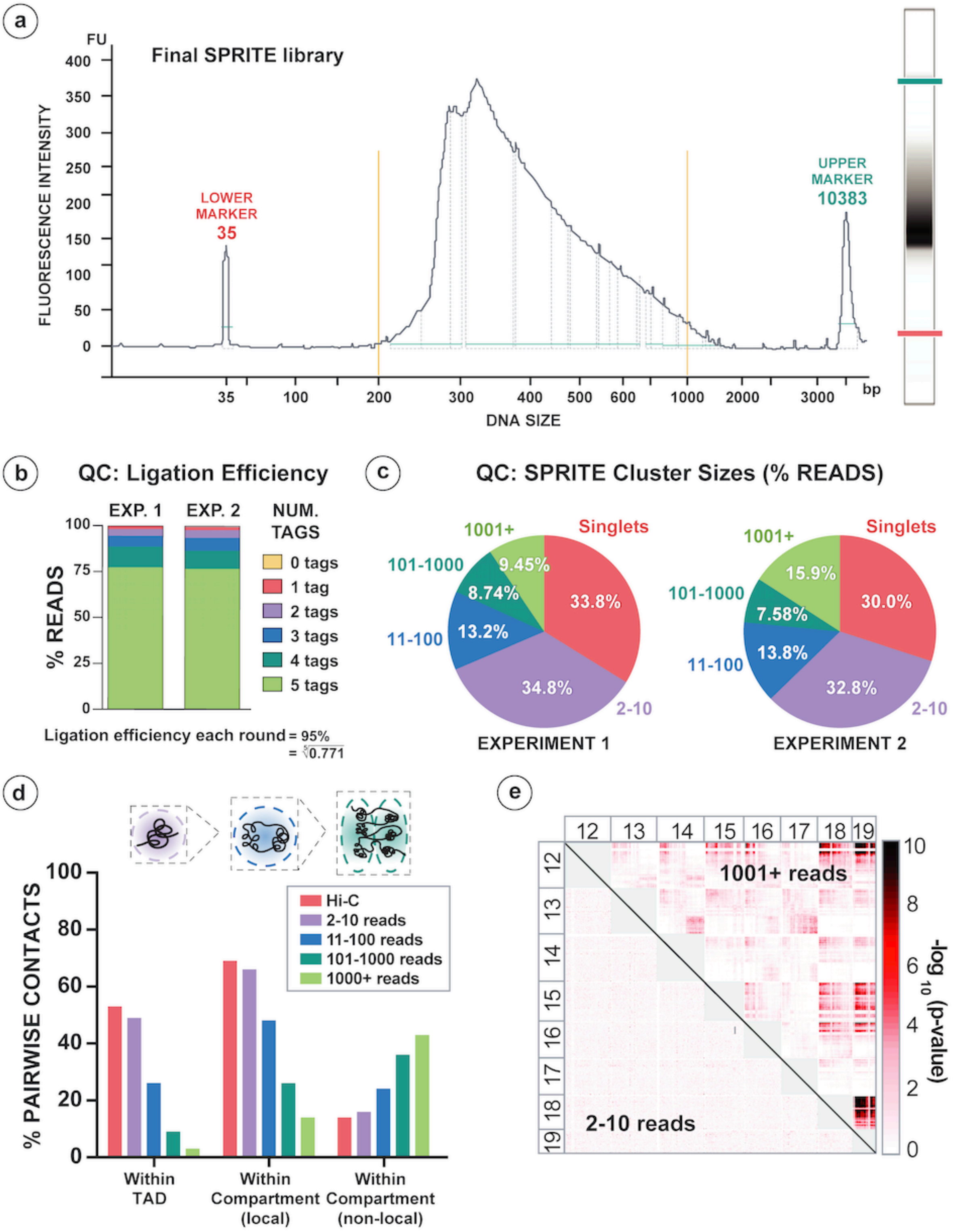
Day 1: (1) Cells are dual crosslinked with DSG and formaldehyde. (2) Cells are lysed and chromatin is fragmented using sonication and DNasing. (3) Lysate is coupled to NHS beads overnight. **Day 2:** (4) DNA is blunt-ended, phosphorylated, and dA-tailed prior to (5) ligation with DPM adapter. (6) 5 rounds of split-and-pool ligations are performed with the DPM, Odd, Even, Odd, and Terminal tags, which we refer to as a barcode. (7) After split-and-pool, samples are split into several aliquots and DNA is reverse crosslinked overnight by addition of Proteinase K enzyme and heat. **Day 3:** (8) Final SPRITE libraries are amplified. **Day 4 onward**: (9) DNA is sequenced and (10) all molecules sharing the same barcodes are matched to generate SPRITE clusters. (11) DNA Interactions occurring in SPRITE clusters are analyzed as multi-way interactions or as pairwise interactions visualized using intra- and inter-chromosomal heatmaps.

Figure 2: Schematic of split-pool procedure.



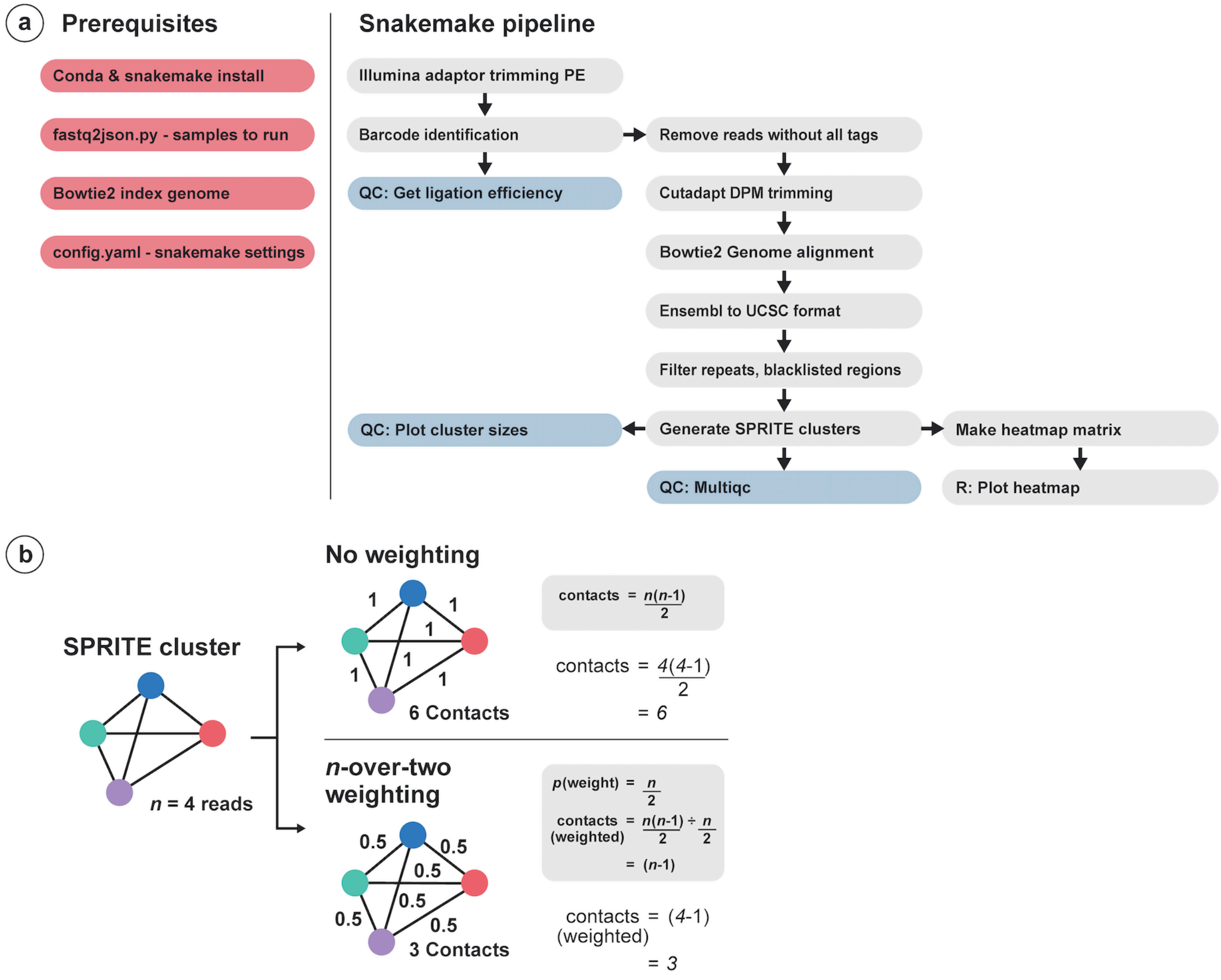
Split-and-pool barcoding works by splitting crosslinked complexes across a 96-well plate containing 96 unique tags, ligating a unique sequence (colored tag) to each DNA molecule, and pooling all crosslinked complexes into a single tube. This split-and-pool process is repeated over multiple rounds, sequentially adding an additional tag each round. Because all molecules within a crosslinked complex are covalently attached, they will sort across the same wells during each round of the split-and-pool process and will obtain the same series of tags, which we refer to as a SPRITE barcode. Genomic DNA and their associated barcodes are then sequenced. All reads sharing the same SPRITE barcodes are matched to generate SPRITE clusters.

Figure 3: Summary of alignment statistics.



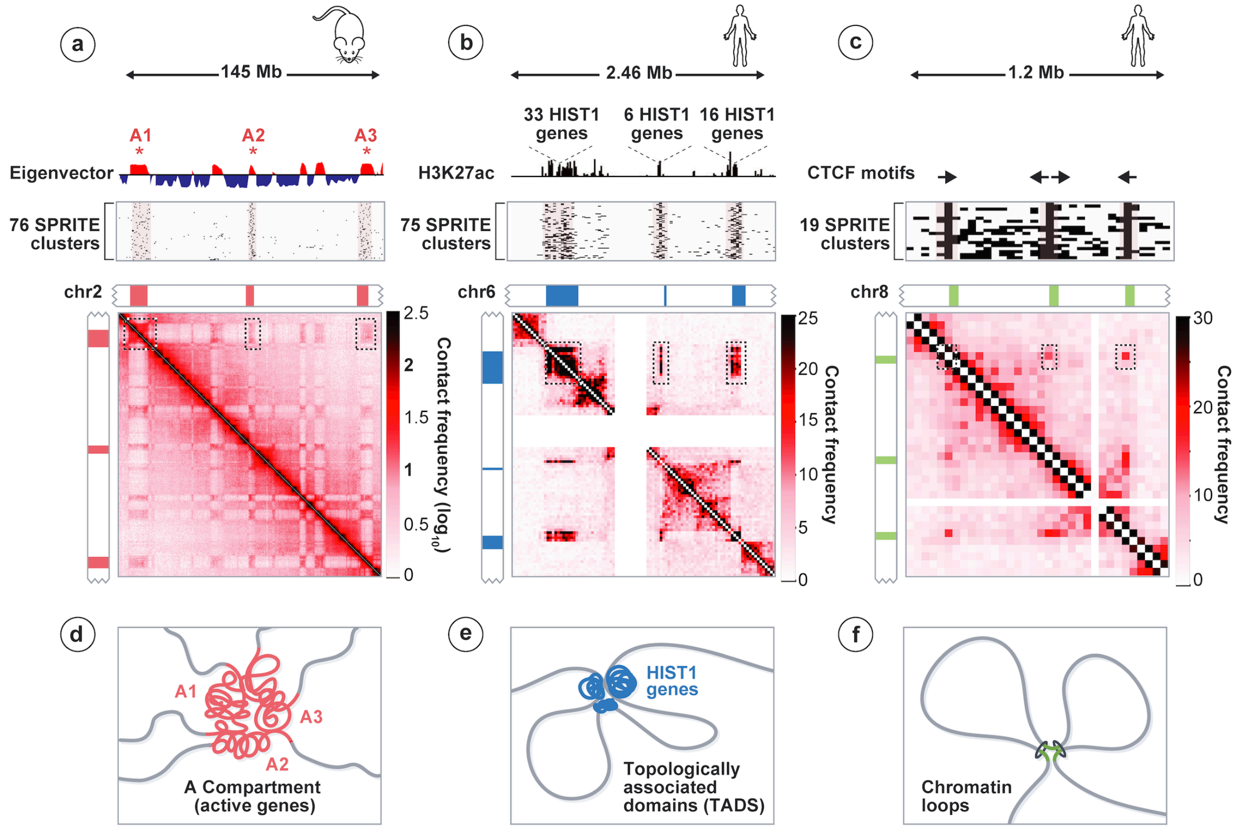
(**a**) An example of a final SPRITE library after PCR amplification. **(b)** Summary of ligation efficiency statistics are outputted as a QC step from the SPRITE pipeline to confirm tags have successfully ligated to each DNA molecule. The distribution of reads containing 0, 1, 2, 3, 4, or 5 SPRITE tags is shown for two independent SPRITE experiments. Ligation efficiency each round (95%) is calculated by taking the 5th root of the fraction of reads containing 5 tags (77.1%). (**c**) SPRITE cluster sizes are outputted as a QC step from the SPRITE pipeline to confirm interactions have been successfully detected. Individual SPRITE clusters contain all reads sharing the same barcode. The number of reads sharing the same barcode within an individual cluster can range from singlets (molecules not interacting with other molecules, red), 2–10 reads per cluster (purple), 11–100 reads per cluster (blue), 101–1000 reads per cluster (dark green), and greater than 1000 reads per cluster (light green). The percentage of reads that correspond to different SPRITE cluster sizes is shown for two independent experiments generated on 3% FA-DSG samples sonicated for 1 minute, 4–5W of sonication as described. Successful experiments typically detect a range of cluster sizes such as those shown here. **(d)** The number of DNA molecules in a SPRITE cluster reflects the distance DNA molecules are interacting in the nucleus. Specifically, smaller SPRITE clusters primarily capture close-range (within TAD) interactions, whereas larger clusters capture longer-distance interactions within A or B compartments (local and non-local). Intra-chromosomal contacts between Hi-C and SPRITE are calculated for various SPRITE cluster sizes. **(e)** Inter-chromosomal contacts were computed for different SPRITE cluster sizes and p-values were generated to identify significant inter-chromosomal interactions. SPRITE clusters containing greater than 1000 reads per cluster are highly enriched for inter-chromosomal interactions that occur between chromosomes that organize around the nucleolus (12, 15, 16, 18, and 19 in mES cells), whereas clusters containing 2–10 reads per cluster are depleted for inter-chromosomal interactions.

Figure 4: Computational SPRITE Pipeline



(**a**) The SPRITE snakemake pipeline works as follows: The prerequisites (red boxes) before starting include installing conda and snakemake. Then, run fastq2json.py to compile the paths of all SPRITE fastq files into a .json file. Generate and index a Bowtie2 genome for genome alignments (mm10 and hg38 are currently supported). For each experiment, modify the config.yaml file to input the total number of tags corresponding to the number of ligation rounds and also the location of the reference genome index for alignment. The pipeline is then launched (white boxes) where (1) adaptor trimming is performed, (2) barcodes are identified, (3) reads without all SPRITE tags are removed, (4) the DPM tag is trimmed from the beginning of Read 1, (5) all reads are aligned to the genome using Bowtie2, (6) chromosome annotations are converted from Ensembl to UCSC format, (6) all regions that do not fall within repeat-masked or blacklisted genomic bins are retained, (7) all reads sharing the same barcodes are matched and collapsed into SPRITE clusters, and (8) heatmaps are generated. Certain QC files are output along the way to quantify ligation efficiency, plot SPRITE cluster sizes, and all summary statistics are outputted in MulitQC. (**b**) Because the number of pairwise contacts scales quadratically with the number of reads (n) contained within a SPRITE cluster, larger clusters will contribute a disproportionately large number of the contacts observed between any two bins. To account for this, we reasoned that a minimally connected graph containing n reads would contain n-1 contacts. Therefore, we down-weighted each of the (n choose 2) = (n(n-1)/2) pairwise contacts in a SPRITE cluster to contribute the same number of contacts of a minimally connected graph. This is achieved by down-weighting each pairwise contact by 2/n. In this way, the total contribution of pairwise contacts from a cluster is proportional to the minimally connected edges in the graph and will have n-1 contacts. This also ensures that the number of pairwise contacts contributed by a cluster is linearly proportional to the number of reads within a cluster. Here, we show an example of SPRITE cluster weighting using a cluster of 4 reads, where the total number of contacts is 6. Our SPRITE cluster weighting scheme computes all 6 possible contacts and down-weights each contact by n/2, such that the total number of contacts sum to 3.

Figure 5: SPRITE Identifies Higher-Order Interactions that Occur Simultaneously (copied with permission from Quinodoz et al. 2018).



(a) Compartment eigenvector showing A (red) and B (blue) compartments on mouse chromosome 2 (top). Individual SPRITE clusters (rows) containing reads mapping to at least three distinct A compartment regions (*) (middle). Pairwise contact map at 200-kb resolution (bottom). **(b)** H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq, ENCODE) signal across a 2.46-Mb region on human chromosome 6 corresponding to three TADs containing 55 histone genes (top). SPRITE clusters containing reads in all three TADs (middle). Pairwise contact map at 25-kb resolution (bottom). **(c)** CTCF motif orientations at three loop anchors on human chromosome 8 (top). SPRITE clusters overlapping all three loop anchors (middle). Pairwise contact map at 25-kb resolution (bottom)**. (d)** Schematic of multiple A compartment interactions. **(e)** Schematic of higher-order interactions of HIST1 genes (green). **(f)** Schematic of higher-order interactions between consecutive loop anchors.

3.17 SUPPLEMENTARY MATERIALS AND LEGENDS

Supplemental Table S1: (Sheet 1) Calculate number of molecules for NHS coupling. To calculate the amount of lysate to couple to NHS-activated beads, enter the average size (bp) and concentration (ng/ul) of DNA in the DNase digested lysate. This spreadsheet will calculate the number of DNA molecules in the DNase digested lysate and determine the µL of lysate to couple. **(Sheet 2) Calculate number of reads required to sequence each SPRITE library to saturation.** To determine the amount of reads required to sequence each SPRITE library aliquot to saturation, estimate the number of unique molecules pre-PCR from the final library concentrations using the library molarity and number of cycles.

Supplemental Table S2: Sequences of the SPRITE barcodes. All sequences of SPRITE barcodes (DPM, Terminal, Odd, Even) and the indexed Illumina sequencing primers are provided.

Supplemental Video 1: The main experimental steps of the SPRITE method are shown on how to execute all of the SPRITE steps, including split-and-pool barcoding and pre-library amplification steps. A conceptual overview of the SPRITE method method is also explained. Please see https://youtu.be/6SdWkBxQGlg for the video.

3.18 REFERENCES

1. Martin, C. *et al.* Genome restructuring in mouse embryos during reprogramming and early development. *Dev. Biol.* (2006). doi:10.1016/j.ydbio.2006.01.009

2. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).

3. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 772–772 (2016).

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

5. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Revelas Folding Principles of the Human Genome. *Science* **326**, 289–294 (2009).

6. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189–191 (2012).

7. Dekker, J. Capturing Chromosome Conformation. *Science (80-.).* **295**, 1306–1311 (2002).

8. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

9. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

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

11. Lawrence, J. B. & Clemson, C. M. Gene associations: True romance or chance meeting in a nuclear neighborhood? *J. Cell Biol.* **182**, 1035–1038 (2008).

12. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* (2018). doi:10.1016/j.cell.2018.05.024

13. Decker, C. J. & Parker, R. P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. *Cold Spring Harbor perspectives in biology* **4**, (2012).

14. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* (2017). doi:10.1038/nmeth.4380

15. Gaublomme, J. T. *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* (2019). doi:10.1038/s41467-019-10756-2

16. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (80-.).* (2017). doi:10.1126/science.aal3327

17. Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*

(2018). doi:10.1101/254797

18. Blecher-Gonen, R. *et al.* High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.* **8**, 539–554 (2013).

19. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* (2011). doi:10.14806/ej.17.1.200

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

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

22. Zhang, J. *et al.* ChIA-PET analysis of transcriptional chromatin interactions. *Methods* **58**, 289–299 (2012).

23. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Research* (2016). doi:10.1038/cr.2016.137

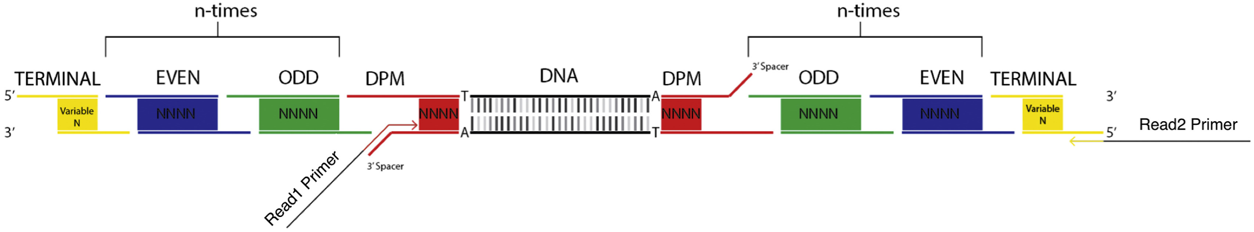
24. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

25. Terranova, C. *et al.* An integrated platform for genome-wide mapping of chromatin states using high-throughput chip-sequencing in tumor tissues. *J. Vis. Exp.* (2018). doi:10.3791/56972

26. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* (2013). doi:10.1038/nmeth.2375

3.19 BOXES

Box 1: Description of SPRITE adaptors and primers



The above figure demonstrates the adaptor and tag scheme that is central to the SPRITE process. SPRITE uses a split-and-pool strategy to uniquely barcode all molecules within a crosslinked complex. The final product contains a series of unique tags (DPM, Odd, Even, and Terminal tags) ligated to each molecule, which we refer to as a barcode.

SPRITE tags are ligated using the following oligos. First, the DPM is ligated to dA-tailed DNA. The DPM tag is then ligated by an “Odd” tag. The “Odd” and “Even” tags were designed so that they can be ligated to each other over multiple rounds, such that after Odd is ligated, then Even ligates the Odd tags, and then Odd can ligate the Even tags. This can be repeated such that the same two plates of 96 tags can be used over multiple rounds of split-pool tagging without self-ligation of the adaptors to each other. Finally, a set of barcoded Terminal tags are ligated at the end to attach an Illumina sequence for final library amplification.

As an example, in our previous study, we performed five rounds total of split-and-pool ligation in the following order: DPM, Odd, Even, Odd, and Terminal tag, but SPRITE can be adapted to perform an arbitrary number of rounds of split-pool barcoding. Additionally, to reduce the cost of adaptors purchased to perform SPRITE, a single DPM and single Terminal tag can be used in exchange for performing more rounds of 96 Odd and Even tags instead.

DPM Adaptor

```
5'Phos AAACACCCAAGATCGGAAGAGCGTCGTGTA   3’ Spcr
       ||||||||||||||||||||
   3' TTTTGTGGGTTCTAGCCTTCTGTACTGTTCAGT  5’Phos
```

The DPM adaptor is the first tag ligated to dA-tailed genomic DNA. Each DPM tag contains: (i) a 5’phosphate group on both ends for ligation to dA-tailed genomic DNA and to the Odd tags, (ii) a single dT-overhang for ligation to dA-tailed genomic DNA (yellow region); (iii) a nine nucleotide sequence unique to each of the 96 DPM tags (purple region); and (iv) a seven nucleotide sticky-end overhang that ligates to the Odd set of adaptors (green region). Each DPM tag also contains a partial sequence that is complementary to the universal Read1 Illumina primer, which is used for library amplification (gray region).

Because the DPM tag will ligate to both ends of the double-stranded DNA molecule, DPM was designed such that the barcode sequence will only be read from one sequencing read (Read2), rather than both reads (Reads 1 and 2). To achieve this, we included a 3′ spacer on the top strand. This prevents the top strand of the Odd tag from ligating to genomic DNA. This modification is also critical for successful amplification of the barcoded DNA by preventing hairpin formation of the single-stranded DNA during the initial PCR denaturation, because otherwise both sides of the tagged DNA molecule would have complementary barcode sequences.

Odd and Even Tags

After DPM ligation, two sets of 96 “Odd” and “Even” tags are ligated. The tags are named as such because Odd tags can be ligated in the 1st, 3rd, 5th,… rounds of the SPRITE process and the Even tags can be ligated 2nd, 4th, 6th,… rounds of SPRITE. The Odd tags contain a seven nucleotide sticky end (5’ CAAGTCA 3’) that anneals to the Even tags (5’ TGACTTG 3’), and the Even tags have a distinct seven nucleotide sticky end (5’ AGTTGTC 3’) that anneals to the Odd tags (5’ GACAACT 3’).

```
5'Phos CAAGTCAAGCTAGATTCCACGAAGAGTTGTCACGTCAGCCGCAGTATC          3′
              |||||||||||||||||||||||||||||||||||||||||
   3'         TCGATCTAAGGTGCTTCTCAACAGTGCAGTCGGCGTCATAGGTTCAGT 5′Phos
```

The above dsDNA molecule is an example of an Odd tag and an Even tag ligated together. The following points are important to note: (i) The 5’ overhang on the top strand of Odd ligates either to the DPM adaptor (green sequence on DPM) or the 5’ overhang on the bottom strand of the Even tag; (ii) the bolded regions on each tag are unique 17 nucleotide sequences for each of the 96 tags (192 total, accounting to Odd tags and Even tags); and (iii) both tags have 5’ phosphate groups to allow for sequential tag ligation. The 5’ overhang on the bottom strands of each of the Odd and Even tags can be ligated to a Terminal tag (designed with a complementary overhang for Odd or Even) to attach the Illumina sequence for library amplification.

Terminal Tag

The Terminal tags contain a seven nucleotide sticky end that ligates to the Odd tags (green region), though a Terminal tag can also be designed to ligate to an Even tag (blue region). The Terminal tag only contains a modified 5′ phosphate on the top strand. The bottom strand contains a priming region (gray) that contains part of the Illumina read 2 sequence, which allows for priming and incorporation of the full-length barcoded Read2 Illumina adaptor. Each Terminal tag contains a unique sequence of nine (or longer) nucleotides (**bold**). Specifically, to avoid any stretches of low-diversity sequences when sequencing the common sticky ends in read 2, we have found that adding a +0, +1, +2, +3 stagger in Terminal sequence length will offset when the sticky ends are read in read 2. This can prevent low-diversity reads during sequencing and makes the SPRITE sequencing compatible on a variety of Illumina Sequencers with one-color, two-color, and four-color chemistry.

Terminal Tag (Ligates Odd):

```
5' Phos   AGTTGTCACCATAATAAGATCGGAAGA             3′
                 ||||||||||||||||||||
3'               TGGTATTATTCTAGCCTTCTCGTGTGCAGAC  5′
```

Terminal Tag (Ligates Even):

```
5' Phos   CAAGTCAACCATAATAAGATCGGAAGA             3′
                 ||||||||||||||||||||
3'               TGGTATTATTCTAGCCTTCTCGTGTGCAGAC  5′
```

Final Library Amplification Primers

DNA libraries are amplified using common primers that incorporate the full Illumia sequencing adaptors. The Read 1 primer amplifies the top strand of the DPM tag on DNA (gray region), and adds the Illumina Read 1 sequence to each molecule. The Read 2 primer amplifies the Terminal tag on DNA (gray region) and adds the Illumina Read 2 sequence to the molecule.

The DPM adaptor is designed with a 3’ spacer to aid in final library amplification. If the 3’ spacer is absent, each strand will form a hairpin loop during the initial denaturation due to reverse complementarity of the barcode sequences on the other side of the target DNA molecule. Instead, the 3’ spacer allows the barcodes to only ligate to the 5’end of each single-stranded DNA sequence, and not the 3’end, preventing these hairpins from forming.

Read 1 Primer: 2P_universal

`5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′`

Read 2 Primer: 2P_barcoded_85

`5′CAAGCAGAAGACGGCATACGAGATGCCTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3′`

The 2P_barcoded primer contains an 8 nucleotide barcode (red region) within the primer. This barcode is read from the Illumina sequencer during the indexing priming step. This barcode effectively serves as an additional round of tag addition during SPRITE. Dilution of the sample into multiple wells is performed at the final step of SPRITE prior to proteinase K elution from NHS beads. Each dilution of the sample prior to proteinase K elution isolates a subset of the tagged complexes into different wells. Each aliquoted sample is amplified with a different 2P_barcoded primer.

Both the First and Second primers are around 30 nucleotides each. Yet the sequences they anneal to initially are ~20 nucleotides. For this reason, we set two different annealing temperatures during the final library PCR. The first annealing temperature (69°C) is for the first four cycles until enough copies are made with fully extended primer regions. After these four cycles, the annealing temperature is raised for a remaining five cycles (72°C).

Sequencing Recommendations

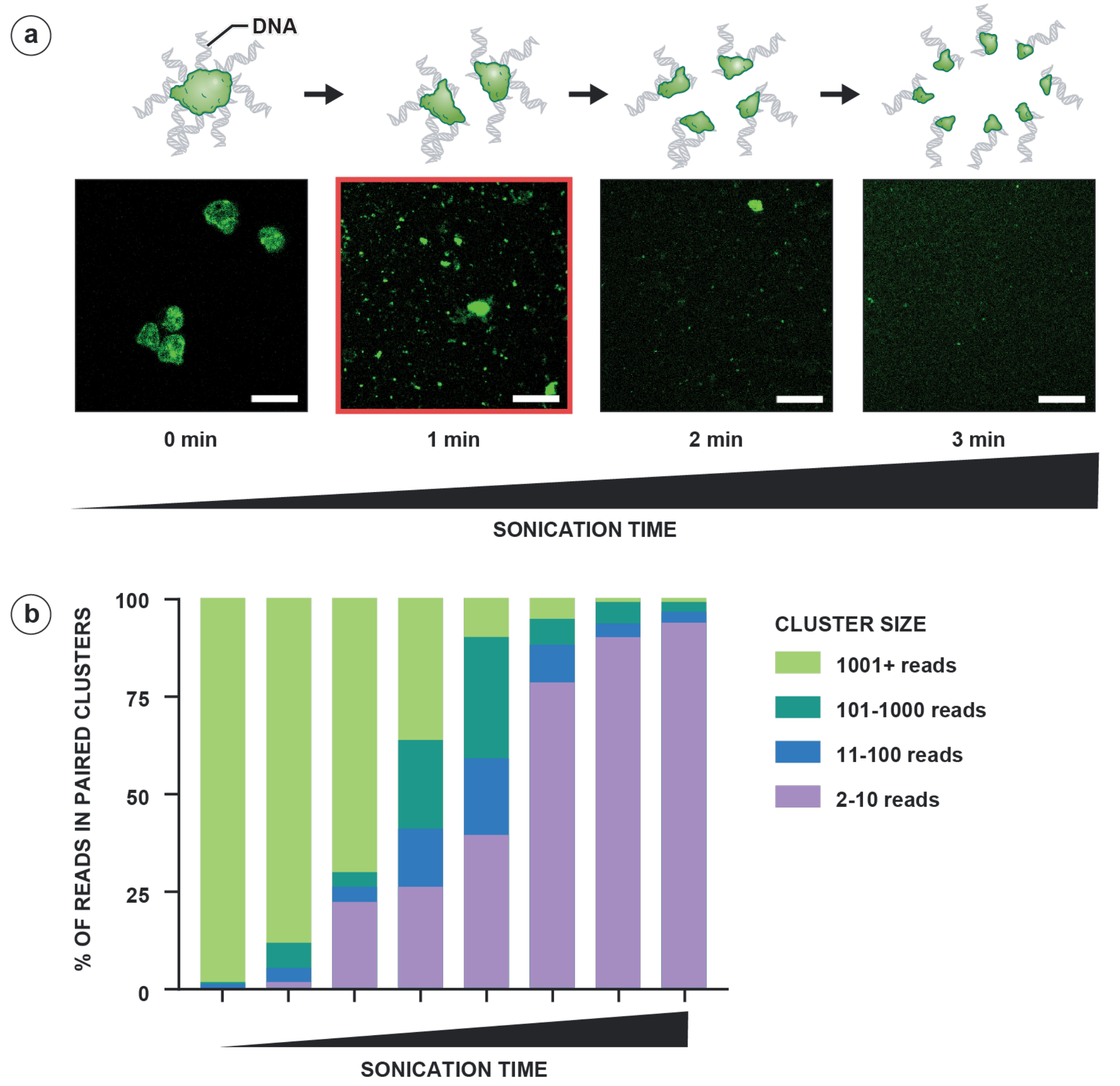
Read 1 begins with the DPM tag and is followed by the associated genomic DNA sequence. Therefore, Read 1 must be sequenced at least 50 bp to allow identification of the 10 bp DPM tag sequence and at least an additional 40 bp for sequencing the genomic DNA sequence. Longer reads will improve genomic DNA and allele-specific alignments.

Read 2 sequences the tags ligated to genomic DNA. Therefore, Read 2 must be long enough to sequence through all ligated tags. For example, if 4 rounds of barcoding were performed after DPM ligation, we recommend sequencing at least 95 bp to sequence through each tag.

Length of SPRITE barcode post-DPM = Odd_{len} + Even_{len} + Odd_{len} + $\text{Terminal}_{\text{len}}$

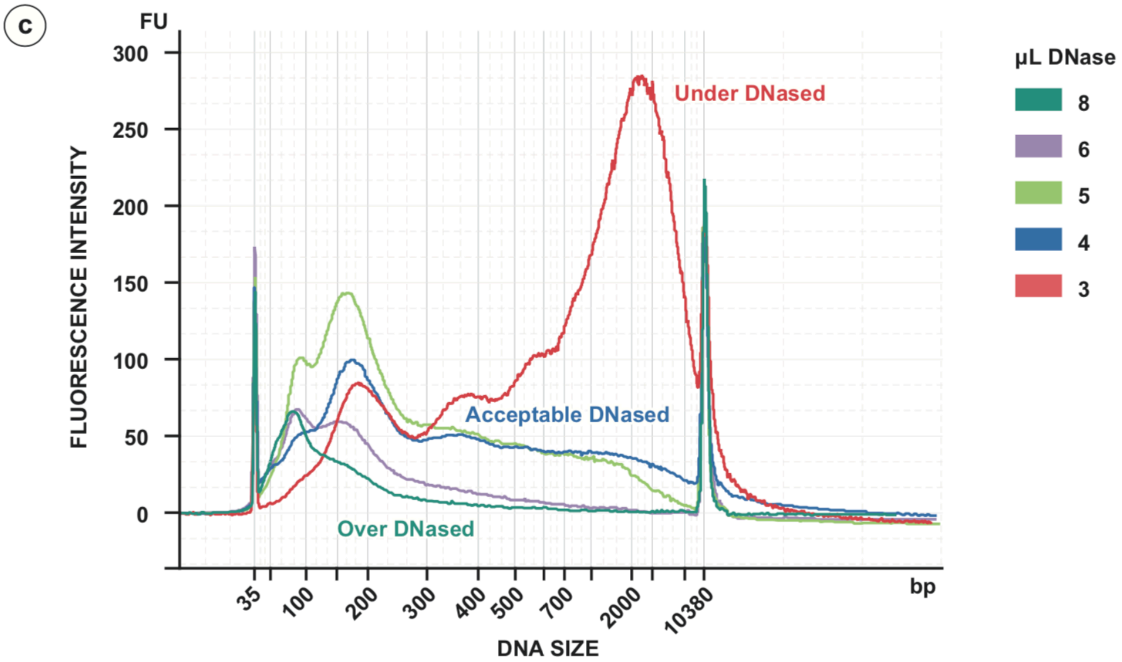
Length of SPRITE barcode post-DPM = 24 bp + 24 bp + 24 bp + (10 to 15 nt) = 82 to 87 nt

Box 2: QC 1: Visualizing sizes of crosslinked complexes after sonication and fragmentation with microscopy - 4 hrs



Optimizing fragmentation of DNA is an important step in the SPRITE protocol. The goal is to break down intact cells into smaller crosslinked complexes of interacting DNA molecules. (**Box 2a, scalebar = 10µm**). Care should be employed to avoid over-fragmentation or under-fragmentation as under-fragmentation using sonication will result in large SPRITE clusters with almost no small crosslinked complexes, or over-fragmentation will result in almost no larger

SPRITE clusters and only small clusters and unpaired interactions (**Box 2b**). One way to assess the distribution of crosslinked complexes is to directly visualize crosslinked complexes by microscopy. Successful SPRITE libraries described here have been obtained after 1 minute of sonication using 4–5 W of power (**Box 2a, red box**) followed by DNasing to digest the DNA down to a sequenceable length (~100bp-1kb) (**Box 2c**). Here we describe a microscopy test that can be employed to qualitatively evaluate successful fragmentation of cells into smaller complexes prior to deciding whether to proceed with the SPRITE procedure.



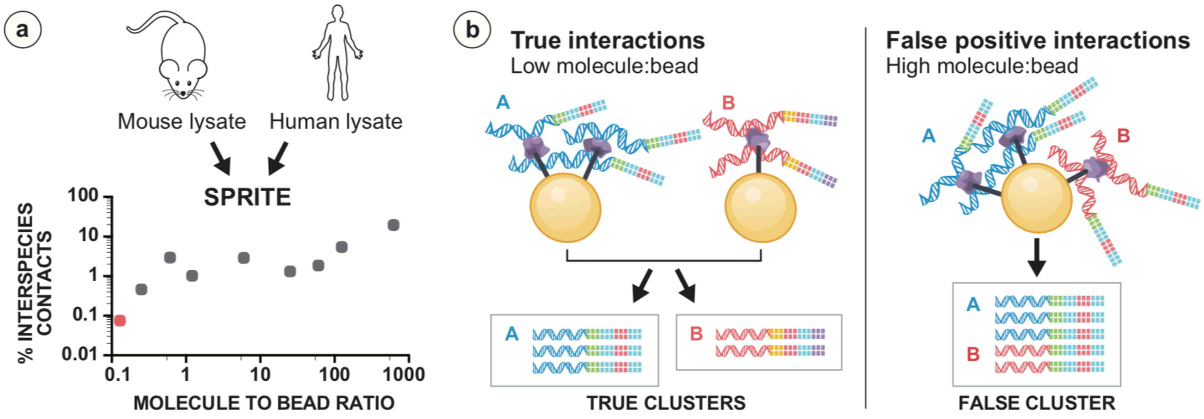
The three samples we recommend preparing for this QC are as follows:

1) No cells (to measure background staining)
2) Positive control: non-sonicated intact cells
3) SPRITE sample cells post-sonication and DNase digestion

DNA Stain

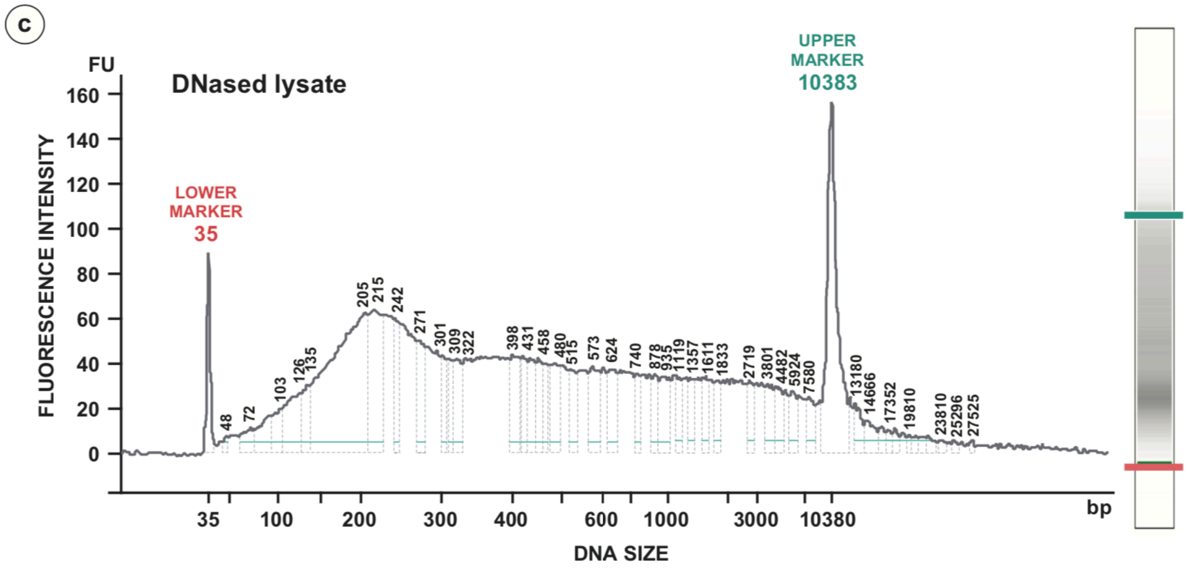
1. Draw three small circles on one side of the mounting glass slide with a fine tip pen. These circles will mark where to spot cells.
2. Dilute 0.1mg/mL Poly-D-lysine stock 1 to 5 in ddH2O to final concentration of 0.02mg/mL.
3. On the other side of the slide, add 5 µL of 0.02mg/uL Poly-D-lysine inside the circles to coat the slide.
4. Dry slides at room temperature for 1 hour or until Poly-D-lysine has completely dried.
5. Place 2 µL of sonicated cell lysate on top of poly-lysine spot.
    a. 2 µL of SPRITE cell lysate corresponds to 33,000 cells. We have successfully imaged down to 2000 cells with this procedure.
6. Dry slides at room temperature for 1 hour or overnight.
7. Incubate dried slides for 1 hour in coplin jars with one of the following DNA dye solutions to visualize the genomic DNA in cells or fragmented cells:
    a. Hoechst 33342 Solution at a recommended dilution of 1:2000 (Ex350/Em461)
    b. SYBR® Gold at a recommended dilution of 1:10.000 (Ex300/Ex494/Em537)
    c. YOYO™-3 Iodide at a recommended dilution of 1:10.000 (Ex612/Em631)
8. Wash slides in a coplin jar filled with 1xPBS for 10–30 minutes.
9. Remove the slide from the coplin jar.
10. Mount with mounting medium (e.g., ProLong Gold Antifade Mountant) and cover glass. Optional: mounting medium containing DAPI may also be used as an additional stain for DNA.
11. Image cells with a confocal microscope.
    a. We image using the LSM800 confocal microscope; objective 100x

Box 3: Molecule-to-bead calculations for NHS coupling with reduced rates of false-positive interactions



Critical to ensuring a successful SPRITE dataset involves coupling the correct number of molecules to beads. When performing SPRITE for the first time, it may be worthwhile to do a human-mouse (or other inter-species) mixing experiment to assess the frequency of cross species pairs as a proxy for non-specific pairing (**Box 3b**). In a human-mouse mixing experiment, we have found that this will result in an increased rate of human and mouse inter-species false-positive interactions (**Box 3a**). Specifically, molecule-to-bead ratios ≥ 1 promotes multiple complexes on the same bead. Instead, low molecule-to-bead ratios (e.g., ≤ 1 molecule-per-bead) ensure that a maximum of one crosslinked complex per bead is coupled and have a reduced number of inter-species pairing.

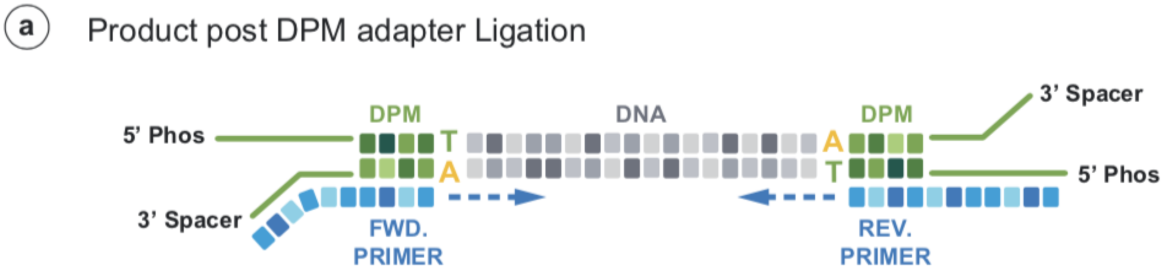
In a typical SPRITE experiment, we aim to bind at a 1:4 to 1:1 ratio of DNA molecules:beads. Assuming that we have 50% NHS coupling efficiency, the coupling ratio is then 1:8 to 1:2 molecules per bead. Generally, we bind around between 9.5 to 19 billion molecules to 19 billion beads, corresponding to 2mL of Pierce NHS-activated beads. We note that we use molecules as an overestimate to reduce noise-per-bead. This is because there should be far more molecules than crosslinked complexes in the SPRITE lysate. As a result, the complex-to-bead ratio should be even lower than stated above.



The number of DNA molecules in the lysate is estimated from the DNA concentration of the lysate after DNase digestion, reverse crosslinked, and column cleaned. We begin the molecule calculation after lysate is DNased, cleaned, and concentrated in 10 ul H2O in Steps 42–47 of the protocol. Steps 48 and 49 from the protocol post-DNasing determine the concentration and size distribution of DNA (**Box 3c**) in each sample using Qubit HS DNA dye and the Agilent DNA Bioanalyzer or Tapestation, respectively. We have provided **Supplemental Table 1** to estimate the number of molecules present in the remaining 10 ul of lysate and calculate the volume of lysate to couple to NHS beads.

Box 4: QC 2: Check to Determine Ligation Efficiency of the DPM Adaptor (Timing 3.5 hours)

After DPM ligation, we recommend removing a 5% aliquot of the ligated material and performing a PCR to amplify DPM-ligated material. This will ensure that DPM was successfully ligated prior prior to proceeding with the subsequent split-and-pool ligation steps.



Product post DPM adapter Ligation

1. Resuspend sample in 1mL SPRITE wash buffer.
2. Vortex well and remove 5% aliquot of beads by transferring 50uL of beads solution to another 1.7mL tube.
3. Resuspend beads in 50uL of SPRITE ProK buffer.
4. Add proteinase K enzyme and additional SPRITE ProK buffer to elute DNA by reversal of crosslinks through heating and proteinase K.

Stock Solution	**Volume**
Sample of Beads in SPRITE ProK Buffer	50uL
SPRITE ProK Buffer	42uL
Proteinase K	8uL
Total	100uL

5. Incubate at 65° C for 2 hours at minimum ✓.

   ✓ As a stopping point, you may reverse crosslink for 55° C for 1 hour, then increasing to 65° C for overnight incubation (>12 hrs). We highly recommend reverse crosslinking for at least 2 hours.

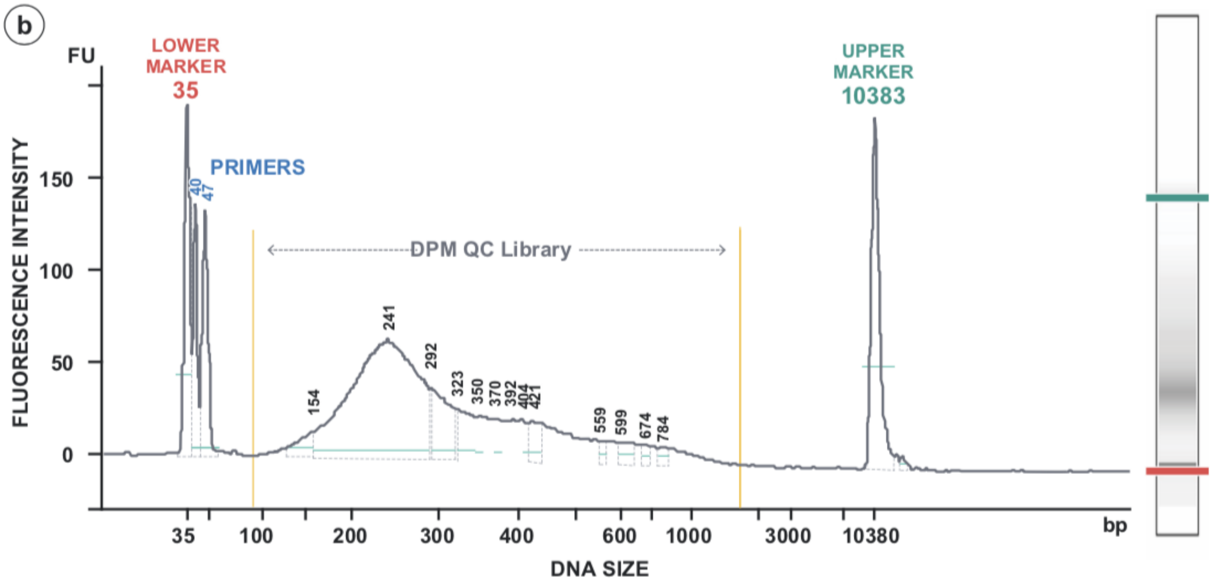
6. Place the microcentrifuge tube on a magnet and capture the beads. Transfer the supernatant containing the DNA to a clean 1.7 mL microcentrifuge tube.
7. Rinse the beads by pipetting 25 uL of H2O into the tube containing the beads. Vortex, briefly centrifuge, and re-capture the beads. Transfer the 25 uL of H2O that now contains any residual nucleic acid and combined with the previous supernatant in the new sample tube. Discard the beads.
8. Clean the DNA by following the protocol provided in the Zymo RNA Clean and Concentrator Kit (>17 nt), binding with 2 volumes of RNA Binding Buffer and 1 total volume of 100% ethanol. Elute in 40 uL of H2O total, eluting twice 20uL volume for a combined eluate for 40uL total.
9. Amplify the DNA molecules that are ligated to DPM. The forward primer should prime off the 5' end of the DPM adaptor and the reverse primer should prime off the 3' end of the DPM adaptor.

Stock Solution	**Volume**
Sample (cleaned)	20 uL
DPMQC Forward Primer (10uM)	2.5 uL
DPMQC Reverse Primer (10uM)	2.5 uL
NEB 2x Q5 Hot Start Master Mix	25 uL
Total	50 uL

PCR Program:

1. Initial denaturation: 98° C - 120 seconds
2. 14–16 cycles:
    a. 98° C -10 seconds
    b. 67° C - 30 seconds
    c. 72° C - 40 seconds
3. Final extension: 72° C - 120 seconds
4. Hold 4C

7. Clean the PCR reaction and size select for your target DNA molecules. Our DPM adaptors are 30 base pairs each and our target DNA molecules no fewer than 100 base pairs. Agencourt AMPure XP beads size select while cleaning the PCR reaction of unwanted products.
8. Add 2.0 x volume (100 uL) of AMPure XP beads to the sample for a total volume of 100uL and mix thoroughly by pipetting up and down at least 5 times.
9. Incubate for 10 minutes at room temperature.
10. Place the beads on an appropriately sized magnet to capture the beads and the bound DNA. Wait a few minutes until all the beads are captured.
11. Remove the supernatant and discard.
12. Wash beads twice with 200uL of 80% ethanol by pipetting ethanol into the tube while beads remain captured on the magnet, moving the tube to the opposite side of the magnet so that beads pass through the ethanol, and then removing the ethanol solution.
13. Quickly spin down the beads in a microcentrifuge, re-capture on magnet, and remove any remaining ethanol.
14. Air-dry beads while the tube is on the magnet until beads start to completely dry.
15. Elute the amplified DNA from the beads by resuspending the beads in 12 uL of H2O. Place the solution back on the magnet to capture the beads. Remove the eluted amplified DNA to a clean microcentrifuge tube.
16. Determine concentration of DNA amplified by measuring 1–2uL of cleaned PCR library following the directions provided with the Qubit dsDNA HS Assay Kit.
17. Determine the size distribution of DNA with the DPM adaptor in each sample following the directions provided with either the High Sensitivity DNA Kit for the Agilent Bioanalyzer or the HS D1000 Screentape for the Agilent 2200 TapeStation. The average size should be roughly similar to the average size of the input lysate (around 200–400 base pairs).



18. Estimate the number of unique molecules pre-amplification using **Supplemental Table 1** to confirm that DPM has been successfully ligated.

✖ CRITICAL: The 5% aliquot should contain, at the very minimum, 15 million unique DNA molecules in order to proceed with SPRITE. If the aliquot contains fewer than this number, there will not be enough unique reads to sequence the SPRITE library. If this is the case, troubleshoot what went wrong by assaying coupling efficiency from the flowthrough saved in Step 64. If lysate was successfully coupled, consider whether a mistake was made during ligation of the DPM adaptor or during one of the critical steps of crosslinking and lysis.

DPM primers for QC of DPM ligation

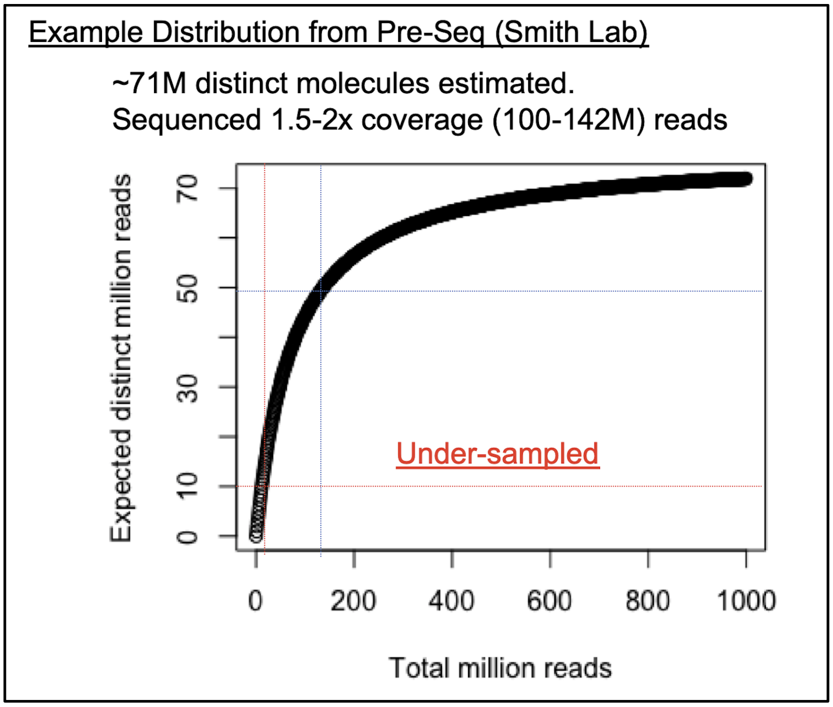
These primers are used to ensure that the DPM adaptor has been successfully ligated to DNA of the lysate. If no visible libraries are obtained at this step after 14–16 cycles of PCR, we strongly recommend to not proceed as subsequent ligation of tags and amplification of tagged DNA during the SPRITE protocol will be unsuccessful.

DPMQCprimerF	5′ TACACGACGCTCTTCCGATCT 3′
DPMQCprimerR	5′ TGACTTGTCATGTCTTCCGATCT 3′

The Forward and Reverse primers amplify the top strand and bottom strand of the DPM adaptor, respectively (see **Box 1**).

Box 5: Sequencing a SPRITE Library to Saturation

Sequencing a SPRITE library is different from sequencing a standard DNA library where it is standard to sequence a fraction of molecules present in the sample in order to avoid PCR duplicates. In contrast, a SPRITE library must be sequenced to saturation in order to sequence most, if not all, molecules that are interacting in a given crosslinked complex. For example, if we assume Poisson sampling during sequencing, a sequencing depth of 1x (e.g., 71 million reads to sequence 71 million unique molecules) would sample 63% of all molecules. For SPRITE libraries, we sequence 1.5 - 2x coverage, which ensures sampling of >77–86% of unique molecules contained within an individual cluster. Additional sampling of unique molecules would require sequencing hundreds of millions of more reads for modest increases in unique molecule sampling and will result in sequencing mostly PCR duplicates.



Example Distribution from Pre-Seq (Smith Lab)

~71M distinct molecules estimated.
Sequenced 1.5-2x coverage (100-142M) reads

To determine the sequencing depth required for each sample, we approximate how many unique molecules are present in the SPRITE library by measuring the number of molecules post-PCR and then account for (i) how many rounds of amplification $2^{(cycles-1)}$ were performed, and (ii) how many SPRI cleanups were performed, each resulting in ~50% loss in molarity.

First, measure the molarity of a SPRITE library to estimate the number of molecules present in the SPRITE library following amplification. We do this in Step 122 by measuring the average size (bp) and concentration (ng/µL) of the SPRITE library. Input the numbers from Steps 120 and 121 into **Supplemental Table 1** to calculate the total number of unique molecules present in the sample prior to library amplification. The step-by-step calculations for estimating this value are as follows:

a = library average size (bp)
c = library concentration (ng/ul)
n = number of cycles amplified
p = molecules post-PCR
m = library molarity (nM)
v = volume of library
u = number of unique molecules pre-PCR

1. First calculate the number of molecules present in the SPRITE library post-amplification.
   $p = 10^6 * c / (649* a)$
2. Calculate the number of unique molecules present in the SPRITE library.
   u = 2 * p / (2^(n-1))
   u = 4p/(2^n)
3. Calculate the sequencing depth required to achieve >70% coverage
   Sequencing depth = 2*u
4. QC: Use Preseq to estimate the complexity of the SPRITE library and whether at least 50% saturation has been successfully achieved.

Box 6: Computational Pipeline

The automated SPRITE computational pipeline allows users to quickly and easily generate contact matrices from raw sequencing reads. The end result of the standard pipeline generates a genome-wide heatmap representing chromosome-level interactions. In the following box, users can see step by step how the SPRITE pipeline goes from raw fastqs to cluster and heatmap generation.

Trimming

1. Remove Illumina sequencing adaptors from paired-end reads using Trim Galore!.

Barcode Identification

2. The next step of the SPRITE pipeline is to identify the barcodes of the sequenced reads. This process is run by a standalone Java program called BarcodeID that accepts two input FASTQ files (paired-end sequencing), and outputs modified versions of these FASTQ files for subsequent alignment. A typical run command looks as follows:

```
java -jar BarcodeIdentification.jar \
--input1 read1.fastq.gz \
--input2 read2.fastq.gz \
--output1 read1.barcoded.fastq.gz \
--output2 read2.barcoded.fastq.gz \
--config configuration_file.txt
```

3. The inputs to this program are the two FASTQ files from a paired-end sequencing run. These files can be gzipped or uncompressed. The outputs are the same FASTQ files with the identified tags appended to the read name. The output files will be gzipped.

   Example input:

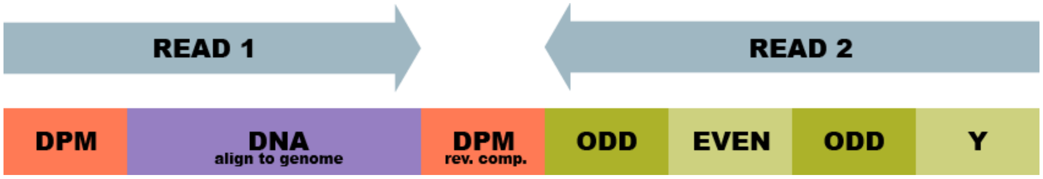
```
@HISEQ:623:HY5KHBCXX:2:2206:7231:7108
ATTGNTAGGTCGGAATTGCACGCTGTAGCGGCATGCTGATGGAGAGGAGAGACTTCTAGCTAGCTACGTGA
CTGATCCGCACACTGCGACACGTGATCGC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Example output:

```
@HISEQ:623:HY5KHBCXX:2:2206:7231:7108::[DPM6A5][NYBot35_Stg][NOT_FOUND]
[Even2Bo19][Odd2Bo69]
ATTGNTAGGTCGGAATTGCACGCTGTAGCGGCATGCTGATGGAGAGGAGAGACTTCTAGCTAGCTACGTGA
CTGATCCGCACACTGCGACACGTGATCGC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

4. In the above example, four of five tags were successfully identified (a DPM sequence named "DPM6A5", a Y-shape sequence named "NYBot35_Stg", an even sequence named "Even2Bo19", an odd sequence named "Odd2Bo69"). No sequence in position three was found, so a NOT_FOUND tag was inserted. This is the standard behavior when a position cannot be identified.

5. The configuration file specifies the base sequence of every possible tag, as well as the ordering of the tags within a valid barcode. An example tag ligation schemes used in our lab is the following:



Not shown are sticky-end sequences. The first thirty lines of configuration file used to analyze this layout is below. The complete file can be found here: https://github.com/GuttmanLab/sprite-pipeline/blob/master/example_config.txt

- The first few lines specify the layout in both read 1 and read 2. If tags are only in read 1 or read 2 then only the corresponding line is kept. Tag categories are separated by pipes. The valid categories (DPM, EVEN, ODD, etc.) have no implicit meaning, and are just used to divide tags into categories. The SPACER category is a special category to represent sticky ends. The SPACER variable defines the length of the sticky end and the LAXITY variable defines the maximum number of bases to proceed in the search if a barcode is not immediately found. In our design, the SPRITE barcodes are 7nt, but we provide a SPACER of 6nt to account for a possible 1nt indel in the sticky-end sequences and a LAXITY of 6nt to account for any issues in barcode identification.

- The tag section consists of many tab-delimited lines. Each line has four fields and represents a unique tag sequence. The fields, in order, are:
  - the tag category (DPM, EVEN, ODD, etc.)
  - the tag name
  - the tag sequence
  - the tag error tolerance
- The tag error tolerance field allows for some tags to have base mismatches or miscalls. A value of zero means to only accept the specified sequence as a match; a value of one means to accept any sequence within one Hamming distance of the specified sequence as a match; and so on for two, three, etc.

6. Reads that do not contain a full complement of barcodes are filtered out at this stage using the get_full_barcodes.py script:

```
python get_full_barcodes.py --r1 file.R1.barcoded.fastq.gz
```

The output of this script is two fastq files, one with reads containing the full complement of barcodes example.R1.barcoded_full.fastq.gz and the other containing everything else example.R1.barcoded_short.fastq.gz

DPM Trimming

In most cases, the DPM identity is derived from read 1, while the remaining barcodes come from read 2. As a result, the residual DPM sequences that would hinder alignment can only be removed after barcode identification with a second round of trimming using Cutadapt[12]. Cutadapt searches for the common 5’ DPM sequence (GATCGGAAGAG), as well as the 96 barcoded 3’ DPM sequence (dpm96.fasta), accounting for cases where read 1 extends through the genomic DNA sequence into the SPRITE tags ligated on the 3’end of the fragment.

Alignment to Genome

7. After the barcodes have been identified, the reads are aligned to the reference genome. In our standard ligation scheme, only read 1 contains genomic DNA and read 2 contains the SPRITE barcodes, so only read 1 is aligned. Therefore, we do not perform a paired-end alignment, despite having paired-end reads.

8. Any aligner will work. We use bowtie2:

```
bowtie2 -p 10 -t --phred33 -x <bowtie2_index> -U
example.R1.barcoded_full.fastq.gz | samtools view -bq 20 -F 4 -F 256 -
> example.DNA.bowtie2.mapq20.bam
```

   - -p specifies the number of cpu threads to use, -t prints the wall clock time required to load the index and perform the alignment, --phred33 specified the read quality encoding, which in this case is the latest one used by Illumina sequences, -x points to the location of bowtie2 indexes and -U specifies the location of the fastq file to be aligned.
   - We filter on MAPQ score of 20, outputting only mapped reads (-F 4) and removing reads that are not the primary alignment (-F 256).

9. If the reference sequences used came from Ensembl, we additionally convert chromosomes to the UCSC style (chr1, chr2, etc.).

```
python ensembl2ucsc.py -i example.DNA.bowtie2.mapq20.bam -o
example.DNA.chr.bam --assembly mm10
```

Repeat masking

10. We use bedtools to discard reads that overlap the annotations in our mask file. The mask file is actually a composite of two separate mask files:

    - UCSC Repeatmasker, millidiv less than 140
    - mm10 blacklisted regions v2

The composite mm10 and hg38 mask file that we use can be found in the main directory of the SPRITE DNA pipeline.

```
bedtools intersect -v -a example.DNA.chr.bam -b mask_file.bed >
example.DNA.chr.masked.bam
```

Cluster generation

11. To define SPRITE clusters, all reads that have the same barcode sequence are grouped into a single cluster. To remove possible PCR duplicates, all reads with the same genomic position and an identical barcode are removed. We generate a SPRITE cluster file for all subsequent analyses, where each cluster occupies one line of the resulting text file containing the barcode name and genomic alignments **Troubleshooting?**.
12. The SPRITE cluster generation script can be run as follows, where N is the number of tags in a barcode:

```
python get_clusters.py --input example.DNA.chr.masked.bam --output
example.clusters --num_tags N
```

13. **Output file format:** The below line represents a single cluster of size three. The first column is the barcode itself ending with the sample name. Each subsequent column contains the type of SPRITE library (DNA), the strand in square brackets, followed by the alignment chromosome, start and end coordinates:

```
DPM6A5.NYBot35_Stg.Odd2Bo71.Even2Bo19.Odd2Bo6.example
DNA[+]_chr1:18884355-18884455 DNA[-]_chr1:18834000-18834100
DNA[+]_chr1:200041887-200041900
```

Heatmap generation

14. The final step of the workflow is to transform a file of clusters into a file of contacts. A contacts file is a text file containing a simple square matrix of values representing the contact strength or contact frequency between any two points on the genome. This matrix is similar to an adjacency matrix, and can be easily plotted with MATLAB or with R. An example of DNA interaction matrices can be seen in **Figure 4**.

15. To compute the pairwise contact frequency between genomic bins *i* and *j*, we counted the number of SPRITE clusters that contained reads overlapping both bins. Specifically, we counted the number of unique SPRITE clusters overlapping the two bins and not the number of reads contained within them. In this way, if a SPRITE cluster contains multiple reads that mapped to the same genomic bin, we only count the SPRITE cluster once to eliminate possible PCR duplicates.
16. **Cluster weighting:** Because the number of pairwise contacts scales quadratically based on the number of reads (*n*) contained within a SPRITE cluster, larger clusters will contribute a disproportionately large number of the contacts observed between any two bins. To account for this, we reasoned that a minimally connected graph containing *n* reads would contain *n*-1 contacts. Therefore, we down-weighted each of the (*n* choose 2) = (*n*(*n*-1)/2) pairwise contacts in a SPRITE cluster such that each pairwise contact is weighted by 2/*n*. In this way, the total contribution of pairwise contacts from a cluster is proportional to the minimally connected edges in the graph and will have *n*-1 contacts. This also ensures that the number of pairwise contacts contributed by a cluster is linearly proportional to the number of reads within a cluster. (See **Figure 5b**).
17. We recommend making heatmaps using “n over 2” weighting to generate heatmaps capturing both long-range (from larger clusters) and short-range (from smaller clusters) interactions.

18. The basic steps are as follows:
    a. Divide the genome or chromosome-of-interest into N bins of a given resolution. Finer resolutions will take longer to run and require much more memory. Initialize a contact matrix with dimensions N-by-N.
    b. **Raw-contacts**: For each cluster in the input clusters file, generate all pairs of reads and record the implied contacts by incrementing the corresponding matrix cells. A value of one-per-contact will be added to the cells if no down-weighting is applied; otherwise the value will be some value less than one that depends on the down-weighting strategy.
    c. **ICE-normalization:** Generate a vector of N bias factors corresponding to each row in the matrix. (The bias factor for a given row will be identical to the bias

factor for a given column since the matrix is symmetric.) For each cell in the matrix, divide by (bias_factor(row_num) * bias_factor(col_num)). These bias factors are generated with Hi-Corrector[18].

d. **Final-heatmaps**: This will transform all values in the matrix to a value between zero and one to make the color scale easier to work with and to compare between samples. Currently this is done by calculating the median value of all cells one-off from the diagonal and dividing by it. Any values which were originally greater than this median value (and so would be greater than one after division) are simply set to one.

e. Run python get_sprite_contacts.py: Example script:

```
python get_sprite_contacts.py --clusters
[sample_name_here].clusters --raw_contacts
[sample_name_here]_raw_contacts.txt --biases
[sample_name_here]_biases.txt --iced [sample_name_here]_iced.txt
--output [sample_name_here]_final.txt --assembly mm9 --chromosome
genome --max_cluster_size 1000000 --min_cluster_size 2 --
resolution 1000000 --iterations 100 --down-weighting n_over_two -
-hicorrector hicorrector/1.2/bin/ic
```

Description of the flags required for heatmap generation:

- --clusters: The input clusters file from the previous step.
- --raw_contacts: The down-weighted output filename.
- --biases: Hi-Corrector outputs a text file of bias factors. Save them to this filename.
- --iced: The ICEd output filename.
- --output: The final, transformed output filename.
- --assembly: Either “mm9”, “mm10”, “hg19”, “hg38”. Used to initialize the contact matrix with the appropriate size.
- --chromosome: For intra-chromosomal heatmaps, one of “chr1”, “chr2”, ..., “chrX”. For inter-chromosomal heatmaps, “genome”.
- --min_cluster_size and --max_cluster_size: Ignore clusters that fall outside these parameters, i.e., clusters that are too big or too small. Default: 2 to 1000.

- --resolution: The binning resolution in nt. Default: 1000000 (1 Mb).
- --down-weighting: The down-weighting strategy, one of “none”, “n_minus_one”, and “n_over_two ”.
    - none: Each contact contributes a value of 1 to the contact matrix in Step 2. For example, a contact from a 5-cluster will contribute 1.
    - n_minus_one: Each contact contributes a value of 1/(N - 1), where N is the size of the corresponding cluster. For example, a contact from a 5-cluster will contribute 1/4.
    - n_over_two: Each contact contributes a value of 2/N, where N is the size of the corresponding cluster. For example, a contact from a 5-cluster will contribute 2/5. We recommend n_over_two weighting.
- --hi-corrector: The path to the Hi-Corrector ic executable. This defaults to the correct location on the Guttman Lab workstation.
- --iterations: The number of iterations to perform when running Hi-Corrector. Default: 100.

f. **Visualize heatmap in R**

a. Run `plot_heatmap.R`.

b. This takes a “final.txt” heatmap file and max value to define the cutoff of a heatmap in R.

Chapter 4

RNA DEMARCATES SPATIAL TERRITORIES IN THE NUCLEUS

4.1 ABSTRACT

It has long been proposed that RNA can play important regulatory and structural roles in the nucleus. While individual RNAs have been identified that demarcate specific nuclear bodies (e.g., pre-ribosomal RNA and nucleolus) and even shape nuclear structure (e.g., Xist and Barr body), it remains largely unclear which RNAs localize in the nucleus and where they localize. For example, a recent study identified that RNA transcribed from CoT1 repeats broadly localize across the entire nucleus in mammalian cells, yet *which* RNAs are within this largely heterogeneous population and *what* specific regions of the nucleus they are localized at remains unclear. The main challenge is that we lack methods to comprehensively define where specific RNAs localize within higher-order structures in the nucleus. Recently, we developed SPRITE, a split-pool barcoding method that enables higher-order mapping of RNA and DNA interactions within the nucleus. Here, we present a dramatically improved implementation of the SPRITE method (SPRITE 2.0) that enables comprehensive mapping of all classes of RNA in the nucleus, from abundant RNAs encoded from DNA repeats to low abundance RNAs such as nascent pre-mRNAs and lncRNAs. We find that RNAs localize broadly across the nucleus, with individual RNAs localizing within discrete territories ranging from nuclear bodies to individual TADs. Because we can measure nascent pre-mRNAs on chromatin, SPRITE allows us to explore how mRNAs are spatially organized during transcription. We uncover that nascent mRNAs interact in structures corresponding to nascent mRNA chromosome territories and compartments. By integrating higher-order measurements of RNA localization, we explore how RNA localization in specific compartments can impact mRNA transcription, providing important insights into RNA function. Together, these results uncover a central and widespread role for non-coding RNA in demarcating 3D nuclear structures within the nucleus.

4.2 INTRODUCTION

Gene regulation entails the precise coordination of RNA transcription and processing in the nucleus. To produce proteins, RNA is transcribed across the genome into mRNAs that are exported to the cytoplasm, where they act as the transient messenger to encode protein sequences. Yet, in addition to this primary role of mRNAs in encoding protein, it has long been proposed that RNA can play structural and regulatory roles in the nucleus. For instance, early studies in the 1980s showed that global disruption of RNA using RNase leads to large-scale morphological defects in nuclear structure. Yet, this result has remained controversial largely because specific RNAs that were involved in such roles remained unknown and therefore their role could not be studied.

Since then, there have been several additional lines of evidence demonstrating an important role for RNA in the nucleus. For instance, recent studies using microscopy have revealed that many RNAs are transcribed in and retained on chromatin, making up the "nuclear matrix" of RNAs retained on chromatin[1]. For example, the Lawrence lab showed that the C0t-1 RNA, which is composed of various RNAs transcribed from **DNA repeats**, localizes broadly across the nucleus, with most regions of the nucleus coated with RNA. Notably, these RNAs, thought to be composed of RNAs encoded from satellite, LINE, and SINE DNA repeats, appear to be stable and highly associated to chromatin despite transcriptional inhibition[1]. Additionally, disruption of these RNAs results in large-scale differences to nuclear organization. While these results suggest a pervasive role for RNA in the nucleus, which specific RNAs are represented in C0t-1 remains uncharacterized and what specific regions of DNA they associate with also remains unknown. This has made it challenging to study to what extent these RNAs encoded from DNA repeats can shape nuclear organization.

Various studies have demonstrated that RNA is a key component of various bodies in the nucleus. Specifically, multiple **nuclear bodies** are demarcated and organized around RNA. For instance, the nucleolus is one of the largest RNA bodies in the nucleus, which is arranged around transcription of ribosomal RNA and contains snoRNAs (small nucleolar RNAs)[2,3]. Indeed, disruption of rRNA transcription leads to disruption of this nuclear body. Other nuclear bodies, nuclear speckles were initially identified based on aggregation of various mRNA splicing proteins. They were subsequently shown to contain a high concentration of various non-coding RNAs including the Malat1 long non-coding RNA (lncRNA) and the classically defined spliceosomal RNAs (U1, U2, U4, U5, U6)[4,5]. Another nuclear body containing RNAs is the histone locus body, which contains histone mRNAs and U7 RNA, which plays a role processing histone mRNAs[6]. Further, the Cajal body is an RNA body enriched with various small Cajal body-enriched RNAs (scaRNAs) that has been implicated in processing the classically defined spliceosomal RNAs (U1, U2, U4, U5, U6) as well as enriched for snoRNAs[7,8].

In addition to demarcating nuclear bodies, RNA has also been proposed to play important **regulatory roles** in the nucleus. Specifically, certain RNAs can guide different chromatin regulatory proteins to genomic DNA in order to form higher-order chromatin structures within the nucleus. For example, the HP1 protein, which is required for heterochromatin formation, has been shown to require RNA to mediate its role in forming higher-order structure and heterochromatin formation[9]. Additionally, lncRNAs can guide chromatin regulatory proteins to specific genomic sites to regulate gene expression and to shape 3D chromatin structure. A key example is the Xist lncRNA, which recruits the SHARP chromatin modifying protein to the inactive X chromosome in order to silence gene expression and to form the Barr body[10]. However, while significant progress has been made to study Xist and Malat1, two highly abundant long non-coding RNAs, the localization patterns of the vast majority of lncRNAs, which are very low abundance, remain unmapped.

Despite these examples of individual RNAs that have been identified within specific nuclear bodies and even shape nuclear structure, it remains largely unclear which RNAs are retained in the nucleus and where the vast majority of RNAs precisely localize in the nucleus. This is largely due to limitations in existing methods for mapping RNA in the nucleus. Specifically, existing methods cannot simultaneously measure both RNA localization and the 3D organization of chromatin - making it challenging to understand how RNA can shape nuclear structure. Additionally, RNA is expressed across at different expression levels making it important, but challenging, to map high abundance and low abundance RNAs simultaneously. Further, compared to DNA, RNA is intrinsically heterogeneous in its localization (e.g., RNAs are transcribed in the nucleus and then can localize in other parts of the cell as mature RNAs to be translated). Further, many nuclear bodies primarily contain RNA rather than DNA, so there is a need to map RNA-RNA interactions in addition to RNA-DNA interactions to map the RNA components of various nuclear bodies.

Here, we present a dramatically improved implementation of the SPRITE method (SPRITE 2.0) that enables comprehensive mapping of all classes of RNA in the nucleus, from abundant RNAs encoded from DNA repeats to low abundance RNAs such as nascent pre-mRNAs and lncRNAs. Because SPRITE does not rely on proximity ligation, it can detect multi-way interactions between multiple RNA molecules (RNA-RNA), RNA and DNA molecules (RNA-DNA), and DNA molecules (DNA-DNA). Here, we report, to our knowledge, the first high-resolution RNA-RNA interaction maps in mESCs. This has enabled identification of many sets of RNA interactions occurring throughout the cell, which we refer to as RNA "hubs". We find that RNAs localize broadly across the nucleus, with individual RNAs localizing within discrete structures ranging from nuclear bodies to individual TADs. Because we can measure nascent pre-mRNAs on chromatin, SPRITE allows us to explore how mRNAs are spatially organized during transcription. We uncover that nascent mRNAs interact in structures corresponding to nascent mRNA chromosome territories and compartments. By integrating higher-order measurements of RNA localization, we explore how RNA localization in specific compartments can impact mRNA transcription, providing important insights into RNA function. Together, these results uncover a central and widespread role for non-coding RNA in demarcating 3D nuclear structures within the nucleus.

4.3 RESULTS

4.3.1 SPRITE GENERATES HIGH-RESOLUTION MAPS OF RNA-DNA CONTACTS IN THE NUCLEUS

To define the role of RNA in nuclear structure, we adapted our SPRITE method to comprehensively map all RNAs in 3D structure relative to DNA and RNA. We previously showed that SPRITE can accurately measure DNA-DNA interactions and measure long-range interactions in the nucleus, including those that are missed by proximity-ligation-based methods. Here we greatly improved the efficiency of RNA tagging molecular biology steps of SPRITE to achieve broad labeling of highly abundant as well as low abundance RNAs and their associated DNA regions in 3D space. This approach enables comprehensive mapping of higher-order DNA-DNA, RNA-DNA, and RNA-RNA interactions all within the same experiment.

Specifically, SPRITE works as follows (**Figure 1A**): (i) All RNA, DNA, and protein contacts are crosslinked in situ to preserve their spatial relationships. (ii) Cells are lysed and fragmented into smaller crosslinked complexes. (iii) A substrate-specific adaptor is ligated to each DNA molecule (DPM) and a distinct adaptor is ligated to each RNA molecule (RPM). This enables accurate assignment of reads as RNA or DNA, with >99.99% of RNA reads aligning to the correct strand and DNA reads aligning uniformly to both strands. (iv) All RNA and DNA molecules within a crosslinked complex are simultaneously barcoded with unique tags over multiple rounds of split-and-pool barcoding. (v) All RNA and DNA interactions are identified by sequencing, aligning each read to the genome, and matching all reads with shared barcodes. (vi) All RNA and DNA reads containing the same barcode sequences are merged into a SPRITE cluster.

We applied this to mouse ES cells and generated ~10 billion reads. We confirmed that we generate clusters containing both DNA and RNA molecules and that we accurately assign reads as RNA or DNA (see **Methods**). We were able to generate SPRITE clusters containing RNAs across a broad expression range, from abundant RNAs species such as 45S pre-ribosomal RNA to low abundance transcripts such as individual nascent pre-mRNAs and lncRNAs.

To ensure that these clusters represent *bona fide* RNA interaction, we explored RNA-DNA interactions that are captured in these data (**Figure 1B-C**). We focused on lncRNAs that had previously been mapped to chromatin and we confirmed that the RNA-DNA interactions we observed accurately reflect known localization patterns across a range of localization patterns.

(i) **Malat1 localizes across the genome at actively transcribed genes**. SPRITE clusters containing the Malat1 lncRNA demonstrate strong enrichments over DNA regions that are actively transcribed RNA polymerase II genes, consistent with the patterns observed previously[11] (**Figure 1C**).

(ii) **Xist localizes across the inactive X chromosome**. SPRITE clusters containing the Xist lncRNA are strongly enriched over the inactive X chromosome, and localize broadly across the chromosome (**Figure 1B**). In contrast, we observe no localization over the active X chromosome. This is consistent with the known localization and function of Xist in coating the inactive X chromosome during X chromosome inactivation[10].

(iii) **Telomerase RNA component (TERC) localizes at telomeric region of various chromosomes**. SPRITE clusters containing the TERC is enriched at telomere-proximal regions of all chromosomes. This is consistent with the known localization of TERC on genomic DNA[12] and its function in guiding extensions of telomeric DNA.

Together these results demonstrate the SPRITE accurately measures known RNA localization patterns on DNA.

4.3.2 MOST LNCRNAS ARE ENRICHED ON CHROMATIN AND LOCALIZE IN 3D PROXIMITY OF THEIR LOCUS

We computed a chromatin enrichment score that reflect the proportion of a given RNA that is within a cluster containing DNA. As expected, we observed that mRNAs and other ncRNAs involved in mRNA translation (e.g., tRNAs, rRNA, 7SL) are depleted on chromatin. In contrast, we observed a strong enrichment for lncRNAs on chromatin suggesting that they may generally function in the nucleus. While these patterns demonstrate general properties of classes of RNAs, there are specific lncRNAs that show depletion within the nucleus, including the NORAD lncRNA which was previously reported to act in the cytoplasm.

We focused on lncRNAs that were enriched in the nucleus and observed that most lncRNAs show strong enrichment within local regions around their transcription loci and in many cases these localization patterns correspond to the known functional targets of these lncRNAs. For example, we found that:

(i) **Kcnq1ot1 localizes to a topologically associated region containing its imprinted targets**. We identified that the Kcnq1ot1 lncRNA localizes over the TAD corresponding to Cdkn1c and other genes that are known to be silenced by the Kcnq1ot1 lncRNA.

(ii) **Tsix localizes to a topologically associated domain within the X inactivation center.** We identified that the Tsix lncRNA localizes over the TAD corresponding to Xist and its activators on the X inactivation center, but is depleted over the Tsix promoter and other repressors of Xist that it is directly adjacent to its transcription locus (**Figure 1B**). Importantly, Tsix is expressed on the active X chromosome where it acts to repress Xist transcription.

(iii) **Airn and imprinting.** We identified that the Airn lncRNA localizes over the TAD containing Igf2r and other imprinted genes within this cluster.

(iv) **Chaserr localizes over the Chd2 gene**. We identified that the Chaserr lncRNA localizes over the promoter of the Chd2 gene. Interestingly, this lncRNA was recently shown to repress transcription of the Chd2 gene[13].

These results uncover the global localization patterns of lncRNAs and confirm a general relationship between 3D genome structure and lncRNA localization. Such a model has long been proposed based on observations from a handful of examples and our data suggests that this represents a more general property of lncRNAs in the nucleus which likely reflects the ability for lncRNAs to form high concentration territories near their transcription foci. These localization patterns may provide critical insights into where different lncRNAs may act.

4.3.3 HIGHER-ORDER RNA AND DNA CONTACTS RELATE 3D SPATIAL ORGANIZATION TO TRANSCRIPTION

In addition to generating comprehensive maps of RNA-DNA contacts, SPRITE can measure higher-order associations of RNA, including nascent pre-mRNAs, and DNA. Accordingly, we reasoned that we should be able to measure the impact of RNA localization, or DNA structure, on mRNA transcription from the SPRITE data. To test this, we focused on the Xist lncRNA which is known to silence transcription of genes on the X chromosome. We explored the number of nascent mRNA transcripts that are associated with DNA of the X chromosome in the presence of Xist relative to the frequency in SPRITE clusters lacking Xist. As expected, we observe a strong depletion of nascent transcripts on DNA in the presence of Xist. The few exceptions correspond to genes that are known to escape XCI.

Using this approach, we sought to explore a central open question related to genome organization and transcription. Specifically, one of the defining features of genome organization is the spatial segregation of chromosomal regions into A and B compartments. These compartments are often referred to as active and inactive compartments, respectively, because A compartments tend to be enriched for genes and the B compartment is comparatively depleted of genes. Yet, there are still actively transcribed genes that are encoded in genomic DNA regions within the B compartments. It has been proposed that active genes may loop out of the B compartment when they are transcribed. Yet, this hypothesis remains untested because genomic structure methods, such as HiC, cannot distinguish between the structure of DNA regions that are actively transcribed from those regions that are not transcribed. Accordingly, it is remains unclear whether A/B compartments represent active versus inactive transcriptional compartments.

To explore this question, we separated our mRNA reads into mature mRNA and nascent pre-mRNAs (see **Methods**). In contrast to mature mRNAs, which showed no structure, we noticed that nascent mRNAs form striking localization patterns that reflect chromosome territories and A/B compartments (**Figure 7**). For example, zooming in on A and B compartments, we found that nascent mRNAs transcribed from these regions are spatially closer to each other than would be expected from linear distance. In fact, we observe individual mRNAs transcribed from B compartments contacting other mRNAs within distinct B compartment regions, while mRNAs transcribed from A compartments are excluded and are more likely to contact mRNAs transcribed from other A compartment regions. Moreover, focusing exclusively on DNA contacts that are contained within SPRITE clusters that contain nascent mRNAs we still observe the same A/B compartments.

This demonstrates that active transcription can occur in both A/B compartments and that active genes contained within B compartments do not loop out of this territory to contact an “active” (A compartment) region. This result argues against the notion that A/B compartments represent active versus inactive chromatin and instead suggests that other features of these regions, such as gene density, may explain compartmentalization. This may also explain why A/B compartments are largely invariant between cell states even though transcription is highly dynamic. This result contrasts with a recent FISH study that reported the absence of “chromosomal” information in nascent pre-mRNA localization.

4.3.4 NON-CODING RNAS FORM HUBS WITHIN THE NUCLEUS

Because SPRITE can generate higher-order RNA and DNA localization in 3D space, we sought to broadly explore RNA localization in nuclear structure. We noticed that numerous sets of RNAs displayed similar localization patterns on genomic DNA and these RNAs cluster into sets of RNAs that display high contact frequencies with each other, but low contact frequencies with other RNAs. This suggests that distinct ncRNAs form nuclear hubs.

To explore these RNA hubs, we computed genome-wide RNA-RNA interactions and clustered the matrix. We identified various sets of RNAs that are highly interacting with each other, which we define as RNA hubs (**Figure 3**). First, we noticed that we could separate the localization of RNAs in different parts of the cell. Specifically, we noticed a clear separation between nuclear RNAs and cytoplasmic RNAs. Specifically, we identify a hub that is enriched for known cytoplasmic RNAs such as mature ribosomal RNAs (5S, 5.8S, 18S, 28S) as well as other RNAs involved in translation such as signal recognition particle RNA (7SL) and transfer RNAs (tRNAs). We define this as a "cytoplasmic hub" because it is well separated from nuclear RNAs. We note that histone mRNAs are also strong enriched within the cytoplasmic hub, consistent with the fact that they are known to be enriched in the cytoplasm and translated.

In contrast to the cytoplasmic hub, we identify several hubs of nuclear-enriched RNAs that are highly interacting with each other corresponding to RNAs in various distinct structures in the nucleus. We explore each of these structures in detail below. These correspond to structures, including nuclear bodies, occurring around regions of specialized RNA transcription and processing.

4.3.5 THE NUCLEOLUS CONTAINS NUMEROUS *TRANS*-NCRNAS ASSOCIATED WITH RIBOSOMAL RNA ASSEMBLY

We identified a hub that includes the 45S pre-ribosomal RNA components and numerous additional ncRNAs, including dozens of snoRNAs, RMRP, and RNase P. Focusing on the genomic DNA regions in proximity to these RNAs, we find that they are enriched for higher-order inter-chromosomal contacts that we previously defined as organized around the nucleolus (**Figure 4**).

The nucleolus is a nuclear body that is associated with transcription of pre-ribosomal RNA, modification and cleavage, and assembly of mature ribosomes. Interestingly, this hub contains all components of the 45S pre-rRNA (Internal Transcribed Spacer 1 and 2 (ITS1, ITS2) and External Transcribed Spacer (ETS)) as well as dozens of snoRNAs that are known to base pair with specific sequences in the 45S pre-rRNA to guide various post-transcriptional modifications and RMRP which guides the cleavage of 45S.

In addition to RNAs involved in ribosomal RNA processing, we observe enrichment of numerous RNA Polymerase III (PolIII) transcribed ncRNAs. Specifically, we observed enrichment of the 5S rRNA within the nucleolus where it is assembled with the other 45S-encoded mature ribosomal RNAs. Surprisingly, we also observed enrichment of tRNAs and RNase P within the nucleolus. Interestingly, RNase P is known to modify tRNAs into their mature and functional forms suggesting that this processing step also occurs within the nucleolus. Similarly, we also observe enrichment of 7SL RNA component of the SRP. SRP is involved in protein sorting in the ER and is enriched along with protein translation machinery. Yet, in addition to this, we observe strong enrichment for the 7SL protein within the nucleolus suggesting that the 7SL RNA may also be processed within the nucleolus prior to export.

These observations highlight another critical aspect of SPRITE, in contrast to DNA structures, ncRNA localization can be heterogeneous by localizing in multiple different nuclear structures within the same cell. Because SPRITE can map RNAs in complexes, we can assign individual RNAs to different nuclear structures. In this way we were able to identify several ncRNAs that appear to have strong associations with the nucleolar hub, but also appear to interact with other RNAs in other nuclear and cytoplasmic compartments.

Together, these results suggest that multiple different RNAs, including ribosomal RNA, tRNAs, and 7SL RNAs, may be processed within the nucleolus. Interestingly, while the pre-rRNA is organized around cis localization, the chromosomes for inter-chromosomal contacts and the other RNAs, including those transcribed by RNA PolII and RNA PolIII, can diffuse and interact in trans to form this high concentration nuclear territory dedicated to transcription and processing of ribosomes.

4.3.6 NCRNAS INVOLVED IN MRNA SPLICING ARE SPATIALLY CONCENTRATED AROUND TRANSCRIBED RNA POLII GENES

We identified another hub that corresponds to multiple ncRNAs contained within the spliceosome complex, including the U1, U2, U4, U5, and U6 spliceosomal RNAs, U11 and U12 minor spliceosomal RNAs, along with the Malat1 lncRNA and 7SK transcriptional regulator (**Figure 3**). Focusing on the DNA associations of these RNAs, we found that these RNAs show a highly correlated localization pattern at actively transcribed RNA PolII genes and corresponds to regions that organize around nuclear speckles (**Figure 4**). In addition to these DNA regions, we also identified strong enrichment of introns of pre-mRNAs within this hub. Importantly, these RNAs are strongly depleted of RNA-RNA interactions with RNA components of the nucleolus and other nuclear hubs. Consistent with this observation, we observe a strong exclusion of both Malat1 and 7SK from the nucleolus by microscopy (Figure X).

Interestingly, we also identify RNAs transcribed from SINE B1/B2 within the speckle hub. This may reflect the fact that many SINE B1/B2 elements are contained within the introns of mRNAs, which are also found within the splicing hub. It was previously reported that RNA from SINE B2 can act to repress RNA PolII *in vitro*, the presence of these RNAs at high concentration over highly transcribed PolII genes suggests that this is likely to have additional functions *in vivo*.

These results suggest that actively transcribed RNA PolII genes and their associated nascent pre-mRNA are organized around nuclear speckles. Because nuclear speckles are assemblies of mRNA splicing and processing proteins, these results suggest that spatial organization around nuclear speckles may act to concentrate nascent pre-mRNAs in cis around spliceosomal RNAs and other splicing regulators in trans. This increased concentration may act to increase the kinetic efficiency of mRNA transcription and splicing.

4.3.7 NCRNAS INVOLVED IN SNRNA BIOGENESIS ARE SPATIALLY ORGANIZED AROUND SNRNA GENE CLUSTERS

We identified another hub with several small Cajal body specific RNAs (scaRNAs 1, 2, 5, 6, 7, 9, 10, 12, 13, and 17). scaRNAs were defined because they are enriched within the Cajal body, which are sites of snRNP biogenesis and maturation in the nucleus. It has also been previously been reported that snoRNAs may also localize with these Cajal bodies. In fact, snoRNAs have been proposed to be trafficked through the Cajal body alongside spliceosomal RNAs prior to localization within nucleoi and nuclear speckles, respectively. To test this, we compared all spliceosomal RNA interactions and snoRNA interactions, and observed a strong enrichment for scaRNAs shared between both RNAs. In fact, we also identify 2 unannotated scaRNAs which strong enrichment with both snoRNAs and snRNAs that are encoded within the Trrap and Gon4l mRNAs that appear to be orthologues of two human RNAs, SCARNA28 and SCARNA26A, respectively. These results suggest that the scaRNA hub corresponds RNA interactions occurring within Cajal bodies in the nucleus.

While it has been shown which RNAs are enriched are Cajal bodies, the DNA contacts of scaRNAs within Cajal bodies have not been previously mapped. Focusing on the DNA associations of these RNAs, we found that these scaRNAs are highly enriched at 4 discrete regions in the genome (**Figure 5**). These correspond to both the Histone 1 and Histone 2 gene clusters on mouse chromosome 3 and 13, respectively, as well as multiple genomic sites on mouse chromosome 11 that have a high density of snRNA genes as well as tRNA genes. Interestingly, these scaRNAs primarily localize around sites of snRNA transcription but not snoRNAs. Unlike most PolII genes, snRNAs are transcribed by PolII with distinct CTD modifications are recruitment of the specialized Integrator complex. This suggests that scaRNAs may localize in a specialized PolII compartment within the genome, separate from mRNA transcriptional loci.

Based on these observations, we hypothesized the nuclear body we defined reflects the Cajal body or “gems”. To confirm this, we performed FISH combined with IF on scaRNA17 and scaRNA2 with both Coilin and SMN. Coilin is a marker of Cajal bodies, SMN is a marker of “Gemini of the CB” or “gems”. We observe a clear co-localization between scaRNAs within

SMN foci. Surprisingly, we were unable to identify focal localization of Coilin, a marker that has previously been reported as the defining marker of Cajal bodies. This focal localization is known to be absent in mouse ES cells and, accordingly, has led to reports of the absence of Cajal bodies within these cell types. Our results indicate that Coilin foci may not be the best determinant of Cajal bodies and instead RNA localization of scaRNAs more accurately define these nuclear bodies, as the scaRNA bodies show many hallmarks of Cajal bodies in terms of their RNA interactions and genomic DNA localization. Consistent with this idea, scaRNAs have a clearly defined functional role in snRNA biogenesis whereas Coilin has no known function and Coilin is dispensable for snRNA biogenesis.

Our scaRNA-DNA interactions reveal that scaRNAs localize to non-canonical RNA PolII genes (spliceosomal RNAs and histone loci). These results suggest actively transcribed snRNA PolII genes and their associated nascent pre-mRNA are processed immediately at their transcriptional loci. These results suggest that spatial organization of processing RNAs and proteins around Cajal bodies may act to concentrate nascent pre-snRNAs around scaRNAs and other regulators in trans. While it has been previously hypothesized that snRNA gene density may be a key feature of Cajal body formation on DNA, directly mapping scaRNA localization on DNA enables direct measurement of scaRNA localization on chromatin. This increased gene density may be an important feature of nuclear body formation to recruit a high concentration of processing RNAs and proteins to process snRNAs immediately upon transcription and increase the kinetic efficiency of snRNP biogenesis. Concentrate processing factors near their substrates prevents inappropriate processing or inefficient snRNP biogenesis.

We also note a strong interaction between the Terc RNA and scaRNAs. Although Terc RNAs do not function is snRNA modification, with Terc RNA has been reported to be a scaRNA because it contains H/ACA and CAB box sequences that mark scaRNAs[14]. Notably, when we look at the scaRNA-DNA interactions, scaRNAs 12 and 17 are highly enriched the Terc RNA locus, which is immediately adjacent to the Hist2 locus on chromosome 3. This suggests that these RNAs may all co-localize around this genomic site.

4.3.8 THE HISTONE PROCESSING U7 SNRNA IS SPATIALLY ENRICHED AROUND HISTONE GENE LOCI

Having identified scaRNAs near both snRNAs and histone gene loci, we considered whether the histone locus body is a separate or same structure as the scaRNA bodies. Specifically, it has been unclear from other reports in multiple organisms whether Cajal bodies and histone locus bodies are distinct structures in the nucleus as they have been observed together in the nucleus. To test this, we compared the localization of the U7 sRNA, a histone locus body marker, as well as histone mRNAs (Hist1 mRNAs) on chromatin with scaRNAs. The U7 snRNA processes histone mRNAs within histone locus bodies. Interestingly, we found some overlap of scaRNAs with U7 snRNA and histone mRNAs in their DNA localization, but they do not share all sites of DNA localization with scaRNAs. Further, in terms of their RNA-RNA interactions, the U7 snRNAs seem to form a hub with histone mRNAs (Hist1 and Hist2). We see that in addition to the tight interaction between U7 and histone mRNAs that this cluster also interacts with scaRNAs within the Cajal body hub. This indicates that this interaction occurs together near the Cajal body. This may be consistent with previous reports suggesting that the Histone Locus Body and Cajal body are two separate nuclear bodies, but that they are very close to each other and may share critical components. To test this, we performed FISH for scaRNAs 2 and 17 combined with immunofluorescence for NPAT, a marker of the histone locus body. We observe that scaRNAs are consistently adjacent to the histone locus body, but do not directly overlap, consistent with the results showing that they form separate RNA hubs as measured with SPRITE. These results suggest that the histone locus body may form at nascently transcribed histone loci and recruit the U7 snRNA in trans to immediately process the histone mRNAs upon their transcription.

4.3.9 CENTROMERES FORM INTER-CHROMOSOMAL CLUSTERS AROUND SATELLITE-DERIVED NCRNAS

Finally, we identified another hub of RNAs corresponding to Minor and Major satellite RNAs transcribed from centromeric and peri-centromeric DNA, respectively. Exploring the DNA interactions of these RNAs, we identified a similar localization pattern for both RNAs. Major and minor satellite RNAs localize precisely over centromere-proximal DNA. These DNA regions correspond to DAPI-dense regions of the chromosome that are often referred to as chromocenters. Interestingly, when we focus on SPRITE clusters contain either of these RNAs, we identify higher-order DNA interactions that occur between multiple centromere-proximal regions from different chromosomes (**Figure 6**). This suggests that these RNAs demarcate a nuclear body where centromeric regions of chromosomes interact with each other. To confirm this, we performed DNA FISH on the major and minor satellite DNA and observed higher-order structures where multiple centromeres from distinct chromosomes interact simultaneously.

These chromocenters represent constitutive heterochromatin that are enriched for various chromatin modification and enzymes associated with this state. For example, HP1 protein and H3K9me3 are enriched at these DNA sites. Interestingly, previous studies have shown that global disruption of RNA, using RNase A, leads to disruption of HP1 localization at chromocenters. Yet, the RNA component has remained uncharacterized. Our results suggest that these major and minor RNAs might play a role in this process. To test this hypothesis, we performed LNA treatment on the major and minor satellite RNAs. We find that disruption of both major and minor satellite RNA recapitulates the phenotype observed in RNase treatment, with a disruption in chromocenter structures and a depletion of HP1 protein at chromocenter structures.

Consistent with this idea, several previous studies have shown that disruption of the major satellite RNA components prior to the formation of chromocenters during preimplantation development leads to disruption of chromocenter formation, lack of heterochromatin formation, and embryonic arrest. Interestingly, despite this repressed transcriptional state, these regions are actively transcribed to produce high levels of these RNA species. Such a role for RNA is reminiscent of known RITS mechanisms in yeast and have also been proposed to play important roles in the establishment and maintenance of heterochromatin at various locations in higher eukaryotes.

4.4 DISCUSSION

Using SPRITE, we can precisely define which RNAs are on chromatin and map where each RNA goes in the cell. This enables us to comprehensively define the RNA and DNA components of nuclear bodies and measure the spatial organization of transcription in the cell.

SPRITE provides a new view of DNA structure whereby RNA plays a central role in organization and processing of around different transcriptional hubs in the nucleus (**Figure 8**). For instance, we identify various nuclear bodies that may correspond to a structure forming around the high density or concentration of transcription of shared genes—for example, PolI and nucleolus, PolII and mRNA compartment, non-canonical PolII snRNA genes and the HLB, and PolIII at Cajal bodies. Our results highlight the value of adding RNA generally into 3D structure because we can now more accurately define the RNA components of nuclear bodies as well as the structural relationships between transcription and recruitment of processing RNAs to these transcriptional loci.

These results identify a common theme where transcription and RNA processing are spatially coupled in the nucleus. In fact, we observe the nucleolus, speckle, Cajal bodies, and histone locus bodies all as specialized transcription and processing centers. Transcription appears to create high local concentrations in space and can guide other components such as processing RNAs and regulatory proteins in the nucleus. Notably, by mapping the genomic loci where these RNAs localize, we observe a clear relationship between the linear density of genes results in spatial clustering. Specifically, in each of these structures, there is a clear relationship where RNAs are concentrated at loci with a high linear density of genes—rRNA as multi-copy rDNA repeats, satellite RNAs transcribed from satellite repeats, snRNAs from snRNA gene clusters on multiple chromosomes, and histone mRNAs from histone gene clusters. We propose that linear density may be a defining characteristic of defining spatial clustering of molecules into nuclear bodies.

These findings have several implications for understanding the role of nuclear bodies in regulating functions in the nucleus. Specifically, we observe that the scaRNAs responsible for snRNA biogenesis are highly concentrated at the snRNA transcriptional loci where these RNAs produced. This increased concentration of processing RNAs at the transcriptional loci of snRNA genes would enable the immediate processing of snRNAs prior to their trafficking to the rest of the nucleus as functional snRNPs. Without this immediate processing, processing may be far less efficient at the sub-stoichiometric concentrations these RNAs are found at in the nucleus. In fact, scaRNAs are likely able to achieve these roles without resorting to super-stoichiometric localization driven by compartmentalization.

We can also begin to observe functional relationships where RNA might impact chromatin structure and function. Specifically, because RNA is the functional readout, it provides a measure of the transcriptional output for different nuclear structures (arrangements). For example, our results clearly demonstrate that A/B compartments cannot simply reflect transcriptional activity, which is a conclusion that is in contrast to current descriptions about compartments, but which could not be observed in the absence of simultaneous measurements of RNA and DNA.

The ability of SPRITE to map the localization of RNA on chromatin can provide testable hypothesis—for example, the role of RNA in HP1 recruitment has been long hypothesized. Here, observing the major and minor satellite RNAs on the peri-centromeric and centromeric proximal DNA, we identified RNAs transcribing from heterochromatin structures. In doing so, we found that major and minor satellite RNAs are required for chromocenter structure formation as well as in recruitment of HP1 to these chromocenters.

This approach can be applied in any cell type. Although SPRITE currently requires many reads to reach high-resolution maps of lower abundance RNAs, this is because most RNA reads are ribosomal RNAs. By applying known rRNA depletion methods, such as DASH or ribodepletion, we expect that we can readily achieve higher resolution with far fewer sequencing reads.

4.5 REFERENCES

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

2. Dieci, G., Preti, M. & Montanini, B. Eukaryotic snoRNAs: A paradigm for gene expression flexibility. *Genomics* (2009). doi:10.1016/j.ygeno.2009.05.002

3. Jarrous, N., Wolenski, J. S., Wesolowski, D., Lee, C. & Altman, S. Localization in the nucleolus and coiled bodies of protein subunits of the ribonucleoprotein ribonuclease P. *J. Cell Biol.* (1999). doi:10.1083/jcb.146.3.559

4. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology* (2014). doi:10.1038/nrm3742

5. Gerbi, S. A. & Lange, T. S. All small nuclear RNAs (snRNAs) of the [U4/U6.U5] tri-snRNP localize to nucleoli; identification of the nucleolar localization element of U6 snRNA. *Mol. Biol. Cell* (2002). doi:10.1091/mbc.01-12-0596

6. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* (2010). doi:10.1101/cshperspect.a000653

7. 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

8. Kaiser, T. E., Intine, R. V. & Dundr, M. De novo formation of a subnuclear body. *Science (80-.).* (2008). doi:10.1126/science.1165216

9. Maison, C. & Almouzni, G. HP1 and the dynamics of heterochromatin maintenance. *Nature Reviews Molecular Cell Biology* (2004). doi:10.1038/nrm1355

10. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* (2015). doi:10.1038/nature14443

11. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* (2014). doi:10.1016/j.cell.2014.08.018

12. Mumbach, M. R. *et al.* HiChIRP reveals RNA-associated chromosome conformation. *Nat. Methods* (2019). doi:10.1038/s41592-019-0407-x

13. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nat. Commun.* (2019). doi:10.1038/s41467-019-13075-8

14. Venteicher, A. S. & Artandi, S. E. TCAB1: Driving telomerase to Cajal bodies. *Cell Cycle* (2009). doi:10.4161/cc.8.9.8288

4.6 FIGURES

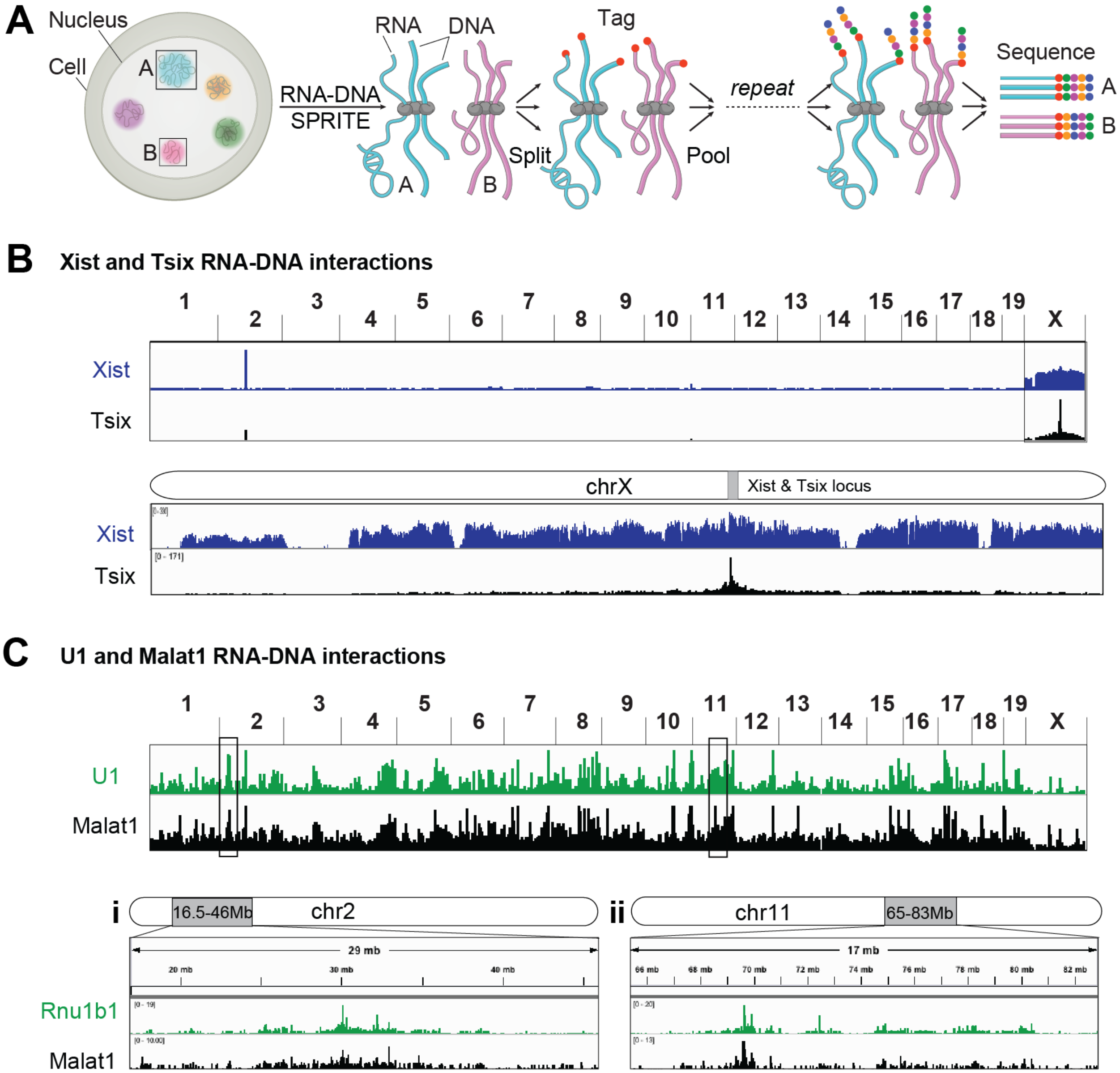


Figure 1. RNA-DNA SPRITE: A method to map higher-order RNA and DNA interactions in the nucleus.

(a) Overview of SPRITE: All RNA, DNA, and protein contacts are crosslinked in situ to preserve their spatial relationships. Cells are lysed and fragmented into smaller crosslinked complexes. All RNA and DNA molecules within a crosslinked complex are simultaneously barcoded with unique tags over multiple rounds of split-and-pool barcoding. All RNA and DNA interactions are identified by sequencing, aligning each read to the genome, and matching all reads with shared barcodes. All RNA and DNA reads containing the same barcode sequences are merged into a SPRITE cluster. **(b)** SPRITE clusters containing the Xist and Tsix lncRNA are strongly enriched over the X chromosome. The Xist lncRNA localizes broadly across the X chromosome. In contrast, the Tsix lncRNA localizes over the TAD corresponding to Xist. **(c)** The U1 snRNA and Malat1 lncRNA is strongly enriched over DNA regions that are actively transcribed RNA polymerase II genes.

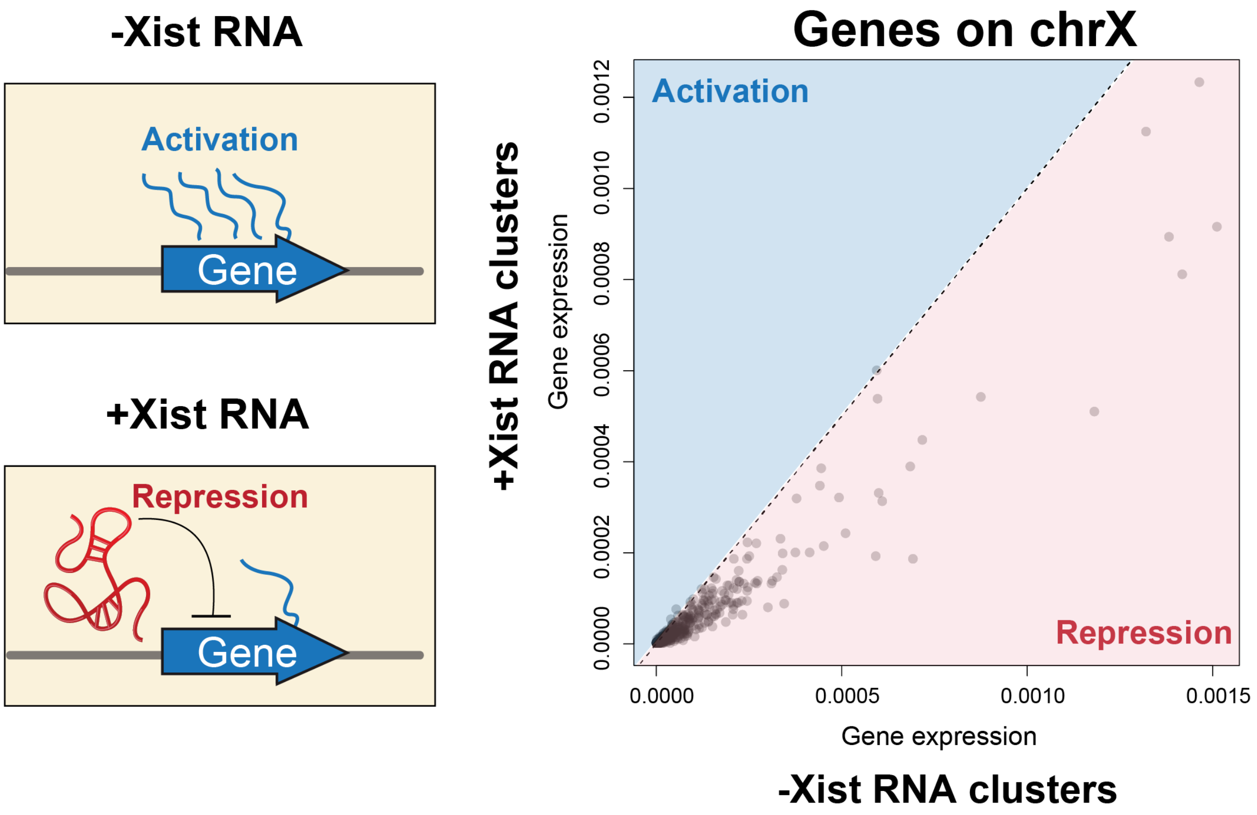


Figure 2. SPRITE can detect RNA-induced changes in gene expression.

Xist is only expressed on one copy of the X chromosome in the female cell nucleus, and only one copy of the X chromosome is silenced. Using SPRITE, we can separate Xist-containing (+Xist) and Xist-depleted (-Xist) chromosomes and measure whether gene expression from mRNAs on the X chromosome are reduced (red) or enriched (blue) in these SPRITE clusters.

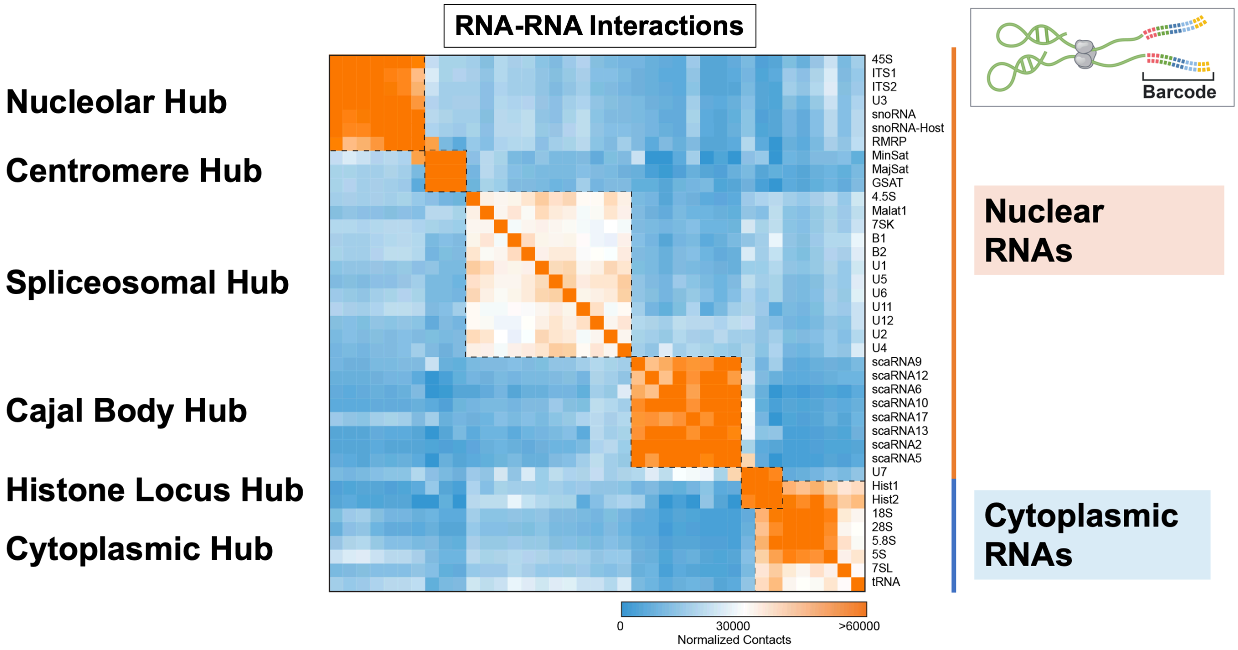


Figure 3. SPRITE detects multiple hubs of RNA that are highly interacting with each other in the cell.

Enumerating all RNA-RNA interactions in the nucleus, we observed several tightly interacting hubs of RNAs that correspond to distinct regions of RNA localization in the cell. Specifically, we can separate cytoplasmic RNAs involved in translation from those primarily localized in the nucleus. In the nucleus, we identify contacts from nucleolar RNAs, centromeric RNAs, spliceosomal/speckle RNAs, small Cajal body RNAs (scaRNAs), and histone locus body RNAs.

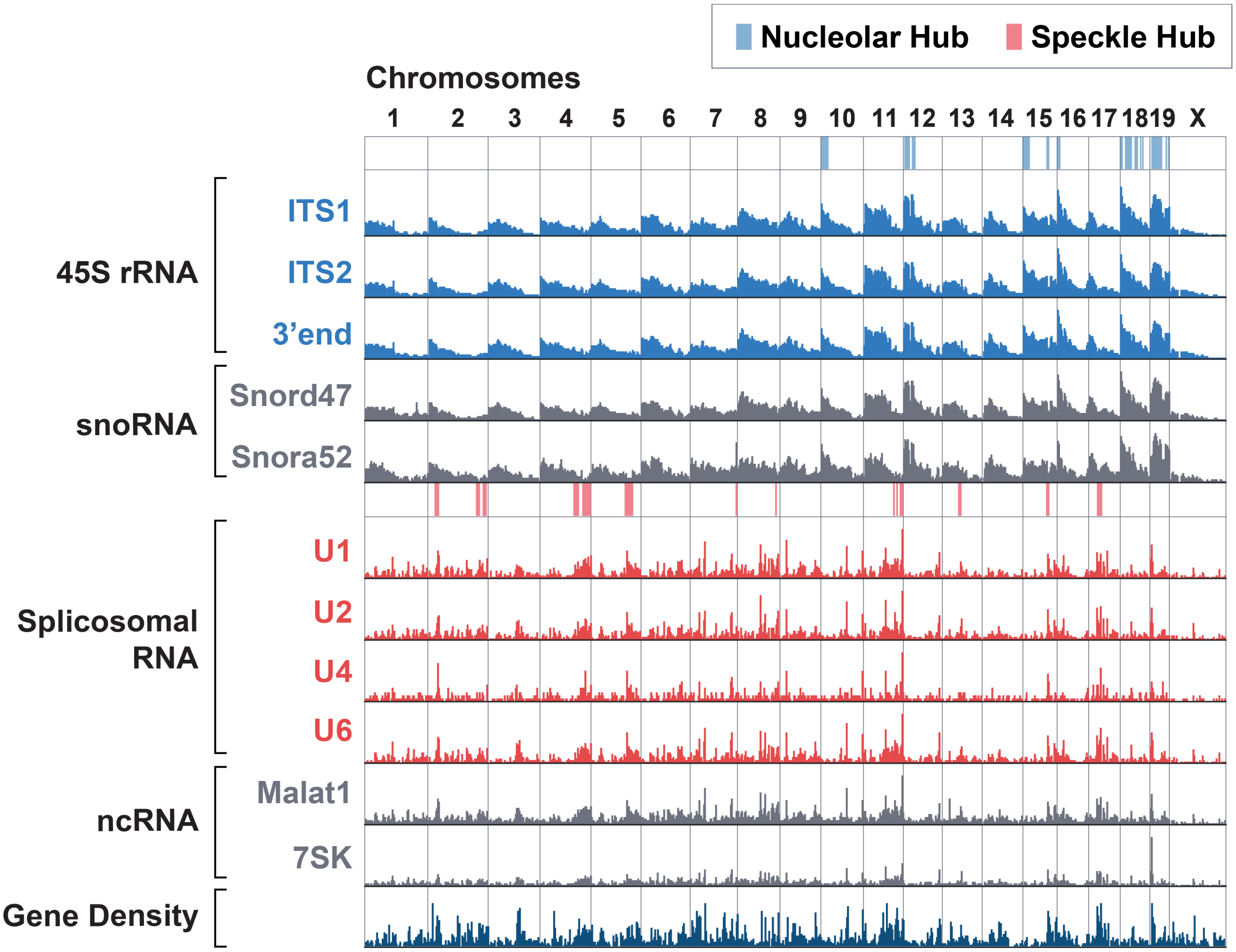


Figure 4. Nucleolar and spliceosomal RNAs have distinct localization patterns on chromatin.

Nucleolar hub RNAs, including 45s rRNA (blue) and snoRNAs (gray) are enriched over nucleolar hub genomic loci, which are positions close to nucleoli. spliceosomal hub RNAs, including the RNA components of the major spliceosome (red) and non-coding RNAs Malat1 and 7SK (gray) are enriched over the gene-dense regions (dark blue) and speckle hub regions, genomic regions positioned close to nuclear speckles.

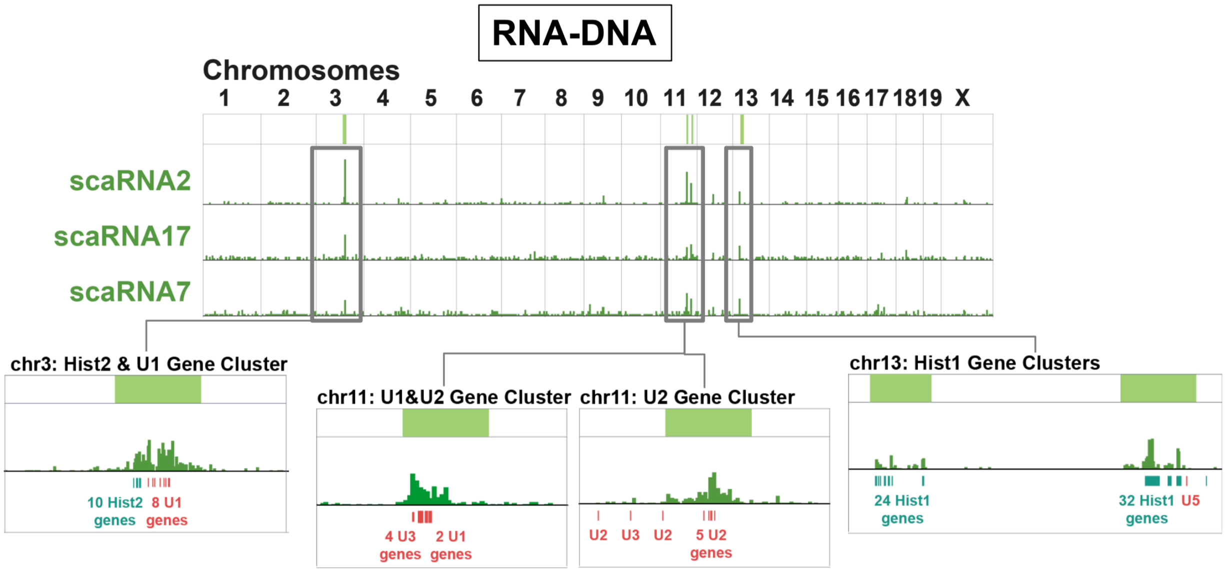


Figure 5. Small Cajal body RNAs (scaRNAs) are highly enriched at snRNA and histone gene clusters.

Multiple scaRNAs are highly enriched at histone gene clusters on chromosomes 3 and 13 in mESCs. They are also highly enriched at two gene-dense snRNA gene clusters on chromosome 11.

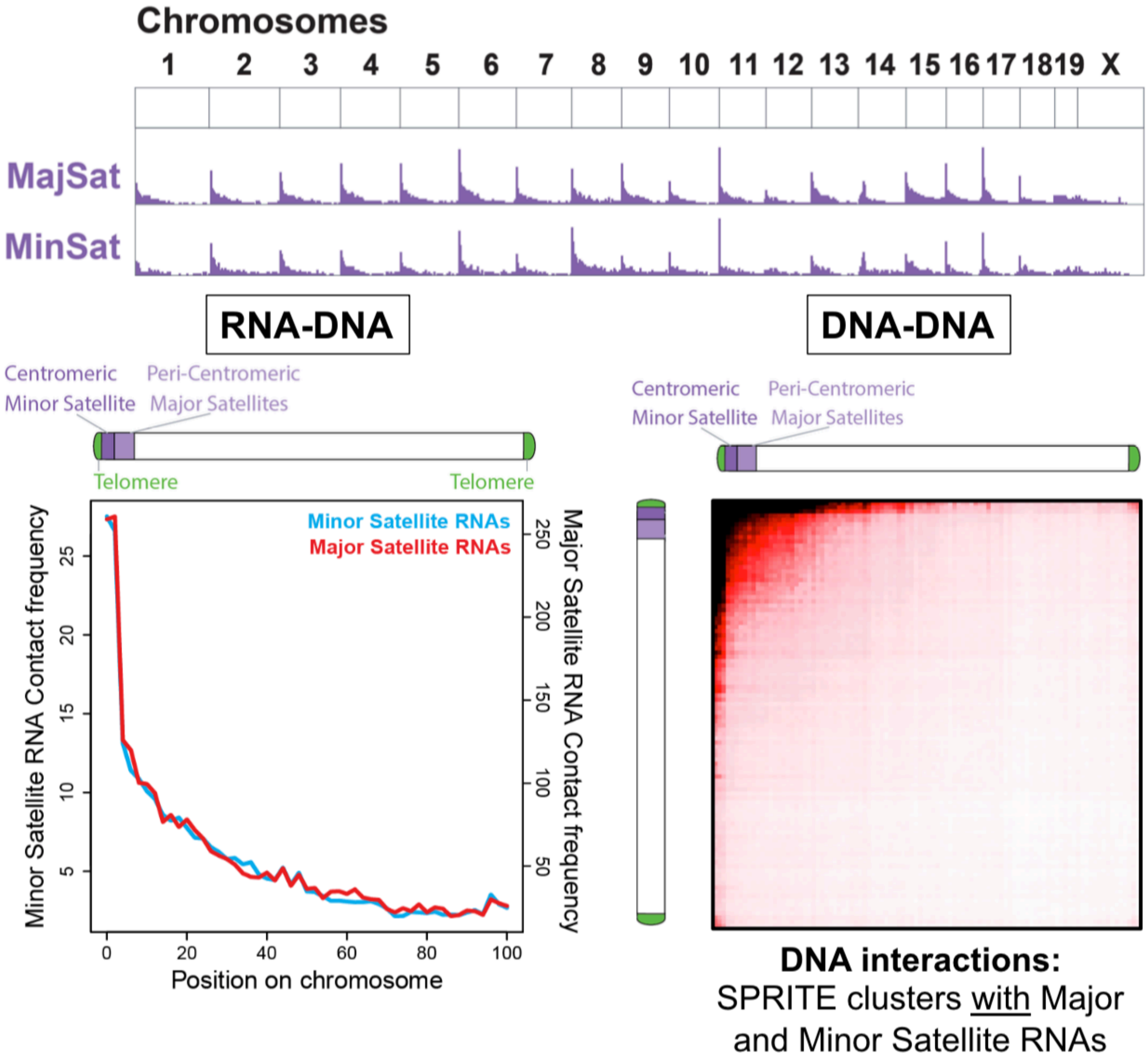


Figure 6. Major and Minor Satellite RNAs localize proximal to the peri-centromeric and centromeric chromatin.

Major and Minor Satellite RNAs are enriched at centromere-proximal DNA on all chromosomes using RNA-DNA SPRITE. In SPRITE clusters containing Major or Minor Satellite RNAs, we observe an enrichment for DNA-DNA interactions between centromere-proximal DNA on different chromosomes.

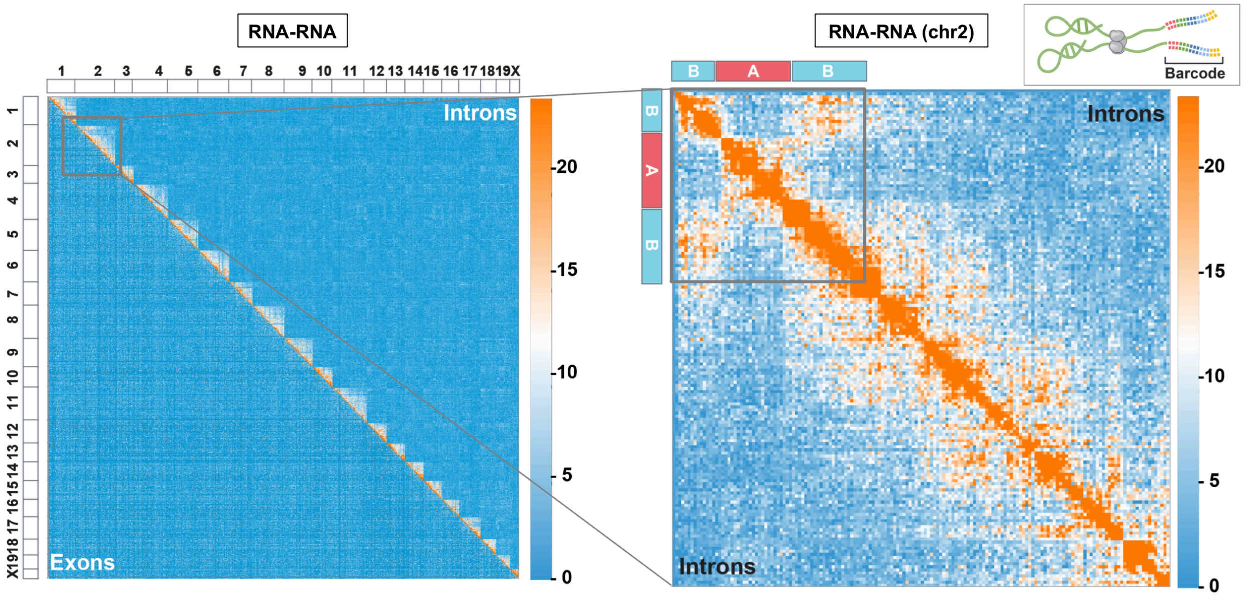


Figure 7. Nascent pre-mRNAs form transcriptional "territories" and "compartments"

We separated our mRNA reads into mature mRNA (exons) and nascent pre-mRNAs (introns). In contrast to mature mRNAs (bottom diagonal), which showed no structure, we noticed that nascent mRNAs (upper diagonal) form striking localization patterns that reflect chromosome territories and A/B compartments. For example, zooming in on A and B compartments, we found that nascent mRNAs transcribed from these regions are spatially closer to each other than would be expected from linear distance. In fact, we observe individual mRNAs transcribed from B compartments contacting other mRNAs within distinct B compartment regions, while mRNAs transcribed from A compartments are excluded and are more likely to contact mRNAs transcribed from other A compartment regions.

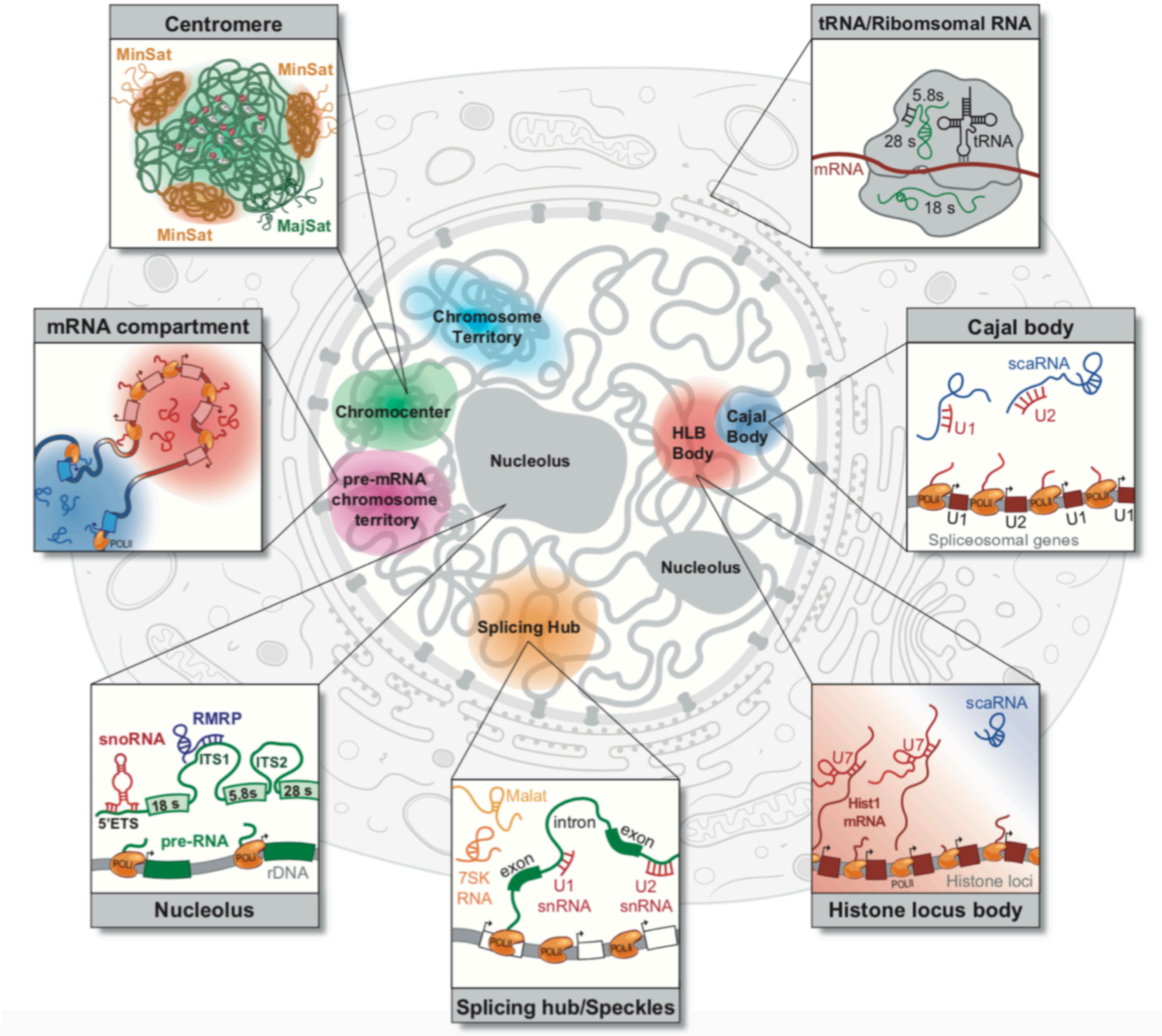


Figure 8. RNA demarcates spatial territories in the nucleus.

We find that RNAs localize broadly across the nucleus, with individual RNAs localizing within discrete territories ranging from nuclear bodies to individual topologically associated domains.

Chapter 5

CONCLUSION

S. Quinodoz and M. Guttman

A part of this chapter was published as “Long noncoding RNAs: an emerging link between gene regulation and nuclear organization” in Trends Cell Biology 24:651–63. 2014. doi: 10.1016/j.tcb.2014.08.009

5 FUTURE DIRECTIONS

It has become increasingly clear that gene expression entails the spatial coordination of transcription and RNA processing. Here, we have uncovered a central role for RNAs localizing in both *cis* and *trans* at different transcriptional loci and as a key component of nuclear bodies. However, many questions remain about the individual roles of RNA, proteins, and DNA in seeding and establishing these nuclear bodies and processing centers. For instance, it remains unclear how various processing RNAs are recruited in *trans* to various transcriptional gene clusters such as histone locus bodies, Cajal bodies, nucleoli, and nuclear speckles. Whether the act of transcription itself, protein recruitment, or RNA expression at each of these transcriptional loci initiates the formation of these bodies requires further study. Emerging observations have revealed that various nuclear bodies can form by “self-assembly” where individual proteins or RNAs can seed a compartment and recruit other processing factors. Additionally, many nuclear bodies have been shown to form liquid nuclear condensates in the nucleus once proteins reach a high local concentration within the nucleus, which can help establish these nuclear bodies with high concentrations of RNA, protein, and DNA of similar function within localized regions in the nucleus.

Our results reveal that the genome is highly organized in the nucleus, where different genomic loci are preferentially positioned proximal to various nuclear bodies. This raises various questions that require further study. For instance, how DNA is arranged and positioned proximal to different nuclear bodies is unknown. In fact, the nucleolus is a nuclear body whereby multiple chromosomes are spatially arranged and positioned around a large nuclear body. Emerging evidence suggests that liquid nuclear condensates can push apart nearby chromatin as they form and do so within low density, euchromatic regions. However, engineered systems recruiting high concentrations of intrinsically disordered proteins to genomic loci can bring together genomic loci within the nucleus. This provides an attractive model whereby genomic loci of shared functions can come together or push apart genomic loci to shape genome organization.

We have also observed that while some lncRNAs localize in *trans,* many lncRNAs can localize in *cis* near their transcriptional loci. This raises an interesting hypothesis whereby lncRNAs can seed local compartments immediately upon their sites of transcription. However, while the role of lncRNAs in establishing nuclear organization is attractive, many questions remain. Currently, there are only a few examples of lncRNAs that organize nuclear domains and even for these few lncRNAs, how they organize these nuclear domains is largely unknown. Future studies will be required to determine whether this role may be a more general role for nuclear-retained lncRNAs and whether there may be general mechanistic principles by which lncRNAs act to shape nuclear domains. In particular, it will be important to identify additional lncRNA-mediated nuclear domains and characterize the dynamics of their formation across various cellular conditions. Such examples will allow us to dissect the precise mechanisms by which lncRNAs can organize nuclear domains and determine the various components required for domain assembly. To address these questions, it will be important to develop experimental systems, such as inducible lncRNA systems that enable precisely controlled formation of the associated nuclear domain, to dissect dynamic nuclear organization at the molecular level. Such experimental systems will enable the systematic perturbation of a lncRNA, including deletion of specific protein binding regions, and the measurement of their roles in the establishment and maintenance of nuclear domains. Finally, it will be essential to determine the role that lncRNA-mediated regulation of nuclear organization plays in the control of gene expression. While much work remains to be done, it is now clear that the roles of lncRNAs in regulating gene expression and establishing nuclear organization may be more tightly linked than previously appreciated.