NAR Genomics and Bioinformatics, 2025, **7**, lqaf010
https://doi.org/10.1093/nargab/lqaf010
Standard Article

OXFORD

Comprehensive analysis of RNA–chromatin, RNA–, and DNA–protein interactions

Daniil A. Khlebnikov [1,2,*], **Arina I. Nikolskaya** [2], **Anastasia A. Zharikova** [1,2,3], **Andrey A. Mironov** [1,2]

[1]RTC Bioinformatics, Kharkevich Institute for Information Transmission Problems of RAS, Bolshoy Karetny per. 19, build.1, 127051 Moscow, Russia
[2]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 1-73 Leninskie Gory, 119991 Moscow, Russia
[3]National Medical Research Center for Therapy and Preventive Medicine, Ministry of Healthcare of the Russian Federation, Petroverigsky per. 10, Bld. 3, 101000 Moscow, Russia

[*]To whom correspondence should be addressed. Email: dkhlebn@gmail.com

Abstract

RNA–chromatin interactome data are considered to be one of the noisiest types of data in biology. This is due to protein-coding RNA contacts and nonspecific interactions between RNA and chromatin caused by protocol specifics. Therefore, finding regulatory interactions between certain transcripts and genome loci requires a wide range of filtering techniques to obtain significant results. Using data on pairwise interactions between these molecules, we propose a concept of triad interaction involving RNA, protein, and a DNA locus. The constructed triads show significantly less noise contacts and are more significant when compared to a background model for generating pairwise interactions. RNA–chromatin contacts data can be used to validate the proposed triad object as positive (Red-ChIP experiment) or negative (RADICL-seq NPM) controls. Our approach also filters RNA–chromatin contacts in chromatin regions associated with protein functions based on ChromHMM annotation.

Introduction

It is well recognized that eukaryotic RNAs play multiple roles beyond their direct role in transporting transcribed genetic information in the nucleus: major advances include understanding the functions of noncoding RNAs such as XIST, NEAT1, MALAT1, HOTAIR, FIRRE, etc. [1, 2]. The study of new potential functions of chromatin-associated RNAs is of interest to molecular biology in the context of searching for DNA loci and genes whose interaction with RNA leads to gene expression regulation [3].

Currently, there are many experimental protocols for determining pairwise contacts between RNA, proteins, and chromatin DNA loci. Such protocols solve the problems of identifying RNA–protein, DNA–protein, and RNA–DNA contacts. These methods are based on short-read sequencing technologies and therefore have different shortcomings.

When studying the pool of RNA–chromatin contacts, there are two main approaches to designing the interaction detection experiment. “One-to-all” experiments, such as RAP (RNA antisense purification) [4], CHART-seq (capture hybridization analysis of RNA targets) [5], and ChIRP-seq (chromatin isolation by RNA purification) [6], focus on single RNA contacts across the genome by pulling down RNA–chromatin complexes of a particular RNA. These experiments are considered to be the gold standard for a further class of methods of the “all-to-all” type. These methods address the challenge of identifying pairwise contacts between all RNAs and every genomic locus. The key step in the experimental protocol is to fix interacting RNAs and DNA loci by bridge ligation of interacting nucleic acids. This fixation predominantly occurs through proteins. Methods to search for RNA–chromatin interactions include iMARGI (enhanced mapping of RNA–genome interactions assay) [7, 8], RADICL-seq (RNA and DNA interacting complexes ligated and sequenced) [9], GRID-seq (global RNA interactions with DNA by deep sequencing) [10–12], and Red-C (RNA ends on DNA capture) [13]. This study analyses data from GRID-seq, RADICL-seq, and Red-C.

The all-to-all experiments show that most of the RNA–chromatin contacts (RD-contacts) detected are those of protein-coding RNAs (Supplementary Table S1) [14]. The presence of regulatory functions for these RNAs has not been well established, and this observation reflects a high level of nonspecific interactions. For most RNAs, the density of contacts near the gene of that RNA peaks and then declines in a power law fashion [13, 14], a phenomenon we call RD-scaling. This is caused by the contacts of incompletely transcribed nascent RNA, as well as by RNA that diffuses away from the site of transcription. Importantly, RD-scaling is a feature of both all-to-all and one-to-all data. The RD-scaling phenomenon together with other factors specific to interactomics protocols also contributes to the bias in the ratio of *cis*/*trans* contacts in the resulting data. Following the conventions of Hi-C experiments, we refer to RNA–DNA pairs where the DNA locus chromosome and the RNA chromosome of origin are different as *trans*-contacts. However, the RNA–chromatin interaction data reveal several other types of biases. First, contacts may be nonspecific, as RNA can contact chromatin proteins such as histones, albeit with lower affinity. The probability of observing contact depends on chromatin accessibility, as heterochromatin regions are less

Received: May 8, 2024. **Revised:** January 22, 2025. **Editorial Decision:** February 3, 2025. **Accepted:** February 5, 2025

accessible to nucleases. Normalization to background is applied to compensate for the aforementioned biases. Various peak calling methods are used to reduce the influence of these biases and partially suppress the noise. The most commonly used approaches include MACS2 peak calling [15] and custom dataset-specific peak calling algorithms (GRID-seq [10–12]). These approaches do not take into account the bias introduced by RD-scaling. A recently developed BaRDIC (binomial RNA–DNA interaction calling) peak-calling algorithm is designed to account for these biases [16]. The RNA–protein interaction data can also contain a significant amount of nonspecific interactions [17, 18], and appropriate peak calling algorithms are also used to suppress the noise. Red-ChIP [19] detection of RNA–chromatin contacts enriched for EZH2 and CTCF proteins by immunoprecipitation was developed to isolate contacts mediated by these proteins. A ChRD-PET (chromatin associated RNA–DNA interactions followed by paired-end-tag sequencing) [20] protocol that includes H3K4me3 immunoprecipitation and ChIP-seq filtering of contacts has been proposed for the study and analysis of RNA–chromatin interactions in *Oryza sativa*. Despite the higher specificity of such protocols, they are still bound to the protocol disadvantages described earlier. For example, in Red-ChIP, the dominant proportion of contacts are mRNA (messenger RNA) contacts (Supplementary Table S1).

There are different models explaining how an RNA can regulate biological processes on a certain DNA locus. One of the possible mechanisms is binding to a regulatory protein. The conformational changes the protein molecule experiences afterward may lead to its recruitment to the genomic site, where it’s involved in regulatory events. We therefore speculate that studying the RNA–chromatin interactome in light of protein immunoprecipitation data may yield functional RNA–protein pairs interacting with certain DNA loci. These objects are what we call triads here.

We compared three types of data: RNA–chromatin contacts (RD-contacts) captured by the aforementioned protocols, protein–chromatin contacts (PD-contacts) captured by chromatin immunoprecipitation and sequencing (ChIP-seq) experiments, and protein–RNA contacts (PR-contacts) from various RNA and protein immunoprecipitation experiments. We utilized this approach to understand which proteins mediate the contacts of which RNAs with chromatin. The data used in this study include RNA–chromatin interactions from RADICL-seq [9], GRID-seq [10–12], and Red-C [13], DNA–protein interactions from ChIP-seq [21], and RNA–protein interactions from RIP-seq (RNA immunoprecipitation and sequencing) [22], fRIP-seq (formaldehyde RNA immunoprecipitation and sequencing) [23], and eCLIP (enhanced crosslinking and immunoprecipitation) [24, 25]. The intersection of these data is referred to as interaction triads, as shown in Fig. 1A. Pairwise contacts within the triads are denoted as RDt-contacts, PDt-contacts, and PRt-contacts. The proposed concept of interaction triads involves a two-stage filtering process of RNA–DNA interaction data using available RNA–protein and DNA–protein interactomes.

As was mentioned before, the RNA–DNA data are very noisy; on the other hand, the comparison of replicates in such experiments shows that a small fraction of contacts is actually consistent between replicates [26], suggesting that the data are substantially incomplete. Our proposed approach can’t increase data completeness, but it significantly reduces noise. Therefore, while we might lose functional contacts when constructing triads, we guarantee a significant reduction in noise levels.

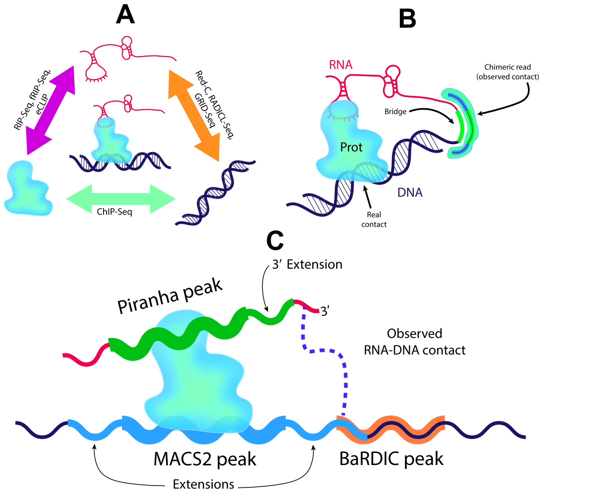

Figure 1. (**A**) The concept of constructing RNA–protein–DNA triads using pairwise interaction data. (**B**) The observed contacts in the all-to-all data may be located away from the actual contacts. This creates inaccuracy in the data. (**C**) To obtain more accurate data on protein-mediated RNA–DNA interactions, peak expansion is required.

This approach enabled us to decrease the aforementioned noise level. We assessed the statistical significance of the triads obtained. When estimating the biological significance of the triads, we observed an association between the triads and the ChromHMM and SPIN putative chromatin states. Additionally, we analyzed the evolutionary stability of the identified triads by comparing human and mouse data.

Materials and methods

Initial data

We have selected 47 eukaryotic nuclear proteins that have available RNA and DNA immunoprecipitation data. These proteins are involved in chromatin remodeling and modification, regulation of transcription, and regulation of RNA processing and splicing. The selected proteins include PRC2 proteins (EZH2 and SUZ12), several ribonucleoproteins (hnRNPC, H, K, L, U, and UL1), various splicing factors (SRSF and U2AF family proteins), and chromatin remodelers (HDAC1, DNMT1, PCAF, and LSD1). The proteins were divided into functional groups (Supplementary Table S2) to simplify the analysis and draw meaningful conclusions for protein groups rather than individual proteins. Some proteins are multifunctional and therefore cannot be unambiguously assigned to one of the groups. These are RNA processing and transcription regulation proteins such as FUS and hnRNP components. Nevertheless, we decided to categorize the proteins into both overlapping and nonoverlapping sets. The data for the overlapping sets of proteins are shown here.

To construct triads of RNA–protein–DNA interactions mediated by one of the 47 nuclear proteins, we used various types of data. These included processed ChIP-seq experiment data that defined DNA loci interacting with the target protein, raw data from the RIP-, fRIP-, and eCLIP-seq experiments that defined sites on RNA that interact with the target protein, and our previously processed data for RNA–chromatin interactomes of the Red-C, GRID-seq, and RADICL-seq experiments for human (K562) and mouse (mESC) cell lines from RNA-Chrom database [14]. All IDs used from the databases can be found in Supplementary Table S3. RNA biotype annotation was derived from the RNA-Chrom [14] database, as this was the source we obtained the data from. To avoid misunderstandings, we refer to the RNA of "X-RNA" biotype as transcripts that are consistently assembled from RNA contacts mapped to unannotated genomic regions in gene deserts.

ChIP-seq data were obtained from the public biological databases ENCODE, ReMap [27], and GTRD [28] as DNA peaks derived from the MACS2 peak caller with a *q*-value threshold of 0.05. This way, we made sure ChIP-seq peaks data were obtained in the exact same way using a common standard set of parameters of the peak-calling algorithm.

RNA–protein interactions data processing

We processed data from RNA–protein immunoprecipitation experiments to determine RNA interactions for target proteins in a unified pipeline. Raw RNA reads from experimental and control replicates were mapped to the human genome version GRCh38.p13 using the hisat2 aligner (v2.2.1) [29] and then binned across the human genome (bin size 300 bp) for more accurate peak searching. The reads were binned and then analyzed using the Piranha [30] RNA–protein peak-calling software. The replicates were double-checked for consistency, and consistent replicates were merged. The input data were specified as the background distribution for normalization, and a filtering *q*-value threshold of 0.05 was applied. Piranha peaks were called using standard software parameters. The resulting peaks were annotated using the gene annotation from the RNA-Chrom database. Mapped reads were annotated with the same annotation. For each PR interaction, the fold change of contact was calculated as the ratio of the fraction of gene counts in the experimental replicate to the fraction of gene counts in the control replicate, with an additional pseudocount of 0.01. Genomic intervals were manipulated using the bedtools package (v2.27.1) [31]. Tables with statistics of peak counts for RNA– and DNA–protein interactions can be found in Supplementary Table S4.

Triads construction

BaRDIC [16] peak calling software was used to process data from RNA–chromatin interactions (BaRDIC parameters are listed in Supplementary Table S5). Only contacts falling within BaRDIC peaks were selected, as we are interested in individual contacts.

RNA–DNA interactome extraction protocols involve the use of restrictases and proximity ligation (Fig. 1B). It is therefore possible that the observed RNA–DNA contacts may not be in direct spatial proximity to the actual contacts. In order to account for these discrepancies, DNA–protein contacts were extended by 2 kbp to either side of the ChIP-seq peak (to try capturing the restriction site occuring approximately once in 256 bp), while RNA–protein contacts were extended by 100 bp toward the 3′ end of the Piranha peaks (Fig. 1C) to try to capture the RNA part that might actually form an RD-contact. Furthermore, the aforementioned parameters were varied by the constructing sets of triads at alternative peak broadening values. The powers of such sets are presented in Supplementary Table S6. We left the RNA–DNA contact data unchanged, as it was the distance between RNA– or DNA–protein peaks and RNA–DNA interaction tracks that was of interest. This is why it is not important which data to augment. Here, we show the results obtained on the widest extensions (2 kbp for DNA- and 100 bp for RNA–protein data). The objective was to select a subset of triads that would minimize the impact of data incompleteness, i.e. to lose as few real functional contacts of the original RNA–DNA contacts set as possible. The RD-contacts were then intersected with PD-contacts and PR-contacts (RIP-seq/fRIP-seq/eCLIP) for each protein separately.

Additionally, we discarded the RD-contacts if the length of intersection with PD- or PR-peaks was <19 bases. Filtering was necessary to reduce the number of false-positive triads resulting from the intersection of close contacts with a peak. Such a threshold was chosen by combined analysis of the intersection size behavior of the RNA-parts of RD-contacts with extended RNA peaks, as well as the DNA-parts of contacts with extended ChIP-seq peaks (Supplementary Fig. S1). Due to the inaccuracies and noise associated with RNA–chromatin contact data, this filtering may discard some true-positive triads along with the false-positive ones. However, we assume that for datasets with extended peaks, the proportion of true-positive triads discarded is minimal. The triads were visualized using the circos-v0.69.9 suite [32]. Red-ChIP data were processed according to the protocol of the study presenting the method [19].

Data analysis was performed mostly in Python3, and scripts can be found in the GitHub repository (github.com/dkhlebn/triads_article_scripts).

Triads simulation

To assess the randomness of the obtained triad data, we performed a permutation test to create a background model. Subsequently, the results obtained on real data were evaluated in relation to this model. It should be noted that the DNA–RNA data are not homogeneous and contain bias (RD scaling). Therefore, these data cannot be shuffled adequately. On the other hand, we analyze the mutual position of RNA– or DNA–protein and DNA–RNA contacts. Therefore, which data to shuffle is not important. When simulating ChIP-seq peaks of proteins, we aimed to preserve the intrinsic structure of PD-contacts, which consisted of two components. First, we had to preserve the profile of interactions, i.e. the height of ChIP-seq peaks. Second, we had to consider and preserve each peak annotation by nuclear A/B-compartment. This is due to a bias in the data introduced by differences in open and repressed chromatin. The observed RD-contacts are depleted in closed chromatin because it is less nuclease-accessible. There may also be a bias in the ChIP-seq data. We retained the annotation of the peak by A- and B-compartments as it contains crucial biological information about protein–DNA interactions. To obtain the shuffled ChIP-seq tracks, each protein peak was randomly shifted up- or downstream by a number of megabases (1, 3, 5, 7, or 10) predetermined independently for each chromosome. If the ChIP-seq peak in the original data was in the A-compartment, its shifted version in the generated data would also be in the A-compartment (Fig. 3A, upper panel). Regarding PR-interactions, we aimed to preserve the biotype-specific content of the PR interaction repertoire for each protein by considering its relationships with other proteins in the dataset. This implies that a protein's shuffled PR-interactions should be estimated by the set of PR-interactions of proteins that share numerous

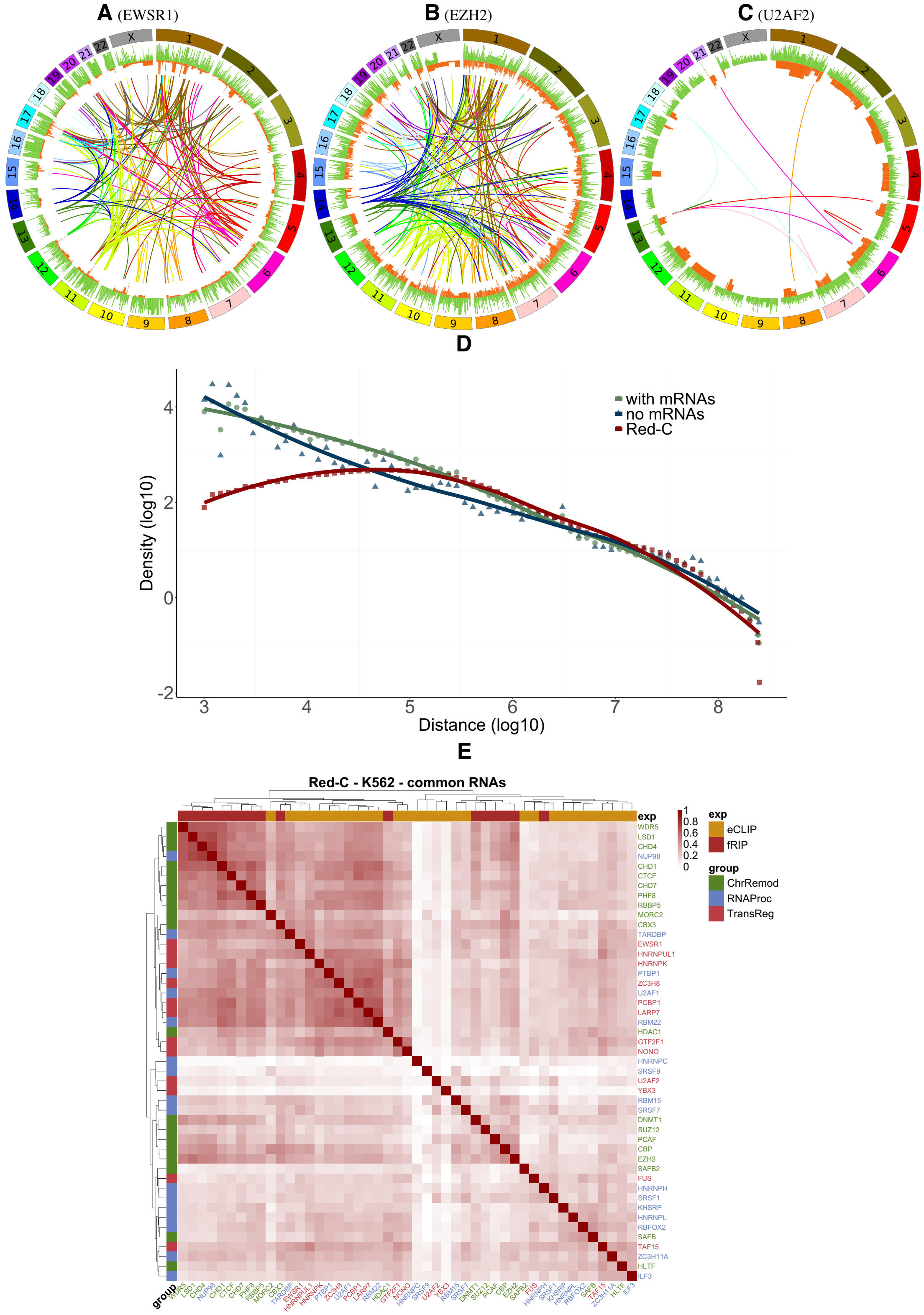

Figure 2. RNA–DNA interaction triads for three proteins: (**A–C**) Schematic diagrams show the interaction triads of EWSR1 (A), EZH2 (B), and U2AF2 (C) proteins (see Supplementary Fig. S3 for the remaining proteins). The outer histogram represents PR-peaks, while the inner histogram track represents PD-peaks. The arc connecting the locations of the RNA and DNA involved in the triad is colored according to the chromosome on which the RNA-part is located. Only triads that are not mediated by protein-coding RNAs are shown. (**D**) RDt scaling in *cis*-triads based on data from the Red-C experiment. (**E**) Jaccard's score was used to measure the similarity of triad datasets based on shared RNAs between each pair of datasets.

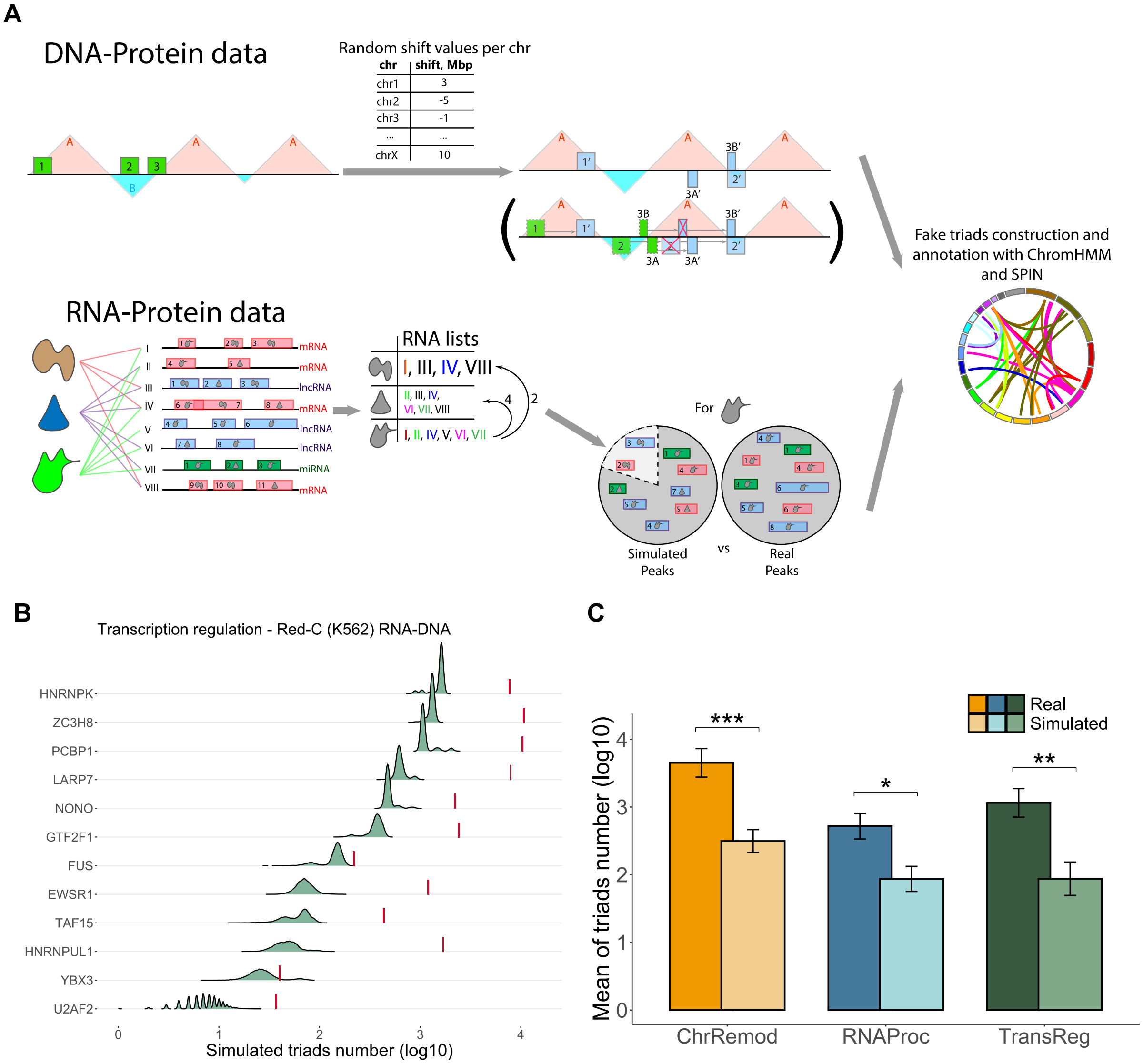

Figure 3. Simulation scheme of the background model used to generate random data for assessing the randomness of the triads. (**A**) The upper panel displays the shift value in megabases generated for each chromosome at every step (from −10 to 10 Mbp). Then, the ChIP-seq peaks on each chromosome are shifted by the corresponding number of megabases. Peak must remain in the same type of chromatin compartment. If the shift causes the peak to move to the opposite compartment type, it is relocated to the nearest compartment of the required type. The bottom panel presents proteins as lists of RNAs that proteins interact with, based on real RNA peak data. For each protein, a pool of "related" and "unrelated" proteins is defined by the distance between their RNA lists. The pool of PR-interaction peaks for a particular protein is formed from the set of all peaks of "related" proteins (including the protein's original PR-peaks themselves) and the set of peaks of "unrelated" proteins in a 4:1 ratio. This preserves the number and RNA biotype composition of the original RNA–protein peaks. The triad assembly pipeline is fed with resulting data, and this process is repeated 10 000 times to generate a sample of sufficient magnitude for drawing meaningful statistics. The number of simulated triads for proteins functionally described as transcriptional regulators in the K562 cell line is shown in the distributions in panel (**B**), with the real triad counts denoted by vertical lines. For information on other protein groups and all cell lines, see Supplementary Figs S4–6. (**C**) Barplots displaying the number of simulated triads for each functional group of proteins, asterisks denote statistical significance for the Wilcoxon test, with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. For protein grouping without intersections, see Supplementary Fig. S7.

qualitative PR-interactions with it. As experimental data defining PR-interactions may contain some level of noise, we also included PR-interactions from other proteins to account for this (Fig. 3A, lower panel). The dataset's proteins were considered as the set of peaks of their PR-interactions and the list of RNAs on which their peaks are located. A distance matrix was obtained based on the magnitude of intersections of the RNA lists between each pair of proteins. Proteins whose distance to some protein was in the top third of all distances to that protein were labeled as "related" for that target protein. The pool of PR-contact peaks for each protein was then formed as follows. The PR-peaks of both "related" and "unrelated" proteins were merged into separate pools. From these, a pool of shuffled PR-peaks for the target protein was drawn with replacement in a 4:1 ratio, while preserving the total number and RNA biotype composition of the original peak set. This is equivalent to having no >20% of the contacts in the final set of simulated RNA–protein data originating from a set of unrelated proteins. This ratio allows us to conduct more rigorous simulations, thereby increasing the reliability of the statistical results. It is important to note that the protein "closeness" obtained from RNA lists resembles the actual protein clustering by function (Supplementary Fig. S2). The data of DNA– and RNA–protein interactions were shuffled and then passed to a triad constructing pipeline for analysis (Fig. 3A, right of both panels). This was done 10 000 times for each protein.

Triad annotation using ChromHMM and SPIN

The triads obtained in the previous step were annotated with grouped annotations of ChromHMM chromatin states and SPIN chromatin compartments. The annotations were taken from the corresponding articles of their origin [33–35]. To simplify the analysis, we combined annotation states into groups (refer to Supplementary Table S7 for ChromHMM and Supplementary Table S8 for SPIN).

RNA–chromatin contacts orthologs search

The ortho2align [36] software was used to select bidirectional best hits from the coordinates of RNAs involved in triads formed by each protein in human and mouse cell lines. Next, the DNA contacts of each RNA were extracted in both cell lines, and the nearest DNA loci [37] among the contacts were searched for orthologous RNA pairs from different organisms. Finally, bedtools was used to annotate the nearest genes. The correlation of genomic intervals was carried out using StereoGene [38]. The parameters for the runs can be found in Supplementary Table S9.

Data visualizations and analyses not previously mentioned were performed using in-house scripts in Python, bash, and R. Data simulations were conducted to verify the randomness of the resulting data using Python scripts. All scripts are available in the corresponding GitHub repository (github.com/dkhlebn/shift_IP_peaks).

Results

Constructed triads demonstrate less noise than original data

Figure 2A–C displays examples of RD-triad graphic visualizations that we constructed. It is evident that *trans*-RDt-contacts make up the majority of triad contacts. Supplementary Table S10 displays the distribution of *cis*- and *trans*-contacts numbers for protein triads. It is clear that, with the exception of the triads mediated by the FUS protein, most proteins exhibit predominantly *trans*-contacts in the filtered contacts.

Any existing *cis*-contacts have RDt-scaling similar to that of the original RNA–chromatin interaction data (Fig. 2D). A paired Wilcoxon test was performed for each RD-experiment (Table 1). The statistics for select RNAs are shown in the Supplementary Table S11. The ratio of *cis*-to-*trans*-contacts decreases when comparing triads to RD-experiment data for all protocols except GRID-seq, specifically for Red-C and RADICL-seq. The GRID-seq result may appear surprising at first glance; however, it can be attributed to peak-calling effects. It has been demonstrated that the peaks of RNA–DNA interactions obtained with BaRDIC are of inferior quality compared to those obtained in other RD experiments. Therefore, it is reasonable to assume that the distribution of *cis*-to-*trans*-contacts in GRID-seq and its triads may differ from that observed in other experiments (Supplementary Table S1).

Table 1. Test results for changes in the ratio of *cis*- to *trans*-contacts

Experiment	*P*-value
Red-C (K562 cells)	7.45×10^{-19}
GRID-seq (mESC cells)	0.537
RADICL-seq 2FA (mESC cells)	1.07×10^{-5}
RADICL-seq NPM (mESC cells)	7.02×10^{-57}

Nevertheless, the result suggests that the interaction triad data we obtained are qualitatively similar and may be worth further investigation.

The intersections of the sets of RNAs forming triads with each of the proteins were analyzed to identify common contacts (Fig. 2E). Proteins with similar functions were found to be clustered together, which can be attributed to their participation in a common process and a larger pool of common RNAs forming triads with them.

Table 2 presents the number of triads found in total and by RNA and protein types. It is known [13, 16, 26] that contacts of protein-coding RNAs can account for up to 80% of all contacts found in RNA–chromatin interactome experimental data. However, the share of mRNA contacts decreases significantly when filtering the data to construct triads (Table 2 and Supplementary Table S10). It is evident that the proportion of noncoding RNAs has increased significantly. However, the total number of contacts mediated by any one of the 47 proteins accounts for no >10% of the original dataset of RNA–chromatin interaction contacts. The unavailability of data for all RNA–chromatin associated proteins, incomplete and noisy RNA–chromatin interactome data, and other pairwise protein to nucleic acid interaction data explain this. Our proposed method of triad construction can also filter out noisy PD- and PR-contacts in other pairwise data, as demonstrated in Supplementary Tables S12–16.

We attempted to compare triads assembled on one-to-all and all-to-all RNA–DNA interaction data of Malat1, Hdac2, and Meg3 RNAs for mESC cells using data from RAP and ChIRP-seq experiments (Supplementary Note S1). While no definitive conclusions could be drawn regarding the Malat1 and Hdac2 RNAs, the comparison of one-to-all triads and all-to-all triads of EZH2 and SUZ12 proteins for lncRNA (long non-codingRNA) Meg3 revealed consistency across different experimental protocols.

The descriptive statistics for the content of the constructed triads suggest that the results obtained provide a clearer sample of RDt contacts mediated by the proteins studied. This sample can be characterized by a much lower level of noisy cis contacts and RD-contacts formed by mRNAs.

Constructed triads are statistically significant

To evaluate the statistical significance of the obtained triads, we created a background model for PR- and PD-contact data that maintains the biological structure of the original data. To achieve this, we shuffled the PD-peaks and PR-interactions of the studied proteins and constructed triads using these datasets. As our focus was on studying RNA–DNA interactions, we did not alter the RD-contacts data during the simulation.

The triad counts on shuffled data from the background model are lower than those on real data for most of the proteins, as shown in Fig. 3B and C (with *P*-values for each pro-

Table 2. Triads composition by RNA biotypes for all proteins split by function. pseudo, pseudogenes; *cnt*, number of observed contacts; *tri*, number of triades; *Z*, z-score of $\frac{tri}{cnt}$ distribution over the RNA biotypes

Biotype	$\frac{cnt}{10^6}$	%	All proteins			Remodelling		Tr.Reg		RNA proc.	
			$\frac{tri}{10^3}$	%	Z	$\frac{tri}{10^3}$	Z	$\frac{tri}{10^3}$	Z	$\frac{tri}{10^3}$	Z
mRNA	156.4	66.1	471.9	19.61	-1.26	409.3	-1.13	32.7	-0.88	29.9	-0.84
rRNA	33.0	13.9	853.6	35.47	-0.10	850.2	0.03	3.2	-0.98	0.2	-1.02
snRNA	5.5	2.3	216.5	8.99	0.59	186.6	0.44	15.3	1.40	14.6	1.53
snoRNA	3.6	1.5	217.2	9.03	1.62	206.8	1.59	5.2	0.22	5.2	0.35
lncRNA	12.5	5.3	113.4	4.71	-0.95	96.3	-0.87	9.1	-0.41	8.0	-0.41
X-RNA	21.0	8.9	431.2	17.92	-0.37	414.9	-0.27	8.6	-0.70	7.7	-0.67
miRNA	0.4	0.2	20.3	0.84	1.40	20.2	1.52	0.0	-0.97	0.0	-1.02
pseudo	1.1	0.5	20.2	0.84	-0.51	11.8	-0.74	2.8	1.14	2.4	1.03
other	3.2	1.4	62.5	2.60	-0.42	44.4	-0.57	8.1	1.17	6.9	1.03
TOTAL	236.6	100.0	2406.7	100.00	-0.90	2240.5	-0.79	85.0	-0.74	74.9	-0.72

tein provided in Supplementary Table S17). The only exceptions are the triads based on data for hnRNPL, YBX3, KHSRP, RBFOX2, SUZ12 (K562 cell line), and WDR5 (mESC cell line). For YBX3 and KHSRP proteins, this could be due to unreliable RNA and/or DNA interaction data, which may contain a lot of noise. It is important to note that in some simulations, the number of triads obtained for hnRNPL, RBFOX2, and SUZ12 was even higher than the actual number. This may indicate either significant noise in the data or the presence/absence of some association of these proteins with the RNA–chromatin interactome. The behavior of the number of simulated triads for the WDR5 protein appears to be influenced by the quality of the original ChIP-seq data. The analysis of strand cross-correlations reveals a poor pattern of read shift correlation, indicating noisy data (see Supplementary Fig. S8).

Taken together, the simulation results demonstrate that the data obtained from real triads is statistically significant. It would be practically unattainable to obtain such a number of triads if the nature of pairwise interactions were different.

RADICL-seq NPM: protein absence affects the construction of triads

Formaldehyde crosslinking primarily fixes protein contacts. Therefore, RD-contacts in the RADICL-seq NPM (nonprotein mediated) [9] experiment, where samples were treated with proteinase K before RNA–DNA ligation, can serve as a negative control for triads. The resulting triads can then be compared to triads obtained from RD-contacts constructed from standard 2% formaldehyde (2FA) crosslinking.

In the case of NPM, the triads will only reflect noise since there are no proteins present. We analyzed RDt-contacts of EZH2, hnRNPK, SUZ12, and WDR5 proteins. For SUZ12, we used triads constructed from fRIP and eCLIP data (see& section "Pairwise interaction data can influence the construction of triads"). We compared common triads derived from these data (Supplementary Figs S9A–C and S10) by selecting those in which the DNA parts overlap and the RNA parts belong to identical transcripts. The metrics obtained indicate that the SUZ12 and EZH2 proteins have overlapping contacts, as they are both functional units of the PRC2 complex. Additionally, the fRIP and eCLIP data for SUZ12 show a significant number of shared contacts. In contrast, the WDR5 protein in both 2FA and NPM data exhibits a large number of common contacts with other proteins. However, this effect disappears in the 2FA data with contacts common with NPM data removed (this set of contacts is labeled as clean). The "CLEAN" dataset for RADICL-seq triads represents only protein-mediated contacts. The correlation of DNA parts was also examined using StereoGene [38] and a similar pattern was observed (Supplementary Fig. S9D–F).

The NPM data show a different RD-scaling pattern, peaking at distances up to 1 kb and then declining to uniformly distributed values as the distance from the gene increases. This effect seems to be partially present in the RADICL-seq 2FA data as well, and is eliminated if the data are cleaned from contacts shared between the RADICL-seq 2FA and RADICL-seq NPM experiments (Supplementary Fig S9G and H).

The distribution of DNA-parts over ChromHMM annotation states is of interest, as shown in Supplementary Fig. S9I. Different trends can be observed in the ratio change of DNA parts falling into a certain ChromHMM state. A pattern arises where this ratio changes drastically when filtering 2FA-based contacts from those shared with NPM-based dataset for some states and some proteins. For further information, see section "Constructed triads are associated with genomic annotations."

It is important to note that RNA–DNA contacts are not solely mediated by proteins; various chromatin structures, such as R-loops or RNA–DNA triplexes, also play a role. While such data are of interest, they are not the focus of this study. Therefore, when searching for protein-mediated interactions, it is possible to filter out RNA–DNA contacts that are not protein-mediated (defined as part of RADICL NPM contacts).

The NPM data contains more contacts in total, which suggests low specificity and potentially higher noise. This observation is supported by the number of simulated triads distribution for RADICL-seq 2FA and RADICL-seq NPM data (see Supplementary Fig. S11). To increase the specificity of the data, subtracting the NPM contacts from the 2FA data is necessary.

Pairwise interaction data can influence the construction of triads

The availability of data for various pairwise interactions of RNA, protein, and DNA allows us to compare the triads that can be constructed from them. As a source of RNA–protein interactions, we used data from different experimental protocols, namely fRIP-seq and eCLIP-seq. Additionally, GRID-seq and RADICL-seq data were obtained for the same cell line, mouse embryonic stem cells (mESCs). The sets of triads constructed from the different data were compared. RADICL-seq 2FA data were used for this analysis. The data from the NPM experiment were filtered out (as described in the “RADICL-seq NPM: protein absence affects the construction of triads” section).

The triads for the SUZ12 protein, constructed from different initial RNA–protein interaction data, are consistent. There are more RNAs forming triads in a shared set of contacts than in either of the two separate datasets (Supplementary Fig. S12A). In addition, the volume of the shared contacts set is greater (Supplementary Fig. S12B). We compared the biotype content of RNAs involved in the fRIP- and eCLIP-based triads (Supplementary Fig. S12C). Comparison of RNA biotype representation in triads revealed that RNAs from eCLIP-based triads are primarily ribosomal, whereas RNAs in fRIP-seq-based triads exhibit greater diversity in composition.

The RDt-contacts of fRIP- and eCLIP-based triads were divided into four groups:

- RDt-contacts of fRIP-based triads
- RDt-contacts of eCLIP-based triads
- RDt-contacts shared between fRIP- and eCLIP-based triads
- The union of RDt-contacts only present in either fRIP- or eCLIP-based triads (essentially the symmetrical difference of the first two sets)

The relationship between the DNA-parts of these RDt-contacts and the ChromHMM annotation states was examined (Supplementary Fig. S12D). The resulting distributions show that although the RNA forming the triads are less consistent between fRIP- and eCLIP-based triads, the DNA loci they interact with and the loci left when constructing triads are consistent.

The aim of this comparison is to determine whether triads constructed on eCLIP or fRIP-seq data are more biologically relevant. The choice of experimental protocol for constructing triads should be based on this criterion. Assuming that many regulatory functions of noncoding or unannotated RNAs are of great interest for further investigation, we selected triads with a higher proportion of biotypes of long noncoding RNAs and XRNAs. Although the replicates of fRIP-seq and eCLIP experiments are consistent with one another within the same protocol, the dissimilarities in protein–RNA data across different experiments can also be attributed to a batch-effect, which is not accounted for in our analysis. Based on the stated condition, the optimal choice among all the sets is the triads shared between those constructed using fRIP-seq and eCLIP (Supplementary Fig. S13). This is in accordance with common sense, as using a consensus of two experiments adds validity to the objects in that consensus, thereby increasing confidence in data obtained from noisy experiments. An additional argument in favor of the already demonstrated consistency of the discussed triads is the high correlation between the tracks of their DNA parts. This correlation coefficient of 0.89 (P-value ~0), obtained using StereoGene, is supported by the cross-correlation function declining within 5 kb (Supplementary Fig. S14). Since the number of peaks and its contents (Supplementary Table S4) for these experiments are the same, fRIP-seq and eCLIP can be considered almost equivalent for the construction of triads.

Table 3. The correlation between DNA parts of triads tracks for GRID-seq and RADICL-seq-based triads

Protein	Correlation	*P*-value
EZH2	0.6525	2.7×10^{-50}
hnRNPK	0.6597	1.6×10^{-3}
SUZ12	0.039	1.1×10^{-243}
WDR5	– 0.078	0.6724

We compared the effects of different types of RNA–chromatin interactome sequencing experiments on the construction of the RDt-contacts set. We used the triads of EZH2, hnRNPK, SUZ12, and WDR5 proteins constructed from GRID-seq and RADICL-seq RD-interactions data in cells of the mESC cell line. The RNA parts in triads of each protein were not consistent, as there were no common RNA parts between them. This was observed in both data filtered from NPM contacts (2FA-no-NPM) and the original 2FA data (Supplementary Fig. S15). Previous research has shown that all-to-all RD-data, specifically its DNA parts, begin to exhibit consistency over relatively large distances (starting from 100 kb) for different experimental protocols [26]. These RNA–chromatin interactions may not represent the complete and accurate sample of functional RNA contacts at DNA loci due to their noisy nature and significant incompleteness. To address this, we extended the DNA parts of the contacts by 500 kb upstream and downstream from the 5′- and 3′-ends of the DNA part of the contact because the data are very sparse. We then used StereoGene to verify the consistency of these extended intervals (Fig. 4A–D). We observed a significant correlation of DNA part tracks for EZH2 and hnRNPK proteins (Table 3). However, the distribution of correlation coefficients is bimodal, with modes at 0 and 1. This indicates that there are consistent regions between the GRID- and RADICL-seq DNA parts, which are of main interest to study in light of triads construction. The remaining contacts, tending toward 0 and not resulting in a non-zero correlation, appear to be noise. However, it is important to note that RNA–chromatin interactome data is incomplete, and uncorrelated loci may indicate a contact that is detected in one experiment but lost in another. The most reliable loci are those with a high correlation between data from both experiments. Regrettably, due to the limited number of WDR5 protein-mediated RDt-contacts and the quality of the PR- and PD-interactions data, it is not possible to draw any meaningful conclusions regarding the consistency of triads mediated by this protein.

The distribution of DNA parts by ChromHMM states was characterized by RDt-contacts constructed on RD-contacts data from RADICL-seq and GRID-seq experiments and PR-/PD-contacts for EZH2, hnRNPK, SUZ12, and WDR5 proteins (Fig. 4E–H). The distributions showed higher consistency when comparing GRID-seq and RADICL-seq based triad DNA parts.

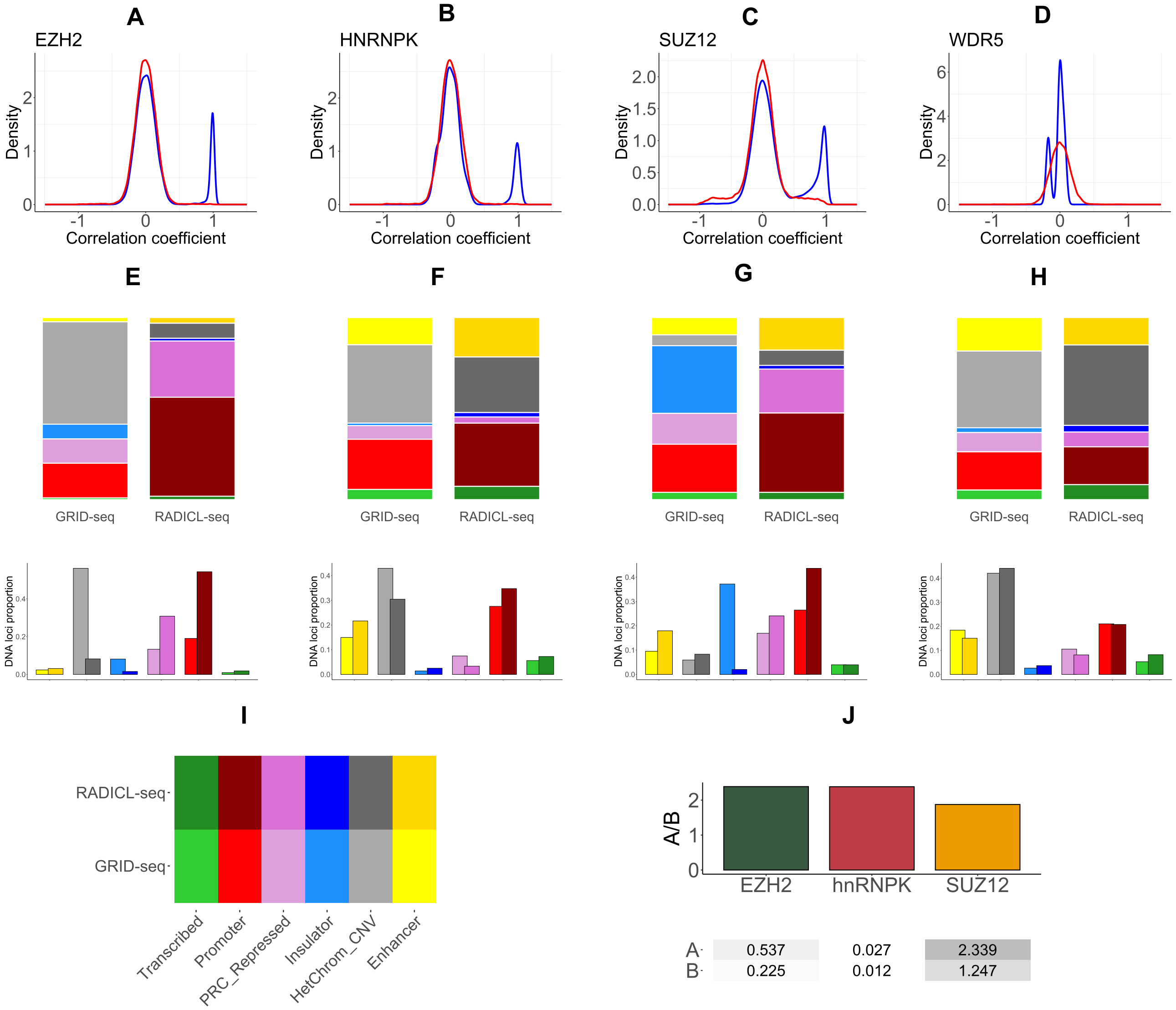

Figure 4. Comparison of EZH2, hnRNPK, SUZ12, and WDR5 protein triads constructed from RADICL- and GRID-seq RNA–chromatin interaction data. (**A**–**D**) The correlation coefficient distribution between the extended DNA parts of triads for EZH2 (A), hnRNPK (B), SUZ12 (C), and WDR5 (D) proteins. (**E**–**H**) The distribution of DNA-part proportions of triads on ChromHMM states for triads constructed on RADICL-seq and GRID-seq RDt-data for EZH2 (E), hnRNPK (F), SUZ12 (G), and WDR5 (H) proteins. (**I**) A color legend for previous panels (E–H). (**J**) The density of consistent triads in RADICL-seq and GRID-seq falling into A/B compartments.

The regions that were consistent between the two experiments were characterized in terms of their A/B-compartments annotation. It was demonstrated that the density of triads falling into A-compartments is approximately twice that of triads falling into B-compartments. Density is defined as the number of DNA loci within A- or B-compartments normalized to the total compartment length (Fig. 4J and Table 4). This suggests that when the chromatin is more open at the site of contact ligation, the contact representation is more likely to be consistent across different protocols. It is also possible that GRID-seq captures more consistent RD-contacts due to their proportion of total GRID-seq contacts. A more comprehensive examination of the impact of chromatin accessibility on RNA–chromatin contact data and triads requires the generation of triads for proteins that are enriched in closed chromatin, such as HP1. However, in the absence of RNA immunoprecipitation data and DNA loci occupied by this protein, this is currently not feasible.

Table 4. Ratio of contacts falling into regions of high correlation based on StereoGene results; cnts, contacts

Protein	# consistent cnts	% RADICL cnts (total)	% GRID cnts (total)
EZH2	910	3.79 (23 981)	21.96 (4144)
hnRNPK	46	2.9 (1585)	27.22 (169)
SUZ12	4348	1.84 (23 6673)	1.91 (22 7292)

The values are given as percentages to represent the ratio of consistent contacts for each experiment. See Supplementary Fig. S16 for further explanation.

The consistency of results from methods used to capture pairwise protein, RNA, and DNA interactions is a critical issue. This is particularly true for the RNA–chromatin interactions data generated by different protocols. Although unfiltered all-to-all data have been described as inconsistent [26], our findings suggest that there is a proportion of all-to-all data that is consistent between various protocols. The triad construction method we propose may extract these data.

Constructed triads are associated with genomic annotations

The RNA–chromatin interactome data are filtered by triad construction, which also selects RNA-associated PDt-contacts (see also Supplementary Tables S12–16). We analyzed the annotation of triads PDt-contacts by different chromatin states and compared them with the annotation of ChIP-seq peaks of the same proteins in the K562 cell line (Fig. 5A–C and F; for all proteins and SPIN annotation see Supplementary Figs S17 and S18). We used the χ^2 criterion to statistically assess the difference in enrichment (see Supplementary Table S18 for test results for all proteins, and Supplementary Table S19 for SPIN annotation).

For proteins that do not participate in regulating certain RNAs and their interactomes, it is reasonable to assume that the distribution of ChromHMM states in DNA regions will remain unchanged. This is because the process of constructing triads for these proteins essentially involves randomly sampling from ChIP-seq loci. However, this may not be the case for proteins involved in RNA–interactome reactions. For certain proteins, we can observe a notable enrichment pattern of states that correspond to the protein's functional association (e.g. PRC2 repressed chromatin for EZH2 and SUZ12 or transcribed chromatin for SRSF9 or U2AF2). This pattern is in addition to significant statistical test results. The distribution of the DNA parts of triads constructed on real data was compared to the distribution of the DNA parts of triads constructed on our simulated data (Fig. 5E and see Supplementary Figs S19–S22 for all proteins and SPIN annotation).

The proteins were divided into two groups based on the median P-value of the χ^2 goodness-of-fit test to determine if there were any differences in the characteristics of RDt-contacts mediated by the two groups. To address the bias in contact density in the RNA–chromatin interactome, the weight of the contact was defined as:

$$W_{\text{scaling}}(RDt_{\text{contact}}) = \frac{N(\text{bin}_i)}{L(\text{bin}_i) * \sum N(\text{bin}_i)} * 10^9, \qquad (1)$$

where $N(\text{bin}_i)$ represents the number of contacts of the target RNA falling into a bin, and $L(\text{bin}_i)$ is the bin length. To estimate the difference between two groups, we introduce a significance metric for triads as the harmonic mean of q-values, HMQ (2):

$$HMQ(Triad) = \frac{3}{\frac{1}{q_{Piranha}} + \frac{1}{q_{MACS2}} + \frac{1}{q_{BaRDIC}}}, \qquad (2)$$

a metric defined as the aggregated significance—q-values—of all pairwise peaks from which the interaction triad is assembled. We used HMQ to estimate importance weights of the contacts.

The comparison of contact weights and significance revealed differences in the significance distributions of the triads. Triads mediated by proteins significantly associated with the RNA–chromatin interactome showed greater significance. However, the distribution of contact scaling weights for the two groups was not different (Fig. 6A and B), indicating that while grouping proteins into two classes helps differentiate more significant contacts, the underlying nature of RDt-contacts present in both groups is the same. The proposed division of proteins based on their association with potential functions of mediating RNA–interactome regulation may be useful for investigating functional regulatory processes in the nucleus.

To determine the specificity of PR-contacts, we identified RNAs that form triad interactions with at least 20 proteins (Supplementary Fig. S23). We referred to these RNAs as the "common" RNAs, while the remaining RNAs were referred to as the "specific" RNAs. Next, we divided all RD-contacts from the original RD-dataset into four classes for each protein:

- The RNA in the contact is specific and the DNA part of the contact overlaps with at least one ChIP-seq locus of the protein.
- The RNA in the contact is common and the DNA part of the contact overlaps with at least one ChIP-seq locus of the protein.
- The RNA in the contact is specific and the DNA part of the contact does not overlap with any ChIP-seq locus of the protein.
- The RNA in the contact is common and the DNA part of the contact does not overlap with any ChIP-seq locus.

Fisher's exact test was applied to the contingency table obtained from these categories to assess the ability of each protein to form specific PR-contacts (Supplementary Table S18).

A group of 25 proteins associated with the RNA interactome was selected based on the consensus of two tests. RNAs forming triads with these proteins were also selected. Each RNA was represented as a vector of its interaction signals with these proteins, defined as the fold change of PR-contact. The vectors of both the RNAs and proteins were then clustered. The results indicate (Fig. 6C) that proteins cluster based on their functional roles as chromatin remodelers, transcription regulators, and RNA processing and splicing regulators. Similarly, selected RNAs are grouped by function, with snoRNA (small nucleolar RNAs) and lncRNA clustering separately.

The identified proteins associated with the RNA–chromatin interactome may have functionally relevant RNA–chromatin interactions. A detailed analysis of their contacts could prove useful.

RD-contacts precipitation on CTCF produces results similar to triads

The Red-ChIP experiment is of interest as it validates the interaction triads constructed for the CTCF protein in K562 cells. The experiment involves immunoprecipitation of RD-complexes on the protein. It is important to understand how the results of this experiment relate to the results obtained from constructing the triads on the Red-C experiment RD-contacts data. DNA parts from the following contact groups were correlated using StereoGene: RD-contacts from the Red-C experiment, input from the Red-ChIP (essentially another Red-C replicate), RD-contacts from those experiments after filtering the data with CTCF peaks, RD-contacts precipitated on CTCF (RedChIP Signal), and DNA parts of the interaction triads that were constructed. The consistency of the sets of interacting DNA loci is high between real and predicted CTCF-mediated RNA–DNA contacts (Fig. 7A).

It is also important to reiterate that the construction of triads does not ensure the completeness of the data obtained for

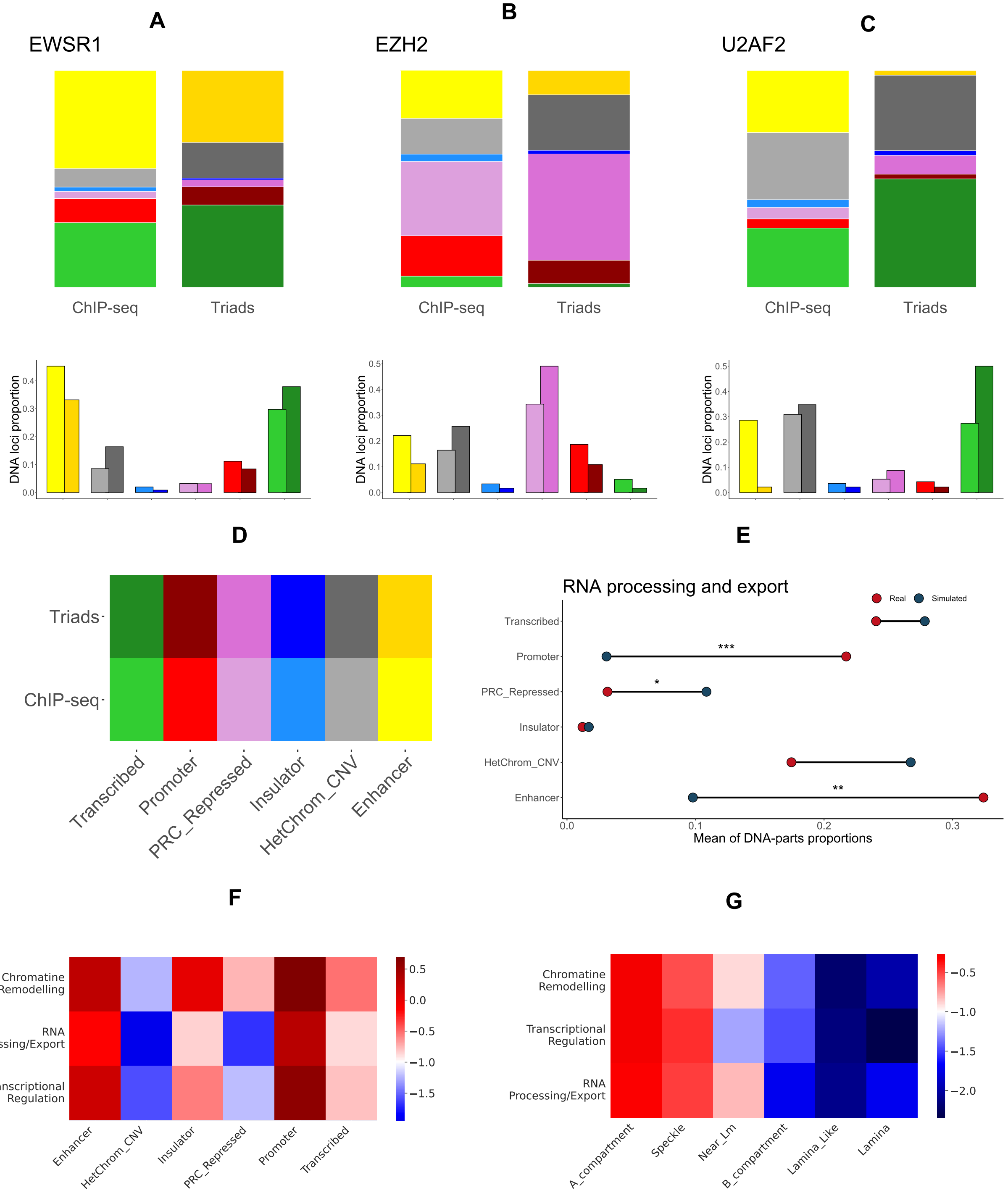

Figure 5. Triads and ChromHMM/SPIN genome annotation. (**A**–**C**) The distribution of DNA parts of triads mediated by EWSR1 (A), EZH2 (B), and U2AF2 (C) proteins across ChromHMM states is presented. (**D**) The figures are accompanied by a color legend. (**E**) The distribution of DNA parts of triads by ChromHMM states constructed on real and simulated DNA– and RNA–protein interaction data for RNA processing and exporting proteins. Asterisks denote statistical significance for the Wilcoxon test, with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. (**F** and **G**) Heatmaps for the distribution of DNA parts of protein triads, averaged by functional groups, over ChromHMM (F) and SPIN (G) annotations. $-\log_{10}$ of normalized densities is shown. The proportion of DNA parts in the annotation state is calculated by dividing the total length of the annotated regions in nucleotides by the total length of the sequence.

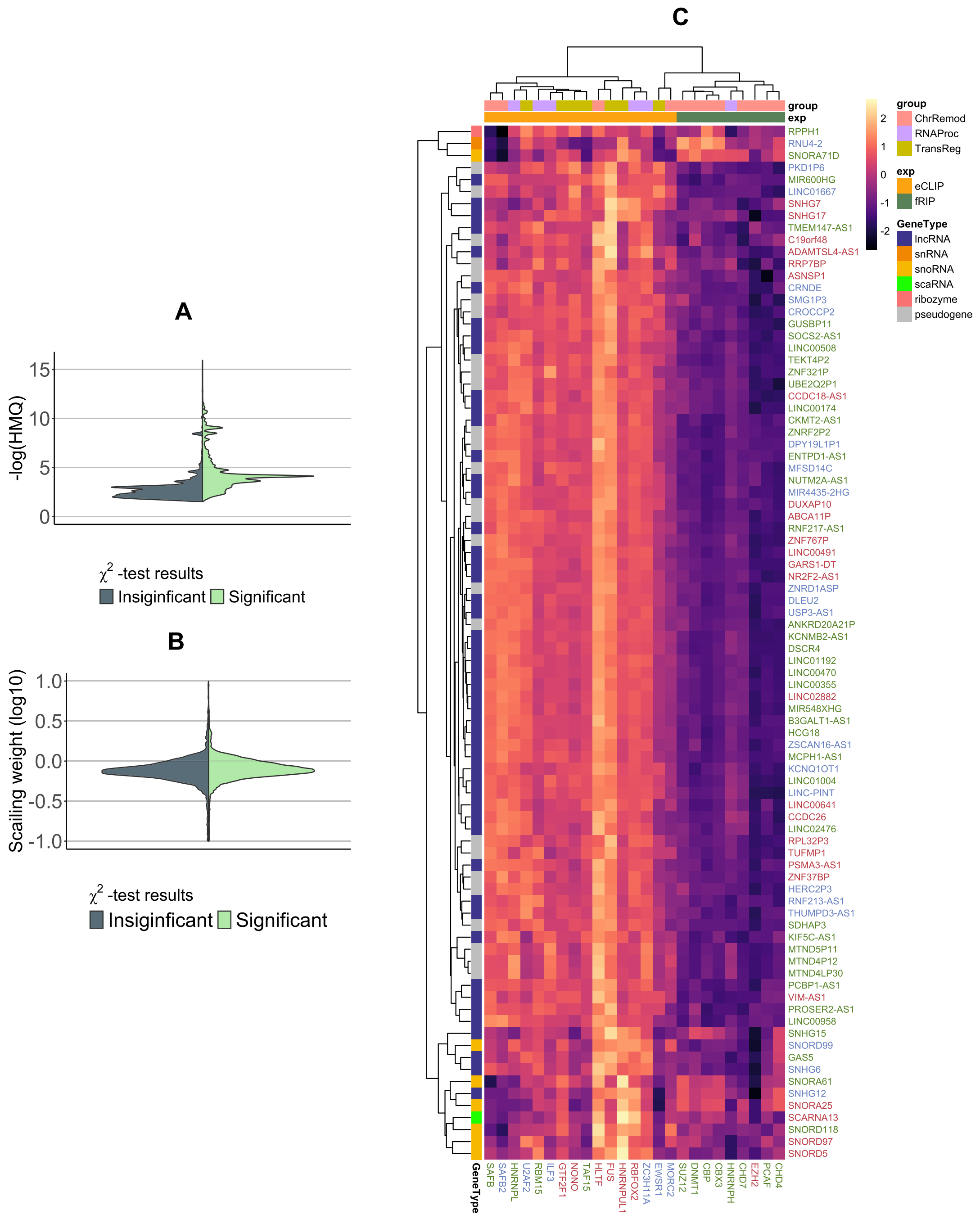

Figure 6. The association of proteins with RD-contacts. (**A**) Differences in HMQ significance between triads formed by proteins that were deemed significant by χ^2 goodness-of-fit test and those that were not. (**B**) Difference in scaling weights for mentioned groups of proteins. (**C**) A heatmap clusterisation of RNA and proteins by PR-contact fold changes.

CTCF. This is due to the initial data on pairwise interactions between RNA and chromatin being incomplete. There are numerous contacts lost in the Red-ChIP and Red-C data. However, by intersecting the incomplete data with each other when constructing triads, we ensure that we obtain more significant interactions than before. This allows us to reduce noise. It is also possible that, when constructing triads, we may lose real contacts, even if they were in the original data.

It is clear that the data filtered by ChIP-seq are well-clustered separately from the unfiltered RD-contacts of the Red-C experiment, alongside the Red-ChIP input data. The set of RNAs involved in triads interactions and in the Red-ChIP RD-contacts overlap well (Fig. 7B). These results suggest that the computationally predicted RD-contacts in triads represent a subset of real contacts, as the RD-contacts data from the Red-ChIP experiment also exhibit some level of noise.

Discussion

RNA–chromatin interactome data is known to be noisy in sequence bioinformatics and requires careful processing to filter out nonspecific and noisy contacts. We carefully studied and processed the data of pairwise interactions to ensure that only relevant and specific contacts were included. We used a

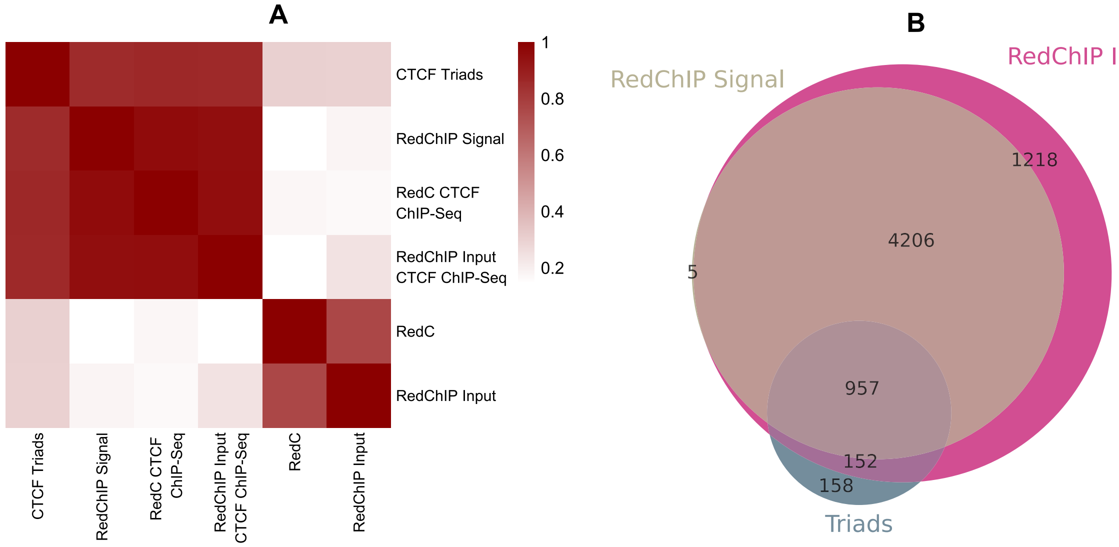

Figure 7. Red-ChIP data supports with predicted triads. (**A**) Stereogene correlations of DNA-parts of triads. (**B**) Shared RNAs between Red-C-based CTCF triads, Red-ChIP input data, and Red-ChIP signal data.

range of statistical peak calling methods designed for pairwise interactions between RNA, DNA, and proteins to search for functional contacts of the RNA–chromatin interactome mediated by one of the 47 chromatin proteins, using data of PR- and PD-interactions. Our approach enables us to choose a significance level for each level of pairwise interactions used to construct the object of study based on an accumulated significance metric. This significance selection also applies to triads that have already been constructed.

Despite the high levels of noise and large number of *cis*-contacts and contacts formed by protein-coding RNA in the RD-contacts data, our approach appears to effectively filter out this flawed data. Filtering out too much data can have drawbacks, and it is important to consider statistical assessment of thresholds to work with the data accurately.

The obtained data are meaningfully different from those obtained by chance and contain less noise than the original RNA–chromatin contacts. There are (Supplementary Table S20 and Supplementary Note S2) a significant number of conserved triads between mouse and human organisms, and the functional meaning of these interactions remains a question.

We demonstrated that RADICL-seq NPM phase data can serve as a negative control for the triads construction method, which is a protocol for nonprotein-mediated RNA–chromatin contacts. Although the nature of these NPM data is not fully understood, they are useful when searching for RDt-contacts mediated by a particular protein, rather than determining the genome-wide RNA–chromatin interactome. It is demonstrated here that the correlation between RDt-contacts mediated by different proteins decreases when the contacts shared between RADICL 2FA and RADICL NPM sets are filtered out.

We investigated whether there are differences in the constructed triads when starting with data from different pairwise interaction detection protocols for PR-interactions (fRIP-seq and eCLIP-seq) and for RD-interactions (namely, GRID-seq and RADICL-seq). Triads built on different datasets appear consistent for PR-contacts and almost consistent for RD-contacts. A range of regions can be identified that is conserved between experiments. To draw more data, other types of pairwise interactions need to be incorporated into the analysis. It might be valuable to see if PD-contacts originating from different experiments produce consistent triads for target proteins.

We demonstrated that constructing interaction triads for chromatin proteins enhances RNA–chromatin contacts in regions where DNA loci functional annotation inferred from ChromHMM is associated with protein function. For instance, EZH2-mediated triads DNA-parts are more abundant in PRC2-repressed chromatin regions than the ChIP-seq loci of this protein. In addition, we verified that the randomly simulated triads were not related to protein function, in the sense of ChromHMM. Triads with a DNA locus located in chromatin states that correspond to protein functions are of interest for studies on regulatory interactions. Additionally, we observed that the enrichment of DNA parts in triads is higher in open chromatin states such as Promoter, Enhancer, and Transcribed states from ChromHMM annotation or A-compartment and Speckle states from SPIN annotation compared to other states. The annotation of A/B-compartment of triads is consistent between RADICL-seq and GRID-seq experiments, supporting the hypothesis that open chromatin–RNA contacts are better represented in the interactome data.

In addition, the interaction triads constructed are consistent with the results of the Red-ChIP experiment, which serves as a positive control for our approach. While the Red-ChIP experimental procedure allows for the presence of contacts of noisy origin, the number of RNAs shared between the set of RD-contacts of Red-ChIP and RDt-contacts is high, and the set of DNA loci with which these RNAs interact is also shared.

It has been demonstrated that the proposed approach is effective. The main challenge in identifying RNA–chromatin interactions mediated by intranuclear proteins is the lack of relevant data. In this study, we analyzed various biological data sources, including ENCODE, NCBI GEO, ReMap, 4D Nucleome, and GTRD databases, and identified a dataset of triads of 47 proteins as successful. To scale this approach, machine learning techniques can be used for data imputation.

Acknowledgements

The authors thank Asya Mendelevich for their valuable comments on manuscript.

Author contributions: D.A.K. was responsible for data curation, formal analysis, investigation, methodology, software development, part of visualization, writing the original draft, and editing the manuscript. A.I.N. processed the Red-ChIP data for the analysis. A.A.Z. contributed to visualization. A.A.M. contributed to conceptualization, supervision, and funding acquisition. All authors contributed to the review process at each stage of the manuscript's development.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

The research was supported by the Russian Science Foundation [Grant/Award Number: 23-14-00136].

Data availability

The data that support the findings of this study are available from the corresponding author, D.A.K., upon reasonable request.

References

1. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* 2016;**17**:47–62. https://doi.org/10.1038/nrg.2015.10
2. Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat Rev Mol Cell Biol* 2016;**17**:756–70. https://doi.org/10.1038/nrm.2016.126
3. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011;**12**:861–74. https://doi.org/10.1038/nrg3074
4. Engreitz JM, Pandya-Jones A, McDonel P *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 2013;**341**:1237973. https://doi.org/10.1126/science.1237973
5. Simon MD, Wang CI, Kharchenko PV *et al.* The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci USA* 2011;**108**:20497–502. https://doi.org/10.1073/pnas.1113536108
6. Chu C, Qu K, Zhong FL *et al.* Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. *Mol Cell* 2011;**44**:667–78. https://doi.org/10.1016/j.molcel.2011.08.027
7. Yan Z, Huang N, Wu W *et al.* Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proc Natl Acad Sci USA* 2019;**116**:3328–37. https://doi.org/10.1073/pnas.1819788116
8. Oh J.-M, Venters CC, Di C *et al.* U1 snRNP regulates cancer cell migration and invasion in vitro. *Nat Commun* 2020;**11**:1. https://doi.org/10.1038/s41467-019-13993-7
9. Bonetti A, Agostini F, Suzuki AM *et al.* RADICL-seq identifies general and cell type-specific principles of genome-wide RNA–chromatin interactions. *Nat Commun* 2020;**11**:1018. https://doi.org/10.1038/s41467-020-14337-6
10. Li X, Zhou B, Chen L *et al.* GRID-seq reveals the global RNA–chromatin interactome. *Nat Biotechnol* 2017;**35**:940–50. https://doi.org/10.1038/nbt.3968
11. Li L, Luo H, Lim DH *et al.* Global profiling of RNA–chromatin interactions reveals co-regulatory gene expression networks in Arabidopsis. *Nat Plants* 2021;**7**:1364–78. https://doi.org/10.1038/s41477-021-01004-x
12. Li J, Xiang Y, Zhang L *et al.* Enhancer–promoter interaction maps provide insights into skeletal muscle-related traits in pig genome. *BMC Biol* 2022;**20**:136. https://doi.org/10.1186/s12915-022-01322-2
13. Gavrilov AA, Zharikova AA, Galitsyna AA *et al.* Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res* 2020;**48**:6699–714. https://doi.org/10.1093/nar/gkaa457
14. Ryabykh GK, Kuznetsov SV, Korostelev YD *et al.* RNA-Chrom: a manually curated analytical database of RNA-chromatin interactome. *Database (Oxford)* 2023;**2023**:baad025. https://doi.org/10.1093/database/baad025
15. Zhang Y, Liu T, Meyer CA *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137. https://doi.org/10.1186/gb-2008-9-9-r137
16. Mylarshchikov DE, Nikolskaya AI, Bogomaz OD *et al.* BaRDIC: robust peak calling for RNA–DNA interaction data. *NAR Genom Bioinform* 2024;**6**:lqae054. https://doi.org/10.1093/nargab/lqae054
17. Jankowsky E, Harris ME. Specificity and nonspecificity in RNA–protein interactions. *Nat Rev Mol Cell Biol* 2015;**16**:533–44. https://doi.org/10.1038/nrm4032
18. Kristofich J, Nicchitta CV. Signal-noise metrics for RNA binding protein identification reveal broad spectrum protein–RNA interaction frequencies and dynamics. *Nat Commun* 2023;**14**:5868. https://doi.org/10.1038/s41467-023-41284-9
19. Gavrilov AA, Sultanov RI, Magnitov MD *et al.* RedChIP identifies noncoding RNAs associated with genomic sites occupied by Polycomb and CTCF proteins. *Proc Natl Acad Sci USA* 2022;**119**:e2116222119. https://doi.org/10.1073/pnas.2116222119
20. Xiao Q, Huang X, Zhang Y *et al.* The landscape of promoter-centred RNA–DNA interactions in rice. *Nat Plants* 2022;**8**:157–70. https://doi.org/10.1038/s41477-021-01089-4
21. Johnson DS, Mortazavi A, Myers RM *et al.* Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* 2007;**316**:1497–502. https://doi.org/10.1126/science.1141319
22. Zhao J, Ohsumi TK, Kung JT *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* 2010;**40**:939–53. https://doi.org/10.1016/j.molcel.2010.12.011
23. G Hendrickson D, Kelley DR, Tenen D *et al.* Widespread RNA binding by chromatin-associated proteins. *Genome Biol* 2016;**17**:28. https://doi.org/10.1186/s13059-016-0878-3
24. Van Nostrand EL, Pratt GA, Shishkin AA *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 2016;**13**:508–14. https://doi.org/10.1038/nmeth.3810
25. Martin G, Zavolan M. Redesigning CLIP for efficiency, accuracy and speed. *Nat Methods* 2016;**13**:482–3. https://doi.org/10.1038/nmeth.3870
26. Ryabykh G, Vasilyev A, Garkul L *et al.* Comparative analysis of the RNA-chromatin interactions data. Completeness and accuracy. bioRxiv, https://doi.org/10.1101/2023.09.21.558854, 23 September 2023, preprint: not peer reviewed.
27. Hammal F, de Langen P, Bergon A *et al.* ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* 2021;**50**:D316–25. https://doi.org/10.1093/nar/gkab996
28. Kolmykov S, Yevshin I, Kulyashov M *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* 2020;**49**:D104–11. https://doi.org/10.1093/nar/gkaa1057

29. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15. https://doi.org/10.1038/s41587-019-0201-4
30. Uren PJ, Bahrami-Samani E, Burns SC *et al.* Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* 2012;**28**:3013–20. https://doi.org/10.1093/bioinformatics/bts569
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033
32. Krzywinski M, Schein J, Birol I *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;**19**:1639–45. https://doi.org/10.1101/gr.092759.109
33. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 2017;**12**:2478–92. https://doi.org/10.1038/nprot.2017.124
34. Pintacuda G, Wei G, Roustan C *et al.* HnRNPK recruits PCGF3/5-PRC1 to the Xist RNA B-repeat to establish polycomb-mediated chromosomal silencing. *Mol Cell* 2017;**68**:955–969. https://doi.org/10.1016/j.molcel.2017.11.013
35. Wang Y, Zhang Y, Zhang R *et al.* SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biol* 2021;**22**:36. https://doi.org/10.1186/s13059-020-02253-3
36. Mylarshchikov DE, Mironov AA. ortho2align: a sensitive approach for searching for orthologues of novel lncRNAs. *BMC Bioinform* 2022;**23**:384. https://doi.org/10.1186/s12859-022-04929-y
37. Hinrichs AS. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006;**34**:D590–8. https://doi.org/10.1093/nar/gkj144
38. Stavrovskaya ED, Niranjan T, Fertig EJ *et al.* StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data. *Bioinformatics* 2017;**33**:3158–65. https://doi.org/10.1093/bioinformatics/btx379

Received: May 8, 2024. **Revised:** January 22, 2025. **Editorial Decision:** February 3, 2025. **Accepted:** February 5, 2025