BaRDIC: robust peak calling for RNA-DNA interaction data

Dmitry E. Mylarshchikov [1], Arina I. Nikolskaya [1,*], Olesja D. Bogomaz [1], Grigory K. Ryabykh [1,2], Anastasia A. Zharikova [1,2] and Andrey A. Mironov [1,2] *

[1]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye Gory, Moscow, 119234, Russia and [2]Kharkevich Institute for Information Transmission Problems RAS, Bolshoy Karetny per., Moscow, 127051, Russia

ABSTRACT

Chromatin-associated non-coding RNAs play important roles in various cellular processes by targeting genomic loci. Two types of genome-wide NGS experiments exist to detect such targets: "one-to-all", which focuses on targets of a single RNA, and "all-to-all", which captures targets of all RNAs in a sample. As with many NGS experiments, they are prone to biases and noise, so it becomes essential to detect "peaks" – specific interactions of an RNA with genomic targets. Here we present BaRDIC – Binomial RNA-DNA Interaction Caller – a tailored method to detect peaks in both types of RNA-DNA interaction data. BaRDIC is the first tool to simultaneously account for the two most prominent biases in the data: chromatin heterogeneity and distance-dependent decay of interaction frequency, or scaling. Since RNAs differ in their interaction preferences, BaRDIC adapts peak sizes according to the abundances and contact patterns of individual RNAs. These features enable BaRDIC to make more robust predictions than currently applied peak-calling algorithms and better handle the characteristic sparsity of all-to-all data. BaRDIC package is freely available at https://github.com/dmitrymyl/BaRDIC.

INTRODUCTION

Chromatin-associated non-coding RNAs play important roles in various cellular processes (1, 2, 3). Many sequencing techniques are now available to localize RNAs on the whole chromatin, which can be broadly classified into "one-to-all" (OTA) and "all-to-all" (ATA) methods (4, 5). OTA methods hybridize biotinylated RNA-specific oligonucleotide probes to find the contacts of a single target RNA (6, 7, 8, 9). High-throughput ATA methods detect genome-wide contacts of all potential chromatin-associated RNAs using RNA-DNA proximity ligation (10, 11). Identifying specific interactions of an individual RNA with a particular genomic location, a process called "peak-calling", has been a focus of all RNA-DNA interaction studies. However, RNA-DNA interaction data contain biological and technology-specific biases, which hinder the detection of real RNA-binding sites on the chromatin. Similar to peak-calling algorithms in ChIP-seq data and chromatin loop detection in Hi-C data, there is a need to develop tailored methods for identifying specific interactions in RNA-DNA interaction data.

Specific interaction loci can be defined as genomic regions where the number of contacts is significantly higher than expected in a given background model. Ideally, the background model captures all possible sources of biases. One of these biases is chromatin heterogeneity, which includes chromatin accessibility, amplification bias, and copy number variation. This bias is also present in ChIP-seq datasets. And since the nature of OTA data is also similar to that of ChIP-seq data due to common experimental steps, a ChIP-seq specific peak caller MACS2 is usually applied to OTA data, using an input track, which evaluates chromatin heterogeneity. For ATA data, chromatin heterogeneity can be inferred from the dataset itself, according to GRID data processing protocol (11). The endogenous background here is based on *trans* chromosomal interactions of mRNAs, assuming such contacts are mostly non-specific. In contrast, RADICL protocol (10) proposed a uniform binomial model that does not account for the heterogeneous nature of the chromatin.

Another source of bias is the decay of contact density with increasing distance from the RNA transcription site. This effect was described in several studies dedicated to RNA-DNA interaction data analysis (10, 12). We observed this distance-dependent effect in both OTA and ATA datasets (Figure 1A) with our previously developed RNA-Chrom database (13). Analogous to Hi-C, we will refer to this phenomenon as scaling (14). MACS2, GRID, and RADICL approaches do not account for scaling, therefore, these algorithms might overestimate the statistical significance of contacts neighboring an RNA source gene.

Finally, functionally diverse classes of RNAs have different patterns of genome occupancy. For instance, an enhancer RNA – promoter interaction may exhibit a narrow peak profile, while a small nuclear RNA can appear to be broadly distributed along gene bodies. Furthermore, we expect that peak width would be affected by the amount of the data, which refers to resolution. In ATA datasets, RNAs differ in contact frequency, while limited sequencing depth leads to data

*To whom correspondence should be addressed. Email: ar-nikolya@fbb.msu.ru; mironov@bioinf.fbb.msu.ru

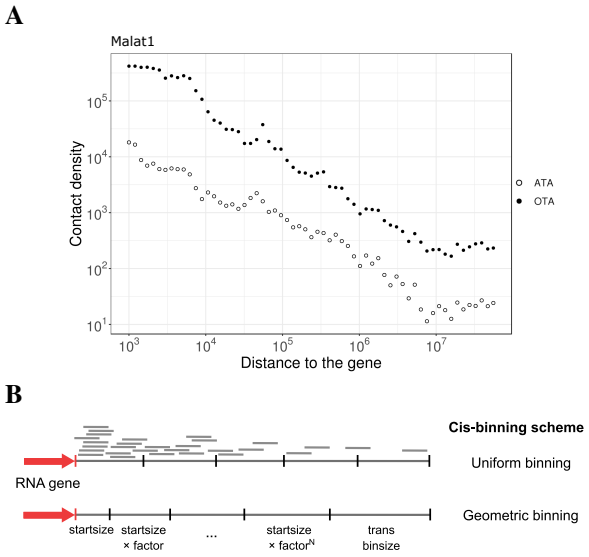

Figure 1. (A) Dependency of contact density on the distance between the RNA source gene and chromatin target loci (scaling) in double logarithmic coordinates for Malat1. Black points: ChIRP OTA data on mESC (Exp. ID 182), white points: ATA GRID data on mESC. **(B)** *Cis* binning strategy with geometrically increasing bins provides more even contact coverage in the vicinity of the source gene.

sparsity as a result of incomplete contact capture. Consequently, there is a ground to introduce a variable peak size for different RNAs depending on their abundances and distribution patterns along the chromatin. Existing GRID and RADICL procedures propose fixed-size bins for all RNAs, which can reduce the resolution for RNAs with narrow binding sites, and lower the statistical power for RNAs with more diffuse binding sites.

In this work, we introduce a versatile tool BaRDIC (Binomial RNA-DNA Interaction Caller), that utilizes a binomial model to identify genomic regions significantly enriched in RNA-chromatin interactions, or "peaks", in ATA and OTA data. Our approach accounts for several key features of RNA-DNA interaction data:

- chromatin heterogeneity;
- scaling;
- data sparsity and distinct RNA contact patterns.

BaRDIC combines approaches from Hi-C and ChIP-seq data analysis and yields accurate predictions of RNA-DNA interactions from both OTA and ATA data.

MATERIALS AND METHODS

Peak-calling on OTA data

Pre-processed contacts and MACS2 peaks for OTA experiments were extracted from the RNA-Chrom database (13). The data identifiers (Exp. ID) are as follows: 94 (RAP-RNA Malat1), 102 (ChIRP Halr1) for mESC cell line, 90 (CHART Paupar) for mouse neuroblastoma cell line N2A. Details on the experiments and data processing protocol are described in the RNA-Chrom database.

Background track for OTA experiments was constructed from input data and converted to begGraph format in 1kb bins using bedtools package (15). The type of background used by BaRDIC was set to a pre-constructed background track (-bt custom). Due to relatively high resolution of OTA experiments, the following parameters were provided for peak-calling on OTA data: 400 nt minimum *trans* bin size (–trans_min), 100 nt initial *cis* bin size (–cis_start 100), 50 nt step to find the optimal *trans* bin size (–trans_step). For Halr1 (Haunt), the peaks for the retinoic acid-induced mESC sample were used for analysis as the peaks from the control mESC sample do not intersect HOXA cluster genes.

Peak-calling on ATA data

Contacts for the GRID-seq ATA experiment on mESC cell line (Exp. ID 6) and RNA gene annotation for mouse genome (except X-RNA biotype) were extracted from the RNA-Chrom database. Peaks were called with default parameters. Peaks by RADICL procedure were called according to the algorithm description from the original study (10). Since we did not succeed in reproducing the original GRID peak-calling method, the peaks published by authors were used in analysis (11). To compare BaRDIC with the GRID peak caller, the publicly available peaks were converted from mm9 to mm10 using LiftOver (16). It is worth noting that the peak comparison may be skewed due to differences in the data processing protocol of the GRID-seq authors and our group.

RESULTS

BaRDIC algorithm

The BaRDIC algorithm consists of three steps:

1. Binning: for each RNA, chromosomes are partitioned into non-overlapping genomic intervals (bins), and the number of contacts is calculated within bins.
2. Statistical modeling: for each RNA, parameters of the background model are calculated in every bin. P-values are calculated based on them.
3. Multiple testing correction with Benjamini-Hochberg procedure (17).

To increase the statistical power and speed up the algorithm's performance, only RNAs with more than 1000 contacts are selected by default. Steps 1 and 2 are atomic with regard to individual RNAs, so we describe them for one RNA only.

Binning. Bin sizes are estimated for each RNA separately. The binning strategy differs for *cis* interactions, which occur on a chromosome harboring the RNA gene, and *trans* interactions – with other chromosomes.

In the case of random ligation, we expect that *trans* contacts are distributed uniformly along chromosomes. Therefore, we apply a uniform binning strategy for *trans* interactions.

For binning in *cis*, scaling must be taken into account. Also, we assume long-range *cis* interactions are similar to *trans* interactions, analogous to observations in Hi-C data analysis (18). In our *cis* binning strategy, chromosomes are partitioned into non-uniform bins of size increasing in a

geometric progression from the source gene. *Cis* bin size increases until it exceeds the *trans* bin size; all subsequent *cis* bins are uniform and equal to the *trans* bin in size (see Figure 1B):

$$cis\ binsize(i) = \min(startsize \cdot factor^{i},\ trans\ binsize)$$

To find optimal bin sizes, we adapted RSEG (19) and JAMM (20) approaches that were developed for ChIP-seq peak-calling. This approach finds a balance between resolution and sufficient bin coverage.

Statistical modelling. To model the background distribution of RNA-DNA contacts in bins, we introduce a frequentist model similar to those used in ChIP-seq and Hi-C data analysis (21, 22). Assuming that contacts arising from random binding are independent, we consider the number of contacts X_{ij} of RNA i in bin j to be binomially distributed:

$$X_{ij} \sim Bin\left(N_i, p_{ij}\right),$$

where N_i is the total number of contacts produced by RNA i (except the gene body), p_{ij} is the background probability.

Note the number of observed contacts O_{ij} is a realization of the random variable X_{ij}. Statistical estimation comes down to inferring the only model parameter p_{ij} from the observed data.

For *trans* bins, only chromatin heterogeneity plays a role. In ATA experiments, we estimate the parameter of the background model by counting mRNA *trans* contacts as proposed in the GRID procedure. The background probability of a single contact of RNA i to appear in the j-th *trans* bin is assumed to be as follows:

$$\hat{p}_{ij} = \hat{p}_j^{bg} = \frac{N_j^{bg}}{N^{bg}},$$

where N_j^{bg} is the number of contacts from the background in a bin j, N^{bg} is the total number of background contacts. For OTA experiments, we use contacts from the input sample similar to ChIP-seq analysis.

For *cis* bins, we additionally consider scaling. To do that, we define a scaling factor $f(d_{ij})$ that depends on the distance between the source gene of RNA i and bin j: $\hat{p}_{ij} = f(d_{ij}) \cdot \hat{p}_{ij}^{bg}$. We estimate the scaling factor as $f\left(d_{ij}\right) = \frac{O_{ij}/N_i}{\hat{p}_j^{bg}}$ and smooth its profile using a cubic smoothing spline in double logarithmic coordinates. As $\hat{f}\left(d_{ij}\right)$ values vary substantially for different RNAs, splines are calculated for every RNA separately. To exclude putative *cis* peaks from the background model, we perform a two-step scaling estimation procedure similar to Fit-Hi-C (21) and HiC-DC (23), removing bins with statistically high contact coverage.

To hold $\sum_j \hat{p}_{ij} = 1$, we renormalize probability estimates for each RNA i. As a result, the sum of background probabilities for *cis* bins equals the fraction of *cis* contacts of each RNA, the same is true for *trans* bins.

We expect that a specific binding event results in a high enrichment of contacts relative to the background. The resulting P-value is then computed with the right-sided binomial test using estimated background parameters. P-values from non-zero bins are subjected to the multiple testing correction using the Benjamini-Hochberg procedure, simultaneously for all RNAs.

Implementation. The algorithm is implemented as a package for Python 3 (24) with a command line interface and is available at https://github.com/dmitrymyl/BaRDIC. It is based on numpy (25), pandas, scipy (26) and statsmodels packages (27). The bioframe package (28) is used for operations with genomic intervals. Since binning and statistical evaluation are performed for each RNA separately, these steps are parallelized.

To organize the data storage and speed up the data access, we developed two HDF5-based data formats (29), which were inspired by the cool (30) format. Detailed algorithm description and format schemas are given in Supplementary Note.

Overview of evaluation and use cases

Evaluating the peak caller performance on RNA-DNA interaction data is still an open problem, given the lack of a gold standard or a "true" dataset. To address this, we will assess peak calling results based on biological principles of RNA-chromatin interactions:

- We assess the quality of BaRDIC peaks based on well-studied RNAs with known binding preferences;
- We compare BaRDIC performance with the three current approaches to identify RNA-binding sites — MACS2 peak caller for OTA and original methods from ATA studies, GRID-seq and RADICL-seq, hereafter referred to as GRID-peak and RADICL-peak.

Importantly, these three current approaches do not account for all typical biases in RNA-DNA interaction data, e.g. scaling, so none of them can be considered the gold standard.

General assessment of BaRDIC ATA peaks

We evaluated BaRDIC by using GRID-seq data on the mESC, a cell line for which many OTA experiments are also available (See Supplementary Tables S1, S2). Only RNAs with at least 1000 contacts in the ATA experiment were selected for peak calling. In total, BaRDIC identified specific interactions mediated by 7741 unique RNAs in GRID-seq data on mESC (Supplementary Figures S1, S3). For the majority of RNAs, the size of *trans* bins ranges from 30 to 80 Kb, with a median of 50 Kb (Supplementary Figure S2). RNAs of distinct biotypes were found to have different preferences for *cis* and *trans* interactions. In general, small non-coding RNAs have a large fraction of *trans* peaks, while long non-coding RNAs can generate both *cis* and *trans* peaks (Figure 2A). Notably, RNA preferences do not correlate with the number of RNA peaks, which suggests BaRDIC is robust against data sparsity.

For further analysis, 1007 unique RNAs were selected: all RNAs with at least one BaRDIC peak, except for ribosomal RNAs and mRNAs. To generally assess specific interactions in ATA data, we focused on RNAs with well-studied binding patterns. As expected, Malat1, SNORD3B-1, TERC, Neat1,

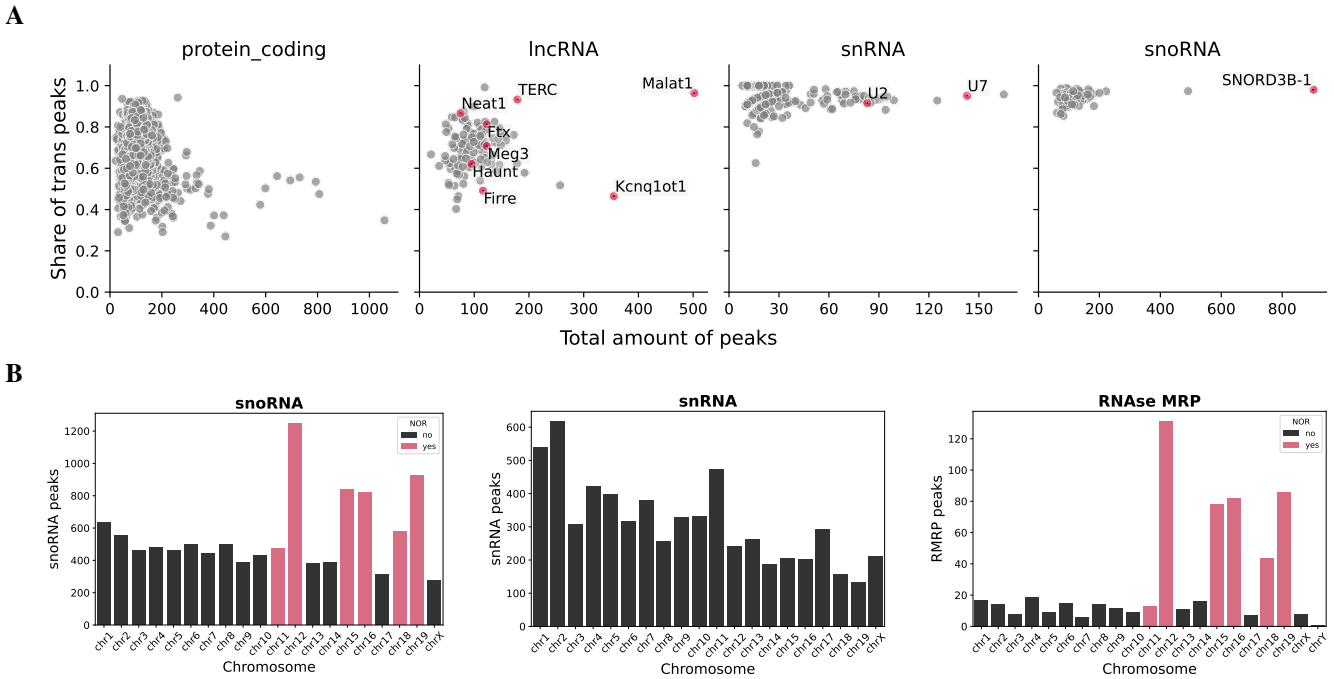

Figure 2. Global assessment of specific RNA interactions identified by BaRDIC for GRID data on mESC data. **(A)** Fractions of *trans* peaks and the number of ATA peaks for select RNA biotypes. **(B)** Distribution of snoRNAs, snRNAs, and MRP RNase peaks across mouse chromosomes. Chromosomes carrying nucleolus organizer regions (NORs) are highlighted in blue.

and RMRP prefer to contact in trans, while Kcnq1ot1, Firre, and Haunt (Halr1) favor *cis* interactions (Figure 2A, Supplementary Figure S4). Long ncRNAs as a class have variable preferences because they can act both as *cis* and *trans* regulators (31). SnoRNAs generate many *trans* peaks as they interact with chromosomes carrying the nucleolus organizer regions (NORs): chromosomes 12, 15, 16, 18, 19 except for chromosome 11 (Figure 2B) (32). A similar distribution is observed for RNase MRP (RMRP) consistent with its localization in the nucleus (33). Apparently, in this particular mESC sample, chromosome 11 does not carry active NOR loci. In contrast, snRNAs do not have specific chromosomal preferences: the number of peaks on chromosomes correlates with their lengths ($r_S = 0.71$, $p-value < 3.1e-4$).

Taken together, the coincidence of *cis* and *trans* peaks distribution with known RNA binding preferences validates BaRDIC peaks at the coarse-grain level.

Detailed comparison of peaks of individual RNAs

To assess the quality of BaRDIC peaks at the local level, we obtained peaks for several RNAs with known binding patterns, for which OTA experiments are available (Supplementary Tables S1, S2). When comparing OTA and ATA experiments for a particular RNA, it is crucial to keep in mind that ATA contacts are much sparser (Supplementary Figure S5) and located at varying distances from defined OTA peaks due to the differences in experimental procedures (Supplementary Figure S6). Therefore, in addition to the direct comparison of peaks, we compare genes intersecting peaks, similar to the RADICL approach (10).

To highlight features of BaRDIC, we chose three non-coding RNAs (ncRNAs) with distinct distributions of contacts. ncRNA Paupar exhibits virtually no scaling; ncRNA Malat1 is abundant in ATA datasets and exhibits a pronounced scaling; ncRNA Halr1 is unabundant RNA in ATA datasets, regulates gene expression in *cis* and therefore should have peaks very close to the source gene.

Paupar: ncRNA with no scaling

Paupar is a vertebrate-conserved lncRNA whose expression is restricted to the central nervous system. CHART-seq revealed that Paupar is predominantly associated with promoters and 5'-UTR regions of protein-coding genes and is almost evenly distributed across chromosomes, except for the lack of contacts on the X chromosome. We have shown that the scaling effect is not pronounced for Paupar (Supplementary Figure S7E). Thus, we can directly compare the performance of BaRDIC and other algorithms given the virtually absent influence of scaling. As ATA data for mESC lacks Paupar interactions, we compared only OTA peaks.

Sets of genes that intersect BaRDIC peaks and MACS2 peaks for OTA were very similar (Figure 3A). We confirmed spatial co-localization of BaRDIC and MACS2 peaks using GenometriCorr (34) metrics (Figure 3B; Supplementary Figure S8). Therefore, for an RNA with a diffuse genome-wide distribution and minor scaling effects, BaRDIC peaks are in agreement with the publicly available results. This proves BaRDIC captures chromatin heterogeneity similarly to other peak calling algorithms.

Malat1: abundant ncRNA with pronounced scaling

Malat1, an abundant lncRNA, is a global regulator of transcription and pre-mRNA splicing. Malat1 has multiple binding sites in *trans*, predominantly in the bodies of

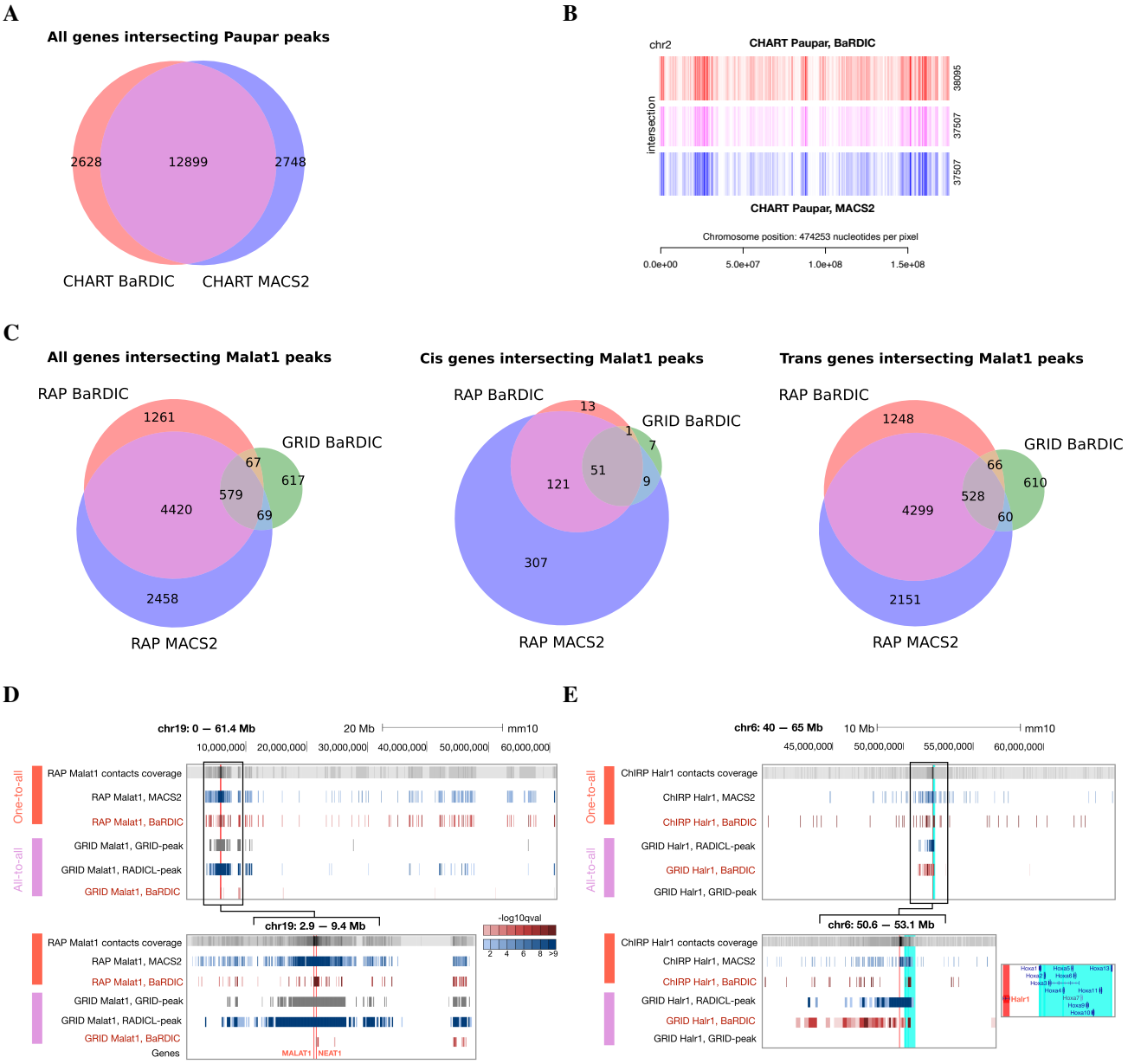

Figure 3. Comparison of peaks obtained by different algorithms for OTA and ATA data on mESC cell line **(A)** Genes intersecting Paupar peaks from CHART data obtained by BaRDIC and MACS2. **(B)** Global patterns of peak distribution and peaks intersections on "parental" chromosome 2, CHART data. **(C)** Genes intersecting Malat1 peaks from RAP and GRID mESC data. From left to right: all genes, *cis* genes, *trans* genes. **(D)** Representative genome browser view of Malat1 peaks identified by different algorithms in RAP OTA data (red) and ATA data (purple). Scales are given: the entire length of mouse chromosome 19 and the vicinity of the Malat1 gene on chromosome 19. Vertical red lines represent Malat1 and Neat1 source genes. **(E)** Representative genome browser view of Halr1 peaks identified by different algorithms in ChIRP OTA data (red) and ATA data (purple). The Halr1 gene located on chromosome 6 is highlighted in red, HOXA gene cluster in its vicinity is highlighted in cyan. Tracks are colored by $-log_{10}Q$ statistical significance when appropriate.

actively transcribed genes (35). Moreover, its possible *cis* regulatory role (36, 37) in controlling Neat1 expression has been discussed. Remarkably, Malat1 generates over 527.000 contacts in the selected ATA experiment, being an RNA with the largest number of contacts (Supplementary Table S1; Supplementary Figure S9). Given a sufficient number of Malat1 *cis* contacts, we investigated the influence of the scaling effect (Supplementary Figure S7A, B) on the performance of peak calling algorithms.

BaRDIC on the OTA data reproduces the set of genes that intersect MACS2 peaks (Figure 3C), in addition, peaks are spatially co-localized (Supplementary Figure S10). Importantly, BaRDIC peaks are more evenly distributed along chromosome 19 that harbors the Malat1 source gene. MACS2 does not take into account scaling and overestimates the significance of peaks near the Malat1 gene boundaries, which is corrected by the BaRDIC algorithm (Figure 3D).

When compared with GRID-peak, BaRDIC shows partial reproducibility of genes intersecting peaks (Supplementary

Figure S11). Genes intersecting Malat1 *cis* peaks called with BaRDIC form a subset of those intersecting peaks called with GRID-peak. GRID-peak and RADICL-peak calls are clustered near the Malat1 source gene (Figure 3D). This suggests that BaRDIC filters *cis* peaks that arose from scaling but preserves previously described specific interactions at the Neat1 locus (35).

Due to scaling correction, BaRDIC improves previous strategies as it does not overestimate the statistical significance of sites proximal to the RNA source gene. At the same time, BaRDIC preserves consensus target genes, which are likely to be true targets of chromatin-associated RNAs.

Halr1: cis-acting unabundant ncRNA

Long ncRNA Halr1 (HOXA upstream noncoding transcript) regulates the HOXA gene cluster during ESC differentiation (38). In the mouse genome, the HOXA gene cluster is located ~40 Kb downstream of the Halr1 transcription site. Halr1 is an unabundant RNA in ATA experiments, with 1724 contacts (23rd percentile) (Supplementary Table S1; Supplementary Figure S9) and pronounced scaling (Supplementary Figure S7C, D) in GRID mESC data. We therefore included a Halr1 case study to examine the quality of the BaRDIC algorithm on low-contacting RNAs, which form the vast majority of ATA data.

Peaks called with BaRDIC in OTA and ATA data indeed covered the genes of the HOXA gene cluster (Figure 3E). In contrast, GRID-peak identifies no Halr1 peaks. Additionally, the geometric binning procedure in *cis*, implemented in BaRDIC, increases the resolution near the boundaries of the Halr1 gene (Supplementary Figure S12). This results in a more accurate localization of peaks near the RNA source gene.

Overall, BaRDIC outperforms other peak callers in detecting peaks of RNAs with a small number of contacts in the ATA data. In addition, despite accounting for scaling, BaRDIC finds biologically relevant specific interactions near the RNA source gene.

DISCUSSION

In this paper, we present BaRDIC, a first-of-its-kind versatile algorithm for identifying specific RNA-chromatin interactions in both "one-to-all" (OTA) and "all-to-all" (ATA) data. BaRDIC takes into account typical biases and other features of RNA-DNA interaction data, including chromatin heterogeneity, scaling, and distinct contact patterns of individual RNAs. In the virtual absence of scaling effects, BaRDIC performs comparably with other algorithms. In the presence of scaling, BaRDIC does not overestimate the statistical significance of findings near the source gene of an individual RNA, unlike previously published approaches. The adaptive bin size selection for each RNA and geometric binning near source genes in ATA data allows for more accurate identification of specific binding sites. In addition, BaRDIC is robust to data sparsity, which is a hallmark of ATA experiments.

We found that when calling peaks in OTA data, BaRDIC and the commonly used MACS2 peak caller incompletely reproduce each other's results on specific *trans* interactions, even with no apparent influence of scaling. This may be due to different strategies of genome binning: BaRDIC relies on non-overlapping fixed-size bins, while MACS2 determines regions (peak candidates) dynamically. Perhaps the combined approach of binning with offsets in BaRDIC would increase the reproducibility of the results. In addition, the discrepancies may arise from different statistical models — a dynamic Poisson model in MACS2 and a binomial model in BaRDIC. Finally, BaRDIC peaks for RNAs in this study are wider than MACS2 peaks (Supplementary Table S2). Consequently, BaRDIC bins may include more non-specific interactions, which also decreases the resolution. Therefore, we controlled FDR more stringently for BaRDIC findings compared to MACS2. However, it is not our goal to achieve the full reproducibility of results between BaRDIC and MACS2 in OTA data, as there is currently no gold standard for specific RNA interactions. Besides, MACS2 does not account for scaling – the prominent bias in RNA-DNA interaction data, particularly in OTA data. Nevertheless, partial, but substantial reproducibility of peaks implies that BaRDIC accounts for chromatin heterogeneity similarly to other algorithms used in this area of research.

For all RNAs studied in this paper, peaks from ATA experiments intersect only a fraction of genes corresponding to OTA experiments, and may as well capture some non-target genes. This may be due to the insufficient sequencing depth in ATA experiments and, as a consequence, dropout. In this process, many true RNA binding sites are lost before the sequencing step, and the remaining nonspecific sites distort the statistical significance of detected peaks. This assumption is supported by the fact that the reproducibility of target genes is better for RNAs with a higher number of contacts in ATA data, such as Malat1. Differences in experimental procedures may also contribute to the inconsistency between OTA and ATA peaks (Supplementary Figure S6). OTA experiments are similar to ChIP-seq with a DNA shearing step that allows a relatively accurate identification of a contact region. In contrast, ATA experiments are similar to Hi-C experiments with DNA restriction and proximity ligation steps, which limit observed contact regions to restriction sites. Collectively, OTA and ATA peaks fundamentally cannot be compared properly, so we do not expect an exact match of called interactions due to the differences and limitations of the two types of experiments.

In this work, we highlighted the importance of the scaling effect when detecting RNA peaks on chromatin, which has already been around in the field. A recent preprint on ChAR-seq (39) proposes a generative model for predicting peak-like interactions based on mRNA *trans* contacts, with an account for scaling. However, that model suggests fix-sized bins for all RNAs, while BaRDIC proposes bins of variable size, individually for each RNA. This feature allows BaRDIC to model non-uniform contact coverage and take into account distinct distribution patterns of RNAs along the chromatin.

BaRDIC is not only applicable to pairwise RNA-DNA interaction data but can also be applied to multiway RD-SPRITE (40) data. To achieve this, the multiway data needs to be expanded into virtual pairwise RNA-DNA contacts similar to the strategy proposed for analyzing multiway DNA-DNA interactions obtained with Pore-C (41).

BaRDIC package is an easy-to-use tool since it requires only three files as input: a BED file with DNA parts of

contacts, an RNA gene annotation, and information on the background — a list of background mRNAs or an input track. This simplicity allows a user to parallelize peak calling over multiple RNAs with an ATA dataset or streamline BaRDIC application over multiple isolated datasets. Calling peaks on all available OTA and ATA experiments with scaling correction could help to refine known target genomic loci for each RNA of interest. Collectively, the resulting peaks can be organized in a database, such as RNA-Chrom. The unified approach will enable unbiased comparisons of specific RNA-DNA interactions between samples and studies as well as with other high throughput omics data.

Scaling correction during peak calling makes RNA-DNA interaction data reconcilable with specific protein-DNA interactions (ChIP-seq peaks), for which scaling does not exist, and specific DNA-DNA interactions (Hi-C loops), for which scaling is observed and accounted for in dedicated algorithms. This harmonization opens a way for unbiased comparisons of these three types of interactomic data for predicting functions of RNA-chromatin interactions. In particular, these comparisons will lead to predictions about how RNAs bind to their genomic targets: either due to a spatial proximity of corresponding DNA loci or due to recruitment by DNA-binding proteins. In turn, GO and GSEA enrichment analysis of genes interacting with novel chromatin-associated RNAs will create a basis for predicting regulatory functions of RNA-chromatin interactions. Taken together, simultaneously accounting for chromatin heterogeneity and scaling in RNA-DNA interaction data enables unbiased and uncomplicated downstream and comparative analyses.

ACKNOWLEDGEMENTS

The research was carried out using facilities of the Makarich HPC cluster provided by the Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University.

FUNDING

The research was supported by RSF (project No. 23-14-00136).

Conflict of interest statement. None declared.

REFERENCES

1. Xiaopei Zhang, Wei Wang, Weidong Zhu, Jie Dong, Yingying Cheng, Zujun Yin, and Fafu Shen. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *International Journal of Molecular Sciences*, 20(22):5573, November 2019.
2. Sergey V. Razin and Alexey A. Gavrilov. Non-coding RNAs in chromatin folding and nuclear organization. *Cellular and Molecular Life Sciences*, June 2021.
3. Xiao Li and Xiang-Dong Fu. Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nature Reviews Genetics*, 20(9):503–519, June 2019.
4. G K Ryabykh, D E Mylarshchikov, S V Kuznetsov, A I Sigorskikh, T Y Ponomareva, A A Zharikova, and A A Mironov. RNA–chromatin interactome: What? Where? When? *Mol. Biol.*, 56(2):210–228, April 2022.
5. Masaki Kato and Piero Carninci. Genome-wide technologies to study RNA–chromatin interactions. *Non-Coding RNA*, 6(2):20, May 2020.
6. Peter D. Mariner, Ryan D. Walters, Celso A. Espinoza, Linda F. Drullinger, Stacey D. Wagner, Jennifer F. Kugel, and James A. Goodrich. Human alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Molecular Cell*, 29(4):499–509, February 2008.
7. Ci Chu, Jeffrey Quinn, and Howard Y. Chang. Chromatin isolation by rna purification (ChIRP). *JoVE*, (61):e3912, Mar 2012.
8. Matthew D. Simon. Capture hybridization analysis of RNA targets (CHART). *Current Protocols in Molecular Biology*, 101(1), January 2013.
9. Jesse Engreitz, Eric S. Lander, and Mitchell Guttman. RNA antisense purification (RAP) for mapping RNA interactions with chromatin. In *Methods in Molecular Biology*, pages 183–197. Springer New York, December 2014.
10. Alessandro Bonetti, Federico Agostini, Ana Maria Suzuki, Kosuke Hashimoto, Giovanni Pascarella, Juliette Gimenez, Leonie Roos, Alex J. Nash, Marco Ghilotti, Christopher J. F. Cameron, Matthew Valentine, Yulia A. Medvedeva, Shuhei Noguchi, Eneritz Agirre, Kaori Kashi, Samudyata, Joachim Luginbühl, Riccardo Cazzoli, Saumya Agrawal, Nicholas M. Luscombe, Mathieu Blanchette, Takeya Kasukawa, Michiel de Hoon, Erik Arner, Boris Lenhard, Charles Plessy, Gonçalo Castelo-Branco, Valerio Orlando, and Piero Carninci. RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions. *Nature Communications*, 11(1), February 2020.
11. Xiao Li, Bing Zhou, Liang Chen, Lan-Tao Gou, Hairi Li, and Xiang-Dong Fu. GRID-seq reveals the global RNA–chromatin interactome. *Nature Biotechnology*, 35(10):940–950, September 2017.
12. Alexey A Gavrilov, Anastasiya A Zharikova, Aleksandra A Galitsyna, Artem V Luzhin, Natalia M Rubanova, Arkadiy K Golov, Nadezhda V Petrova, Maria D Logacheva, Omar L Kantidze, Sergey V Ulianov, Mikhail D Magnitov, Andrey A Mironov, and Sergey V Razin. Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Research*, 48(12):6699–6714, June 2020.
13. G K Ryabykh, S V Kuznetsov, Y D Korostelev, A I Sigorskikh, A A Zharikova, and A A Mironov. RNA-chrom: a manually curated analytical database of RNA–chromatin interactome. *Database*, 2023, January 2023.
14. Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.
15. Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, January 2010.
16. W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, , and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, May 2002.
17. Yoav Benjamini, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, 125(1-2):279–284, November 2001.
18. Hamid Alinejad-Rokny, Rassa Ghavami Modegh, Hamid R. Rabiee, Ehsan Ramezani Sarbandi, Narges Rezaie, Kin Tung Tam, and Alistair R. R. Forrest. MaxHiC: A robust background correction model to identify biologically relevant chromatin interactions in hi-c and capture hi-c experiments. *PLOS Computational Biology*, 18(6):e1010241, June 2022.
19. Qiang Song and Andrew D. Smith. Identifying dispersed epigenomic domains from ChIP-seq data. *Bioinformatics*, 27(6):870–871, February 2011.
20. Mahmoud M. Ibrahim, Scott A. Lacadie, and Uwe Ohler. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, 31(1):48–55, September 2014.
21. Arya Kaul, Sourya Bhattacharyya, and Ferhat Ay. Identifying statistically significant chromatin contacts from hi-c data with FitHiC2. *Nature Protocols*, 15(3):991–1012, January 2020.
22. Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-seq (MACS). *Genome Biology*, 9(9), September 2008.
23. Mark Carty, Lee Zamparo, Merve Sahin, Alvaro González, Raphael Pelossof, Olivier Elemento, and Christina S. Leslie. An integrated model

for detecting significant chromatin interactions from high-resolution hi-c data. *Nature Communications*, 8(1), May 2017.

24. Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.
25. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
26. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza and. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, February 2020.
27. Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the Python in Science Conference*. SciPy, 2010.
28. Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, and Sergey V. Venev. Bioframe: Operations on genomic intervals in pandas dataframes. February 2022.
29. The HDF Group. Hierarchical data format version 5, 2000-2010.
30. Nezar Abdennur and Leonid A Mirny. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1):311–316, 2020.
31. Noa Gil and Igor Ulitsky. Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews Genetics*, 21(2):102–117, November 2019.
32. J Britton-Davidian, B Cazaux, and J Catalan. Chromosomal dynamics of nucleolar organizer regions (NORs) in the house mouse: micro-evolutionary insights. *Heredity*, 108(1):68–74, November 2011.
33. Xing Li, Daniel N. Frank, Norman Pace, Janice M. Zengel, and Lasse Lindahl. Phylogenetic analysis of the structure of RNase MRP RNA in yeasts. *RNA*, 8(6):740–751, June 2002.
34. Alexander Favorov, Loris Mularoni, Leslie M. Cope, Yulia Medvedeva, Andrey A. Mironov, Vsevolod J. Makeev, and Sarah J. Wheelan. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Computational Biology*, 8(5):e1002529, May 2012.
35. Jason A. West, Christopher P. Davis, Hongjae Sunwoo, Matthew D. Simon, Ruslan I. Sadreyev, Peggy I. Wang, Michael Y. Tolstorukov, and Robert E. Kingston. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Molecular Cell*, 55(5):791–802, September 2014.
36. Bin Zhang, Gayatri Arun, Yuntao S. Mao, Zsolt Lazar, Gene Hung, Gourab Bhattacharjee, Xiaokun Xiao, Carmen J. Booth, Jie Wu, Chaolin Zhang, and David L. Spector. The lncRNA malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell Reports*, 2(1):111–123, July 2012.
37. Moritz Eißmann, Tony Gutschner, Monika Hämmerle, Stefan Günther, Maïwen Caudron-Herger, Matthias Groß, Peter Schirmacher, Karsten Rippe, Thomas Braun, Martin Zörnig, and Sven Diederichs. Loss of the abundant nuclear non-coding RNAiMALAT1/iis compatible with life and development. *RNA Biology*, 9(8):1076–1087, August 2012.
38. Yafei Yin, Pixi Yan, Jinlong Lu, Guang Song, Yangyang Zhu, Zhaohui Li, Yi Zhao, Bin Shen, Xingxu Huang, Heng Zhu, Stuart H. Orkin, and Xiaohua Shen. Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. *Cell Stem Cell*, 16(5):504–516, May 2015.
39. Charles Limouse, Owen K. Smith, David Jukam, Kelsey A. Fryer, William J. Greenleaf, and Aaron F. Straight. Global mapping of rna-chromatin contacts reveals a proximity-dominated connectivity model for ncrna-gene interactions. *bioRxiv*, 2022.
40. Sofia A. Quinodoz, Joanna W. Jachowicz, Prashant Bhat, Noah Ollikainen, Abhik K. Banerjee, Isabel N. Goronzy, Mario R. Blanco, Peter Chovanec, Amy Chow, Yolanda Markaki, Jasmine Thai, Kathrin Plath, and Mitchell Guttman. RNA promotes the formation of spatial compartments in the nucleus. *Cell*, 184(23):5775–5790.e30, November 2021.
41. Aditya S. Deshpande, Netha Ulahannan, Matthew Pendleton, Xiaoguang Dai, Lynn Ly, Julie M. Behr, Stefan Schwenk, Will Liao, Michael A. Augello, Carly Tyer, Priyesh Rughani, Sarah Kudman, Huasong Tian, Hannah G. Otis, Emily Adney, David Wilkes, Juan Miguel Mosquera, Christopher E. Barbieri, Ari Melnick, David Stoddart, Daniel J. Turner, Sissel Juul, Eoghan Harrington, and Marcin Imieliński. Identifying synergistic high-order 3d chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology*, 40(10):1488–1499, May 2022.