ISSN 0006-2979, Biochemistry (Moscow), 2025, Vol. 90, No. 11, pp. 1698-1710 © Pleiades Publishing, Ltd., 2025.
Russian Text © The Author(s), 2025, published in Biokhimiya, 2025, Vol. 90, No. 11, pp. 1816-1829.

Comparative Analysis of RNA-Chromatin Interactome Data: Resolution, Completeness, and Specificity

Grigory K. Ryabykh[1,2,a*], Arina I. Nikolskaya[1,2], Lidia D. Garkul[1], and Andrey A. Mironov[1,2]

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119234 Moscow, Russia*
[2] *Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991 Moscow, Russia*
[a] *e-mail: ryabykhgrigory@gmail.com*

Received July 3, 2025
Revised October 25, 2025
Accepted October 28, 2025

Abstract—Two types of experiments are used to study RNA-chromatin interactions: the interactome search for individual RNAs (“one-to-all” or OTA) and genome-wide contact mapping for all RNAs (“all-to-all” or ATA). Comparative analysis of ATA and OTA data revealed fundamental differences in resolution, completeness, and specificity. OTA data exhibit high resolution (~1000 bp) and reproducibility (>90%), serving as a “gold standard”. ATA data, however, have lower resolution (~5000 bp), and their reproducibility (<10%) is critically dependent on the protocol, with two-step fixation using disuccinimidyl glutarate and formaldehyde (GRID-seq) showing a clear advantage over formaldehyde alone. The introduced “chromatin potential” metric and BaRDIC peak filtering effectively isolate the specific signal. This study proposes a strategy for reliable interactome analysis: combining RNA selection based on chromatin potential with the use of concordant contacts from peaks.

DOI: 10.1134/S0006297925601923

Keywords: RNA-chromatin interactome, non-coding RNAs, RNA-Chrom database, RNA-chromatin interactome data concordance

INTRODUCTION

Non-coding RNAs (ncRNAs) in animals and plants are involved in a wide range of biological processes, including cell differentiation, gene expression regulation, chromatin remodeling, chromatin structure maintenance, splicing, RNA processing, and biomolecular condensate formation. Disruptions in the ncRNA-mediated regulatory pathways are associated with the development of various diseases, emphasizing importance of understanding their mechanisms of action [1]. A significant portion of ncRNA functions is realized in the cell nucleus, necessitating a detailed study of the RNA-chromatin interactome.

RNA molecules interact with numerous proteins, chromatin, and other RNAs. Experimental methods to identify DNA loci in contact with ncRNAs can be divided into two groups: “one-to-all” (OTA) and “all-to-all” (ATA). The first group (RAP [2], CHART-seq [3], ChIRP-seq [4], dChIRP-seq [5], ChOP-seq [6], CHIRT-seq [7]) identifies contacts of a known RNA with chromatin, while the second group (MARGI [8], GRID-seq [9], ChAR-seq [10, 11], iMARGI [12], RADICL-seq [13], Red-C [14]) aims to determine all possible RNA-DNA contacts in the cell [15].

Both groups of methods are actively used in research, but typically not in combination thus preventing development of the unified standards to enhance reliability and significance of the RNA-chromatin interactome data analysis. To date, there has been no systematic comparison of ATA and OTA data in terms of key characteristics such as accuracy, completeness, and specificity.

Despite the rapid development of these technologies, the resulting data are characterized by

* To whom correspondence should be addressed.

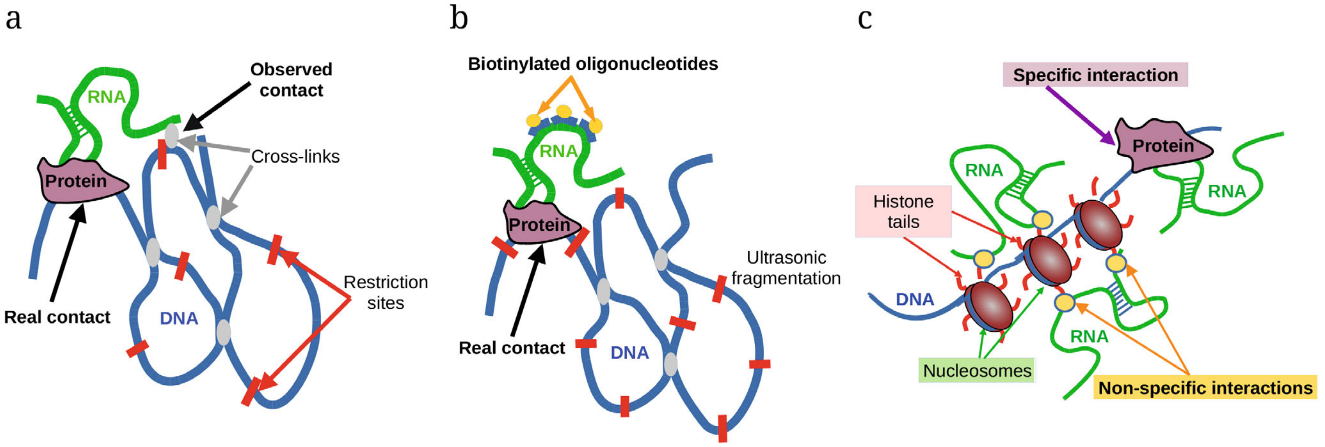

Fig. 1. Accuracy of determining position of the real contact differs in ATA and OTA protocols. a) Source of position bias in ATA data – chromatin structure. b) In OTA, the observed contact position offset from the real position is determined only by the size of DNA fragments. c) Possible source of non-specific interactions.

significant methodological issues and systematic biases. First, the density of RNA contacts depends on the distance between the RNA source gene and the target DNA loci on the same chromosome [9, 11-14, 16]. This bias, termed “RNA-DNA scaling” (RD-scaling), is analogous to scaling in the DNA-DNA interactome data (Hi-C method) [17]. Second, chromatin accessibility significantly influences the data, referred to as “background.” Background is assessed using “input” data in OTA experiments or contacts of the protein-coding RNAs (mRNAs) in the ATA experiments [9]. Additionally, these experiments inherently have limited precision in determining contacts. In the ATA experiments, RNA crosslinking with chromatin can occur at a distance from the real contact (Fig. 1a), whereas in the OTA experiments, precision of the contact position determination depends only on the size of DNA fragments (Fig. 1b). Presence of non-specific interactions poses a particular problem. A significant portion of the observed contacts could be explained by electrostatic attraction between the negatively charged RNA and positively charged histone tails, as well as by preferential crosslinking of amino groups present on the lysines and arginines of histones by formaldehyde [18]. Although affinity of such non-specific interactions is relatively low, their cumulative contribution is substantial due to the large number of potential binding sites (Fig. 1c). On the other hand, technical limitations of the existing experimental methods result in the loss of some true contacts. These factors collectively raise questions about the specificity of the detected interactions and accuracy, completeness, and specificity of the RNA-chromatin interactome data.

The aim of this study is a systematic comparative analysis of data obtained by OTA and ATA methods to assess their accuracy, completeness, and specificity. The following objectives were addressed:

- development of metrics for assessing interaction specificity (chromatin potential, chP) and data reproducibility (concordance);
- comparative analysis of replicate consistency within each method;
- cross-validation of data obtained by different methods;
- development of recommendations for improving the reliability of RNA-chromatin interactome analysis.

MATERIALS AND METHODS

Data. Human and mouse RNA-chromatin interactome data were obtained from the RNA-Chrom database [19]. Only ATA data with corresponding RNA-seq data from the same cell line were used. When more than two replicates were available in the ATA data, the two most complete replicates were selected. RNA-seq data were obtained from the GEO database and processed similarly to the ATA data processing procedure described in the RNA-Chrom. List of the used data is provided in Tables S1 and S2 in the Online Resource 1. Only RNAs demonstrating more than 1000 contacts with chromatin in each replicate were included in analysis to ensure sufficient statistical power for identifying “peaks” (genomic regions enriched with RNA-chromatin contacts) using the BaRDIC program [20]. Ribosomal RNAs were excluded from the analysis. For example, applying this filter in the “RADICL, ES (NPM)” and “RADICL, ES (ActD)” experiments left fewer than 1000 RNAs and less than 50% of contacts from the initial size of the selected replicates (Fig. S1, a and b in the Online Resource 1). Considering significant overrepresentation of the proximal contacts (RD-scaling), interactions located within 1 Mb of the genes encoding the corresponding RNAs were excluded from further analysis in this study.

Use of BaRDIC and threshold selection. Like most genome-wide data, RNA-chromatin interactome data are characterized by high level of non-specific signals ("noise"). Specialized peak-calling algorithms are used to identify significant interactions by detecting statistically significant clusters of interactions in the specific genomic loci.

In this study, we used the BaRDIC algorithm [20], which accounts for RD-scaling and chromatin openness. This algorithm uses a probabilistic estimate of the likelihood that the contacts in a chromatin locus belong to a peak or noise. The Benjamini–Hochberg multiple testing correction (FDR, false discovery rate) is next applied to control proportion of false positives based on the background distribution. However, in our case, significant overlap of signal and noise distributions leads to the loss of a substantial proportion of true interactions when using a strict FDR threshold. To avoid this problem, we used a flexible selection criterion: for each RNA, we selected top 10% of the peaks with the lowest FDR. Since peak sizes could reach tens of kilobases due to the data sparsity, all comparisons were conducted at the level of individual contacts intersecting these peaks.

For the ATA data analysis, BaRDIC was run with default parameters. For the OTA experiments, which have better contact coverage, the following parameters were set: *--trans_min* 400 bp; *--cis_start* 100 bp; *--trans_step* 50 bp. Background was calculated using input data converted to BedGraph, with a window size of 1000 bp.

Chromatin potential. In nearly all studies involving ATA experiments, it has been noted that the number of RNA contacts with chromatin linearly depends on the expression level of the corresponding RNA [9-11, 13, 14]. Normalization by expression level allows us to identify RNAs that demonstrate an increased tendency to interact with chromatin, i.e. the molecules with contact frequency significantly exceeding what would be expected at a given expression level.

To assess tendency of RNAs to contact chromatin, we introduce the concept of "chromatin potential." Let us consider RNAs that have more than 1000 contacts in each replicate. Let N_c be the total number of contacts of selected RNAs in the ATA experiment accounting for the RD-scaling filter; N_e be the total number of uniquely mapped and gene-annotated reads in the RNA-seq experiment; $n_c{}^i$ be the number of contacts accounting for the RD-scaling filter of a specific *i*-th RNA in the ATA experiment; and $n_e{}^i$ be the number of reads of a specific *i*-th RNA in the RNA-seq experiment. To compare these observations, we apply a Z-test for proportions. For each *i*-th RNA, we calculate the Z-statistic (Z_i) (1):

$$Z_i = \frac{p_c^i - p_e^i}{\sqrt{p_i(1-p_i)\,(1/N_c + 1/N_e)}};$$
$$p_c^i = \frac{n_c^i}{N_c};\ p_e^i = \frac{n_e^i}{N_e};\ p_i = \frac{n_c^i + n_e^i}{N_c + N_e}. \quad (1)$$

The Z-statistic follows a standard normal distribution, allowing us to estimate *p*-value and the Benjamini–Hochberg FDR. We refer to the Z-statistic value as the chromatin potential. Chromatin potential surpasses the simple ratio of the number of contacts to the expression level because it accounts for the statistical significance of the deviation. The ratio of the number of contacts to the expression level is heavily biased toward RNAs with low coverage in the RNA-seq data, where the denominator (expression level) is estimated with significant error, leading to a wide scatter of values (up to six orders of magnitude; see Fig. S2 in the Online Resource 1).

However, the following circumstances must be considered. First, strand-specific total RNA sequencing with rRNA depletion is required for such analysis. Second, this analysis is applicable only to long RNAs, as standard RNA-seq data do not allow for an adequate assessment of the expression level of RNAs shorter than 100 nucleotides [21].

RESULTS

Chromatin potential. In all genome-wide studies of RNA-chromatin interactions (ATA experiments), there is a significant predominance of mRNA contacts. This is because mRNAs generally have higher expression levels compared to ncRNAs. Chromatin potential (chP) addresses the question of whether the proportion of contacts for a given RNA is statistically significantly different from what would be expected if all RNAs contacted chromatin non-specifically and proportionally to their expression levels. If most mRNA contacts with chromatin are non-specific, it can be expected that ncRNAs will demonstrate higher affinity for chromatin. As expected, most ncRNAs exhibited a chromatin potential greater than zero (Fig. 2a; Fig. S3 in the Online Resource 1), but a large number of mRNAs also had a positive chromatin potential. As the chP threshold increased, the proportion of mRNAs among the RNAs passing the threshold decreased (Fig. 2b; Table S3 in the Online Resource 1) with a sharp drop in almost all experiments at chP ≥ 20.

The fact that even at high chromatin potential thresholds, a significant number of protein-coding RNAs remain may be due to several factors.

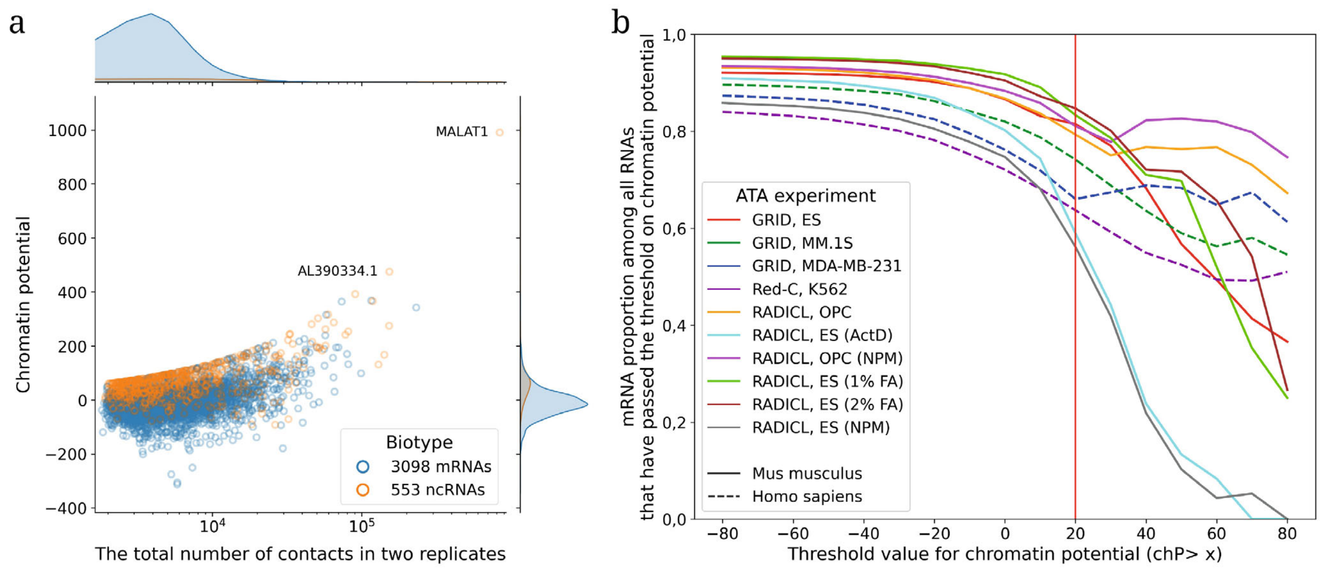

Fig. 2. Characteristics of RNA chromatin potential. a) Dependence of chromatin potential on the number of RNA contacts in the Red-C K562 experiment. Blue – protein-coding RNAs, orange – non-coding RNAs. b) Proportion of mRNAs depending on the chromatin potential threshold (chP > x) for different ATA experiments. The proportion of ncRNAs corresponds to 1 minus the proportion of mRNAs. ActD – actinomycin D treatment; NPM – proteinase K treatment; 1% FA – fixation with 1% formaldehyde; 2% FA – fixation with 2% formaldehyde.

For example, some protein-coding genes contain functional ncRNAs in their intronic regions [22, 23], among which a significant number of unannotated ncRNAs can be expected. Positive chromatin potential of some mRNAs may be associated with these ncRNAs. On the other hand, non-coding isoforms of mRNAs may themselves play a role in chromatin regulation [24].

Comparison of replicates in ATA data. To assess consistency of the RNA-chromatin interactions between replicates, we evaluated proportion of the reproducible contacts. Since the exact contact coordinate in the ATA methods could be shifted due to the protocol specifics, we introduced a genomic distance parameter (L), within which contacts belonging to the same RNA but detected in different replicates were considered concordant. To determine the L threshold that adequately reflects the method's resolution, we calculated, for each RNA, proportion of its contacts for which at least one contact of the same RNA was detected in another replicate within the specified distance L. Analysis of this proportions dependence on L for the "GRID, ES, *Mus musculus*" data showed that the median proportion of concordant contacts ceased to increase significantly at $L \geq 5000$ bp (Fig. S4 in the Online Resource 1), reaching a plateau. This indicates that 5000 bp is an empirical estimate of contact positioning accuracy in the ATA methods. Based on this result, for subsequent analysis, we divided the genome into non-overlapping fragments (bins) of a fixed size (*bin* bp). The main analysis was conducted with the bin size of 5000 bp, corresponding to the empirically estimated positioning accuracy. To test robustness of the results and simulate a "high-resolution" scenario, a bin size of 1000 bp was also used. A bin was considered concordant for a given RNA if at least one contact in that bin was detected in both replicates, and discordant if contacts were present in only one replicate. This approach allows data aggregation and quantitative assessment of interaction reproducibility at the level of genomic loci.

To assess randomness of the matches, we used a simple model. Assuming that, all RNAs contact genomic DNA uniformly, the probability of at least one contact falling into a bin in one experiment can be estimated as $p_{bin}{}^{e}(i) = n_i{}^{e}/N_{bin}$, where i is the RNA index; e is the experiment (replicate) number; $n_i{}^{e}$ is the number of bins with contacts of the i-th RNA; N_{bin} is the total number of bins into which the corresponding genome was divided. Here, we neglect biases in the data, particularly chromatin accessibility, and assume that the bin size is sufficiently small. To avoid the influence of RD-scaling, we selected bins located more than 1 Mb away from the gene source of the i-th RNA. The probability that contacts of the i-th RNA from two experiments (a and b) fall into the same bin is $p_{bin}(i) = p_{bin}{}^{a}(i)\cdot p_{bin}{}^{b}(i)$. A rough estimate of the probability of observing k matching bins can be made using the Bernoulli distribution (2):

$$P^{i}_{obs} = C^{k}_{N_{bin}} p_{bin}(i)^{k} \cdot (1 - p_{bin}(i))^{N_{bin}-k}. \qquad (2)$$

This allows for a probabilistic assessment of the correspondence between the replicates or experiments. We define $\lambda(i) = (n_i{}^{a} n_i{}^{b}/N_{bin})$. For $\lambda \geq 10$, a normal approximation can be used to estimate the probability of such event assuming that replicates have independent contacts (3):

$$P^{i}_{obs} \simeq \mathcal{N}\left(\lambda, \sqrt{\lambda}\right);\ \ P(X \geq k) \approx 1 - \Phi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right). \quad (3)$$

For λ < 10, a Poisson approximation is used (4):

$$P(X \geq k) \approx 1 - F_{Poisson}(k-1, \lambda). \quad (4)$$

To assess consistency of the replicates in the ATA experiments, we analyzed a set of RNAs with more than 1000 contacts with chromatin in each replicate. When selecting RNAs, the filter to exclude RD-scaling regions (within 1 Mb of the RNA source gene) was not applied; it was applied when selecting contacts. Importantly, we assessed not just presence of at least one concordant bin but statistical significance of the overall level of concordance for each RNA as a whole. For this, we counted total number of the concordant and discordant bins for each RNA and applied a statistical criterion to test the hypothesis of non-randomness of the observed level of matches (see above for details). An RNA was considered concordant if its calculated FDR was less than 0.05. As seen in Fig. S5 in the Online Resource 1, presence of the individual concordant bins does not guarantee passing this strict significance threshold.

Table 1 and Fig. S6 in the Online Resource 1 show the number of RNAs with concordant bins between the replicates with FDR < 0.05. The analysis was conducted under four conditions to assess the influence of two factors: genome bin size (1000 bp vs. 5000 bp) and contact filtering (all contacts vs. contacts from BaRDIC peaks). For the GRID experiments, neither of these conditions had an effect: the number of concordant RNAs remained unchanged and almost always equal to the initial number of selected RNAs (discussion below).

For other ATA data, as expected, when using all contacts, the number of statistically significantly

Table 1. Number of RNAs with concordant bins in the replicates (FDR < 0.05)

Experiment	Initial number of mRNAs (ncRNAs)	Number of concordant mRNAs (ncRNAs), all contacts		Number of concordant mRNAs (ncRNAs), contacts from peaks	
		Bin 1000 bp	Bin 5000 bp	Bin 1000 bp	Bin 5000 bp
Red-C, K562, *H. sapiens*	3230 (636)	1571 (341)	2418 (486)	2779 (556)	3188 (628)
GRID, MM.1S, *H. sapiens*	3771 (413)	3771 (413)	3771 (413)	3771 (413)	3771 (413)
GRID, MDA_MB_231, *H. sapiens*	4844 (653)	4844 (653)	4844 (653)	4844 (653)	4844 (653)
GRID, ES, *M. musculus*	4706 (436)	4706 (429)	4706 (427)	4706 (432)	4706 (435)
RADICL (2% FA), ES, *M. musculus*	2758 (162)	1829 (87)	2552 (124)	2226 (131)	2704 (158)
RADICL, OPC, *M. musculus*	2580 (197)	1954 (136)	2484 (175)	2203 (161)	2555 (191)
RADICL (ActD), ES, *M. musculus*	657 (87)	345 (42)	576 (76)	512 (74)	646 (86)
RADICL (NPM), OPC, *M. musculus*	3734 (275)	504 (40)	1464 (103)	2128 (136)	2839 (200)
RADICL (1% FA), ES, *M. musculus*	2079 (117)	1533 (66)	1986 (80)	1811 (102)	2056 (115)
RADICL (NPM), ES, *M. musculus*	643 (149)	643 (149)	643 (149)	643 (148)	643 (148)

Note. Only RNAs with more than 1000 contacts in each replicate were selected. A filter was applied to exclude contacts within 1 Mb of the RNA source gene.

concordant RNAs was substantially lower under the strict condition (bin size = 1000 bp) compared to the condition corresponding to the method's resolution (bin size = 5000 bp). This confirms that a larger bin better aggregates technical variations and more accurately reflects interaction reproducibility. The most important observation was that preliminary selection of the contacts belonging to BaRDIC peaks significantly increased replicate consistency. This filtering either increased the number of concordant RNAs or allowed achieving a comparable level of concordance even when using a strict bin size of 1000 bp compared to analyzing all contacts with the bin size of 5000 bp. Thus, identifying RNA-chromatin interaction peaks using BaRDIC effectively filters out random interactions and highlights the most reliable, reproducible RNA-chromatin contacts, significantly increasing consistency between the replicates.

After identifying statistically significant concordant RNAs, we assessed completeness of the ATA data by calculating the median proportion of contacts that fall into concordant 5000-bp bins, which corresponds to the estimated resolution of the ATA methods. This metric reflects proportion of interactions reproducible between the replicates out of the total number of detected contacts (Tables S4 and S5 in the Online Resource 1).

Fundamental differences between the methods were identified. Completeness of the data for the Red-C and RADICL-seq did not exceed 2% when analyzing all contacts and 5% for the contacts filtered by the BaRDIC peaks. In contrast, completeness of the GRID-seq data was significantly higher, reaching 29% and 82% for all contacts and contacts from the peaks, respectively.

High reproducibility of the GRID data could likely be explained by the specifics of the fixation protocol. Unlike the methods that use only formaldehyde (such as Red-C and RADICL-seq), the GRID-seq protocol employs a two-step fixation with disuccinimidyl glutarate (DSG) and formaldehyde. DSG is a crosslinking agent with a long spacer (7.7 Å), which effectively crosslinks protein-protein interactions, stabilizing protein complexes before the chromatin structure is fixed with formaldehyde [25]. This allows for more effective "sealing" of protein-mediated RNA-chromatin interactions, which constitute majority of the specific contacts.

This hypothesis is quantitatively supported. Despite the comparability of the medians of the total number of contacts of concordant RNAs across all ATA data (Fig. S7a in the Online Resource 1), the median number of reproducible (concordant) contacts in the GRID-seq data was an order of magnitude higher than the corresponding indicators for the Red-C and RADICL-seq data (Fig. S7b in the Online Resource 1). This indicates that the DSG protocol not only increases the volume of data but fundamentally enhances proportion of the specific, reproducible signals in the overall dataset. Thus, the protocol with additional DSG treatment ensures more complete and stable capture of multiprotein complexes, leading to the significant reduction in technical noise and increased reproducibility between the replicates. Meanwhile, fixation with formaldehyde alone may inadequately stabilize large supramolecular complexes, which, in turn, increases the proportion of random, unstable interactions and reduces overall concordance.

Despite the fundamental differences in the absolute level of concordance between the methods, we found a common pattern for all ATA experiments: reproducibility of contacts positively correlates with the total number of RNA interactions and its chromatin potential (Fig. 3; Fig. S8 in the Online Resource 1). The discovered dependency allows us to draw two important conclusions about the nature of RNA-chromatin interactome data:

1. Completeness of the data is a function of sequencing depth for a specific RNA. Low reproducibility of RNAs with a small number of contacts (<10,000) indicates that for such molecules, the data are substantially incomplete and contain a high level of noise. Sufficient completeness is achieved only with large number of interactions, indicating the need for deep sequencing to reliably identify interactome of the individual RNAs.

2. Reproducibility is a marker of biological significance. Positive correlation between the chromatin potential and the proportion of concordant contacts suggests that the more specific is the interaction (higher chromatin potential), the more stable and reproducible it is between replicates. This strengthens the position of chromatin potential not only as a measure of specificity but also as a predictor of reliability and reproducibility of interactions.

It is also worth noting that the median proportions of concordance of contacts between mRNAs and ncRNAs were practically indistinguishable (Tables S4 and S5, columns 1 and 4 in the Online Resource 1), indicating that reproducibility does not depend on the RNA biotype. To interpret the unexpectedly high level of concordance of the mRNA contacts, comparable to that of ncRNAs, two non-exclusive hypotheses have been proposed:

1. Existence of non-specific but statistically reproducible interactions, where electrostatic or other weak forces may lead to massive yet stable binding of RNA to chromatin.

2. Presence of unknown specific functions in some mRNAs related to direct interaction with chromatin (e.g., mediated by non-coding isoforms or unannotated intronic ncRNAs).

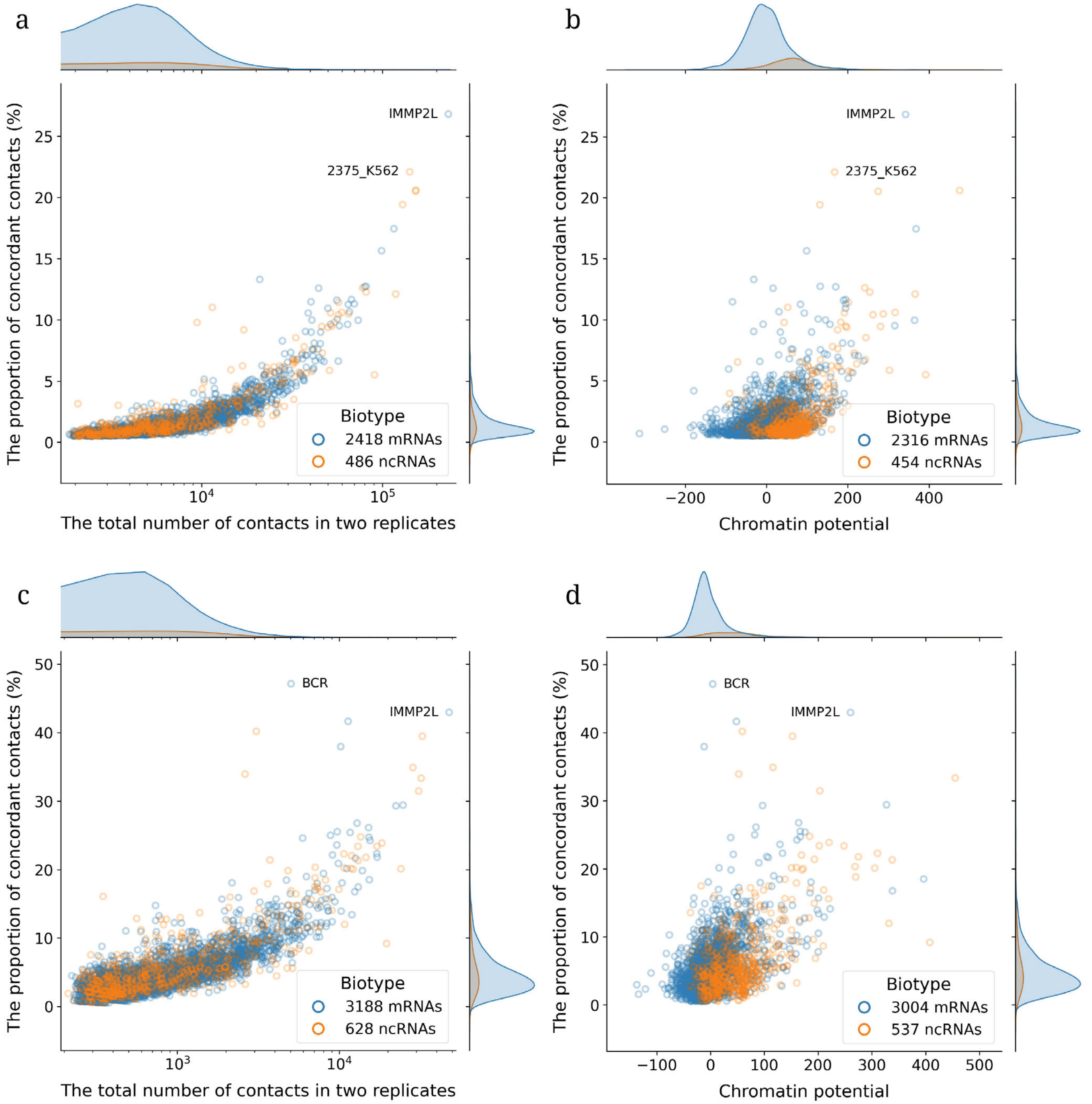

Fig. 3. Dependence of replicate concordance on the number of contacts and chromatin potential. a and b) Concordance calculated for all contacts. c and d) Concordance calculated for contacts from BaRDIC peaks. Data from Red-C on K-562 cells, bin size = 5000 bp. MALAT1 is not shown in the graph because this RNA has extreme values of chromatin potential and proportion of concordant contacts: 991 and 58.2% – panels (a) and (b); 740 and 71.9% – panels (c) and (d).

Comparison of replicates in the OTA data. To assess reproducibility of the experiments with individual RNAs, we analyzed consistency of the replicates in the corresponding datasets (Table 2; Fig. S9 in the Online Resource 1). First, the expected high level of reproducibility was confirmed: in the complete OTA dataset, proportion of the concordant contacts between the replicates exceeded 90% even at a bin size of 1000 bp. This indicates that the OTA data have resolution of 1000 bp and high completeness. Second, a critically important aspect of signal specificity was identified. When the analysis was limited to only those contacts that fall into peaks identified by the BaRDIC program (which filters out rare, single contacts in favor of statistically significant clusters), the level of concordance dropped by nearly half.

Table 2. Proportion of concordant contacts in the OTA replicates (%)

RNA	Experiment	Bin = 1000 bp		Bin = 5000 bp	
		All contacts, %	Contacts from peaks, %	All contacts, %	Contacts from peaks, %
JPX	CHART, ES d0 (GSM4278791, GSM4278795)	99.5	53.3	100.0	78.2
JPX	CHART, ES d3 (GSM4278799, GSM4278803)	99.3	36.8	100.0	70.6
JPX	CHART, ES d7 (GSM4278807, GSM4278811)	99.2	44.5	100.0	75.0
MALAT1	ChIRP, ES, genotype: Ythdc1-cKO (conditional); treatment: DMSO, (GSM4669091, GSM4669092)	79.9	26.9	99.6	50.6
MALAT1	ChIRP, ES, genotype: Mettl3-WT, (GSM4875651, GSM4875652)	92.4	40.9	99.9	66.3

Note. d0, d3, and d7 correspond to 0, 3, and 7 days of cell differentiation, respectively; $p < 0.05$.

This sharp decline suggests that the significant proportion (more than half) of all detected contacts, including the concordant ones, in the OTA experiments are likely non-specific.

Comparative analysis of reproducibility between the ATA and OTA methods. Comparative analysis of reproducibility between the ATA and OTA methods allows us to draw the following conclusions:

1. ATA data (except GRID) are characterized by low reproducibility between the replicates (median proportion of concordant contacts <5%), indicating their substantial incompleteness.

2. OTA data, on the other hand, demonstrate high reproducibility (>90%), confirming their completeness and allowing them to be considered as a reliable reference ("gold standard") for validating interactions identified in the genome-wide approaches (ATA data).

Comparison of ATA and OTA experiments. High reproducibility of the OTA data, demonstrated in the previous section, allows their use as a reference to assess the degree of consistency between the genome-wide approach data (ATA) and this reference. Conducting such comparative analysis comes with the significant limitations, as it requires availability of both types of data for the same RNAs in the identical cell lines and under similar cultivation conditions, as well as sufficient number of contacts in the ATA data to ensure statistical power. The publicly available OTA data matching the conditions of ATA experiments were found only for two RNAs: MALAT1 and JPX.

For the ncRNAs MALAT1 and JPX, we conducted comparison using bins with 5000-bp size. As a measure of consistency, we calculated proportion of the contacts from the ATA data that fell into bins enriched with the contacts from the BaRDIC peaks of the corresponding OTA experiment. The analysis was performed for all ATA contacts as well as for the subset filtered by the BaRDIC peaks. The sets of contacts from the ATA replicates were combined to increase data power. The results for the ncRNA MALAT1 are presented in Table 3 and Fig. S10 in the Online Resource 1, and for the ncRNA JPX in Table 4.

For the ncRNA MALAT1, which exhibits an extremely high level of interactions in the ATA data, a significant proportion (~50%) of overlaps with the OTA data was identified, indicating good consistency between the methods. However, approximately half of the MALAT1 contacts detected solely by the ATA method are not confirmed by the independent OTA method. This allows us to estimate proportion of non-specific signals in the ATA data for this RNA as ~50%. Importantly, in this case, we do not observe a significant advantage of the GRID-seq method, which was so evident in the analysis of the ATA replicate consistency. This is likely because the total number of contacts and their reproducibility for MALAT1 are so high in all ATA experiments that the effect of the more specific fixation protocol is overshadowed by the dominant signal.

The situation for the ncRNA JPX, characterized by low level of contacts in the ATA data, is fundamentally different. Overlap with the OTA data was ~60%, allowing us to roughly estimate proportion of non-specific contacts as ~40%. The low absolute number of contacts makes this estimate less reliable. As expected, based on the known association of JPX with XIST [26], most of its contacts are localized

Table 3. Percentage of the consistent MALAT1 RNA contacts with chromatin in the ATA data when compared with the MALAT1 RNA contacts from the OTA experiments (contacts from BaRDIC peaks) in mouse embryonic stem cells

Experiment	Number of contacts in ATA data	RAP		ChIRP		
		pSM33 ES, DMSO 1 hour, %	V6.5 ES, %	ES, Ythdc1-cKO; DMSO, %	ES, Mettl3-WT, %	E14 ES, %
GRID, ES, *M. musculus*	522,741 (109,371)	38.0 (46.5)	55.8 (61.6)	50.6 (51.0)	58.8 (58.3)	53.8 (57.0)
RADICL (1% FA), ES, *M. musculus*	636,802 (138,422)	42.9 (56.2)	61.5 (74.1)	50.4 (49.6)	58.9 (57.2)	55.1 (58.6)
RADICL (2% FA), ES, *M. musculus*	484,878 (99,985)	41.5 (51.0)	59.7 (69.0)	50.7 (49.6)	59.2 (58.2)	54.9 (56.8)

Note. Values in parentheses represent results for the ATA contacts from BaRDIC peaks. Bin size = 5000 bp.

Table 4. Percentage of the consistent JPX RNA contacts with chromatin in the ATA data (all contacts) when compared with the JPX RNA contacts from OTA experiments (contacts from BaRDIC peaks) in mouse embryonic stem cells

Experiment	Number of JPX contacts in ATA data	CHART, ES		
		d0	d3	d7
GRID, ES, *M. musculus*	459	57.1 (0.05)	61.9 (0.22)	63.6 (0.002)
RADICL (1% FA), ES, *M. musculus*	341	57.2 (0.09)	62.8 (0.15)	61.0 (0.03)
RADICL (2% FA), ES, *M. musculus*	332	54.5 (0.24)	56.6 (0.65)	63.9 (0.005)

Note. Values in parentheses represent the *p*-value of concordance. Bin size = 5000 bp; d0, d3, and d7 correspond to 0, 3, and 7 days of cell differentiation, respectively.

on the X chromosome. The low number of contacts did not allow application of the BaRDIC peak filtering, which would likely have increased specificity of the analysis.

The conducted analysis demonstrates fundamental possibility of cross-validation but also highlights its limitations. Unfortunately, for the RNAs with average level of interactions – which are of the greatest interest for assessing specificity of the ATA methods – comparative analysis with OTA was not possible due to the lack of paired data under consistent biological conditions. Thus, the OTA data serve as a reliable reference primarily for the highly interacting RNAs, while assessing specificity of ATA for the rest of interactome requires development of alternative approaches.

Comparison of the OTA experiments. To assess consistency of the OTA data, we conducted comparative analysis of the RNA-chromatin interaction maps for various ncRNAs in human cells (Fig. 4) and mouse cells (Fig. S11 in the Online Resource 1). As a measure of similarity, we used the ratio of concordant contacts to the total number of detected interactions in the compared OTA experiments (Jaccard index).

The analysis revealed clusters of high functional consistency, as well as overlaps, likely related to the common principles of chromatin organization. In the heatmap for human cells, distinct clusters were observed, corresponding to the specific RNAs such as MALAT1, NEAT1, and HOTAIR. The most striking example of the expected similarity were the profiles of MALAT1 and NEAT1. High concordance of their chromatin contacts aligns well with their known co-localization in the nucleus: NEAT1 serves as a structural basis of paraspeckles, while MALAT1 is a key

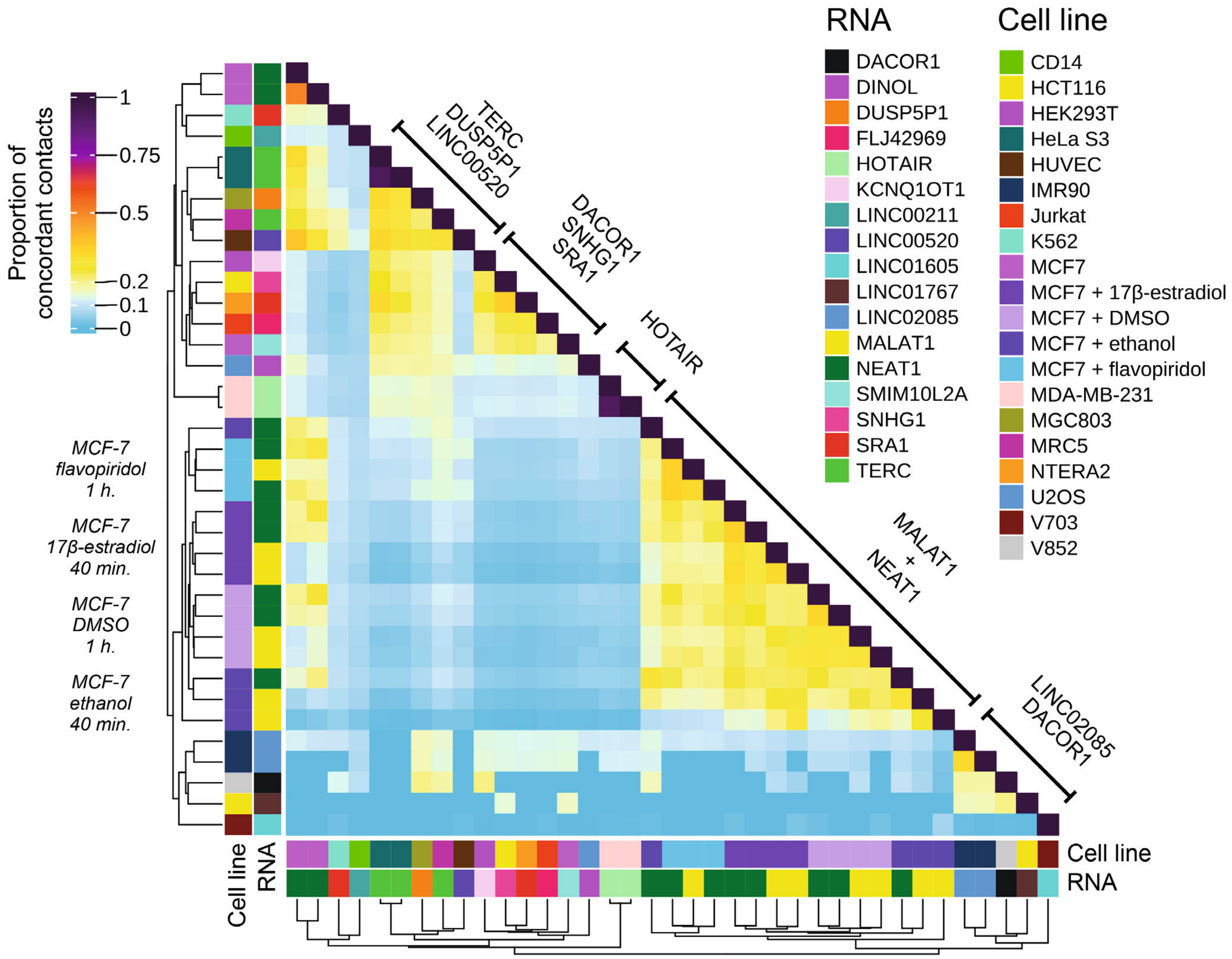

Fig. 4. Heatmap reflecting proportion of the concordant contacts (from BaRDIC peaks, FDR < 0.05) from "one-to-all" experiments for the human cell lines. Non-significant enrichments ($p > 0.05$) are set to zero. Clustering is performed by cell types and RNAs used in the experiment. Bin size = 1000 bp.

component of nuclear speckles [27, 28]. Both RNAs are associated with active genes and are involved in splicing regulation [16], which explains similarity in their chromatin landscape.

Overlaps were also detected, for example, between the contacts of LINC02085 and DACOR1. LINC02085 is involved in the NF-κB-dependent regulation [29], while DACOR1 is implicated in maintaining DNA methylation patterns [30], which may reflect their joint involvement in epigenetic control.

At the same time, the analysis identified similarity clusters lacking an obvious functional explanation. For instance, profile of the telomerase RNA TERC showed significant concordance with the RNAs such as SRA1, SNHG1, and KCNQ1OT1, for which direct functional links are unknown. This result suggests presence of the background noise. If we assume that most of the detected RNA-chromatin interactions are protein-mediated, low specificity of these contacts could be attributed not to the experimental methods themselves but to the relatively low specificity of the RNA-binding domains in the proteins [31, 32]. This leads to similar association patterns for the functionally unrelated RNAs.

Unlike the human data, the mouse OTA data primarily focus on the study of XIST. The observed high concordance of the XIST profiles with the RNAs such as its known activator JPX [26] serves as an additional internal quality control for the data and confirms specificity of the method for the functionally related pairs (Fig. S11).

CONCLUSION

This study conducted comparative analysis of the RNA-chromatin interactome data obtained using "all-to-all" (ATA) and "one-to-all" (OTA) methods, focusing on evaluating their accuracy, completeness, and specificity.

We compared the genome-wide RNA-chromatin interactome data (ATA) with the RNA-seq data and introduced the concept of chromatin potential – a numerical characteristic of individual RNAs that indicates the extent to which the number of contacts of a given RNA exceeds the expected number based on the RNA-seq data. This metric allows filtering out RNAs with predominantly non-specific interactions due to the high expression levels. Setting a threshold for chP significantly reduces proportion of mRNAs in the interactome, effectively isolating RNAs with higher affinity for chromatin. It is important to note that some mRNAs exhibit high chromatin potential, which could indicate presence of unknown specific functions or expression of the non-coding isoforms and unannotated intronic ncRNAs. Positive correlation of chP with the contact reproducibility confirms that chromatin potential is not only a measure of specificity but also a predictor of interaction reliability.

Comparison of the methods revealed fundamental differences in resolution (~5000 bp for ATA vs. ~1000 bp for OTA) and reproducibility. The developed metric for replicate consistency showed that the OTA data have high reproducibility (>90%), allowing them to be used as a "gold standard." In contrast, the ATA data (except for GRID-seq) were characterized by low concordance (<5-10%), indicating substantial incompleteness.

It was found that completeness of the ATA data is a function of sequencing depth for each specific RNA. To achieve statistically significant reproducibility, the number of RNA contacts must exceed 10,000, indicating the need for exceptionally deep sequencing in the genome-wide experiments to reliably identify interactome of the individual RNAs.

Critical influence of the fixation protocol was demonstrated. It was shown that the use of the two-step fixation with DSG/formaldehyde (GRID-seq) compared to the fixation with formaldehyde alone (Red-C, RADICL-seq) leads to the significant increase in the proportion of reproducible signals. This suggests critical role of the protein complex stabilization in the quality of ATA data.

In all types of experiments, preliminary selection of contacts belonging to the peaks identified using BaRDIC significantly increased the data consistency. This proves that identifying statistically significant interaction clusters is a powerful tool for separating the biologically significant signals from the background noise.

Based on these results, we recommend the following approach to enhance reliability and significance of conclusions when working with the RNA-chromatin interactome data:

- When analyzing the OTA data, focus on the contacts that have passed peak filtering (e.g., using BaRDIC), as they demonstrate significantly higher specificity. The high overall reproducibility of OTA data confirms their reliability as a reference.
- Note that chromatin potential selects promising RNAs, while concordance analysis and peak searching select significant RNA-chromatin contacts. Therefore, when analyzing the ATA data, the strategy should be two-level:

1. First, select RNAs with high chromatin potential (chP > 20), focusing on the molecules with increased probability of specific interactions with chromatin.
2. Second, select RNAs with more than 10,000 contacts and use only those contacts that both fall into the BaRDIC peaks and are reproducible between the replicates.

Thus, the combined use of chromatin potential (for RNA selection) and concordant contacts from the peaks (for genomic locus selection) maximizes filtering of non-specific noise and highlights the most reliable interactions. The proposed approach enhances reliability of bioinformatics analysis and interpretation of the RNA-chromatin interactome data, which is particularly important for identifying functionally significant associations.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1134/S0006297925601923.

Contributions

G. K. Ryabykh – analysis of chromatin potential, concordance analysis of replicates, and comparison of "all-to-all" (ATA) and "one-to-all" (OTA) data; A. I. Nikolskaya – calculation of BaRDIC peaks and concordance analysis of "one-to-all" (OTA) data; L. D. Garkul – processing of RNA sequencing data; A. A. Mironov – conceptualization and supervision of the study; G. K. Ryabykh, A. I. Nikolskaya, L. D. Garkul, and A. A. Mironov – writing the manuscript; G. K. Ryabykh, A. I. Nikolskaya, L. D. Garkul, and A. A. Mironov – editing the manuscript.

Funding

This work was financially supported by the Russian Science Foundation (grant no. 23-14-00136).

Ethics approval and consent to participate

This work does not contain any studies involving human and animal subjects.

Conflict of interest

The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., Mercer, T. R., Moore, K. J., Nakagawa, S., Rinn, J. L., Spector, D. L., et al. (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations, *Nat. Rev. Mol. Cell Biol.*, **24**, 430-447, https://doi.org/10.1038/s41580-022-00566-8.
2. Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, j., Goren, A., Lander, E. S., Plath, K., and Guttman, M. (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome, *Science*, **341**, 1237973, https://doi.org/10.1126/science.1237973.
3. Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., Borowsky, M. L., Kuroda, M. I., and Kingston, R. E. (2011) The genomic binding sites of a noncoding RNA, *Proc. Natl. Acad. Sci. USA*, **108**, 20497-20502, https://doi.org/10.1073/pnas.1113536108.
4. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., and Chang, H. Y. (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions, *Mol. Cell*, **44**, 667-678, https://doi.org/10.1016/j.molcel.2011.08.027.
5. Quinn, J. J., Ilik, I. A., Qu, K., Georgiev, P., Chu, C., Akhtar, A., and Chang, H. Y. (2014) Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification, *Nat. Biotechnol.*, **32**, 933-940, https://doi.org/10.1038/nbt.2943.
6. Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A. R., Hoberg, E., Moustakas, A., Gyllensten, U., Jones, S. J., Gustafsson, C. M., Sims, A. H., Westerlund, F., Gorab, E., and Kanduri, C. (2015) MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures, *Nat. Commun.*, **6**, 7743, https://doi.org/10.1038/ncomms8743.
7. Chu, H. P., Cifuentes-Rojas, C., Kesner, B., Aeby, E., Lee, H.-G., Wei, C., Oh, H. J., Boukhali, M., Haas, W., and Lee, J. T. (2017) TERRA RNA antagonizes ATRX and protects telomeres, *Cell*, **170**, 86-101, https://doi.org/10.1016/j.cell.2017.06.017.
8. Sridhar, B., Rivas-Astroza, M., Nguyen, T. C., Chen, W., Yan, Z., Cao, X., Hebert, L., and Zhong, S. (2017) Systematic mapping of RNA-chromatin interactions *in vivo*, *Curr. Biol.*, **27**, 602-609, https://doi.org/10.1016/j.cub.2017.01.011.
9. Li, X., Zhou, B., Chen, L., Gou, L. T., Li, H., and Fu, X. D. (2017) GRID-seq reveals the global RNA-chromatin interactome, *Nat. Biotechnol.*, **35**, 940-950, https://doi.org/10.1038/nbt.3968.
10. Bell, J. C., Jukam, D., Teran, N. A., Risca, V. I., Smith, O. K., Johnson, W. L., Skotheim, J. M., Greenleaf, W. J., and Straight, A. F. (2018) Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts, *Elife*, **7**, e27024, https://doi.org/10.7554/eLife.27024.
11. Limouse, C., Smith, O. K., Jukam, D., Fryer, K. A., Greenleaf, W. J., and Straight, A. F. (2023) Global mapping of RNA-chromatin contacts reveals a proximity-dominated connectivity model for ncRNA-gene interactions, *Nat. Commun.*, **14**, 6073, https://doi.org/10.1038/s41467-023-41848-9.
12. Yan, Z., Huang, N., Wu, W., Chen, W., Jiang, Y., Chen, J., Huang, X., Wen, X., Xu, j., Jin, Q., Zhang, K., Chen, Z., Chien, S., and Zhong, S. (2019) Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs, *Proc. Natl. Acad. Sci. USA*, **116**, 3328-3337, https://doi.org/10.1073/pnas.1819788116.
13. Bonetti, A., Agostini, F., Suzuki, A. M., Hashimoto, K., Pascarella, G., Gimenez, J., Roos, L., Nash, A. J., Ghilotti, M., Cameron, C. J. F., Valentine, M., Medvedeva, Y. A., Noguchi, S., Agirre, E., Kashi, K., Samudyata, Luginbühl, J., Cazzoli, R., Agrawal, S., Luscombe, N. M., Blanchette, M., Kasukawa, T., Hoon, M., Arner, E., Lenhard, B., et al. (2020) RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions, *Nat. Commun.*, **11**, 1018, https://doi.org/10.1038/s41467-020-14337-6.
14. Gavrilov, A. A., Zharikova, A. A., Galitsyna, A. A., Luzhin, A. V., Rubanova, N. M., Golov, A. K., Petrova, N. V., Logacheva, M. D., Kantidze, O. L., Ulianov, S. V., Magnitov, M. D., Mironov, A. A., and Razin, S. V. (2020) Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics, *Nucleic Acids Res.*, **48**, 6699-6714, https://doi.org/10.1093/nar/gkaa457.
15. Ryabykh, G. K., Mylarshchikov, D. E., Kuznetsov, S. V., Sigorskikh, A. I., Ponomareva, T. Y., Zharikova, A. A., and Mironov, A. A. (2022) RNA-chromatin interactome: What? Where? When? *Mol. Biol.*, **56**, 210-228, https://doi.org/10.31857/S002689842202015X.
16. West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., Tolstorukov, M. Y., and Kingston, R. E. (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites, *Mol. Cell*, **55**, 791-802, https://doi.org/10.1016/j.molcel.2014.07.012.
17. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009) Comprehensive mapping of

long-range interactions reveals folding principles of the human genome, *Science*, **326**, 289-293, https://doi.org/10.1126/science.1181369.

18. Hoffman, E. A., Frey, B. L., Smith, L. M., and Auble, D. T. (2015) Formaldehyde crosslinking: a tool for the study of chromatin complexes, *J. Biol. Chem.*, **44**, 26404-26411, https://doi.org/10.1074/jbc.R115.651679.
19. Ryabykh, G. K., Kuznetsov, S. V., Korostelev, Y. D., Sigorskikh, A. I., Zharikova, A. A., and Mironov, A. A. (2023) RNA-Chrom: a manually curated analytical database of RNA–chromatin interactome, *Database*, **2023**, baad025, https://doi.org/10.1093/database/baad025.
20. Mylarshchikov, D. E., Nikolskaya, A. I., Bogomaz, O. D., Zharikova, A. A., and Mironov, A. A. (2024) BaRDIC: robust peak calling for RNA-DNA interaction data, *NAR Genom. Bioinform.*, **6**, lqae054, https://doi.org/10.1093/nargab/lqae054.
21. Alberti, A., Belser, C., Engelen, S., Bertrand, L., Orvain, C., Brinas, L., Cruaud, C., Giraut, L., Silva, C., Firmo, C., Aury, J.-M., and Wincker, P. (2014) Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data, *BMC Genomics*, **15**, 912, https://doi.org/10.1186/1471-2164-15-912.
22. Lin, S.-L., Miller, J. D., and Ying, S.-Y. (2006) Intronic microRNA (miRNA), *J. Biomed. Biotechnol.*, **2006**, 026818, https://doi.org/10.1155/JBB/2006/26818.
23. Bergeron, D., Faucher-Giguère, L., Emmerichs, A.-K., Choquet, K., Song, K. S., Deschamps-Francoeur, G., Fafard-Couture, E., Rivera, A., Couture, S., Churchman, L. S., Heyd, F., Elela, S. A., and Scott, M. S. (2023) Intronic small nucleolar RNAs regulate host gene splicing through base pairing with their adjacent intronic sequences, *Genome. Biol.*, **24**, 160, https://doi.org/10.1186/s13059-023-03002-y.
24. Nam, J.-W., Choi, S.-W., and You, B.-H. (2016) Incredible RNA: dual functions of coding and noncoding, *Mol. Cells*, **39**, 367-374, https://doi.org/10.14348/molcells.2016.0039.
25. Machyna, M., and Simon, M. D. (2018) Catching RNAs on chromatin using hybridization capture methods, *Brief. Funct. Genomics*, **17**, 96-103, https://doi.org/10.1093/bfgp/elx038.
26. Oh, H. J., Aguilar, R., Kesner, B., Lee, H.-G., Kriz, A. J., Chu, H.-P., and Lee, J. T. (2021) Jpx RNA regulates CTCF anchor site selection and formation of chromosome loops, *Cell*, **184**, 6157-6173, https://doi.org/10.1016/j.cell.2021.11.012.
27. Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., and Lawrence, J. B. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles, *Mol. Cell*, **33**, 717-726, https://doi.org/10.1016/j.molcel.2009.01.026.
28. Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G., and Prasanth, K. V. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation, *Mol. Cell.*, **39**, 925-938, https://doi.org/10.1016/j.molcel.2010.08.011.
29. Cai, D., and Han, J.-D. J. (2021) Aging-associated lncRNAs are evolutionarily conserved and participate in NFκB signaling, *Nat. Aging*, **1**, 438-453, https://doi.org/10.1038/s43587-021-00056-0.
30. Merry, C. R., Forrest, M. E., Sabers, J. N., Beard, L., Gao, X.-H., Hatzoglou, M., Jackson, M. W., Wang, Z., Markowitz, S. D., and Khalil, A. M. (2015) DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer, *Hum. Mol. Genet.*, **24**, 6240-6253, https://doi.org/10.1093/hmg/ddv343.
31. Stitzinger, S. H., Sohrabi-Jahromi, S., and Söding, J. (2023) Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains, *NAR Genom. Bioinform.*, **5**, lqad057, https://doi.org/10.1093/nargab/lqad057.
32. Khlebnikov, D. A., Nikolskaya, A. I., Zharikova, A. A., and Mironov, A. A. (2025) Comprehensive analysis of RNA-chromatin, RNA-, and DNA-protein interactions, *NAR Genom. Bioinform.*, **7**, lqaf010, https://doi.org/10.1093/nargab/lqaf010.

Publisher's Note. Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. AI tools may have been used in the translation or editing of this article.