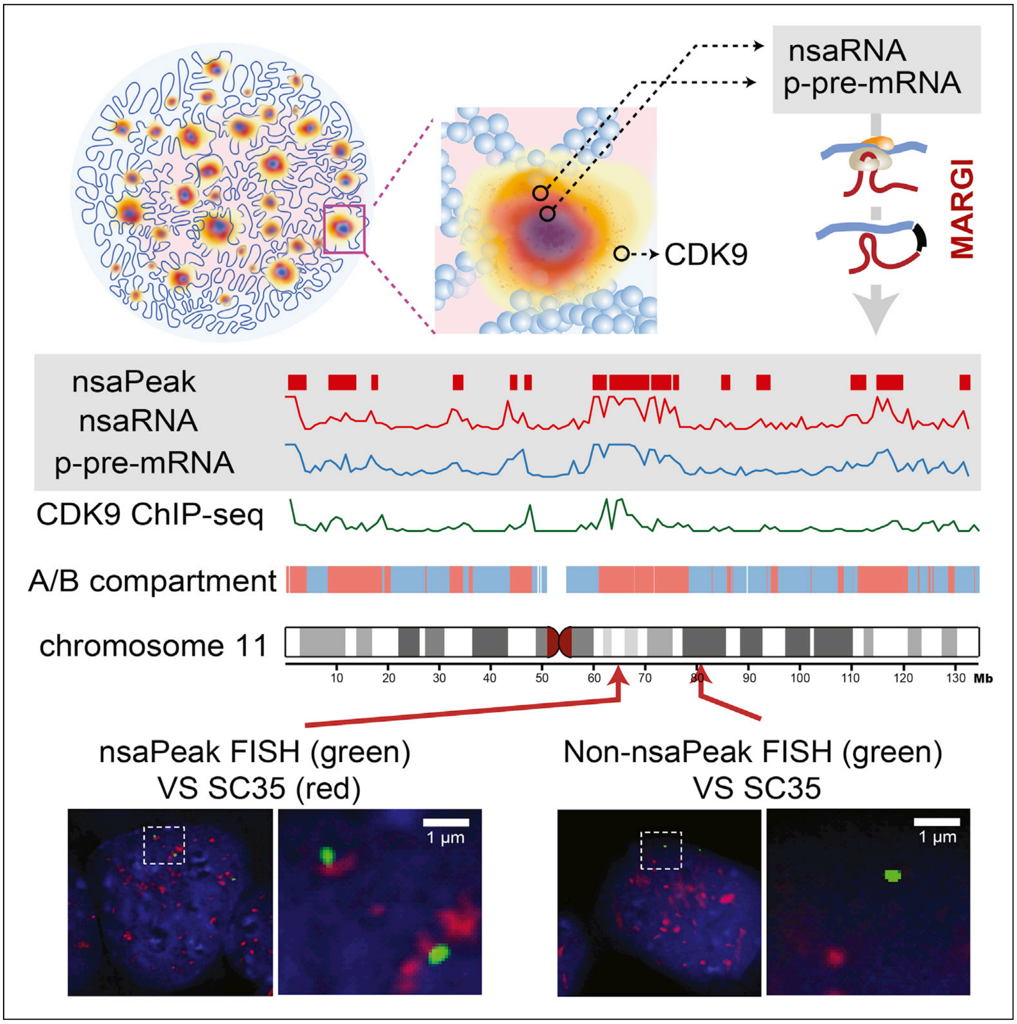
iScience

CellPress

Article

RNAs as Proximity-Labeling Media for Identifying Nuclear Speckle Positions Relative to the Genome



Weizhong Chen, Zhangming Yan, Simin Li, Norman Huang, Xuerui Huang, Jin Zhang, Sheng Zhong

jiz175@ucsd.edu (J.Z.)
szhong@ucsd.edu (S.Z.)

HIGHLIGHTS

MARGI captures interactions of nuclear speckle-associated RNAs (nsaRNA) and DNA

nsaRNA-interacting genomic sequences were clustered (nsaPeaks) in the genome

Posttranscriptional pre-mRNAs and CDK9 proteins exhibited proximity to nsaPeaks

Single-cell images confirmed proximity of nuclear speckles to an nsaPeak

Chen et al., iScience 4, 204–215
June 29, 2018 © 2018 The Author(s).
https://doi.org/10.1016/j.isci.2018.06.005

iScience

CellPress

Article

RNAs as Proximity-Labeling Media for Identifying Nuclear Speckle Positions Relative to the Genome

Weizhong Chen,[1,4] Zhangming Yan,[1,4] Simin Li,[2,4] Norman Huang,[1] Xuerui Huang,[3] Jin Zhang,[2,*] and Sheng Zhong[1,5,*]

SUMMARY

It remains challenging to identify all parts of the nuclear genome that are in proximity to nuclear speckles, due to physical separation between the nuclear speckle cores and chromatin. We hypothesized that noncoding RNAs including small nuclear RNA (snRNAs) and Malat1, which accumulate at the periphery of nuclear speckles (nsaRNA [nuclear speckle-associated RNA]), may extend to sufficient proximity to the genome. Leveraging a transcriptome-genome interaction assay (mapping of RNA-genome interactions [MARGI]), we identified clusters of nsaRNA-interacting genomic sequences (nsaPeaks). Posttranscriptional pre-mRNAs, which also accumulate to nuclear speckles, exhibited proximity to nsaPeaks but rarely to other genomic regions. Our combined DNA fluorescence *in situ* hybridization and immunofluorescence analysis in 182 single cells revealed a 3-fold increase in odds for nuclear speckles to localize near an nsaPeak than its neighboring genomic sequence. These data suggest a model that nsaRNAs are located in sufficient proximity to the nuclear genome and leave identifiable genomic footprints, thus revealing the parts of genome proximal to nuclear speckles.

INTRODUCTION

It is increasingly evident that positioning and organization of various subnuclear structures are critical for regulating gene expression, and therefore resolving the spatial organization of nuclear components has become a central task to nucleome research (Dekker et al., 2017). Nuclear bodies, previously known as interchromatin structures, typically exhibit non-overlapping spatial distributions with the genome (Brasch and Ochs, 1992; Sternsdorf et al., 1997). With an exception of nucleoli, which are positioned near ribosomal DNA (rDNA) (O'Sullivan et al., 2013), it remains challenging to identify the genomic sequences near most of the nuclear bodies, especially nuclear speckles (Lamond and Spector, 2003). Chromatin immunoprecipitation sequencing (ChIP-seq) targeting nuclear speckle core proteins rarely produces reproducible peaks (Kim et al., 2018), likely due to the lack of stable physical interactions between nuclear speckle core proteins and chromatin (Spector and Lamond, 2011; Chen, 2016; Kim et al., 2018).

Advanced imaging technologies including super-resolution imaging have started to reveal the multilayer structure of nuclear speckles, with the proteins SC35 and SON at the center (Fei et al., 2017) and nuclear speckle-associated noncoding RNAs (nsaRNA) including small nuclear RNA (snRNA) and Malat1 (Fei et al., 2017) as well as posttranscriptional precursor mRNAs (pre-mRNA) accumulated at the periphery (Misteli and Spector, 1997; Cmarko et al., 1999). In addition, distribution of Cdk9-cyclin T1 complex correlates with nuclear speckles (Dow et al., 2010; Herrmann and Mancini, 2001) but more often extends beyond the periphery of nuclear speckles (Spector and Lamond, 2011) (Figure 1A). A number of other proteins are associated with nuclear speckles (Spector and Lamond, 2011); however, it remains unclear whether their distribution corresponds to specific layers. The microscopic observation that noncoding RNAs are located at the outer layer of nuclear speckles (Fei et al., 2017) led us to hypothesize that these peripheral noncoding RNAs may be present in sufficient proximity to nuclear genome, leaving identifiable proximal sequences as their genomic footprints. Hereafter, we call this hypothesis the "nsaRNA proximity" hypothesis.

The recent technology on global mapping of RNA-genome interactions (MARGI) enabled the identification of interacting genomic sequences of chromatin-interacting RNAs (Sridhar et al., 2017). After cross-linking

[1]Department of Bioengineering, University of California San Diego, San Diego, CA 92093, USA

[2]Department of Pharmacology, University of California San Diego, San Diego, CA 92093, USA

[3]Division of Biological Sciences, University of California San Diego, San Diego, CA 92093, USA

[4]These authors contributed equally

[5]Lead Contact

*Correspondence: jiz175@ucsd.edu (J.Z.), szhong@ucsd.edu (S.Z.)

https://doi.org/10.1016/j.isci.2018.06.005

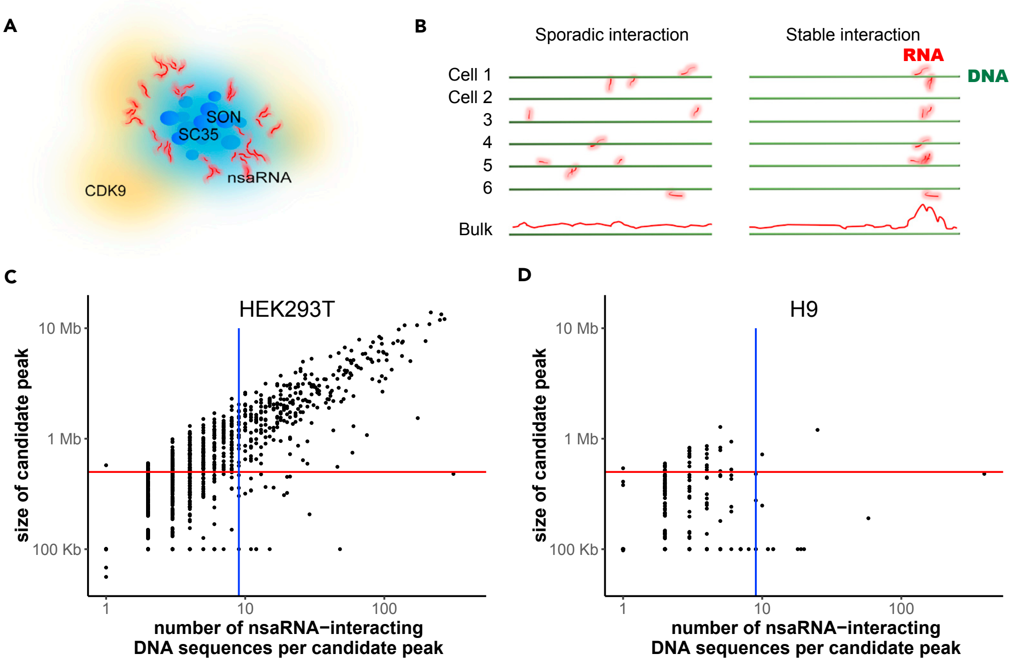


Figure 1. DNA Interaction Sites of nsaRNAs

(A) A cartoon of multilayer structure of nuclear speckles.

(B) Models of RNA-chromatin interaction in single cells, including sporadic interaction model and stable interaction model.

(C and D) Candidate peaks of nsaRNA-interacting DNA sequences in the genome. The number of nsaRNA-interacting DNA sequences (x axis) is plotted against cluster size (y axis) for every candidate peak in HEK293T (C) and H9 hES cells (D). Vertical line, 9 reads; horizontal line, 500 Kb.

and genome fragmentation, MARGI ligates RNA, a linker sequence, and proximal DNA to form an RNA-linker-DNA chimeric sequence, which is subsequently converted to double-stranded DNA and subjected to paired-end sequencing (see Figure 1 of Sridhar et al. (2017)). Because MARGI simultaneously assayed thousands of noncoding RNAs, including nsaRNAs, we will leverage MARGI data to test the nsaRNA proximity hypothesis.

Resolving spatial organization of nuclear components requires connecting information through different length scales and data types. Microscopic analyses have revealed non-uniform three-dimensional (3D) distribution of several types of RNAs in the nucleus. Prominent examples include Xist RNA cloud in adult female cells (Jonkers et al., 2008), accumulation of rRNA in nucleoli (Beven et al., 1996), and accumulation of snRNAs, Malat1, and posttranscriptional pre-mRNAs (p-pre-mRNAs) in nuclear speckles (Prasanth et al., 2010; Nakagawa et al., 2012; Tripathi et al., 2010; Galganski et al., 2017; Misteli and Spector, 1997; Cmarko et al., 1999; Girard et al., 2012). However, it remains a challenge to connect these microscopic findings with the latest information on 3D genome organization derived from genomics assays (Dekker et al., 2017). This challenge lies partially in the different length scales that vary in orders of magnitudes. For instance, the protein core of a nuclear speckle varies from one to several micrometers in diameter (Spector and Lamond, 2011), which is approximately 20%–50% of the spread of metaphase chromosomes (Lemke et al., 2002; Cremer and Cremer, 2001) or the diameters of chromosome territories (Cremer and Cremer, 2001; Bolzer et al., 2005). These relative sizes suggest that genomic regions in proximity to nuclear speckles may be significantly larger than the typical sizes of ChIP-seq or ATAC-seq peaks. Nevertheless, the enrichment of Xist RNA on X chromosome revealed by imaging (Jonkers et al., 2008) was successfully corroborated by genomics technologies including RAP-seq (Engreitz et al., 2014) and MARGI (Sridhar et al., 2017), offering an example of convergent findings from imaging and genomics approaches. In this work, we tested our "nsaRNA proximity" hypothesis by combining microscopic information and genomics data and aimed for establishing an RNA-based approach for identifying the relative positions of the folded genome and subnuclear structures.

RESULTS

MARGI Captures Proximity of Nuclear rRNA to Ribosomal DNA

We used the co-localization of nuclear rRNA and rDNA (human rDNA complete repeating unit) in nucleoli (Beven et al., 1996) as a test bed system to verify the assumption that RNA-DNA ligation sequencing (MARGI) data reflect spatial co-localization of a group of nuclear body-associated RNAs with specific genomic sequences. We reanalyzed MARGI datasets from human embryonic kidney (HEK) cells (GEO: GSM2427902 and GSM2427903) and human embryonic stem (hES) cells (GEO: GSM2427895 and GSM2427896) (Sridhar et al., 2017), which yielded approximately 9.9 million and 5.6 million RNA-DNA sequence pairs, respectively (Table S1). To test whether rRNAs are enriched in the proximity of rDNA, we categorized the RNA-DNA sequence pairs by the RNA type (rRNA or other types) and by the DNA (rDNA or the rest of the genome [hg38]) (Table S1). Compared with the other types of RNA, rRNA exhibited more than 400-fold increase of odds to ligate with rDNA in HEK cells (odds ratio = 404, p value $< 10^{-16}$) and more than 1,800-fold increase of odds in hES cells (odds ratio = 1,810, p value $< 10^{-16}$), confirming that MARGI data reflected co-localization of nucleolus-associated RNA and DNA.

nsaRNA-DNA Interaction Is Cell Type Specific

We asked which genomic regions are in proximity to nsaRNAs. At the single-cell level, there are four possible answers (models) to this question, which are (1) there is lack of nsaRNA expression; (2) nsaRNAs do not stably locate in the proximity of any specific genomic region in a single cell; (3) nsaRNAs are proximal to different DNA sequences in different single cells; however, none of these DNA sequences are shared by the majority of the cells; and (4) nsaRNAs are proximal to some DNA sequences and at least a fraction of these DNA sequences are shared by the majority of cells. Experiments with bulk cells could potentially differentiate the fourth model (stable interaction model) from its opposite (the first two models, collectively called the sporadic interaction model) (Figure 1B) but cannot further differentiate the first three models. Under the sporadic interaction model, bulk cell analysis (MARGI) is not expected to identify nsaRNA-DNA interactions (bulk lane, Figure 1B).

We used MARGI datasets to test the competing models. We reprocessed MARGI datasets generated from HEK and hES cells using the MARGI analysis pipeline (http://systemsbio.ucsd.edu/margi/) (Sridhar et al., 2017). This pipeline obtains the RNA-DNA read pairs with both ends uniquely mapped to the genome (hg38) and subsequently removes the proximal read pairs that may represent nascent transcripts. HEK and hES cells yielded 559,873 and 211,487 uniquely mapped RNA-DNA read pairs, respectively. In HEK cells, 14,904 pairs (2.5%) were nsaRNA-DNA pairs with the RNA end uniquely mapped to nsaRNAs (U1, U2, U4, U4atac, U5, U6, U6atac, U11, U12 [Carmo-Fonseca et al., 1992; Huang and Spector, 1992; Will and Luhrmann, 2011; Patel and Steitz, 2003; Pessa et al., 2008], 7SK [Peterlin et al., 2012; Prasanth et al., 2010], Malat1 [Tripathi et al., 2010]). In comparison, there were 1,857 nsaRNA-DNA pairs in hES cells, corresponding to only 0.88% RNA-DNA pairs in hES cells. Compared with HEK, hES-derived read pairs exhibited 3-fold reduction in odds of being nsaRNA-DNA repairs (odds ratio = 2.9, chi-square p value $< 10^{-16}$), which is reminiscent of lack of nuclear speckle formation in hES cells, where SC35 proteins and nsaRNAs are diffusely distributed in the nuclei (Butler et al., 2009).

In HEK, the nsaRNA-interacting DNA formed candidate peaks (Figure 1C; red curve in Figure 2). Analysis with Homer (v4.8.3) yielded a total of 295 broad peaks (nsaPeaks, Figure 5), which contained 10,771 (72%) of the nsaRNA-interacting DNA sequences (permutation p value < 0.001). The sizes of nsaPeaks ranged from 100 kb to 13 Mb, on the same scale of nuclear lamina-associated domains that span 10 kb–10 Mb (van Steensel and Belmont, 2017). The clustering of nsaRNA-interacting DNA sequences in the genome is consistent with the stable interaction model. In contrast, nsaRNA-interacting DNA sequences barely exhibited any clustering formation in the genome of hES cells and yielded two broad peaks by Homer analysis. Adjusting for the total amount of candidate peaks and isolated nsaRNA-interacting DNA in each cell type (Figure 1D), hES cells exhibited more than 80-fold reduction in production of nsaRNA interaction peaks when compared with HEK cells (odds ratio = 88.9, p value $< 10^{16}$). The sporadic distribution of nsaRNA-interacting DNA sequences in hES cells is also consistent with the lack of SC35 clusters in hES cells (sporadic interaction, Figure 1B).

We proceeded to test 295 nsaRNA-associated broad peaks (nsaPeaks) identified in HEK cells as the genomic regions close to nuclear speckles. We carried out these tests with two other types of nuclear speckle-associated molecules, namely, p-pre-mRNAs and CDK9 proteins.

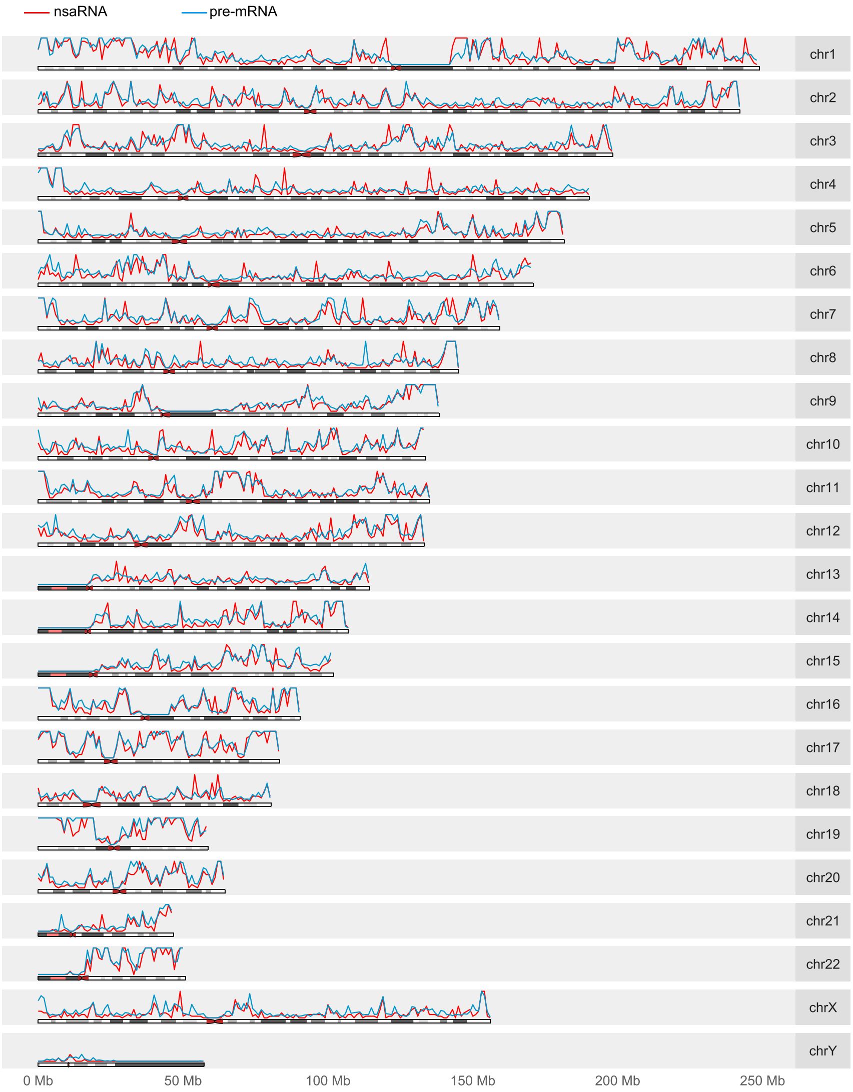


Figure 2. Genome-wide Density Distributions of nsaRNA-Interacting DNA Sequences (Red Curve) and Pre-mRNA-Proximal DNA (Blue Curve) in HEK293T Cells

Binsize = 1 Mbp.

Posttranscriptional Pre-mRNAs Exhibit Spatial Proximity to nsaRNA-Interacting DNA

If nsaPeaks are near nuclear speckles, other nuclear speckle-associated molecules besides nsaRNAs may also exhibit enrichment in spatial proximity of nsaPeaks. Although splicing is generally initiated co-transcriptionally, not all splicing events are completed during transcription. The resulting p-pre-mRNAs are observed to cluster at the nuclear speckle domains (Misteli and Spector, 1997; Cmarko et al., 1999; Girard et al., 2012). The clustering of p-pre-mRNAs offers another characteristic of nuclear speckles. We leveraged this characteristic for testing nsaPeaks as the part of genome proximal to nuclear speckles. The key assumption of this test is that clustering of RNAs in 3D predicts clustering of their interacting genomic sequences in the genome. To test this assumption, we examined whether p-pre-mRNA-interacting genomic sequences exhibit clustering patterns or are sporadically distributed in the genome. We processed MARGI data from HEK cells (Sridhar et al., 2017) to identify p-pre-mRNA-DNA interactions. To identify pre-mRNA reads, we required the RNA end of a MARGI read pair to span across an exon-intron junction and cover the intron by 10 or more nucleotides. To eliminate the reads that were potentially derived from nascent pre-mRNAs, we removed any MARGI read pairs with the RNA end mapped within 2,000 nt to the genomic location where the DNA end was mapped to. The remaining 187,724 uniquely mappable sequence pairs representing p-pre-mRNA-DNA interactions were obtained. More than 93% (175,626) of these 187,724 read pairs represented interchromosomal interactions. The 187,724 DNA ends of these sequence pairs were not uniformly distributed in the genome (blue curve, Figure 2); instead they concentrated to certain genomic regions, yielding 284 broad peaks (Homer v4.8.3, broad peak option) (p value < 0.001, permutation test) (Figure S1). Taken together, p-pre-mRNAs exhibited proximity to remote DNA sequences. These remote interacting sequences exhibited clustering patterns in the genome, which corroborates with the idea that p-pre-mRNAs are clustered rather than diffusively distributed in the nucleus (Misteli and Spector, 1997; Cmarko et al., 1999; Girard et al., 2012).

We compared nsaRNA-interacting DNA and the DNA sequences proximal to p-pre-mRNAs by genome-wide density distributions, broad peaks, and genomic windows. The genome-wide distribution of p-pre-mRNA-proximal DNA sequences exhibited remarkable similarity to the distribution of nsaRNA-interacting DNA (Figure 2). A total of 170 (57.6%) nsaPeaks overlapped with p-pre-mRNA broad peaks (Figures S1 and S2A) (p value < 0.001, permutation test). Finally, we broke the genome into equal-sized windows and calculated the densities of nsaRNA-interacting DNA and p-pre-mRNA-proximal sequences in each window. These two density profiles exhibited a genome-wide correlation (Spearman correlation = 0.957, p value < 10^{-16}) (Figures S2B and S2C). Taken together, p-pre-mRNA-proximal genomic regions exhibited significant overlap with nsaRNA-interacting DNA, supporting the idea that nsaPeaks reflect the parts of genome near nuclear speckles.

Correspondence of Genome-wide Binding Profile of CDK9 and Genome-wide Distribution of nsaRNA-Interacting DNA

We compared genome-wide binding profile of CDK9 to genome-wide distribution of nsaRNA-interacting DNA sequences. ChIP-seq of nuclear speckle core proteins has been regarded a questionable approach for identifying the relative positions of nuclear speckles and the genome (Chen, 2016; Dekker et al., 2017), due to the physical separation of nuclear speckle cores from chromatin (Spector and Lamond, 2011). For example, suppose 95% of copies of a core protein, for instance, SC35, were located at the nuclear speckle cores and the other 5% were sporadically distributed, some of which are attached to chromatin, ChIP would select for the few chromatin-associated SC35 rather than those at the nuclear speckle cores. To alleviate this documented concern, we resorted to CDK9 proteins that are distributed throughout the core and periphery of nuclear speckles (Spector and Lamond, 2011; Dow et al., 2010; Herrmann and Mancini, 2001) for a ChIP-seq analysis. And even so we did not anticipate many overlaps between CDK9 ChIP-seq peaks and nsaRNA-interacting DNA sequences. We identified a total of 6,517 CDK9 peaks from HEK293T cells (GEO: GSM1249897) (Liu et al., 2013) (MACS2) (Zhang et al., 2008), of which only 551 (8.5%) were located within 200 bp of an nsaRNA-interacting DNA sequence. This overlap was statistically significant (p value < 0.001, permutation test), consistent with the idea that CDK9's distribution overlaps with nuclear speckles. However, the relatively small number of actual overlaps is reminiscent of the recognized challenge of using ChIP to identify nuclear speckle-interacting genomic regains (Chen, 2016; Dekker et al., 2017). These data also suggest that nsaPeaks do not precisely overlap with CDK9-bound promoters.

Considering that the 3D distribution of CDK9 is centered at nuclear speckles (Dow et al., 2010; Herrmann and Mancini, 2001; Spector and Lamond, 2011), we tested the possibility that CDK9 ChIP-seq peaks cluster to the same genomic regions as nsaPeaks. Indeed, genome-wide density distribution of CDK9 peaks (green curve, Figure 3) resembled the density distribution of nsaRNA-interacting DNA (red curve, Figure 3). In a control comparison, genome-wide density distributions of H3K9me3 (Encode: ENCFF002AAZ) (Consortium, 2012) and nsaRNA-interacting DNA exhibited a poor correlation (Pearson correlation = 0.03, Spearman correlation = 0.27) (Figure S3). To test whether CDK9 binding sites cluster to the same genomic regions as nsaPeaks, we identified a total of 262 CDK9 broad peaks (sizes range from 514,083 bp to 6,262,520 bp, median size = 1,328,930 bp) (Homer, v4.8.3) (Heinz et al., 2010), of which 206 (78.6%) overlapped with nsaPeaks (p value < 0.001, permutation test) (Figures S4A and S5). Next, we split the genome (hg38) into 3.08 million 1,000-bp windows, of which 0.44 million windows overlapped with CDK9 broad peaks, of which 0.32 million windows also overlapped with nsaPeaks, suggesting strong association (odds ratio = 11.5, p value <10^{-16}, Fisher's exact test) (Figure S4B). Taken together, although CDK9 does not frequently bind to the exact sequences as nsaRNA-interacting DNA, CDK9 binding sites accumulated to nsaPeaks, corroborating with the idea that nsaPeaks reflect the portion of genome closer to nuclear speckles.

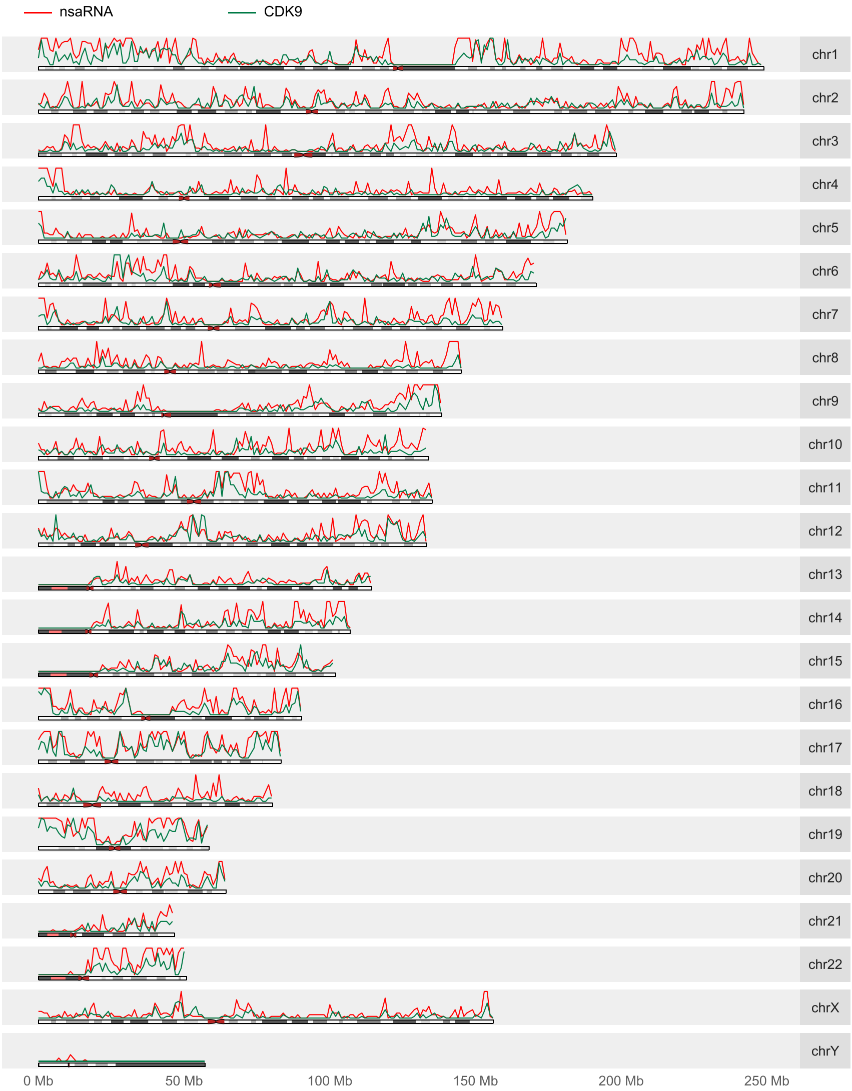


Figure 3. Genome-wide Density Distributions of nsaRNA-Interacting DNA Sequences (Red Curve) and CDK9 ChIP-seq Sequences (Green Curve) in HEK293T Cells

Binsize = 1 Mbp.

Co-localization of SC35 Clusters and nsaPeaks in Single Cells

We examined the proximity of nuclear speckles to nsaPeaks at single-cell resolution using a combination of immunofluorescence staining of a nuclear speckle core protein SC35 and DNA fluorescent *in situ* hybridization (FISH) (Sayegh et al., 2005). We opted to use commercially validated FISH probes, and we wanted the probes to be on the same chromosome arm. We identified a pair of probes satisfying these criteria on chromosome 11 with one probe (bacterial artificial chromosome plasmid DNA) inside an nsaPeak (Empire Genomics: RP11-772K10, hereafter called the nsaPeak probe) and the other probe outside the nsaPeaks (Empire Genomics: RP11-908J16, hereafter called the non-nsaPeak probe) (Figure 4A). We imaged 82 and 100 single cells with nsaPeak probe and non-nsaPeak probe, respectively, co-stained with SC35 antibody. Each cell exhibited 1 to 3 FISH spots, consistent with pseudotriploidy of HEK293T cells, and 20 to 35 SC35 clusters (Figures 4B and 4C).

To minimize the sensitivity of results to image analysis methods, we carried out two sets of analyses based on different analysis methods. First, we identified each FISH spot and its associated pixel on every z stack by particle analysis (ImageJ) (Schneider et al., 2012). A FISH spot was called isolated

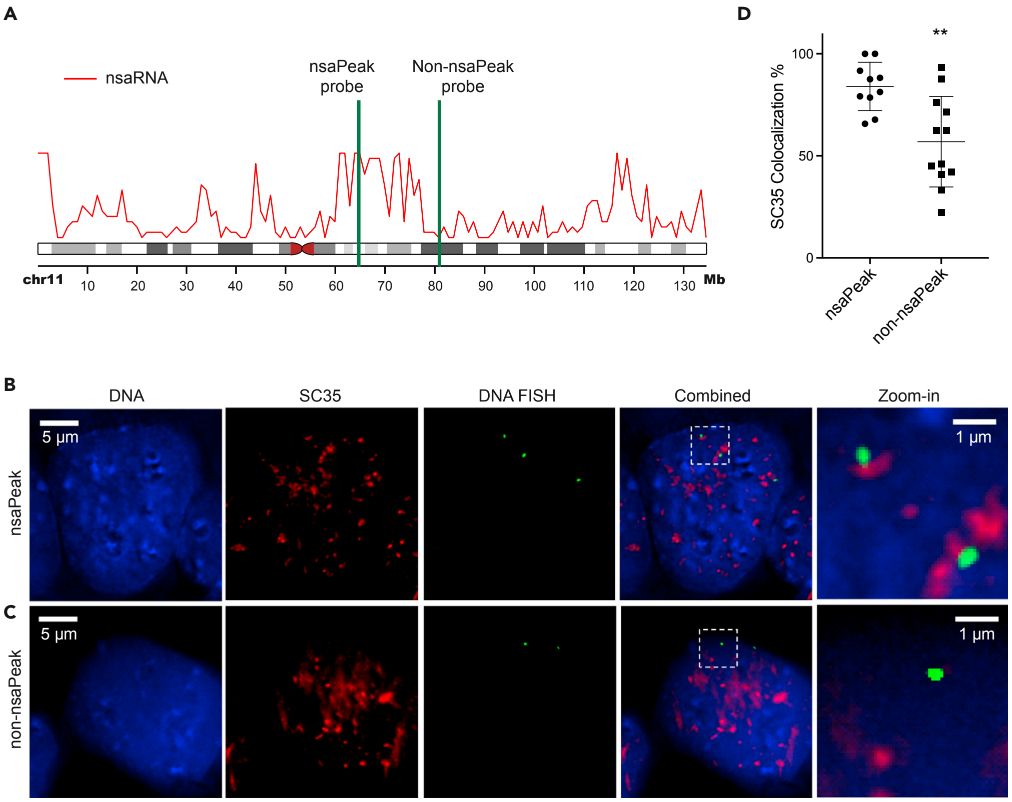


Figure 4. Visualization of Representative nsaPeak and Non-nsaPeak with SC35 Clusters

(A) Genomic positions of nsaPeak probe and non-nsaPeak probe (green) with respect to nsaPeaks (red).

(B and C) Representative images of HEK293T cells co-stained with Hoechst (DNA, blue), SC35 (red), and DNA FISH (green) with nsaPeak probe (B) and non-nsaPeak probe (C). Scale bar: 5 μm. Last column: zoom-in views of the selected regions in the dashed boxes. Scale bar: 1 μm.

(D) The percentage of FISH spots that exhibited overlapping SC35 signal (y axis) in each image (dot) were plotted for samples interrogated with the nsaPeak probe images (left) and the non-nsaPeak probe (right). Error bar: SD. $^{**}p < 0.003$.

from SC35 clusters only when none of its associated pixels exhibited SC35 signal. Otherwise a FISH spot was called co-localized with SC35. This is a conservative approach to call isolated FISH spots. Among the 210 nsaPeak FISH spots identified from 82 individual cells, 170 FISH spots (84.0%) co-localized with SC35 clusters. In comparison, among the 193 non-nsaPeak FISH spots identified from 100 cells, 110 co-localized with SC35 (56.9%), reflecting a 3-fold reduction in odds (odds ratio = 3.1, p value $< 5\times10^{-7}$, chi-square test). We also summarized the proportion of co-localized FISH spots in each image. The 10 images stained with nsaPeak probe exhibited on average 84.0% of their FISH spots co-localized with SC35 (dots in left column, Figure 4D). In comparison, the 12 images (dots in right column) stained with the non-nsaPeak probe had on average 56.9% FISH spots co-localized with SC35 (p value < 0.003, t test) (Figure 4D).

In the second analysis, we compared the FISH-to-SC35 distance distributions between nsaPeak and non-nsaPeak samples. We computed center-to-center distance in 3D from every FISH spot to its nearest SC35 cluster. We summarized the number of center pairs at each distance from 1 to 10 voxels in every image (Figure S6). The nsaPeak images exhibited two to three times more center pairs than non-nsaPeak images at every distance (p value $< 10^{-5}$, Kolmogorov test). For example, the nsaPeak images exhibited 1 to 18 center pairs at a distance of 8 voxels, whereas the non-nsaPeak images exhibited 0 to 3 at this distance (Figure S6). The different distance distributions suggest that the interrogated nsaPeaks are closer to the SC35 clusters than the interrogated non-nsaPeaks among the analyzed single cells. Taken together, the two analyses based on different analysis assumptions both revealed clear differences in relative positions of nuclear speckles to the two interrogated genomic regions. In summary, pre-mRNA data,

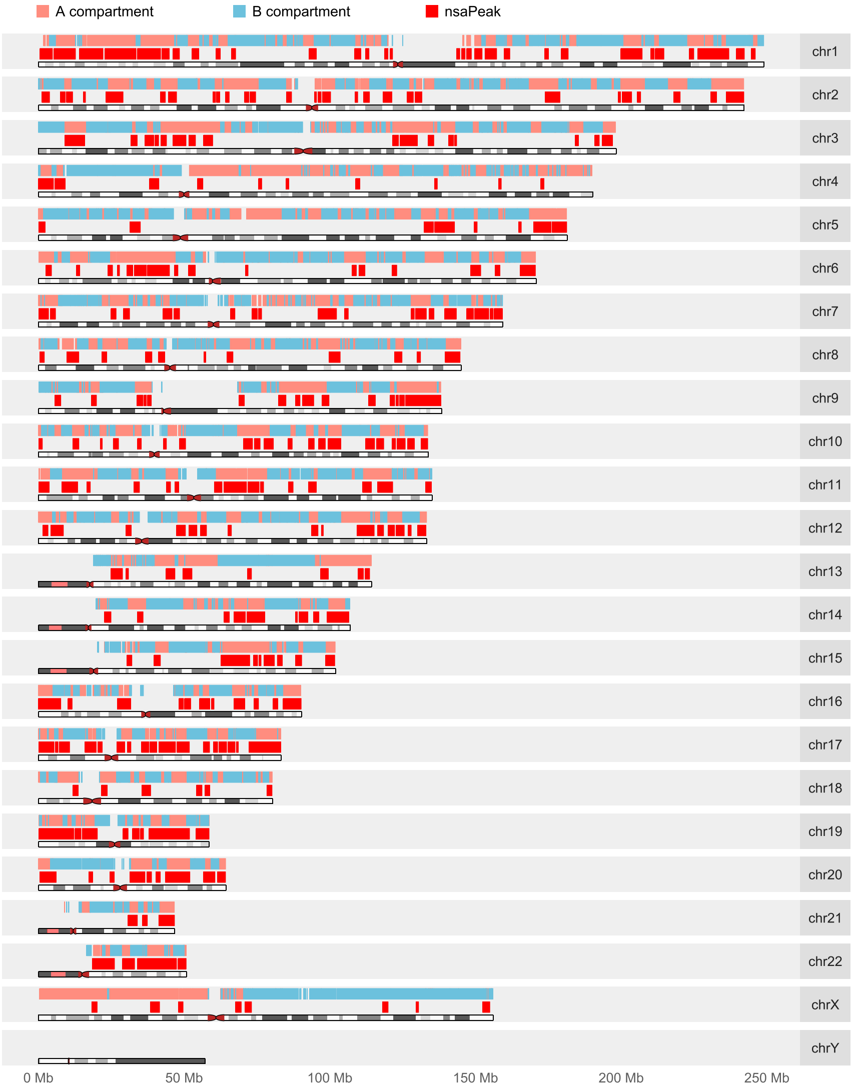


Figure 5. Genome-wide View of A (pink)/B (Light Blue) Compartments and nsaPeaks (Red)

CDK9 data, and single-cell image data supported the nsaPeaks as nuclear speckle-proximal genomic regions.

nsaPeaks Correlate with the A Compartment in HEK Cells but Not in Embryonic Stem Cells

We exploited how nsaPeaks fit into the current knowledge of the 3D structure of genome. Toward this goal, we compared nsaPeaks with nuclear compartments (Cockell and Gasser, 1999) and topologically associated domains (TADs) (Dixon et al., 2012). We called A/B compartments (Lieberman-Aiden et al., 2009) from HEK293T Hi-C data (Zuin et al., 2014) with Homer (v4.8.3) (Heinz et al., 2010; Lieberman-Aiden et al., 2009). Approximately half of the genome was associated with the A compartment (first row, Table S2). Approximately half of the genome in the A compartment and slightly more than 10% of the genome in the B compartment are associated with nsaPeaks (Figure 5), suggesting that nsaPeaks are enriched in but are not a complete subset of the genomic sequences in the A compartment. In line with this observation, nsaPeaks exhibited baseline increase of H3K4me3, H3K27ac, and H3K36me3 (25–75 quantiles, Figure S7B). In contrast, although a sizable proportion of the genome was categorized into the A compartment in hES cells (Dixon et al., 2015), merely two nsaPeaks were detected in hES cells (Figure 1D). These data argue against the possibility of the formation of nsaRNA clusters as a cause of A/B compartmentation of the nucleus, but suggest that nsaRNA clusters were formed in the same nuclear compartment that contains the A compartment of the genome.

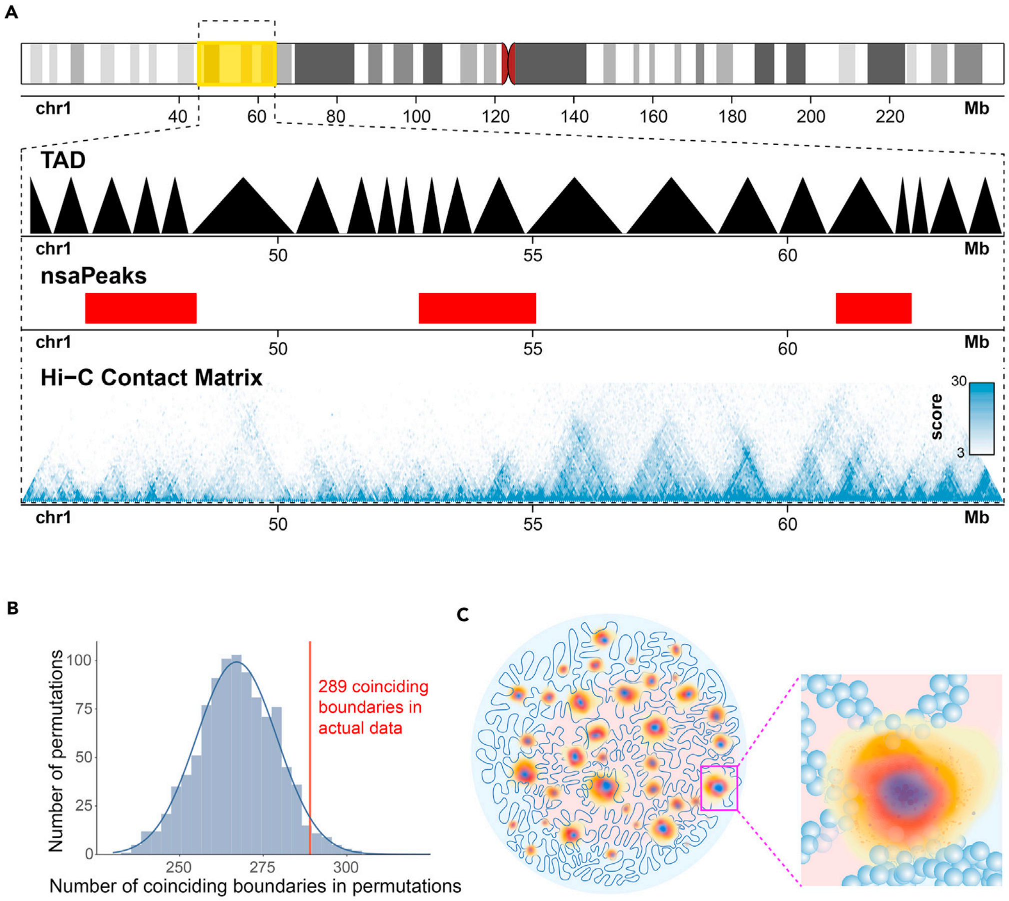


Figure 6. nsaPeaks and TADs

(A) Genome view of TADs, nsaPeaks, and Hi-C contact matrix.

(B) Background distribution of the numbers of TAD boundaries coinciding with TAD boundaries from 1,000 permutations (histogram and fitted curve) versus the number of observed coinciding boundaries in actual data (red line).

(C) A model of boundaryless nuclear speckles and the genome. Nuclear speckle cores are in red. Other nuclear speckle-associated molecules exhibit diffusive patterns centered by nuclear speckle cores (blue, orange, yellow), some of which extend to sufficient proximity to certain TADs (balls, inset). Pink/light blue: A/B compartment.

Genome Sequence in a TAD Tends to Be Either Entirely Close to or Distant from Nuclear Speckles

The notion of TADs was derived from Hi-C experiments (Dixon et al., 2012), and TADs are subsequently proposed as a structural unit of genome organization (Dixon et al., 2016). We reasoned that organizational units should exhibit unity in relative positions to other nuclear components, and therefore proximity of the genome and nuclear speckles may offer an alternative test to this proposition. We compared the 3,258 TADs derived from HEK293T Hi-C data (GEO: GSM1081530) (Zuin et al., 2014) and nsaPeaks. Nearly 50% (289 of 590) of the boundaries of nsaPeaks were aligned with TAD boundaries (p value = 0.03, permutation test) (Figures 6A and 6B). A total of 74 nsaPeaks were aligned with 361 TADs, where each nsaPeak coincided with one TAD or several consecutive TADs (p value = 0.051, permutation test). Recognizing the sensitivity of peak boundaries to noises in data and to algorithm, we did another test with an alternative set of boundaries. Based on the significant overlap of nsaPeaks and CDK9 broad peaks (Figure S5), we merged the two sets of peaks (union) and obtained 334 union-peaks. Approximately 52% (350 of 668) of union-peak boundaries were aligned with TAD boundaries (p value = 0.001, permutation test). A total of 98 union-peaks were aligned with 468 TADs, where each union-peak coincided with one TAD or several consecutive TADs (p value = 0.005, permutation test). Taken together, MARGI data suggest that the genomic sequence of a TAD tends to be either entirely close to or entirely distant from nuclear speckles, supporting the proposition of TADs being structural units.

DISCUSSION

Challenges in Identifying Relative Positions of Nuclear Speckles with Respect to Genomic Sequence

More than 150 proteins were reported to be associated with nuclear speckles (Saitoh et al., 2004), including small nuclear ribonucleoprotein particles and SR proteins essential for RNA splicing (Fu, 1995) and a number of kinases and phosphatases that regulate splicing machinery (Spector and Lamond, 2011). However, most of these proteins are not only present in nuclear speckles, and there is not sufficient data to assess the specificity of their localization to nuclear speckles. Therefore, the small number of proteins localized at the core of nuclear speckles, namely, SC35 and SON, received focal attention and were used as nuclear speckle markers in attempts to identify nuclear speckle-proximal genomic regions (Spector and Lamond, 2011; Chen, 2016; Kim et al., 2018). However, the detachment of nuclear speckle cores to chromatin suggested that ChIP-seq analyses of nuclear speckle core proteins would unlikely reveal the genomic sequences close to nuclear speckles (Chen, 2016). Thus, finding relative positions of nuclear speckles with respect to genomic sequences remains a major challenge in nucleome research (Dekker et al., 2017).

RNAs as Media for Proximity Labeling

The increasing evidence on ''noncoding RNAs functioning as scaffolds in the construction of nuclear bodies'' points to the essential role of RNA in nuclear bodies (Wrighton, 2016). Nuclear speckles exhibit clear centers but showed inconsistent boundary lines when visualized by staining different nuclear speckle markers (Fei et al., 2017). Evidence of nsaRNA locating at the peripheral regions of nuclear speckles (Fei et al., 2017) fostered our hypothesis of this study that nsaRNAs serve as ''proximity labeling'' media, which ''mark'' proximal DNA (Figure 6C). Recently developed MARGI technology (Sridhar et al., 2017) enabled us to further examine this hypothesis by analyzing RNA-chromatin interactions of many noncoding RNAs at the same time.

Cellular Heterogeneity and Assays of Bulk Cells

A rationale of ChIP-seq and ATAC-seq analyses of bulk cells is that the majority of single cells share the same transcription factor binding regions or transposase-accessible regions, and such commonality would be identified as peaks in bulk cell experiments. This rationale was verified by single-cell data produced by subsequently invented single-cell ChIP-seq (Rotem et al., 2015) and single-cell ATAC-seq technologies (Buenrostro et al., 2015). The same rationale is applicable to the MARGI technology in that only the genomic regions shared (relatively invariable) across many single cells would have a chance to appear in a bulk cell assay, whereas single-cell-specific interaction regions can hardly produce significant signals in a bulk cell assay (Figure 1B). Although there does not exist a single-cell version of MARGI technology, single-cell imaging analysis provided data consistent with this rationale.

Genome as a Surrogate Coordinate for Studying Nuclear Organization

Revealing spatial organization of nuclear components has become a central task in nucleome research. This task is hindered by the lack of a 3D coordinate system for the nucleus. Without a coordinate system, spatial data obtained from different single cells cannot be aligned, making it difficult to derive or test for any underlying principles.

Chromosome territories fill sizable portions of interphase nuclei (Bolzer et al., 2005). The correspondence between the any piece of uniquely mappable sequence and its genomic location makes it possible for the nuclear genome to serve as a surrogate coordinate system of the nucleus, given that a 3D location in the nucleus could be approximated by its nearest genomic sequence. Compared with the alternative of not having any 3D coordinate at all, the genome-surrogate-3D coordinate provides a primitive means to record positional information that is potentially comparable across single cells or cell types. This surrogate coordinate has its own limitations, including lack of means to transform the surrogate coordinate into a physical coordinate and lack of power to differentiate chromosome pairs. This work was a test of this genome-surrogate-3D coordinate. Both chromosomes and nuclear bodies could have variable and cell-specific 3D positions; however, our data suggested that the relative positions between nuclear speckles and chromosomes were relatively stable. Thus, accumulated knowledge of relative positions of various nuclear components (van Steensel and Henikoff, 2000; van Steensel and Belmont, 2017) with respect to the nuclear genome may unleash the power of the genome-surrogate-3D coordinate in future analyses of spatial organization of the nucleus.

Limited Resolution of MARGI-Derived nsaRNA-DNA Interactions

A major limitation of the proposed mapping strategy is the small number of MARGI-derived nsaRNA-DNA interaction read pairs. The total number of uniquely mapped nsaRNA-DNA read pairs in HEK293T cells from one MARGI experiment was 14,904. Each snRNA was only reflected by hundreds or thousands of read pairs, making it nearly infeasible to distinguish the proximity regions of major and minor spliceosomes. Another possible caveat of this analysis is the distance threshold for identifying p-pre-mRNA-DNA interactions. The currently applied threshold, 2,000 bp, may not be sufficient to remove all nascent RNA-DNA interactions. However, among the 187,724 identified p-pre-mRNA-DNA read pairs, the vast majority (175,626, 93%) were interchromosomal interactions. Therefore, even if the distance threshold is significantly increased, it is unlikely to result in large changes to the p-pre-mRNA analysis results.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, seven figures, and two tables and can be found with this article online at https://doi.org/10.1016/j.isci.2018.06.005.

ACKNOWLEDGMENTS

Authors thank Dr. Alexandra Bortnick for advices on optimizing the immunofluorescence and DNA FISH assay, Dr. Lucie Hebert for providing HEK293T cells, and Xingzhao Wen for comments to the manuscript. This work is funded by NIH R35CA197622 (J.Z.), DP1HD087990 (S.Z.), and U01CA200147 (S.Z.).

AUTHOR CONTRIBUTIONS

Conceptualization, W.C., Z.Y., and S.Z.; Methodology, W.C., Z.Y., S.L., and N.H.; Investigation, W.C., Z.Y., S.L., N.H.,J.Z., and S.Z.; Writing – Original Draft, W.C., Z.Y., S.L.,N.H., X.H., and S.Z.; Writing – Review & Editing, J.Z. and S.Z.; Funding Acquisition, J.Z. and S.Z.; Resources, J.Z. and S.Z.; Supervision, J.Z. and S.Z.

DECLARATION OF INTERESTS

S.Z. is a co-founder and board member of Genemo, Inc., which does not have financial interests with this work. The authors have no financial interests to declare.

Received: May 7, 2018
Revised: May 30, 2018
Accepted: June 5, 2018
Published: June 29, 2018

REFERENCES

Beven, A.F., Lee, R., Razaz, M., Leader, D.J., Brown, J.W., and Shaw, P.J. (1996). The organization of ribosomal RNA processing correlates with the distribution of nucleolar snRNAs. J. Cell Sci. *109* (Pt 6), 1241–1251.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M.R., and Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol. *3*, e157.

Brasch, K., and Ochs, R.L. (1992). Nuclear bodies (NBs): a newly "rediscovered" organelle. Exp. Cell Res. *202*, 211–223.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Butler, J.T., Hall, L.L., Smith, K.P., and Lawrence, J.B. (2009). Changing nuclear landscape and unique PML structures during early epigenetic transitions of human embryonic stem cells. J. Cell.Biochem. *107*, 609–621.

Carmo-Fonseca, M., Pepperkok, R., Carvalho, M.T., and Lamond, A.I. (1992). Transcription-dependent colocalization of the U1, U2, U4/U6, and U5 snRNPs in coiled bodies. J. Cell Biol. *117*, 1–14.

Chen, Y. (2016). "TSA-Seq": A Novel Proximity Mapping Approach for Studying Three Dimensional Genome Organization and Function (University of Illinois Urbana-Champaign).

Cmarko, D., Verschure, P.J., Martin, T.E., Dahmus, M.E., Krause, S., Fu, X.D., van Driel, R., and Fakan, S. (1999). Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-UTP microinjection. Mol. Biol. Cell *10*, 211–223.

Cockell, M., and Gasser, S.M. (1999). Nuclear compartments and gene regulation. Curr.Opin.Genet. Dev. *9*, 199–205.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. Mol. Cell *62*, 668–680.

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

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.

Dow, E.C., Liu, H., and Rice, A.P. (2010). T-loop phosphorylated Cdk9 localizes to nuclear speckle domains which may serve as sites of active P-TEFb function and exchange between the Brd4 and 7SK/HEXIM1 regulatory complexes. J. Cell. Physiol. *224*, 84–93.

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.

Fei, J., Jadaliha, M., Harmon, T.S., Li, I.T.S., Hua, B., Hao, Q., Holehouse, A.S., Reyer, M., Sun, Q., Freier, S.M., et al. (2017). Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. J. Cell Sci. *130*, 4180–4192.

Fu, X.D. (1995). The superfamily of arginine/serine-rich splicing factors. RNA *1*, 663–680.

Galganski, L., Urbanek, M.O., and Krzyzosiak, W.J. (2017). Nuclear speckles: molecular organization, biological function and role in disease. Nucleic Acids Res. *45*, 10350–10368.

Girard, C., Will, C.L., Peng, J., Makarov, E.M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K., and Luhrmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. Nat. Commun. *3*, 994.

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

Herrmann, C.H., and Mancini, M.A. (2001). The Cdk9 and cyclin T subunits of TAK/P-TEFb localize to splicing factor-rich nuclear speckle regions. J. Cell Sci. *114*, 1491–1503.

Huang, S., and Spector, D.L. (1992). U1 and U2 small nuclear RNAs are present in nuclear speckles. Proc. Natl. Acad. Sci. USA *89*, 305–308.

Jonkers, I., Monkhorst, K., Rentmeester, E., Grootegoed, J.A., Grosveld, F., and Gribnau, J. (2008). Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. Mol. Cell. Biol. *28*, 5583–5594.

Kim, J., Khanna, N., and Belmont, A.S. (2018). Transcription enhancement by nuclear speckle association. Biophys.J. *114*, 246a.

Lamond, A.I., and Spector, D.L. (2003). Nuclear speckles: a model for nuclear organelles. Nat. Rev. Mol. Cell Biol. *4*, 605–612.

Lemke, J., Claussen, J., Michel, S., Chudoba, I., Muhlig, P., Westermann, M., Sperling, K., Rubtsov, N., Grummt, U.W., Ullmann, P., et al. (2002). The DNA-based structure of human chromosome 5 in interphase. Am. J. Hum. Genet. *71*, 1051–1059.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Liu, W., Ma, Q., Wong, K., Li, W., Ohgi, K., Zhang, J., Aggarwal, A., and Rosenfeld, M.G. (2013). Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell *155*, 1581–1595.

Misteli, T., and Spector, D.L. (1997). Protein phosphorylation and the nuclear organization of pre-mRNA splicing. Trends Cell Biol. *7*, 135–138.

Nakagawa, S., IP, J.Y., Shioi, G., Tripathi, V., Zong, X., Hirose, T., and Prasanth, K.V. (2012). Malat1 is not an essential component of nuclear speckles in mice. RNA *18*, 1487–1499.

O'Sullivan, J.M., Pai, D.A., Cridge, A.G., Engelke, D.R., and Ganley, A.R. (2013). The nucleolus: a raft adrift in the nuclear sea or the keystone in nuclear structure? Biomol. Concepts *4*, 277–286.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Pessa, H.K., Will, C.L., Meng, X., Schneider, C., Watkins, N.J., Perala, N., Nymark, M., Turunen, J.J., Luhrmann, R., and Frilander, M.J. (2008). Minor spliceosome components are predominantly localized in the nucleus. Proc. Natl. Acad. Sci. USA *105*, 8655–8660.

Peterlin, B.M., Brogie, J.E., and Price, D.H. (2012). 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. Wiley Interdiscip. Rev. RNA *3*, 92–103.

Prasanth, K.V., Camiolo, M., Chan, G., Tripathi, V., Denis, L., Nakamura, T., Hubner, M.R., and Spector, D.L. (2010). Nuclear organization and dynamics of 7SK RNA in regulating gene expression. Mol. Biol. Cell *21*, 4184–4196.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol. *33*, 1165–1172.

Saitoh, N., Spahr, C.S., Patterson, S.D., Bubulya, P., Neuwald, A.F., and Spector, D.L. (2004). Proteomic analysis of interchromatin granule clusters. Mol. Biol. Cell *15*, 3876–3890.

Sayegh, C.E., Jhunjhunwala, S., Riblet, R., and Murre, C. (2005). Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. Genes. Dev. *19*, 322–327.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb.Perspect. Biol. *3*, a000646.

Sridhar, B., Rivas-Astroza, M., Nguyen, T.C., Chen, W., Yan, Z., Cao, X., Hebert, L., and Zhong, S. (2017). Systematic mapping of RNA-chromatin interactions in vivo. Curr. Biol. *27*, 610–612.

Sternsdorf, T., Grotzinger, T., Jensen, K., and Will, H. (1997). Nuclear dots: actors on many stages. Immunobiology *198*, 307–331.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell *39*, 925–938.

van Steensel, B., and Belmont, A.S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. Cell *169*, 780–791.

van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat. Biotechnol. *18*, 424–428.

Will, C.L., and Luhrmann, R. (2011). Spliceosome structure and function. Cold Spring Harb.Perspect. Biol. *3*, a003707.

Wrighton, K.H. (2016). Nuclear organization: building nuclear bodies with RNA. Nat. Rev. Mol. Cell Biol. *17*, 463.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van, I.W.F., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. USA *111*, 996–1001.

ISCI, Volume *4*

Supplemental Information

RNAs as Proximity-Labeling Media for Identifying Nuclear Speckle Positions Relative to the Genome

Weizhong Chen, Zhangming Yan, Simin Li, Norman Huang, Xuerui Huang, Jin Zhang, and Sheng Zhong

Supplementary figures and legends

Figure S1. Genome-wide view of broad peaks of pre-mRNA proximal DNA (blue) and nsaPeaks (red), related to Figure 2.

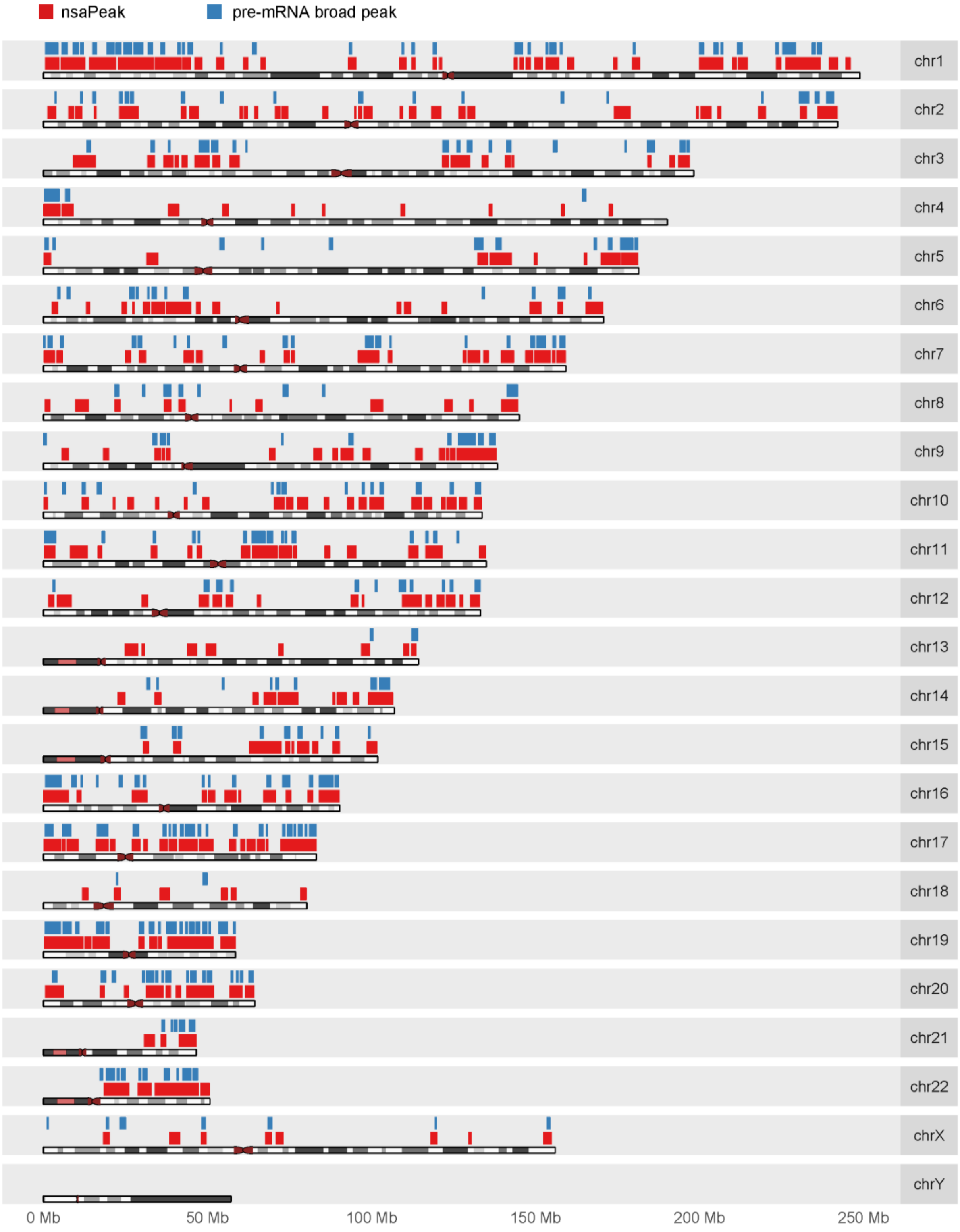


Figure S2. Comparison of nsaPeaks and pre-mRNA broad peaks, related to Figure 2. (A) Venn diagram of numbers of nsaPeaks (red), pre-mRNA broad peaks (blue), and overlaps. (B) Scatter plot of 3,102 genomic windows (1 Mb), with the number of nsaRNA interacting sequences (x axis) and the number of pre-mRNA proximal sequences in each window (y axis). (C) Scatterplot of 311 bins, where each bin is a group of 10 genomic windows, with the average number of nsaRNA interacting sequences (x axis) and the average number of pre-mRNA proximal sequences (y axis) of the 10 genomics windows of each bin.

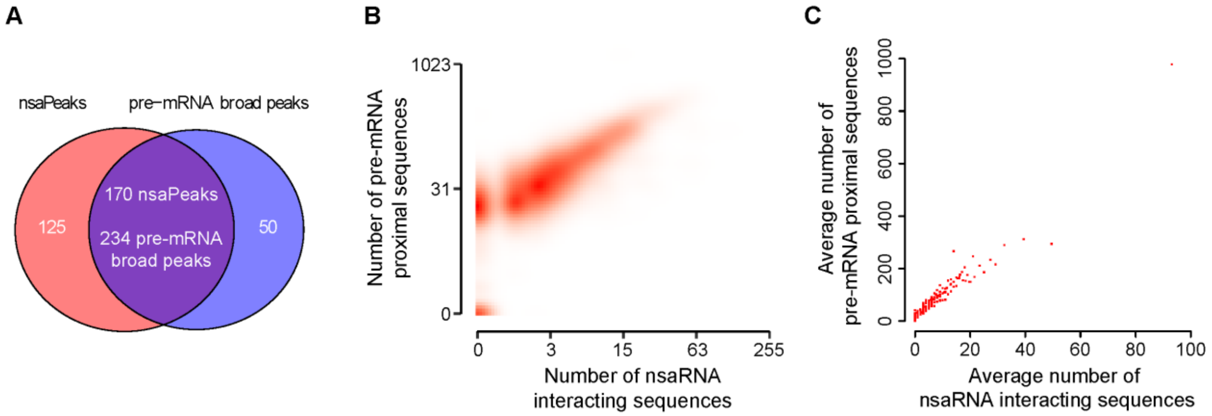


Figure S3. Genome-wide distribution of nsaRNA-interacting DNA sequences from MARGI (red) and H3K9me3 ChIP-seq sequences (blue) in HEK293T cells, related to Figure 3.

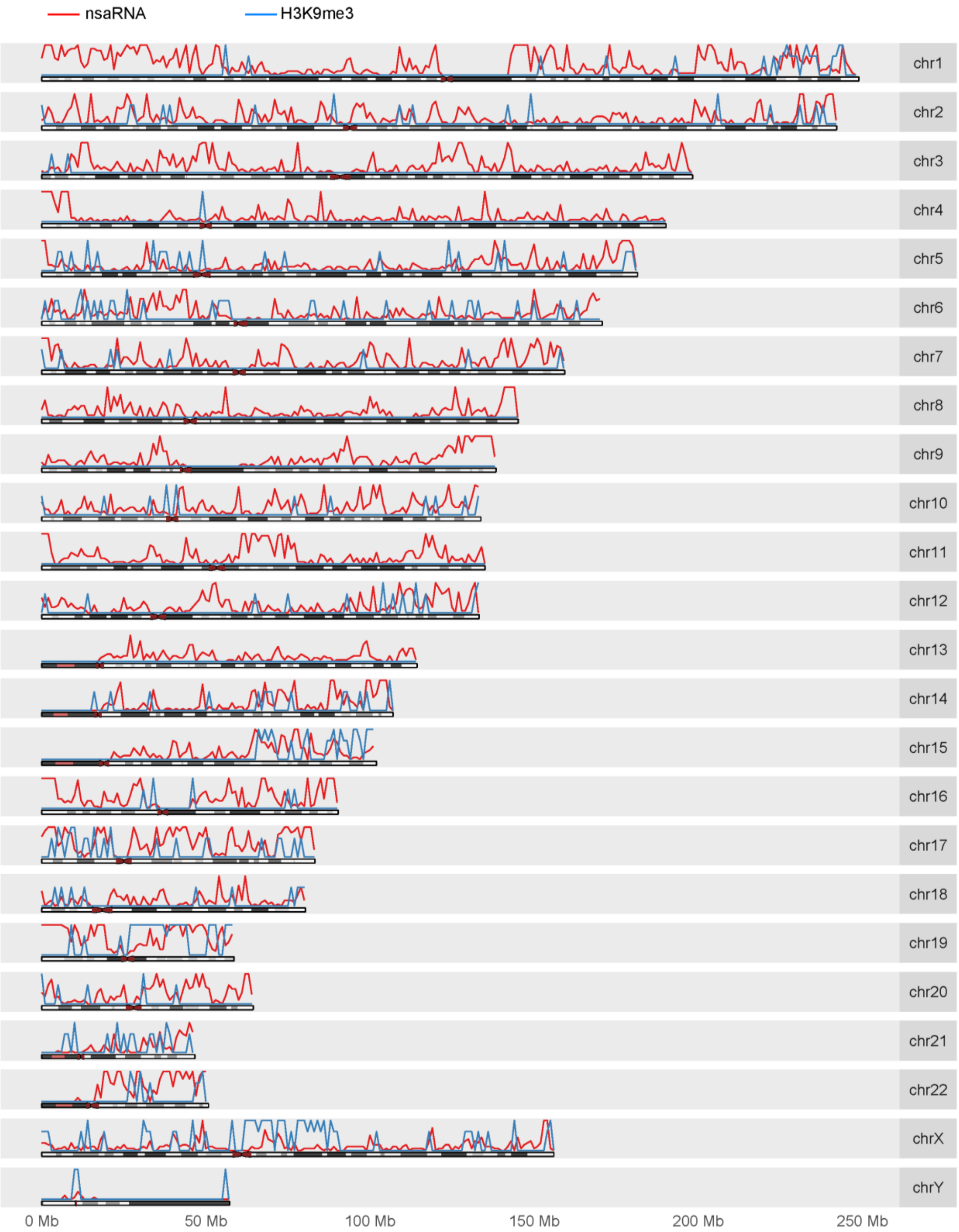
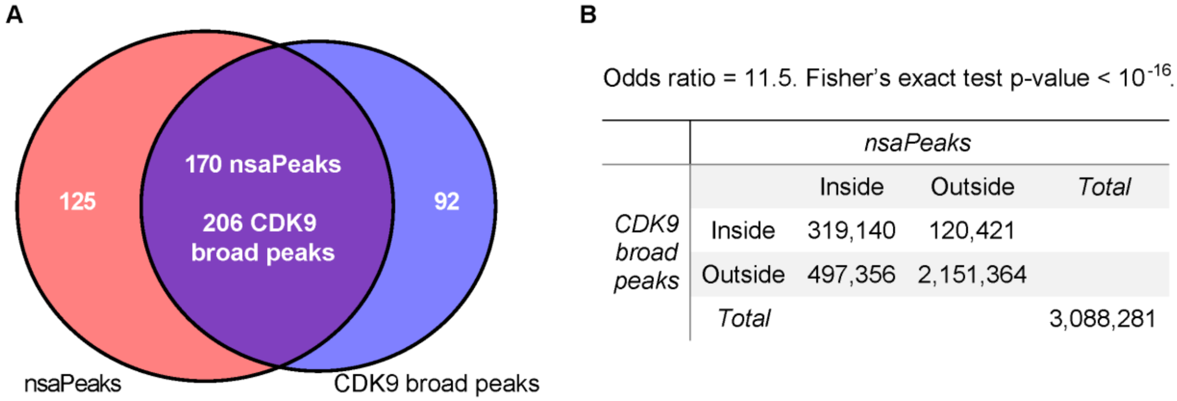


Figure S4. Comparison of nsaPeaks and CDK9 broad peaks, related to Figure 3. (A) Venn diagram. (B) Contingency table between genomic windows covered by CDK9 broad peaks and by nsaRNA-associated broad peaks. The genome (hg38) was split into 3,088,281 windows, with 1,000 bp equal size.



Odds ratio = 11.5. Fisher’s exact test p-value < 10^{-16}.

		nsaPeaks		
		Inside	Outside	*Total*
CDK9 broad peaks	Inside	319,140	120,421	
	Outside	497,356	2,151,364	
	Total			3,088,281

Figure S5. Genome-wide view of CDK9 broad peaks (green) and nsaPeaks (red), related to Figure 3.

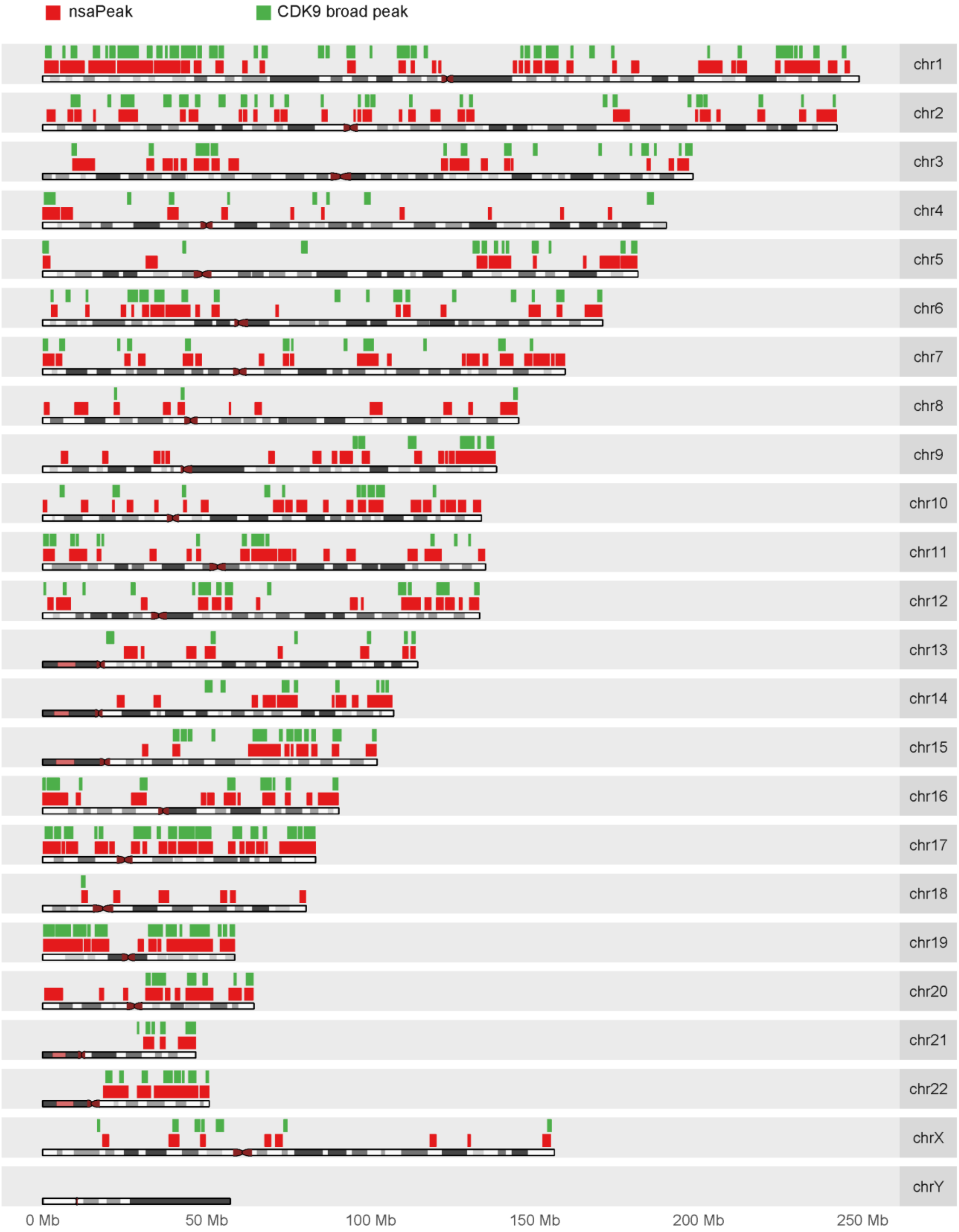


Figure S6. Distances between FISH spots and SC35 clusters, related to Figure 4. The center-to-center distance was calculated between each FISH spot and its nearest SC35 cluster. (A) The number of FISH spots (y axis) that have SC35 clusters at each designated distance (x axis) was plotted in each image (each dot) stained with the nsaPeak (black) or the non-nsaPeak probe (grey). There were 10 black dots and 12 grey dots in each column. (B) The average number of FISH spots that have SC35 clusters at each designated distance (x axis) in the 10 nsaPeak images (black) and 12 non-nasPeak images (grey). Error bar: 95% confidence interval.

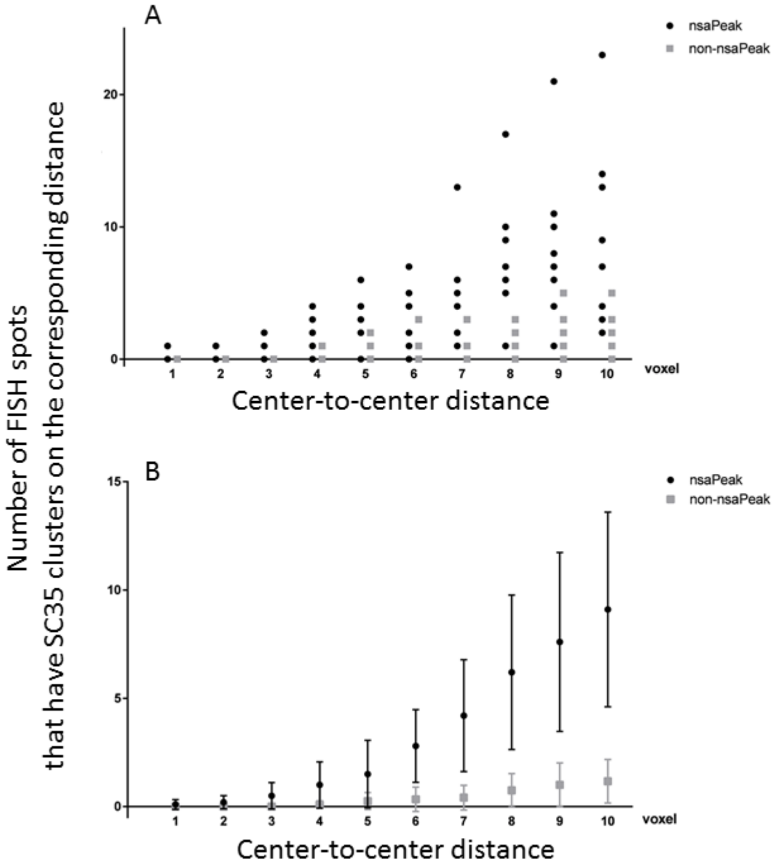
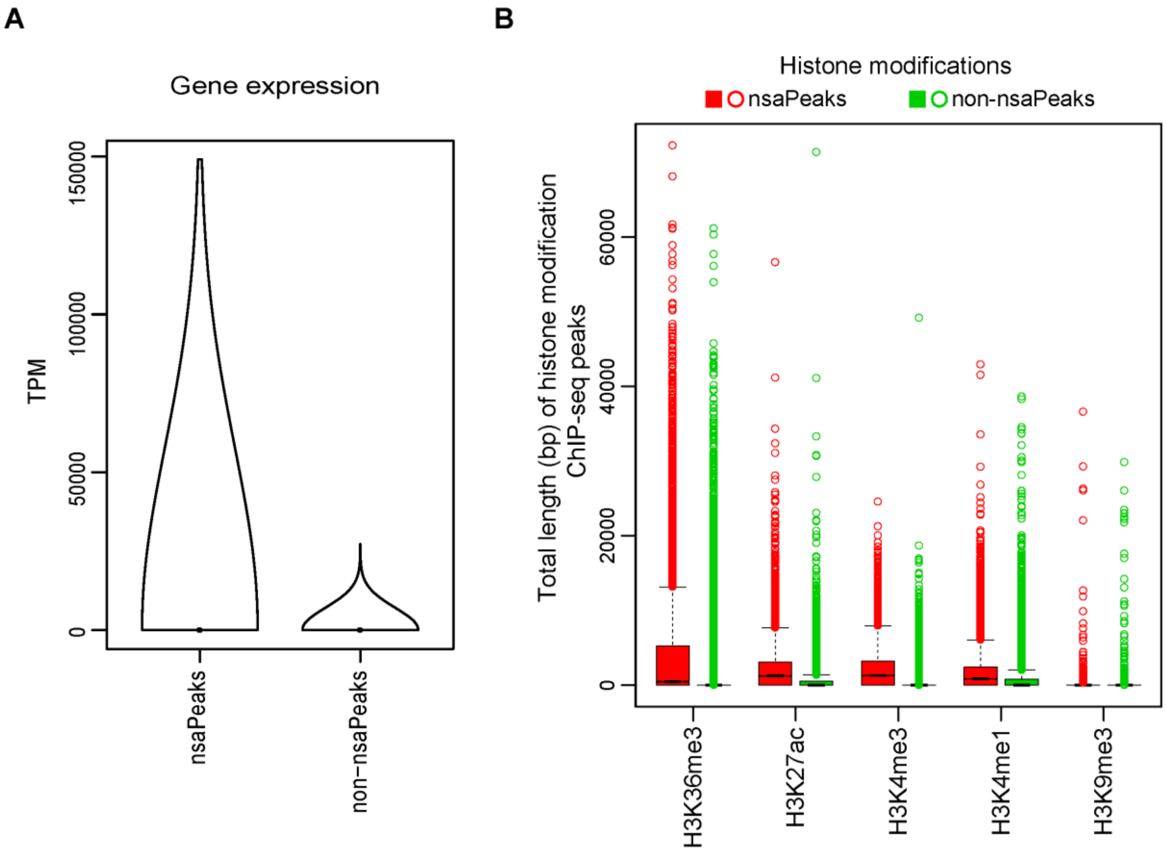


Figure S7. Gene expression (A) and histone modification levels (B) in nsaPeaks and the rest of the genome, related to Figure 5. (A) Violin plots of gene expression levels (y axis). A total of 21,566 and 15,386 genes were inside (nsaPeaks column) and outside of nsaPeaks (non-nsaPeak column), respectively. TPM: transcripts per million. (B) Distribution of the total length (bp) of histone modification ChIP-seq peaks in each 100,000bp genomic window (y axis), for the genomic windows inside (red) and outside nsaPeaks (green).



Supplementary tables

Table S1: Contingency table of RNA-DNA sequence pairs in HEK293T (A) and hES cells (B), related to Figure 1. Each pair is uniquely assigned to a cell based on its RNA-end sequence (columns) and DNA-end sequence (rows). rDNA: human ribosomal DNA complete repeating unit (Genbank ID: U13369.1), which is not assembled into the latest genome assembly (hg38). rRNA: transcripts originated from rRNA genes 5S, 5.8S, 28S, and 18S.

(A) HEK293T, odds ratio = 404, p-value $<10^{-16}$

RNA-end DNA-end	rRNA	Other RNA
rDNA	132,469	18,984
hg38 DNA	167,417	9,604,768

(B) hES, odds ratio = 1,809, p-value $<10^{-16}$

RNA-end DNA-end	rRNA	Other RNA
rDNA	809,763	29,366
hg38 DNA	71,531	4,681,715

Table S2. Contingency table between genomic windows covered by A compartment and by nsaPeaks, related to Figure 5. The genome (hg38) was split into 3,088,281 windows (with 1,000 bp equal size), among which 2750850 windows of either A or B compartment were taken for analysis. Odds ratio = 6.86. Fisher's exact test p-value < 10^{-16}.

	Inside nsaPeaks	*Outside nsaPeaks*	*Total*
A compartment	652,352	789,548	1,441,900
B compartment	140,711	1,168,239	1,308,950
Total	793,063	1,957,787	2,750,850

Transparent Methods

Datasets and accession numbers

Public datasets used in this work are MARGI data from HEK293T cells (GEO: GSM2427902 and GSM2427903) and H9 hES cells (GEO: GSM2427895 and GSM2427896) (Sridhar et al., 2017), CDK9 ChIP-seq (GEO: GSM1249897) (Liu et al., 2013) control ChIP-seq (GEO: GSM2423406) (Consortium, 2012) and Hi-C data from HEK293T cells (GEO: GSM1081530) (Zuin et al., 2014), RNA-seq (GEO: GSM2155552) (Ustianenko et al., 2016), H3K4me3 ChIP-seq (GEO: GSM945288, Encode: ENCFF001FJZ) and control ChIP-seq (GEO: GSM945256, Encode: ENCFF001HNC) from HEK293 cells (Thurman et al., 2012), H3K4me1 ChIP-seq (Encode: ENCFF002AAV) (Frietze et al., 2012), H3K9me3 ChIP-seq (Encode: ENCFF002AAZ) (Consortium, 2012), H3K27ac ChIP-seq (Encode: ENCFF002ABA) (Frietze et al., 2012), H3K36me3 ChIP-seq (Encode: ENCFF002ABD) (Consortium, 2012), control ChIP-seq (GEO:GSM935586, Encode: ENCFF000WXY) (Consortium, 2012) from HEK293 cells.

Mapping MARGI data

After removing RCR duplicates, the RNA-end and the DNA-end of a read pair were separately mapped to the genome (hg38) using STAR (Version 2.5.1b) (Dobin et al., 2013). Splice junction was allowed in mapping the RNA-end, by feeding the junction information (gtf file from ENSEMBL, hg38 release 84) to STAR. Splice junction was not allowed in mapping the DNA-end. Only the read pairs with both the RNA-end and the DNA-end uniquely mapped to the genome were used for downstream analysis.

Identifying rRNA-DNA read pairs

Human rRNA genes include 45S (18S, 5.8S and 28S) in rDNA (human ribosomal DNA complete repeating unit, GenBank: U13369.1) as well as 5S and 5.8S in the human genome assembly (hg38) (Stults et al., 2008). A MARGI read pair is categorized as an rRNA-DNA pair when the RNA-end is uniquely mapped to any human rRNA gene and the DNA-end is uniquely mapped to a combined “genome” of hg38 concatenated with rDNA.

Identifying nsaRNA-DNA read pairs

Human U1, U2, U4, U4atac, U5, U6, U6atac, U11, U12, 7SK, and Malat1 genes are considered nsaRNA genes. A MARGI read pair is categorized as an nsaRNA-DNA pair when the RNA-end is uniquely mapped to any human rRNA gene and the DNA-end is uniquely mapped to human genome (hg38). To minimize inclusion of nascent transcripts, the read pairs with the RNA-end and DNA-end mapped to within 2,000 bp in the genome are removed from further analysis.

Identifying pre-mRNA-DNA pairs

A MARGI read pair is categorized as a pre-mRNA-DNA pair when the RNA-end is uniquely mapped to an exon-intron junction with at least 10 bp overlap with the intron and the DNA-end is uniquely mapped to human genome (hg38). To minimize inclusion of nascent transcripts, the read pairs with the RNA-end and DNA-end mapped to within 2,000 bp in the genome are removed from further analysis.

Calling peaks and broad peaks

ChIP-seq and control ChIP-seq reads were mapped to human genome (hg38) and the uniquely mapped reads were fed to MACS2 (Zhang et al., 2008) to call peaks. CDK9 broad peaks, pre-mRNA broad peaks, and nsaPeaks were identified by the findPeaks module in Homer (v4.8.3) (Heinz et al., 2010). Any nsaPeak containing less than 9 MARGI reads was removed from further analysis.

Calling TADs and A/B compartments

HEK293 Hi-C data (GEO: GSM1081530) (Zuin et al., 2014) were aligned to hg38 retaining uniquely mapped reads. TADs were identified using a previously described HMM model (Dixon et al., 2012) automated in the GITAR software (Calandrelli et al., 2018). A/B compartments were called by the runHiCpca module in Homer (v4.8.3) (Heinz et al., 2010).

DNA FISH and immunofluorescence staining

The nsaPeak probe (RP11-772K10, covering chr11:64,663,168-64,947,112) with 5-ROX conjugate and the non-nsaPeak probe (RP11-908J16, covering chr11:80,767,575-80,980,051) with fluorescein conjugate were ordered from Empire Genomics. HEK293T cells were used through this study. In each experiment, cells were seeded on 18 X 18 mm glass coverslips with #1.5 thickness (#12-541A, Fisher Scientific) in 6-well tissue culture plate (Thermo Fisher Scientific) and grown in DMEM high-glucose media containing 10% (v/v) fetal bovine serum and 1% (v/v) penicillin-streptomycin at 37°C with 5% CO_2. Once reaching approximately 80% confluency, the cells were rinsed with PBS and fixed with 4% paraformaldehyde (PFA) in pH 7.2 phosphate-buffered saline (PBS) for 30 min at room temperature. PFA was discarded and residual PFA was quenched by incubation with 0.1 M Tris buffer (pH 7.4) at room temperature for 10 min followed with one wash with PBS. Cells were permeablized with PBS containing 0.1% saponin (#84510-100, Sigma) and 0.1% TritonX-100 for 10 min, then with 20% glycerol in PBS for 20 min at room temperature with gentle shaking. Cells were rapidly frozen in liquid nitrogen and thawed at room temperature for three cycles, and rinsed with PBS. To detect SC35, cells were blocked with 5% bovine serum albumin (BSA) in PBS with

0.1% TritonX-100 (PBST) at 37°C for 30 min, and incubated with mouse monoclonal anti-SC35 antibody (1:250) (#Ab11826, Abcam, RRID: AB_298608) in blocking buffer at 37°C for 1 hour. Cells were washed with PBST for 10 min for twice with gentle shaking, incubated with goat anti-mouse IgG antibody conjugated with Alexa647 (1:200 dilution) (#A21236, Invitrogen, RRID: 141725) in blocking buffer at 37°C for 30 min, and then washed again with PBST for 10 minutes twice while shaking. The cells were fixed again with 2% PFA at room temperature for 10 min, quenched with 0.1 M Tris buffer as previously described and washed with PBS for 5 min. Cells were incubated with 0.1 M HCl for 30 min at room temperature, followed by 1 hr incubation with 3% BSA and 100 μg/mL RnaseA (#EN0531, Thermo Fisher Scientific) in PBS at 37°C. Cells were permeablized again with 0.5% saponin and 0.5% TritonX-100 in PBS for 30 min at room temperature with gentle shaking, and rinsed with PBS. Cells were further denatured by incubation in 70% formamide with 2X saline-sodium citrate (SSC) buffer at 73°C for 2 min 30 sec and then incubation in 50% formamide with 2X SSC at 73°C for 1 min. For each coverslip, 1.2 μL of FISH probes were mixed with 4.8 μL formamide, incubated at 55°C for 15 min and mixed with 6 μL 2X hybridization buffer (8X SSC with 40% dextran sulfate) followed with denaturation at 75°C for at least 5 min until the cells were ready. 12 μL of FISH probe mixture was added onto a glass slide and quickly covered by freshly denatured coverslips with the cell side facing down. The coverslip was sealed with rubber cement and incubated in a humidified chamber at 37°C for 24 hr in the dark. The coverslips were collected the next day, washed twice with 50% formamide with 2X SSC for 15 min each at 37°C, three times with 2X SSC for 5 min each at 37°C, three times with 4X SSC containing 0.1% Tween 20 for 5 min each at room temperature, with gentle shaking, and rinsed with PBS. Cells were then stained with Hoescht 33342 (1:500 dilution) (#62249, Thermo Fisher Scientific) for 15 min followed with 5 min washing in PBS, mounted on slides with 80% glycerol in PBS and sealed with nail polish. Images in size of 1024 X 1024 were acquired on wide-field SIM DeltaVision Deconvolution Microscope using a 100X/1.40 oil objective (GE Healthcare Life Sciences) (pixel size = 0.66 μm). A series of z-stack images across the cells were acquired with thickness of 0.15 μm. Deconvolution was performed on these Z-stacks for subsequent image analysis.

Co-localization analysis

Deconvoluted images for each field of view contain a series of z stacks in three channels, DAPI, FISH and SC35. FISH spots were identified by performing particle analysis on the 2D maximal projection of z stacks of each field of view in the FISH channel with the threshold being set to the minimal value allowing only FISH spots to be recognized as “particles” in the size range of 10 – 500 pixels. These FISH particle regions were saved and applied to the z-stacks of FISH and SC35 channels and z-axis profiles of selected regions (min, max and mean values of fluorescence intensity) in both channels were recorded and examined. Positive co-localization of a given FISH spot with SC35 was defined by the presence of positive SC35 signals above background in any of the FISH-signal containing regions of that FISH spot. In order to determine one FISH region as positive SC35-colocalized, it needs to contain more than one stack with mean intensity above SC35 background, or contain more than half amount of stacks with max intensity over SC35 background. For each analyzed image, SC35 background value was based on the average mean intensities in areas outside of SC35 clusters within the nucleus region. For each field of view, the SC35 co-localization rate represents the ratio of the amount of SC35-colocalized FISH spots over the amount of total FISH spots.

Center-to-center distances were calculated as follows. After deconvolution, each cluster or spot was identified as a connected 3D region such that all voxels within this region are above a threshold. The threshold was determined as described previously (Raj et al., 2008). Briefly, each deconvoluted image was scanned to identify all pixels on every stack that was could not possibly be background. The threshold was chosen such that the number of detected fluorescent clusters would not change within 3% variation of this threshold. The center of a cluster (spot) was calculated as the gravity center. Center-to-center distance was calculated with voxel as the unit.

Supplemental references

CALANDRELLI, R., WU, Q., GUAN, J. & ZHONG, S. 2018. GITAR: An open source tool for analysis and visualization of Hi-C data. *bioRxiv,* https://doi.org/10.1101/259515.

CONSORTIUM, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S. & REN, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature,* 485**,** 376-80.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

FRIETZE, S., WANG, R., YAO, L., TAK, Y. G., YE, Z., GADDIS, M., WITT, H., FARNHAM, P. J. & JIN, V. X. 2012. Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol,* 13**,** R52.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**,** 576-89.

LIU, W., MA, Q., WONG, K., LI, W., OHGI, K., ZHANG, J., AGGARWAL, A. & ROSENFELD, M. G. 2013. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell,* 155**,** 1581-1595.

RAJ, A., VAN DEN BOGAARD, P., RIFKIN, S. A., VAN OUDENAARDEN, A. & TYAGI, S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods,* 5**,** 877-9.

SRIDHAR, B., RIVAS-ASTROZA, M., NGUYEN, T. C., CHEN, W., YAN, Z., CAO, X., HEBERT, L. & ZHONG, S. 2017. Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Curr Biol,* 27**,** 610-612.

STULTS, D. M., KILLEN, M. W., PIERCE, H. H. & PIERCE, A. J. 2008. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res,* 18**,** 13-8.

THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAUGEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B. K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L., VONG, S., WEAVER, M., YAN, Y., ZHANG, Z., ZHANG, Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E. & STAMATOYANNOPOULOS, J. A. 2012. The accessible chromatin landscape of the human genome. *Nature,* 489**,** 75-82.

USTIANENKO, D., PASULKA, J., FEKETOVA, Z., BEDNARIK, L., ZIGACKOVA, D., FORTOVA, A., ZAVOLAN, M. & VANACOVA, S. 2016. TUT-DIS3L2 is a mammalian surveillance pathway for aberrant structured non-coding RNAs. *EMBO J,* 35**,** 2179-2191.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

ZUIN, J., DIXON, J. R., VAN DER REIJDEN, M. I., YE, Z., KOLOVOS, P., BROUWER, R. W., VAN DE CORPUT, M. P., VAN DE WERKEN, H. J., KNOCH, T. A., VAN, I. W. F., GROSVELD, F. G., REN, B. & WENDT, K. S. 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A,* 111**,** 996-1001.