Exploring the Roles of RNAs in Chromatin Architecture Using Deep Learning

Shuzhen Kuang[1], Katherine S. Pollard[1,2,3*]

[1]Gladstone Institute of Data Science and Biotechnology, San Francisco, CA

[2]Department of Epidemiology & Biostatistics, University of California, San Francisco, CA

[3]Chan Zuckerberg Biohub, San Francisco, CA, USA

* Corresponding author. Email: katherine.pollard@gladstone.ucsf.edu

Abstract

Recent studies have highlighted the impact of both transcription and transcripts on 3D genome organization, particularly its dynamics. Here, we propose a deep learning framework, called AkitaR, that leverages both genome sequences and genome-wide RNA-DNA interactions to investigate the roles of chromatin-associated RNAs (caRNAs) on genome folding in HFFc6 cells. In order to disentangle the *cis*- and *trans*-regulatory roles of caRNAs, we compared models with nascent transcripts, *trans*-located caRNAs, open chromatin data, or DNA sequence alone. Both nascent transcripts and *trans*-located caRNAs improved the models' predictions, especially at cell-type-specific genomic regions. Analyses of feature importance scores revealed the contribution of caRNAs at TAD boundaries, chromatin loops and nuclear sub-structures such as nuclear speckles and nucleoli to the models' predictions. Furthermore, we identified non-coding RNAs (ncRNAs) known to regulate chromatin structures, such as MALAT1 and NEAT1, as well as several novel RNAs, RNY5, RPPH1, POLG-DT and THBS1-IT, that might modulate chromatin architecture through *trans*-interactions in HFFc6. Our modeling also suggests that transcripts from Alus and other repetitive elements may facilitate chromatin interactions through *trans* R-loop formation. Our findings provide new insights and generate testable hypotheses about the roles of caRNAs in shaping chromatin organization.

Introduction

The human genome is folded into complex structures within the nucleus with multiple levels of organization, including compartments, topologically associated domains (TADs) and chromatin loops[1,2]. This spatial organization is dynamic and varies across cell types and tissues, and it is interconnected with cellular processes such as gene transcription and DNA replication[3–5]. Recent studies have unraveled the critical roles of CTCF and cohesin in three-dimensional (3D) genome organization, including their involvement in TAD and loop formation via the loop extrusion mechanism[6–8]. Other proteins, such as YY1 and ZNF143, are potentially also regulating chromatin organization[9–12]. However, all these structural proteins are widely expressed, and alone cannot explain the dynamic and cell-type specific aspects of chromatin organization.

A growing number of studies point to transcription as a potential contributor to the dynamic aspects of genome folding[13–17]. While 3D chromatin structures are known to play a role in gene silencing and activation, the process of transcription can in turn affect 3D genome folding in a cell-type- or tissue-specific manner[13,18,19]. For example, TAD boundaries are often located near or at active gene promoters[3]. Furthermore, transcribing RNA polymerases (RNAPs) are reported to act as moving barriers for the loop-extruding cohesins[13]. Thus, some chromatin dynamics are expected to reflect a *cis* effect of nascent transcription.

Transcribed RNA molecules may also contribute to chromatin dynamics. Specifically, RNAs known as chromatin-associated RNAs (caRNAs) have been observed to directly interact with DNA or to bind chromatin-associated proteins[14,16,20]. These caRNAs include nascent RNAs, long non-coding RNAs (lncRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), enhancer RNAs (eRNAs) and repeat RNAs[15–17,21,22]. Most caRNAs bind close to their locus of origin (*cis*-interactions), but many interact with distant genomic loci (*trans*-interactions). Several of the latter *trans*-located caRNAs have been implicated in chromatin regulation. For example,

lncRNA HOTTIP promotes distal TAD formation by forming RNA-DNA hybrid structures (R-loop) in leukemia[23]. Enhancer RNAs have been proposed to mediate promoter-enhancer interactions by forming *trans* R-loops at Alu sequences[24]. Several other *trans*-located caRNAs, such as lncRNAs MALAT1, NEAT1, and Firre, also play critical roles in chromatin organization[25–28].

CaRNAs, particularly non-coding RNAs (ncRNAs), are proposed to shape 3D genome structure via multiple mechanisms[14–17]. First, caRNAs can recruit chromatin regulatory proteins to specific genomic loci. For example, caRNAs have been found to directly bind CTCF and serve as locus-specific factors to recruit CTCF to TAD boundaries and loop anchors[23,29–33]. Perturbing the abundance of RNAs or mutating the RNA-binding region of CTCF weakens the insulation of TAD boundaries or disrupts the formation of chromatin loops[23,29,30,33]. Second, caRNAs can act as scaffolds to organize chromosomal architecture by integrating multiple regulatory proteins. A well-known example is the lncRNA Xist, which initiates and maintains X chromosome inactivation by interacting with proteins[15]. Third, caRNAs can drive phase separation and coordinate the formation of various membrane-less nuclear bodies[16,17]. For example, the lncRNA NEAT1 induces the assembly of paraspeckles via phase separation and is indispensable for this nuclear structure[25,34].

Given the relatively small number of functionally characterized *trans*-located caRNAs in genome folding, we hypothesized that other examples remain to be discovered. To explore this hypothesis, we used machine learning and bioinformatics tools to interrogate RNA-DNA interaction data. Several high-throughput approaches have been developed to globally profile caRNAs, including chromatin-associated RNA sequencing (ChAR-seq)[35], global RNA interaction with DNA sequencing (GRID-seq)[36], RNA & DNA split-pool recognition of interactions by tag extension (RD-SPRITE)[27], and *in-situ* mapping of RNA-genome interaction (iMARGI)[20,37]. These techniques enable genome-scale investigations of the mechanisms through which dynamically expressed caRNAs contribute to nuclear organization.

Modeling 3D genome folding using machine learning offers an efficient way to study chromatin dynamics that complements experimental strategies. Recently, deep learning models, such as Akita[38], DeepC[39] and ORCA[40], have been developed to predict 3D genome structure from DNA sequence. Since these models are highly accurate, they enable researchers to decode sequence determinants of genome folding through computational techniques such as *in silico* mutagenesis and feature importance scores[41]. More recently, models incorporating epigenomic data were built to achieve cell-type-specific predictions[42,43]. Significantly, these models learned the sequence and epigenetic correlates of 3D genome folding. The capability of deep learning models to probe sequence and epigenetic dependencies of genome folding motivated us to use this approach to explore the roles of caRNAs in 3D genome architecture.

We thus extended the Akita model to predict cell-type-specific chromatin contact frequencies using not only DNA sequence but also RNA-DNA interaction data. We call the resulting modeling framework AkitaR. To advance our understanding of the *cis*- and *trans*-regulatory roles of caRNAs in chromatin architecture, we designed AkitaR to use either nascent RNA or *trans*-located caRNA. Comparisons of these models to each other and to models trained on sequence or open chromatin data allowed us to dissect how each of these relate to chromatin interaction frequencies genome-wide. We showed that AkitaR achieved significantly better predictions on regions of the human genome with cell-type-specific genome folding. Particularly, some chromatin interactions were uniquely captured by the model with *trans*-located caRNAs. Examination of the feature importance scores showed not only the general contribution of caRNAs at CTCF peaks, TAD boundaries and loop anchors but also revealed slightly different contributions of different types of caRNAs at nuclear structures, such as snoRNAs in nucleoli and nuclear speckles. This enabled us to develop testable hypotheses about the roles of specific types of caRNAs in genome folding.

Results

In order to characterize the roles of caRNAs in 3D genome folding, we downloaded the genome-wide RNA-chromatin interactions in human foreskin fibroblast cells (HFFc6) and human embryonic stem cells (H1ESC) captured by iMARGI and the corresponding genome-wide DNA-DNA interactions captured by Micro-C from 4DN data portal (https://data.4dnucleome.org/) (Supplementary Table 1)[20,44,45]. We chose iMARGI data over other techniques that map genome-wide RNA-DNA contacts, because iMARGI has been performed in human cell lines that have rich transcriptomic and epigenomic datasets we could use to interpret the high-quality chromatin interaction data. We used HFFc6 for our primary analyses, and leveraged H1ESC for identifying cell-type differences.

To disentangle the roles of nascent transcription versus *trans*-located caRNAs, both of which are involved in 3D genome organization, we broke down the human genome into 2,048-bp bins and defined nascent transcripts as all the RNAs transcribed from a given 2,048-bp DNA bin and *trans*-located caRNAs as all RNAs transcribed from at least 1 Mb away from the bin. We opted for 1 Mb to identify *trans*-located caRNAs instead of the 100 or 200 Kb used in previous studies[46,47] in order to align with the window size of our predictive models and also to remove self-interactions for the vast majority of genes (~99.9%). As different types of caRNAs may engage in different *trans*-interactions (Fig. 1a) and contribute to different chromatin features, we further classified *trans*-located caRNAs into eight groups: snRNAs, snoRNAs, other small RNAs, lncRNAs, misc_RNAs, RNAs from protein coding genes, RNAs from other types of genes and RNAs from regions without known gene annotation.

CaRNAs preferentially locate at open chromatin and many interactions occur in *trans*

To explore how caRNAs are spatially localized inside the nucleus, we examined whether caRNAs identified by iMARGI preferentially interact with any parts of the genome. Similar to DNA-DNA

interactions, we observed that RNAs tend to interact with DNA regions that share the same spatial or functional annotations as the loci from which they are transcribed, such as being in the same compartment or having the same Spatial Position Inference of the Nuclear genome (SPIN) state[48] (Zhang et al. in preparation) or chromatin state identified by chromHMM[49,50] (Fig. 1b and Supplementary Fig. 1). Beyond that, caRNAs interact more frequently with DNA regions with high versus low transcriptional activity (Fig. 1b and Supplementary Fig. 1). This trend is confirmed by the enrichment of caRNAs at open chromatin regions (Fig. 1c). Interestingly, the enrichment was also observed for *trans*-located caRNAs (Fig. 1c), and the amount of *trans*-located caRNA attached to DNA regions was positively correlated with the region's chromatin accessibility (Pearson's R =0.37).

Considering that many RNA-DNA interactions across spatial or functional annotations may be from *trans*-interactions, we assessed the percentage of RNA-DNA interactions occurring *in trans* in HFFc6 both globally and for each annotated gene. We included caRNAs transcribed from the DNA loci on the same chromosome but at least 1 Mb away plus those encoded on different chromosomes. CaRNAs primarily interacted with proximal DNA regions (Fig. 1d), and of all the interactions on the same chromosome, over 90% spanned a distance of less than 1Kb (Fig. 1e). Nevertheless, 38.38% of RNA-DNA interactions occurred *in trans*, including 6.14% within the same chromosome and 32.24% on different chromosomes (Fig. 1d). These results are quite different from DNA-DNA interactions from Hi-C data, where *trans*-interactions on the same chromosome are much more frequent than across chromosomes. This difference suggests that the proximity of most caRNAs to chromatin is not due to their being transcribed from DNA that is closeby in the 3D nucleus.

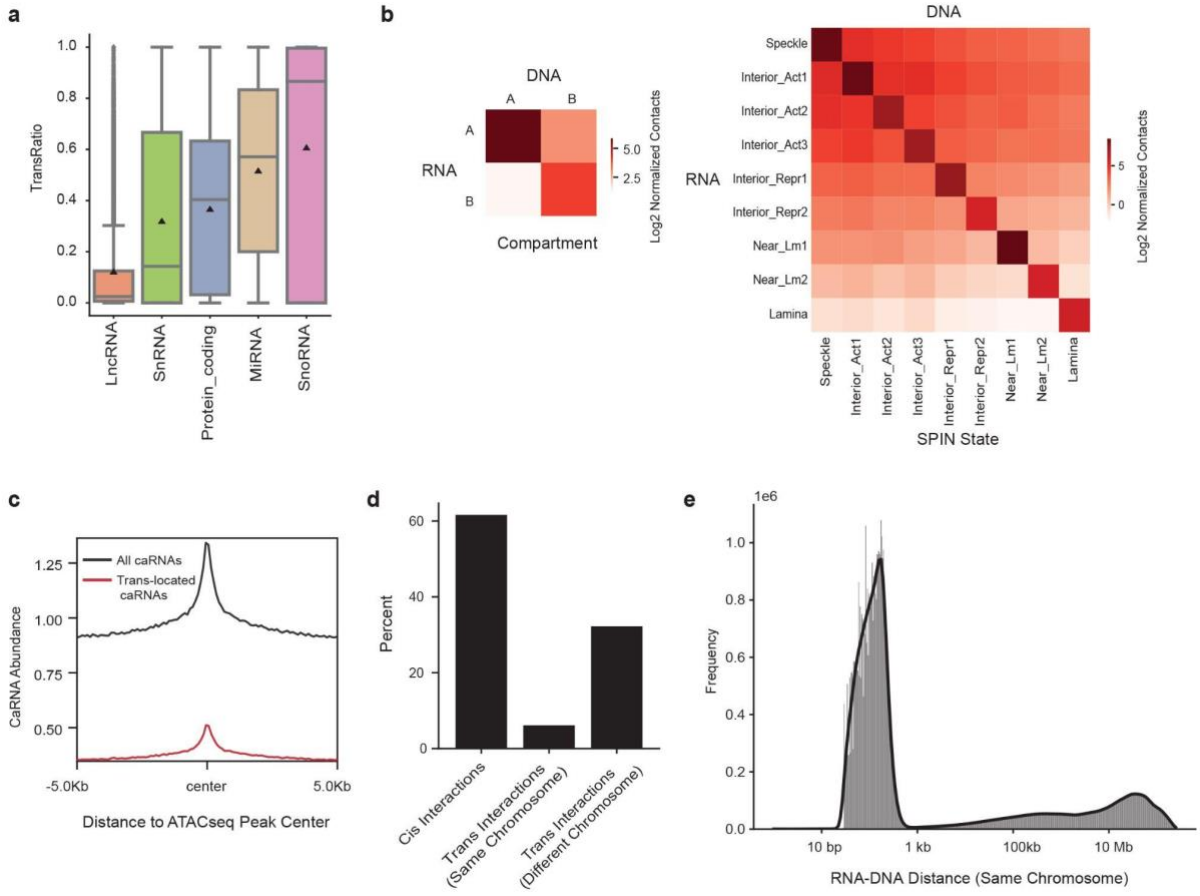

Figure 1. Chromatin-associated RNAs preferentially bind open chromatin and many interactions occur in *trans.* (a) The proportion of *trans*-interactions for RNAs transcribed from each gene that is annotated as lncRNA, snRNA, protein-coding genes, miRNA or snoRNA. (b) The number of RNA-DNA interactions (log2) within and across compartments (left panel) and SPIN states (right panel). The interaction frequencies were normalized to the size of compartments and SPIN states. (c) The abundance of all caRNAs or *trans*-located caRNAs at ATAC-seq peaks and their flanking regions. (d) The percentage of genome-wide *cis*-interactions (RNA-DNA distance 1Mb or less) and *trans*-interactions (RNA-DNA distance > 1Mb on the same chromosome or RNA and DNA on different chromosomes). (e) Histogram of RNA-DNA interaction frequencies as a function of genomic distance between DNA and RNA loci on the same chromosome. Interior_Act 1: Interior Active 1, Interior_Act 2: Interior Active 2, Interior_Act 3: Interior Active 3, Interior_Repr1: Interior Repressive 1, Interior_Repr2: Interior Repressive 2, Near_Lm1: Near Lamina 1, Near_Lm2: Near Lamina 2

Notably, we observed that the majority of the small ncRNAs and a number of lncRNAs and RNAs from protein-coding genes were engaged in *trans*-interactions (Fig. 1a and Supplementary Fig. 2). Given the well-established importance of several snRNAs and snoRNAs in nuclear structures, these results suggest that other ncRNAs and transcripts of some protein-coding genes may also regulate chromatin structures.

Trans-located caRNAs are particularly enriched at TAD boundaries

To investigate whether caRNAs play roles at particular landmarks within the 3D genome, we first used the iMARGI data to examine their abundance at TAD boundaries. Since most TAD boundaries are located in compartment A in HFFc6 (Fig. 2a) and tend to have higher chromatin accessibility compared to surrounding regions (Fig. 2b), we hypothesized that caRNAs would be enriched at TAD boundaries. In order to check whether nascent transcripts and *trans*-located RNAs follow similar patterns, we conducted separate analyses for each of them. As anticipated, *trans*-located caRNAs peaked at TAD boundaries and greatly decreased in flanking regions (± 50 kb) (Fig. 2b). After categorizing TAD boundaries in HFFc6 and H1ESC based on their strength and cell type specificity (see Methods, Fig. 2b), we found that HFFc6 *trans*-located caRNAs are significantly less prevalent at TAD boundaries unique to H1ESC or with higher insulation strength in H1ESC than at TAD boundaries shared with or more prominent in HFFc6. Similar but weaker patterns were also observed for open chromatin signals (ATAC-seq) (Fig. 2b). These results suggest the potential involvement of *trans*-located caRNAs at TAD boundaries and their contribution to TAD dynamics across cell types. Additionally, strong TAD boundaries exhibited significantly higher ATAC-seq and *trans*-located caRNA signals than did weak boundaries, and the association between boundary strength and *trans*-located caRNA abundance held after normalizing to the corresponding ATAC-seq signals (Fig. 2c). These results further indicate that the accumulation of *trans*-located caRNAs at TAD boundaries is not solely driven by DNA accessibility.

Unlike *trans*-located caRNAs that peaked at all HFFc6 TAD boundaries, nascent transcripts in HFFc6 mostly accumulated at TAD boundaries unique to HFFc6 (Fig. 2b). They also tended to be more frequent at strong TAD boundaries than weak ones (Fig. 2c). Overall, these results indicate that nascent transcripts could also contribute to the formation of TAD boundaries,

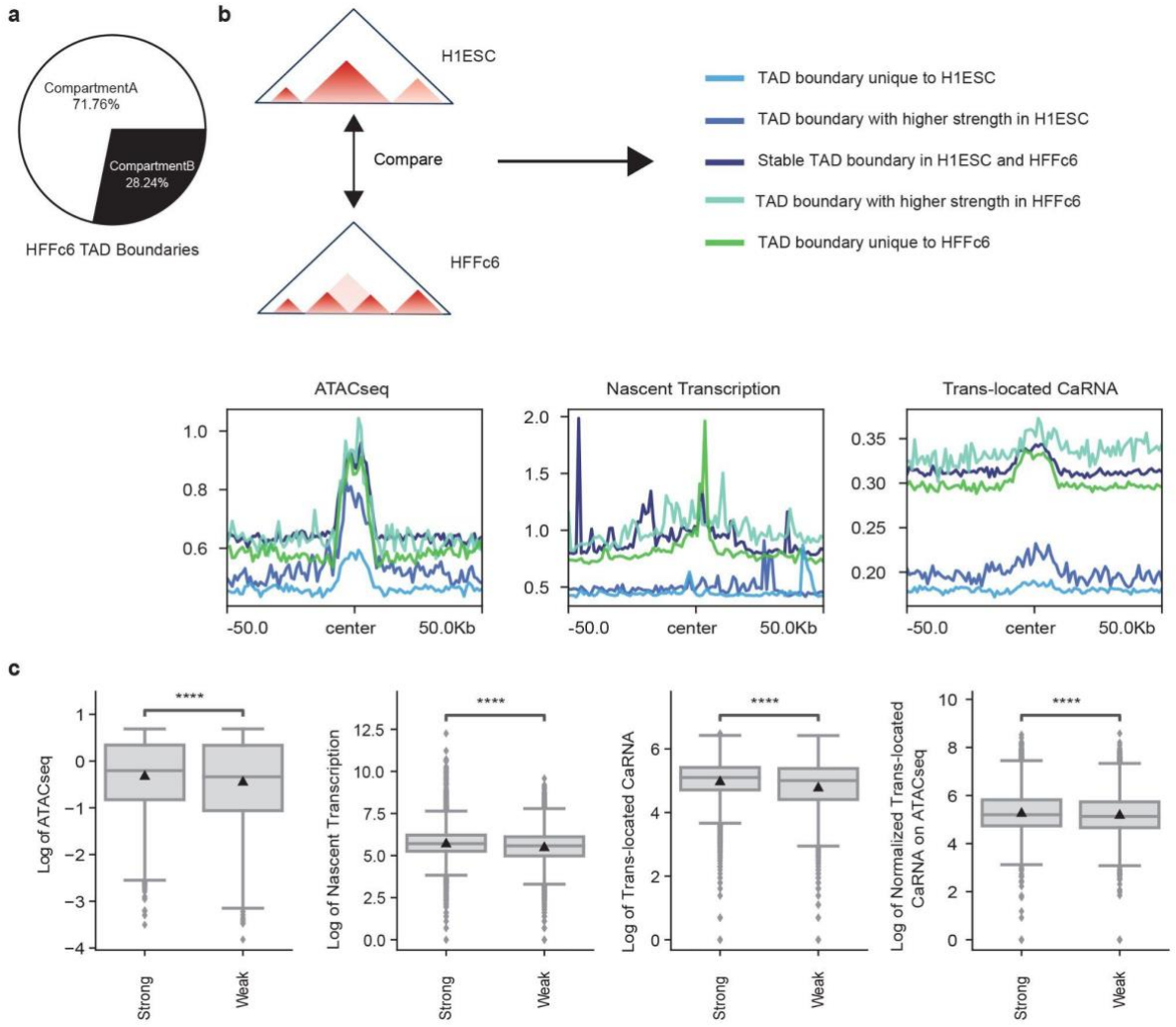

Figure 2. *Trans*-located caRNAs are particularly enriched at TAD boundaries. (a) The percentage of HFFc6 TAD boundaries located in A and B compartments. (b) Chromatin accessibility (ATAC-seq, left panel), and the abundance of nascent transcripts (middle panel) and *trans*-located caRNAs (right panel) at TAD boundaries ("center") and their flanking regions in HFFc6. (c) Chromatin accessibility, the abundance of nascent transcripts and *trans*-located caRNAs, and the abundance of *trans*-located caRNAs normalized to chromatin accessibility, at strong versus weak TAD boundaries. ****: p-value < 0.0001

particularly cell-type-specific ones, aligning with the enrichment of TAD boundaries at active promoters and the barrier function of RNAPs[3,13].

CaRNAs increase the accuracy of 3D genome folding predictions

To learn how caRNAs contribute to 3D genome organization beyond TAD boundaries and in an unbiased way, we developed a deep learning framework called AkitaR. The models we

implemented extend Akita[38] to predict chromatin interaction maps by incorporating both DNA sequence and RNA features extracted from nascent transcripts or *trans*-located RNAs (Supplementary Fig. 3). Similar to the original Akita, AkitaR uses 1D convolution neural networks to learn representations from ~1 Mb DNA sequence segments. The learned representations at the resolution of 2,048 bp were subsequently concatenated with the RNA features, and dilated convolution neural networks were used to learn long-range dependencies. Lastly, 1D representations were averaged to 2D and further processed by dilated 2D convolutional neural networks to predict the ~1 Mb x 1 Mb contact matrices at 2,048 bp resolution (Fig. 3a). These were normalized to observed-over-expected contact frequencies and log transformed to generate the model outputs (see Methods).

We also designed additional models as controls or for comparison with the iMARGI based models (Supplementary Table 1 and Supplementary Fig. 3). For instance, since caRNAs are enriched in open chromatin, one of these models combined DNA sequence with features from chromatin accessibility (ATAC-seq) or ATAC-seq plus *trans*-located caRNAs. To disentangle the expression level of RNAs from their DNA contact frequencies and from nascent transcription, we incorporated steady-state transcription (RNA-seq). A control model with randomized signals from a standardized normal distribution was also built to alleviate the possibility that the improved performance was solely due to more features as input. Natural log transformations were applied on the RNA or open chromatin features before model fitting. Besides the quantitative input features, we also tried sparse binary features, but these compromised the model's performance.

We found that all models with additional informative features achieved better predictions than the model with DNA sequence alone as input (Fig. 3b and c, Supplementary Fig. 4-6). This is consistent with results from models that incorporate epigenetic features such as CTCF binding or histone modifications[42]. Of the three RNA features, *trans*-located caRNA signals led to the

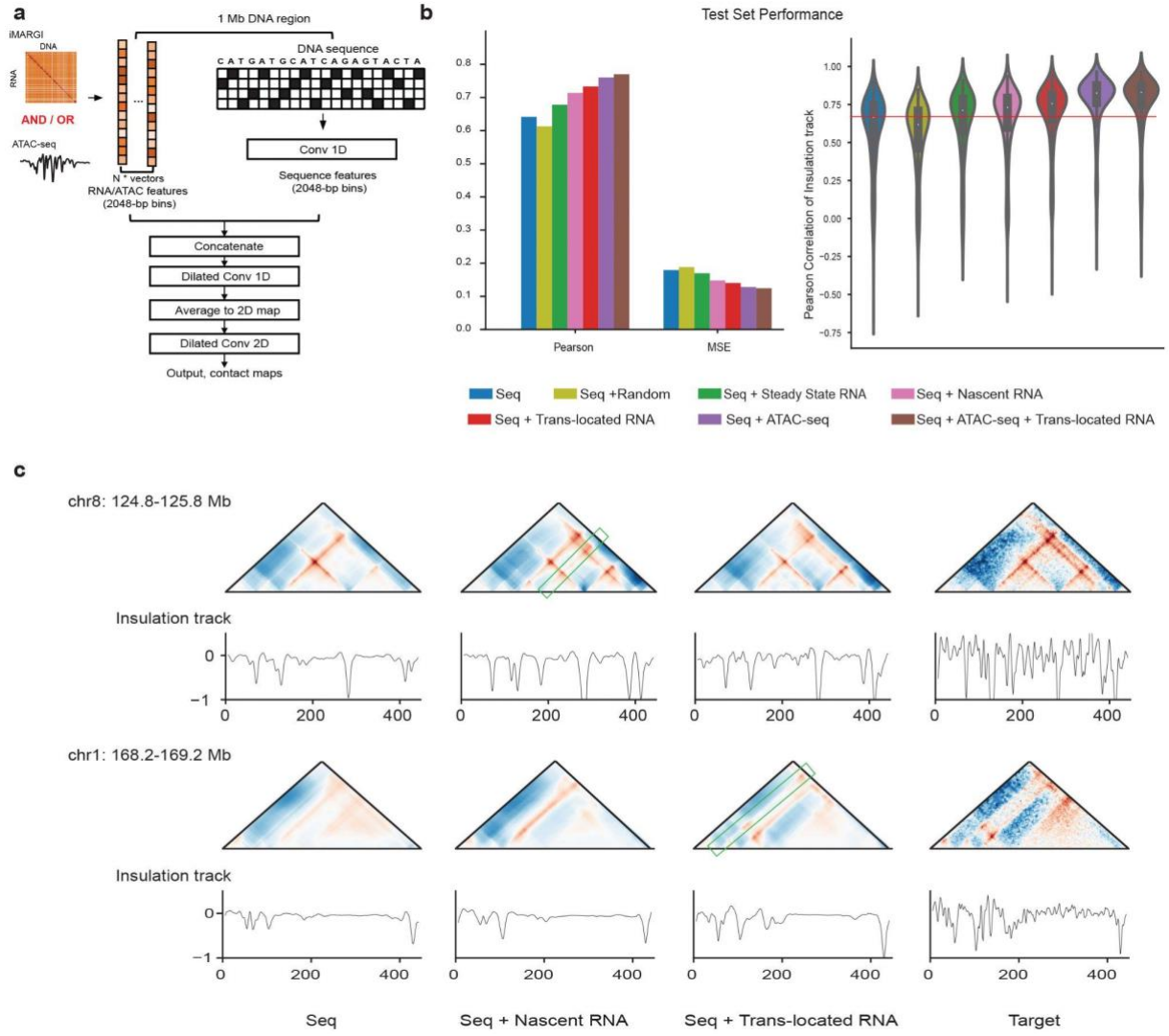

Figure 3. Chromatin-associated RNAs contribute to accurate prediction of 3D genome folding. (a) The architecture of the models in the AkitaR framework. (b) Barplots of Pearson's correlation and MSE (left panel) and violin plot of Pearson's correlation of insulation tracks (right panel) between experimental and predicted contact maps of the held-out test set. (c) Examples showing better prediction of contact maps with nascent transcripts (top panel) or *trans*-located RNAs (bottom panel). The 3D genome contacts with better prediction are highlighted with green rectangles.

AkitaR model with the highest performance, closely followed by nascent RNA, and then steady-state transcription (Fig. 3b, Supplementary Fig. 4 and 6). On the other hand, at some regions, nascent RNA signals contributed to more accurate predictions than *trans*-located RNA inputs did (Fig. 3c, Supplementary Fig. 5). These results suggest that all the RNA features carry useful information about 3D genome folding, particularly *trans*-located caRNAs, though nascent transcription is more helpful at some loci. Adding chromatin accessibility signals yielded better

performance than adding RNA features did (Fig. 3b, Supplementary Fig. 4). However, adding *trans*-located caRNA plus chromatin accessibility signals achieved even better performance than chromatin accessibility signals alone (Fig. 3b, Supplementary Fig. 4), suggesting that RNA-DNA interactions provide additional information beyond marking open chromatin. In support of this hypothesis, we found that incorporating *trans*-located caRNAs into the models increased the correlation between predicted and observed insulation signals at TAD boundaries (Fig. 3b). Thus, deep learning clearly highlights the information that RNA-DNA interactions carry about chromatin organization.

CaRNAs are helpful for predicting cell-type-specific genome folding

Since RNAs, particularly ncRNAs, are often expressed in cell-type-specific ways[51], we hypothesized that the performance boost provided by incorporating RNA features into the AkitaR models would be most notable in regions with cell-type-specific genome folding. To evaluate this hypothesis, we first identified test regions that showed the largest differences in chromatin organization between H1ESC and HFFc6 based on MSE (34 regions) or MSE plus stratum-adjusted correlation coefficient (SCC) and structural similarity index measure (SSIM) (109 regions; see Methods) (Supplementary Fig. 7). We then evaluated the performance of our models in these cell-type-specific regions, and found that they showed a notably larger performance gap between models with additional features and the model with DNA sequence alone as compared to the ensemble of all test regions (Fig. 3b, Fig. 4a and b, Supplementary Fig. 8 and 9). This finding demonstrates the capability of the AkitaR models to capture dynamic chromatin organization. Since genome compartmentalization correlates with RNA-chromatin interaction[20], we further evaluated model performance on cell-type-specific regions with compartment changes. Interestingly, the model with *trans*-located caRNA signals achieved similar or even better performance than the model with chromatin accessibility signals on cell-type-specific regions with a compartment transition from B in H1ESC to the more active A compartment in HFFc6 (Fig. 4a

and b, Supplementary Fig. 8 and 9). This suggests the potential association of *trans*-located caRNAs with compartment transitions. Furthermore, by visually checking the cell-type-specific regions with better predictions from the models incorporating *trans*-located caRNAs, we observed that *trans*-located caRNAs helped capture some cell-type-specific chromatin interactions better than all other RNA and ATAC-seq features (Fig. 4c, Supplementary Fig. 10), prompting us to explore where these interactions mapped and what caRNAs they involved.

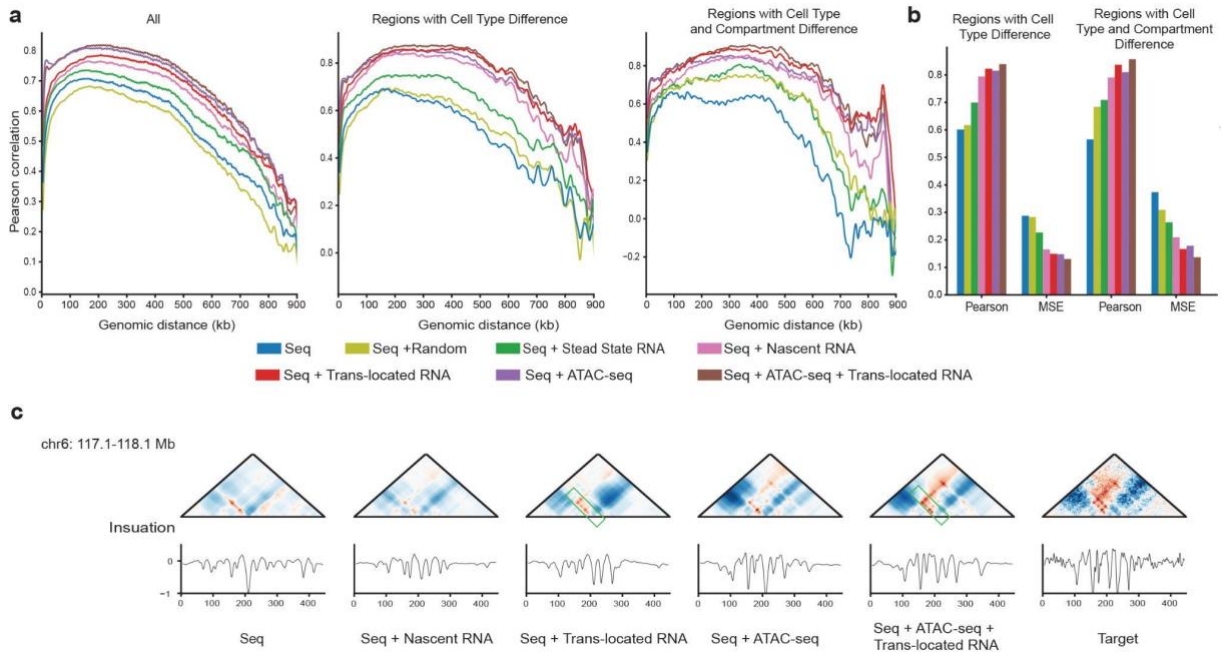

Figure 4. Chromatin-associated RNAs help predict cell-type-specific genome folding. (a) Stratified Pearson's correlation between experimental and predicted contact maps on the held-out test set, cell-type-specific subsets of test regions identified by MSE (MSE>0.3) and cell-type-specific subsets (MSE>0.3) with compartment changes from B compartment in H1ESC to A compartment in HFFc6. (b) Barplots of Pearson's correlation and MSE between experimental and predicted contact maps on the cell-type-specific subsets (MSE>0.3) and cell-type-specific subsets (MSE>0.3) with compartment changes from B compartment in H1ESC to A compartment in HFFc6. (c) An example showing the contribution of *trans*-located RNAs to the prediction of some chromatin interactions. The regions with better prediction are highlighted with green rectangles.

Trans-located caRNAs are associated with TAD boundaries, loop anchors and nuclear structures

To identify the caRNAs that contribute to 3D genome organization and the DNA regions to which they associate, we used DeepExplainer[52,53], which allowed us to quantify the importance of each RNA and ATAC-seq feature to the contact map predictions. DeepExplainer generates a score for each feature at each 2,048 bp bin (see Methods). A negative score indicates that contact frequency decreases when the feature increases at that bin (e.g., caRNA is associated with loss of a chromatin loop or increased insulation at a TAD boundary); a positive score indicates that contact frequency rises when the feature increases. We observed that the distributions of contribution scores for the multiple RNA types were asymmetric, with slightly elongated left tails (Supplementary Fig. 11), hinting that RNA-DNA interactions may be more linked to insulation than to enhancing chromatin interactions. Though the absolute contribution scores of caRNA features showed moderate positive correlation with caRNA signals, many genomic bins with high caRNA signals received low contribution scores, indicating that our models learned where the caRNAs might contribute to genome folding (Supplementary Fig. 12). *Trans*-located caRNAs, which we already showed are enriched at TAD boundaries (Fig. 2b and c), tended to have higher absolute contribution scores at TAD boundaries than at their flanking regions (Fig. 5a). In contrast, the contribution scores of nascent transcripts were less elevated at TAD boundaries and remained high in flanking regions, as we might expect when there is active transcription within TADs.

To test the hypothesis that AkitaR has learned a relationship between caRNAs and insulation, we performed a simulation: we generated 1,000 random sequences of ~1 Mb and introduced TAD boundaries at randomly selected loci by inserting convergent CTCF motifs. Progressively adding 1 to 4 CTCF motifs to increase the insulation strength led the sequence model to predict a decrease in average contact frequency (Supplementary Fig. 13), validating the negative correlation between insulation strength and average contact frequency. Next, we repeated this

computational experiment using AkitaR by inserting *trans*-located caRNAs with large negative scores rather than CTCF motifs. We observed a similar average decrease in contact frequency and increase in insulation strength (Supplementary Fig. 13). These simulation results suggest that AkitaR has inferred a potential causal role of *trans*-located caRNAs in strengthening TAD boundaries.

To further characterize the regions where caRNAs might shape chromatin organization, we ranked the contribution scores for each feature type and identified the genomic regions with scores in the top (positive contribution scores) or bottom (negative contribution scores) 1%. We found that the regions with bottom snRNA scores were preferentially located in nuclear speckles, aligning with their well-established roles in pre-mRNA splicing within nuclear speckles (Fig. 5b)[54,55]. Additionally, we observed that regions with top snoRNA scores were enriched in loci annotated as Interior_Repr2 (Interior Repressive 2) by SPIN (Fig. 5b), which was putatively associated with nucleoli[48] where snoRNAs function[56]. These two expected associations validate the capability of AkitaR to capture the functional roles of caRNAs.

Beyond these cases, we observed that genomic regions with bottom scores across RNA types were enriched at active chromatin, CTCF peaks, active promoters, enhancers, TAD boundaries and loop anchors (Fig. 5b). LncRNAs, snoRNAs and RNAs from unknown genes were particularly enriched at CTCF peaks, stable TAD boundaries between H1ESC and HFFc6, and shared TAD boundaries with higher insulation in HFFc6 (Fig. 5b), suggesting that these RNAs contribute to TAD boundaries, potentially by recruiting or stabilizing CTCF, in active chromatin. Interestingly, snoRNAs showed enrichment in nuclear speckles, consistent with the increasing evidence of the regulatory roles of some box C/D snoRNAs in alternative splicing[57–59]. On the other hand, regions with top scores were predominantly found in heterochromatin, particularly near lamina or at lamina associated regions (Fig. 5b), indicating that caRNAs play different roles at active and repressed chromatin, potentially via different mechanisms. CaRNAs from Protein coding genes, however,

showed very different patterns from all other caRNAs, with bottom rather than top scoring bins being enriched in compartment B and in particular at TAD boundaries with elevated insulation in HFFc6. Extending this analysis to the top and bottom 5% regions produced similar patterns, indicating that the enrichments are robust to the contribution score threshold (Supplementary Fig. 14).

In contrast to these nuanced patterns that differ across caRNA types and between active versus repressed chromatin, ATAC-seq features were significantly enriched in active chromatin, regardless of whether they had top or bottom scores (Supplementary Fig. 15). These findings suggest that AkitaR captures differences between caRNA-DNA interactions and chromatin accessibility, motivating us to explore specific *trans*-located caRNAs. To further disentangle the independent effects of caRNAs beyond their association with open chromatin, we identified specific DNA regions where *trans*-located caRNAs have high absolute contribution scores and ATAC-seq features do not (absolute normalized contribution score > 0.25, |fold change| >5) (Supplementary Fig. 16). These regions showed similar chromatin and SPIN state enrichments as the top and bottom scoring regions more generally (Supplementary Fig. 16), confirming the contribution of *trans*-located caRNAs to chromatin features beyond chromatin accessibility.

CaRNAs may promote TAD and loop formation at Alu sequences via *trans* R-loops

To identify the caRNAs with the largest contributions to AkitaR's chromatin map predictions, we ranked them based on their association with DNA regions that have high absolute contribution scores (top or bottom 5% for each RNA type; Supplementary Table 2). We observed that the top 10 RNAs were all highly prevalent in HFFc6 (Supplementary Fig. 17 and Supplementary Table 2). These included multiple ncRNAs previously known to play roles in chromatin structures, such as lncRNAs MALAT1 and NEAT1 and snRNAs RNU2-2P, RNU12 and RN7SK [16,25,28,34,60,61]

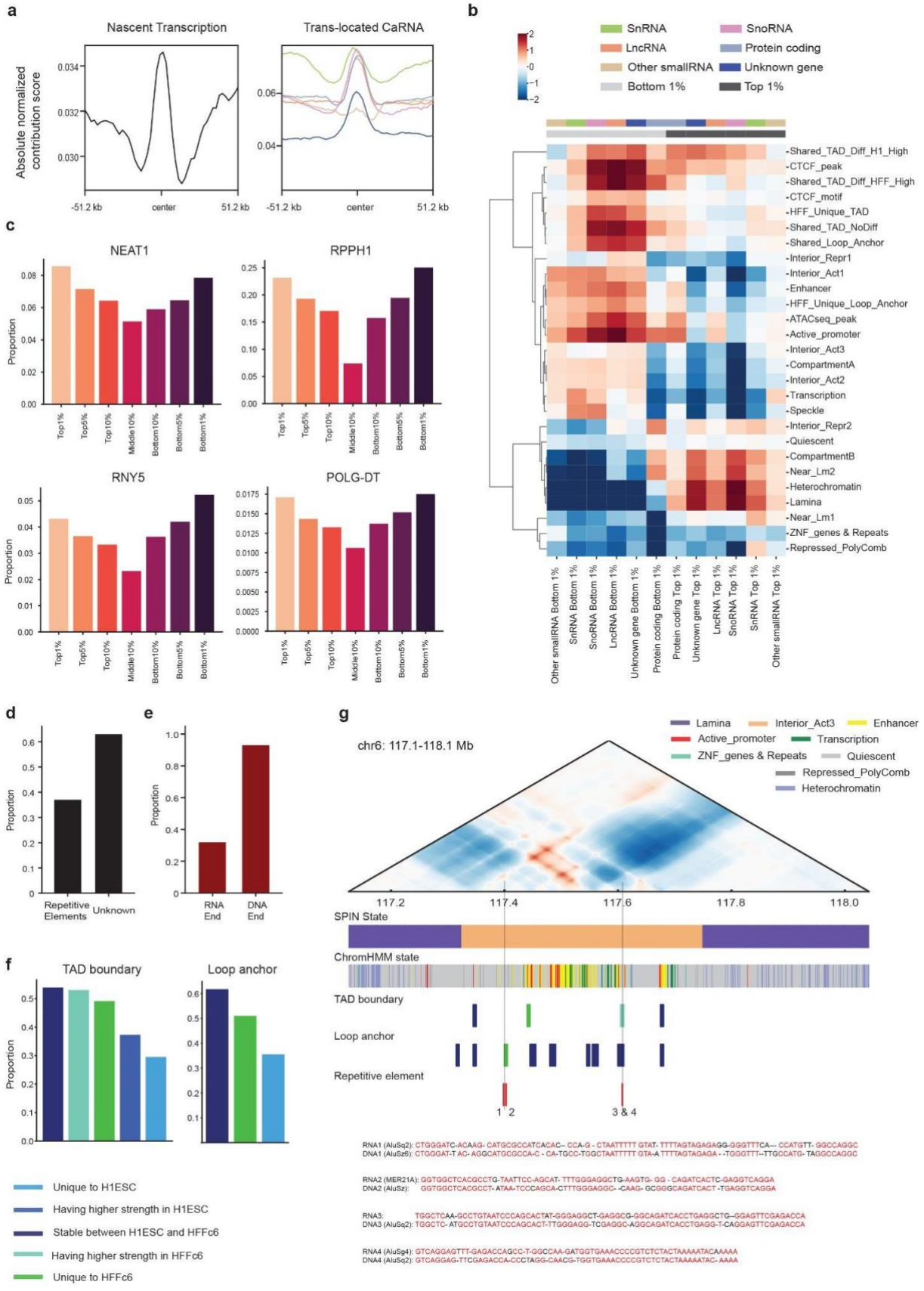

Figure 5. Chromatin-associated RNAs are associated with TAD boundaries, loop anchors and nuclear structures. (a) The absolute contribution scores of nascent transcripts (left panel) and *trans*-located caRNAs (right panel) at TAD boundaries and their flanking regions. (b) Heatmap showing enrichment of genomic regions with top 1% (positive) and bottom 1% (negative) contribution scores of each type of caRNA across TAD boundaries, loop anchors, SPIN and chromHMM states. (c) Example of four RNAs that preferentially interact with genomic regions with high absolute contribution scores (top 1%, top 5%, top 10%, bottom 1%, bottom 5%, bottom 10%) rather than with regions with lower absolute contribution scores (middle 10%). (d) The proportion of RNAs transcribed from repetitive elements for the interactions between DNA and RNAs derived from unknown genes. (e) The proportion of the RNA sequences in the candidate interactions that might form R-loops originated from Alu sequences and the proportion of the corresponding DNA sequences were annotated as Alu elements. (f) The proportion of TAD boundaries (left panel) and loop anchors (right panel) having RNA-DNA interactions that could form R-loops. (g) An example locus showing candidate R-loops at Alu elements illustrating the contribution of *trans*-located RNAs to the prediction of chromatin interactions. The Alu elements that may form R-loops with RNAs at the loop anchors of a chromatin interaction that was specifically predicted by the model with *trans*-located caRNAs are shown in the track "Repetitive elements". The candidate R-loops are numbered as 1, 2, 3 and 4 at the associated Alu elements. The best local alignment (with gaps) of the RNA and DNA sequences of the candidate R-loops (the complementary DNA sequences to RNAs are not shown) is shown. The nucleotides matching between RNA and DNA sequences are highlighted in red.

(Fig. 5c, Supplementary Fig. 18 and Supplementary Table 2). Interestingly, all top 10 snoRNAs are C/D box snoRNAs, including SNORD47, SNORD79 and SNORD27 (Supplementary Fig. 18). Beyond these, many novel RNAs stood out, such as lncRNAs RNY5, RPPH1, POLG-DT and differentially expressed lncRNAs between H1ESC and HFFc6, such as THBS1-IT1 and ENSG00000260772. In addition, these lncRNAs were preferentially associated with regions with high absolute contribution scores compared to the regions with low scores (Fig. 5c, Supplementary Fig. 18). Since the pattern of enrichment of these caRNAs mirrors that of MALAT1 and NEAT1, we hypothesize that these caRNAs also play mechanistic roles in 3D genome organization.

To explore the caRNAs that might shape genome structure over chromatin accessibility, we investigated the caRNAs (top10) that were preferentially associated with genomic regions having higher absolute *trans*-located caRNA contribution compared to chromatin accessibility. We identified a list of RNAs that was nearly identical to the one identified genome-wide (Supplementary Table 2). However, some RNAs were found to preferentially interact with these differentiated regions but were not enriched in top or bottom scoring regions overall. These

included the ncRNAs ZNRF3-AS1, MIR6726 and MIR4796, which not only showed higher interaction ratios with the differentiated regions but also interacted with more than one of these DNA regions (Supplementary Table 3 and Supplementary Fig. 19). These caRNAs are high-confidence candidates for contributing to nuclear structures in specific ways beyond being generally associated with accessible chromatin.

Since genomic regions with bottom contribution scores from RNAs of unknown genes were enriched at TAD boundaries and loop anchors, we further explored the interactions between DNA and RNAs derived from unknown genes. We found that around 37% of these RNAs were transcribed from repetitive elements (Fig. 5d). Since Alu sequences were proposed to promote long-range enhancer-promoter interactions, possibly through R-loops[24,62], we aligned the sequences of each pair of DNA-RNA *trans* interactions in HFFc6 using pairwise2 local alignment in search of potential candidates for R-loop formation. Around 0.3% of *trans* interactions were considered as candidates by exhibiting over 80% identity between RNA and DNA sequences plus continuous, uninterrupted perfect matches exceeding 10 base pairs. We found that 32% of the RNA sequences in these candidate interactions originated from Alu sequences, and 93% of the DNA sequences were annotated as Alu elements (Fig. 5e). Furthermore, these candidate interactions tended to increase at stable TAD boundaries, TAD boundaries having higher insulation strength in HFFc6 or unique to HFFc6 in contrast to TAD boundaries with higher strength in H1ESC or unique to H1ESC (Fig. 5f). The same trend was also observed for loop anchors (Fig. 5f), aligning with the roles of Alu sequences in long-range enhancer-promoter interactions. More importantly, both loop anchors of the cell-type-specific interaction that was captured by the *trans*-located caRNA model but not other models in Fig. 4c could form *trans* R-loops at Alu sequence loci (Fig. 5g). This provides a potential mechanism for loop formation at loci with *trans* Alu RNA-DNA interactions, demonstrating the capability of the AkitaR model to capture these interactions and generate testable, mechanistic hypotheses.

Discussion

In this study, we proposed deep learning models that leverage both DNA sequence and the distribution of caRNAs across the genome to predict chromatin interaction maps. Both nascent transcripts and *trans*-located RNAs contributed to these AkitaR models being able to make more accurate predictions than with sequence alone, especially in regions of the genome with different folding between cell types. The models also learned the importance of caRNAs at chromatin features, such as CTCF peaks, TAD boundaries and loop anchors. Moreover, we identified several novel RNAs that might be involved in the regulation of chromatin organization in HFFc6, including RNY5, RPPH1, POLG-DT and THBS1-IT. Validating these observations, AkitaR highly prioritized the lncRNAs MALAT1 and NEAT1, which have known roles in chromatin structures, while also detecting the importance of snRNAs at regions located in nuclear speckles and of snoRNAs in nucleoli.

Since *trans*-located RNAs tended to be enriched at open chromatin regions and the model with chromatin accessibility signals achieved better performance than the one with *trans*-located RNA signals, it might be argued that *trans*-located RNAs diffused randomly and that their enrichment in these regions solely reflected the accessibility of chromatin. Our results suggest that *trans*-located RNAs play roles in genome folding on top of being randomly diffused to distant DNA regions. Firstly, *trans*-located caRNA signals were found to be higher at strong TAD boundaries compared to weak ones even after being normalized on ATAC-seq signals. Secondly, stable TAD boundaries, TAD boundaries with higher insulation strength in HFFc6 or the ones unique to HFFc6 also tended to have RNA-DNA interactions that might form *trans*-acting R-loops. Moreover, the model with both chromatin accessibility and *trans*-located caRNAs also slightly outperformed the model with only the chromatin accessibility. Additionally, the model with *trans*-located caRNA signals achieved better performance on some subsets of the test regions compared to the one with chromatin accessibility signals. Particularly, we observed that some

local chromatin features were only accurately predicted by the models with *trans*-located caRNA signals as input. Their loop anchors might be mediated by *trans*-acting R-loops formed at Alu sequences. Lastly, genomic regions with high absolute contribution scores from different RNA types also showed enrichment in chromatin and SPIN states both in active and repressed chromatin in contrast to the enrichment of genomic regions with high absolute ATAC-seq contribution scores only in active chromatin. On the other hand, the chromatin being open could also have been the result of the binding of *trans*-located caRNAs.

Although most of the eight RNA types increased the accuracy of the chromatin map predictions, each type was associated with somewhat distinct chromatin structures. For example, snoRNAs, lncRNAs and RNAs from unknown genes showed more enrichment at CTCF peaks, shared TAD boundaries and loop anchors than did snRNAs, whereas lncRNAs were more enriched at promoters and enhancers compared to other RNA types. Moreover, besides snRNAs, we observed the enrichment of snoRNAs in nuclear speckles. While it is well-established that snoRNAs have vital functions within the nucleolus, growing evidence suggests that some snoRNAs, including SNORD27, SNORD88C, and SNORD115, may exert regulatory influence over the alternative splicing of pre-mRNAs that originate from distantly located genomic loci[57–59]. Finally, we noticed that regions with positive contribution from RNAs of different types, particularly RNAs from unknown genes, showed enrichment at heterochromatin and lamina or near lamina regions. This is consistent with the recent evidence that ncRNAs, especially repetitive ncRNAs, play roles in anchoring specific genomic loci to nuclear lamina or recruiting H3K9me3-related methyltransferases to promote heterochromatin[22,63,64].

The high performance of our AkitaR models allowed us to explore the contribution of caRNAs in genome organization in an unbiased and effective way. Leveraging feature importance scores or high-throughput *in silico* screening, we could efficiently prioritize candidate genomic loci that are dependent on caRNAs for accurate genome folding and develop hypotheses for functional

characterization with additional analyses. These hypotheses could be further validated with experimental techniques, such as genome engineering, RNA inhibition or RNA overexpression, in the context of 3D genome folding. We anticipate that this strategy of integrating deep learning models with bioinformatics analyses will drive the generation of novel hypotheses and accelerate wet lab discoveries.

While AkitaR offers us an effective way to unravel the roles of *trans*-located caRNAs in genome folding, our approach has several limitations. First, the genome-wide RNA-chromatin interaction data that we used to extract the *trans*-located RNA features were limited to several cell types, making it difficult to generalize our models and analyses to a wide range of cellular contexts. Secondly, as *trans*-located caRNA signals might somewhat reflect the accessibility of chromatin, models may face challenges in distinguishing which regions *trans*-located caRNAs play a driver role and which regions they act as passengers. Lastly, many of the RNAs might only function *in trans* at limited regions, and our analyses based on genome-wide signals might not be able to capture the contribution of these RNAs.

In summary, we investigated the roles of caRNAs, particularly *trans*-located caRNAs, in regulating 3D genome folding by genome-wide analyses and deep learning models. We showed the contribution of both nascent transcripts and *trans*-located caRNAs to genome organization. These analyses provide new insights and generate testable hypotheses about the roles of caRNAs in chromatin organization.

Methods

Micro-C data and processing

High-quality Micro-C datasets mapped to hg38 in .paris format for HFFc6 and H1ESC were downloaded from the 4DN data portal (https://data.4dnucleome.org/)[44,45] and processed into 2,048-bp (2^{11} bp) bins, followed by normalization, interpolation and smoothing, as previously

described[38]. These data and their paired genome were further divided into training, validation and test examples, each of which was a ~1Mb (2^{20} bp) region.

Annotations of compartment and TAD boundaries for the Micro-C datasets identified by cooltools at the resolution of 25 Kb and 5 Kb, respectively, were also downloaded from the 4DN data portal[65]. TAD boundaries with insulation strength between 0.2 and 0.5 were considered as weak boundaries and the ones with strength larger than 0.5 were defined as strong boundaries. The TAD boundaries in a cell type that were within 20 Kb of the TAD boundaries from the other cell type were defined as shared TAD boundaries, otherwise they were considered as cell type unique TAD boundaries. The $\log_2$ fold change of insulation strength for the shared TAD boundaries were further calculated and used to classify them into stable TAD boundaries between H1ESC and HFFc6 with no insulation difference ($|log2(HFFc6/H1ESC)| <= 1$), shared TAD boundaries with higher insulation strength in H1ESC ($log2(HFFc6/H1ESC) < -1$) and shared TAD boundaries with higher insulation strength in HFFc6 ($log2(HFFc6/H1ESC) > 1$).

Chromatin loops at 5 Kb and 10 Kb resolution for the Micro-C datasets were identified using HiCCUPS[7]. Similar to TAD boundaries, the loop anchors were classified as shared loop anchors and cell type unique loop anchors with distance limit of 20 Kb.

iMARGI data and processing

iMARGI data in .pairs format for HFFc6 and H1ESC on hg38 were obtained from the 4DN data portal and converted into contact matrices at the resolution of 10-bp (for preliminary analyses), 2,048-bp (for model inputs), and 5,000-bp (for analyses at TAD boundaries and loop anchors) after removing low-quality mappings (MAPQ <= 30)[20]. Nascent transcription was estimated as the number of reads with their RNA ends mapped to each bin (10-bp/2,048-bp/5,000-bp) in the contact matrices (log value for model input). In order to get the signals of *trans*-located RNAs at each bin for the *trans*-located caRNA model, the interactions between RNAs and DNAs within ~1

Mb (2**20 bp) linear distances were filtered out. The self interactions between genes that are longer than ~1Mb (2^{20} bp) were also removed. Considering the potentially distinct roles of different RNA types, we annotated the RNA ends of the contact matrices with comprehensive genes from GENCODE (v43). We noticed that many snoRNA genes annotated in Refseq were missed in GENCODE but showed high expression in iMARGI data. We thus incorporated the annotations of snoRNAs from Refseq into GENCODE. We then classified the bins in the RNA end into eight groups based on their overlap with the transcription sites of different types of RNAs, which are snRNAs, snoRNAs, other small RNAs, lncRNAs, miscellaneous RNAs, RNAs from protein-coding genes, RNAs from all other types of genes and RNAs from regions without known gene annotations. The total number of reads from all RNAs in each RNA group with their DNA ends mapped to a bin was calculated as the *trans*-located caRNA signal of that bin from the RNA group. Log transformation was performed for model input.

Besides gene annotations, the RNA and DNA end of the iMARGI interactions were annotated for repetitive elements with data from the RepeatMasker database for downstream analyses[66].

RNA-seq and ATAC-seq data

RNA-seq and ATAC-seq data in .bigWig format for HFFc6 were downloaded from the 4DN data portal, respectively[45,67,68]. Log values of the normalized signals of each 2,048-bp bin on the library size of iMARGI data were extracted from the data to get the input for the model with steady-state transcription level or the model with chromatin accessibility. The signals of ATAC-seq at 5,000-bp bins were also calculated for the analyses at TAD level.

Model architecture, training and evaluation

AkitaR was extended from Akita to predict 3D genome folding by using both DNA sequence and RNA / ATAC-seq signals. We kept the “Head” of Akita and adjusted the “Trunk” architecture by concatenating the above RNA / ATAC-seq features of length 512 to the vector representations of

DNA sequence. The DNA representation was the output of 11 convolution blocks, each of which included convolution, batch normalization and max-pooling layers. Keeping hyperparameters the same as Akita, the model was trained to maximize Pearson's correlation coefficient between experimental maps and predictions. We chose to optimize on Pearson's correlation coefficient over mean squared error (MSE) because we noticed that pixel-wise MSE tends to be very sensitive to noise[69], and the models trained using a loss function based on Pearson's correlation achieved slightly better performance than the models trained on MSE in most cases in a preliminary evaluation.

Model performance was evaluated on the test dataset using MSE, SCC, Pearson's and Spearman's correlations. To examine the capability of the models to capture cell-type-specific regions, two subsets of test regions that had different contact maps between H1ESC and HFFc6 were selected based on MSE, SCC and SSIM. One cell-type-specific subset included the regions with high MSE (>0.3) between H1ESC and HFFc6 experimental maps. The other one consisted of not only regions with high MSE (>0.3), but also those with low SCC (<0.2) or SSIM (<0.08). Here, SCC is the weighted sum of Pearson's correlation for each stratum and shares the similar range as Pearson's correlation coefficients[70]. Since both the predicted and experimental contact map used in this study were normalized against distance dependent decay, SCC is highly consistent with Pearson's correlation. SSIM is a widely used metric in imaging studies that qualifies the similarity between two images[71]. To further evaluate whether RNAs were associated with the compartment changes between cell types, the cell-type-specific subsets were further divided into the ones without compartment change, the ones that switched to compartment B in HFFc6 from compartment A in H1ESC and also the ones that changed to compartment A in HFFc6 from compartment B in H1ESC.

Insulation scores

Insulation profiles of experimental and predicted contact maps were identified by sliding along each diagonal bin of the contact matrix using a diamond-shaped window and calculating the average contact frequency within the window[72]. The bins at the end of the diagonal were ignored for calculation.

Trans-located Ratio of RNAs

iMARGI data in .pairs format was first converted into .bedpe format. The total number of reads with RNA end mapped to each gene was calculated as its nascent transcription. Then the read pairs with their DNA and RNA end within ~1 Mb (2^{20} bp) linear distances were removed and the resulting reads mapped to each gene was regarded as its *trans*-located abundance. The ratio of the *trans*-located abundance to its nascent transcription was calculated as the *trans*-located ratio of each RNA gene. To better distinguish the roles of host genes and the genes within them, the reads mapped to the genes within them were subtracted from the host genes.

Signals at TAD boundaries and flanking regions

Nascent transcription and *trans*-located caRNA signals at the resolution of 10-bp were first converted into bigWig format using bedGraphToBigWig[73]. Then the signals of ATAC-seq, nascent transcription and *trans*-located caRNAs at TAD boundaries and their flanking regions were calculated using deepTools computeMatrix from the bigWig files and plotted using deepTools plotHeatmap[74].

CTCF ChIP-seq and binding sites

CTCF ChIP-seq peaks and signals were downloaded from ENCODE data portal[75]. The genome-wide CTCF sites identified by FIMO using all three CTCF PWMs in JASPAR database with p-value less than 1e-5 were downloaded from the R resources AnnotationHub[76].

ChromHMM state and SPIN state

ChromHMM were employed to annotate the chromatin regions of H1ESC and HFFc6 using six epigenomes (H3K27ac, H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3), which were downloaded from the ENCODE data portal[75]. We defined 18 chromatin states using the default parameters and annotated them by checking their enrichment for gene related regions (transcription start site (TSS), transcription end site, gene body, exon and intron), repetitive elements and epigenetic peaks. The ones annotated as active TSSs, TSS flanking regions or bivalent promoters were extracted as active promoters and the ones annotated as active enhancers, genic enhancers, weak enhancers and bivalent enhancers were obtained as enhancer regions and used for downstream analyses. Annotations of SPIN states were unpublished data that were obtained from Jian Ma's lab[48] (Zhang et al. in preparation).

Contribution scores

DeepExplainer (DeepSHAP implementation of DeepLIFT)[52,53] was employed to compute the contribution scores of the RNA and ATAC-seq features. For the examples in the validation and test dataset, randomly selected 20 examples from the training dataset were used as background. For the training dataset, we divided it into two subsets. For the first half, randomly selected examples from the second half acted as background, and vice versa. The contribution scores for each feature were normalized by dividing into their maximum absolute values.

The genomic regions with their contribution scores located within the top (positive) and bottom (negative) 1% and 5% for each feature were extracted for enrichment analyses. Specifically, their enrichment at CTCF sites, active promoters, enhancers, other chromHMM states, TAD boundaries, loop anchors and SPIN states were measured by calculating their odds ratio against all DNA bins. To avoid the bias caused by the bins with positive signals, only the ones with positive input values were used in the enrichment analyses.

Differential analyses of RNA contribution scores

The normalized scores of *trans*-located signals of each RNA type at each DNA bin were compared to the scores of ATAC-seq signals. The DNA bins with fold change greater than 5 and the absolute value of normalized contribution score larger than 0.25 were considered as the ones with differential contributions between *trans*-located caRNAs and ATAC-seq signals.

Candidate RNA identification

A hypergeometric test was employed to evaluate whether RNA-DNA interactions occur more often than expected by random chance. The test assumes that each DNA bin has an equal probability to interact with any RNA in a random manner and each interaction is independent. The interactions with FDR <= 0.05 were extracted as high-confidence interactions. These high-confidence interactions were then used to identify RNAs that preferentially interacted with selected DNA bins.

Simulations of increasing TAD insulations

One thousand random DNA sequences of 2^{20}bp were first generated using the SimDNA python package (https://github.com/kundajelab/simdna). TAD structures were then introduced by symmetrically inserting forward and reverse CTCF motifs at randomly selected loci between 0.15 and 0.85 of each DNA sequence. Following that, two different simulations were performed. First, one to four convergent CTCF motifs were progressively added to TAD boundaries with distance from previously inserted CTCF motifs at 500 bp and the contact maps of the DNA sequences were generated using the sequence alone model. Second, the randomly generated *trans*-located caRNA inputs at TAD boundaries were replaced by caRNA inputs with top 5% input values and bottom 5% contribution scores (except input from protein-coding genes) and predictions were made with the model incorporating *trans*-located caRNA signals.

Pairwise alignment of DNA and RNA sequences

To search for potential candidates of R-loop formation, we extracted the sequences of each pair of DNA-RNA *trans* interactions in HFFc6 and then aligned them using the pairwise2 sequence alignment module in the Bioptyhon package (local alignment)[77]. The ones with over 80% of RNA sequence matching to DNA sequences and continuous perfect matches exceeding 10 bp were considered as candidates.

Statistical analysis

Two-sided Mann-Whitney U tests were used to compare strong versus weak TAD boundaries and to compare simulation scenarios with different numbers of inserted sequences or RNA features. Hypergeometric tests were employed to identify high-confidence, statistically significant RNA-DNA interactions.

Data Availability

Publicly available data used in this study can be found at: 4D Nucleome Data Portal (https://data.4dnucleome.org/) with accession numbers (1) Micro-C for HFFc6 and H1ESC: 4DNESWST3UBH, 4DNES21D8SP8, (2) iMARGI for HFFc6 and H1ESC: 4DNES9Y1GHK4, 4DNESNOJ7HY7, (3) ATAC-seq for HFFc6: 4DNESMBA9T3L, (4) RNA-seq for HFFc6: 4DNESFH3EHTU; ENCODE data portal (www.encodeproject.org/) with accession numbers (1) ChIP-seq data, H3K27ac for HFFc6 and H1ESC: ENCSR510VXV, ENCSR880SUY, (2) H3K4me1 for HFFc6 and H1ESC: ENCSR340XKM, ENCSR000ANA, (3) H3K4me3 for HFFc6 and H1ESC: ENCSR639PCR, ENCSR000AMG, (4) H3K9me3 for HFFc6 and H1ESC: ENCSR938NXC, ENCSR000APZ, (5) H3K27me3 for HFFc6 and H1ESC: ENCSR129TUY, ENCSR186OBR, (6) H3K36me3 for HFFc6 and H1ESC: ENCSR519CMW, ENCSR000ANB, CTCF for HFFc6: ENCSR163ULN; R resources AnnotationHub for CTCF binding sites

(https://github.com/mdozmorov/CTCF); GENCODE (https://www.gencodegenes.org/human/), NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/, RefSeq); UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/, RepeatMasker). Pretrained Akita models, model input, contribution scores of RNA / ATAC-seq features as well as target map and test set predictions are available at Zenodo (https://zenodo.org/records/10015010). Other data used to generate the figures are available in the CaRNAs_in_Chromatin_Architecture github repository (https://github.com/shuzhenkuang/CaRNAs_in_Chromatin_Architecture).

Code Availability

Custom code for data exploration and downstream analyses and a jupyter notebook for figure generation are available at https://github.com/shuzhenkuang/CaRNAs_in_Chromatin_Architecture. The code of AkitaR, which was modified from Akita, is available upon request.

Funding

This work was supported by the NIH 4D Nucleome Project (grant #U01HL157989) and Gladstone Institutes. K.S.P. is an investigator of the Chan Zuckerberg Biohub San Francisco.

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
2. Rowley, M. J. & Corces, V. G. Organizational Principles of 3D Genome Architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
3. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).
4. Tan, L. *et al.* Changes in genome architecture and transcriptional dynamics progress

independently of sensory experience during post-natal brain development. *Cell* **184**, 741-758.e17 (2021).

5. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **20**, 721–737 (2019).
6. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
7. 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).
8. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**, E6456–E6465 (2015).
9. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **6**, 6186 (2015).
10. Beagan, J. A. *et al.* YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* **27**, 1139–1152 (2017).
11. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).
12. Ye, B. *et al.* ZNF143 in Chromatin Looping and Gene Regulation. *Front. Genet.* **11**, 338 (2020).
13. Banigan, E. J. *et al.* Transcription shapes 3D chromatin organization by interacting with loop extrusion. Preprint at https://doi.org/10.1101/2022.01.07.475367 (2022).
14. Bouwman, B. A. M., Crosetto, N. & Bienko, M. RNA gradients: Shapers of 3D genome architecture. *Curr. Opin. Cell Biol.* **74**, 7–12 (2022).
15. Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat. Rev. Mol. Cell Biol.* **17**, 756–770 (2016).
16. Li, X. & Fu, X. Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nat. Rev. Genet.* **20**, 503–519 (2019).

17. Yeo, S. J., Ying, C., Fullwood, M. J. & Tergaonkar, V. Emerging regulatory mechanisms of noncoding RNAs in topologically associating domains. *Trends Genet.* (2023).
18. Melé, M. & Rinn, J. L. “Cat’s Cradling” the 3D Genome by the Act of LncRNA Transcription. *Mol. Cell* **62**, 657–664 (2016).
19. Steensel, B. van & Furlong, E. E. M. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* **20**, 327–337 (2019).
20. Calandrelli, R. *et al.* Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells. *Nat. Commun.* **14**, 6519 (2023).
21. Tang, J., Wang, X., Xiao, D., Liu, S. & Tao, Y. The chromatin-associated RNAs in gene regulation and cancer. *Mol. Cancer* **22**, 27 (2023).
22. Trigiante, G., Blanes Ruiz, N. & Cerase, A. Emerging Roles of Repetitive and Repeat-Containing RNA in Nuclear and Chromatin Organization and Gene Expression. *Front. Cell Dev. Biol.* **9**, 735527 (2021).
23. Luo, H. *et al.* HOTTIP-dependent R-loop formation regulates CTCF boundary activity and TAD integrity in leukemia. *Mol. Cell* **82**, 833-851.e11 (2022).
24. Bai, X., Li, F. & Zhang, Z. A hypothetical model of trans-acting R-loops-mediated promoter-enhancer interactions by Alu elements. *J. Genet. Genomics* **48**, 1007–1019 (2021).
25. 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**, 717–726 (2009).
26. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* **21**, 198–206 (2014).
27. Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).
28. Tripathi, V. *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell* **39**, 925–938 (2010).
29. Hansen, A. S. *et al.* Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-

Binding Region in CTCF. *Mol. Cell* **76**, 395-411.e13 (2019).

30. Islam, Z. *et al.* Active enhancers strengthen insulation by RNA-mediated CTCF binding at chromatin domain boundaries. *Genome Res.* **33**, 1–17 (2023).
31. Kung, J. T. *et al.* Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol. Cell* **57**, 361–375 (2015).
32. Saldaña-Meyer, R. *et al.* CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes Dev.* **28**, 723–734 (2014).
33. Saldaña-Meyer, R. *et al.* RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol. Cell* **76**, 412-422.e5 (2019).
34. Yamazaki, T. *et al.* Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol. Cell* **70**, 1038-1053.e7 (2018).
35. Bell, J. C. *et al.* Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *eLife* **7**, e27024 (2018).
36. Li, X. *et al.* GRID-seq reveals the global RNA-chromatin interactome. *Nat. Biotechnol.* **35**, 940–950 (2017).
37. Wu, W. *et al.* Mapping RNA–chromatin interactions by sequencing with iMARGI. *Nat. Protoc.* **14**, 3243–3272 (2019).
38. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).
39. Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* **17**, 1118–1124 (2020).
40. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* **54**, 725–734 (2022).
41. Gunsalus, L. M., Keiser, M. J. & Pollard, K. S. In silico discovery of repetitive elements as key sequence determinants of 3D genome folding. *Cell Genomics* 100410 (2023).
42. Tan, J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-

throughput in silico genetic screening. *Nat. Biotechnol.* 1–11 (2023).

43. Yang, R. *et al.* Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. *Genome Biol.* **24**, 134 (2023).
44. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554-565.e7 (2020).
45. Reiff, S. B. *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat. Commun.* **13**, 2365 (2022).
46. Xiao, Q. *et al.* The landscape of promoter-centred RNA–DNA interactions in rice. *Nat. Plants* **8**, 157–170 (2022).
47. Yan, Z. *et al.* Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proc. Natl. Acad. Sci.* **116**, 3328–3337 (2019).
48. Wang, Y. *et al.* SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biol.* **22**, 36 (2021).
49. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
50. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
51. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).
52. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at http://arxiv.org/abs/1705.07874 (2017).
53. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. Preprint at http://arxiv.org/abs/1704.02685 (2019).
54. Fei, J. *et al.* Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. Cell Sci.* **130**, 4180–4192 (2017).
55. Girard, C. *et al.* Post-transcriptional spliceosomes are retained in nuclear speckles until

splicing completion. *Nat. Commun.* **3**, 994 (2012).

56. Maxwell, E. & Fournier, M. THE SMALL NUCLEOLAR RNAs. *Annu. Rev. Biochem.* **64**, 897–934 (1995).
57. Falaleeva, M. *et al.* Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *Proc. Natl. Acad. Sci.* **113**, (2016).
58. Kishore, S. & Stamm, S. The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C. *Science* **311**, 230–232 (2006).
59. Scott, M. S. *et al.* Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Res.* **40**, 3676–3688 (2012).
60. Galganski, L., Urbanek, M. O. & Krzyzosiak, W. J. Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Res.* **45**, 10350–10368 (2017).
61. Sikand, K. & Shukla, G. C. Functionally important structural elements of U12 snRNA. *Nucleic Acids Res.* **39**, 8531–8543 (2011).
62. Liang, L. *et al.* Complementary Alu sequences mediate enhancer–promoter selectivity. *Nature* **619**, 868–875 (2023).
63. Lu, J. Y. *et al.* Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. *Cell Rep.* **30**, 3296-3311.e5 (2020).
64. Yin, Y. & Shen, X. Noncoding RNA-chromatin association: Functions and mechanisms. *Fundam. Res.* S2667325823000870 (2023).
65. Open2C *et al.* Cooltools: enabling high-resolution Hi-C analysis in Python. Preprint at https://doi.org/10.1101/2022.10.31.514564 (2022).
66. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. https://www.repeatmasker.org/.
67. Oksuz, B. A. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat. Methods* **18**, 1046–1055 (2021).
68. The 4D Nucleome Network *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).

69. Gunsalus, L. M. *et al. Comparing chromatin contact maps at scale: methods and insights*. Preprint at http://biorxiv.org/lookup/doi/10.1101/2023.04.04.535480 (2023).
70. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
71. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
72. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
73. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
74. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
75. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
76. Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinforma. Adv.* **2**, vbac097 (2022).
77. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

Competing interests:

None declared.

Author contributions:

S.K. and K.S.P. conceived and designed the work. S.K. conducted all analyses. S.K. and K.S.P. wrote the manuscript.

Acknowledgements

We gratefully acknowledge members of the Pollard lab for project feedback, Jian Ma's group for providing the SPIN state data and Sheng Zhong's group for project suggestions and providing the iMARGI data.