Title

Systematic Mapping of RNA-Chromatin Interactions In Vivo

Journal

Current Biology, 27(4)

ISSN

0960-9822

Authors

Sridhar, Bharat
Rivas-Astroza, Marcelo
Nguyen, Tri C
et al.

Publication Date

2017-02-01

DOI

10.1016/j.cub.2017.01.011

Peer reviewed

Published in final edited form as:

Curr Biol. 2017 February 20; 27(4): 602–609. doi:10.1016/j.cub.2017.01.011.

Systematic mapping of RNA-chromatin interactions *in vivo*

Bharat Sridhar[1,2,3], **Marcelo Rivas-Astroza**[1,3], **Tri C. Nguyen**[1,3], **Weizhong Chen**[1], **Zhangming Yan**[1], **Xiaoyi Cao**[1], **Lucie Hebert**[1], and **Sheng Zhong**[1,4]

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

[2]Department of Molecular and Integrative Physiology, University of Illinois Urbana-Champaign, Urbana, IL, 61801

Summary

RNA molecules can attach to chromatin. It remains difficult to know what RNAs are associated with chromatin and where are the genomic target loci of these RNAs. Here, we present MARGI (Mapping RNA-genome interactions), a technology to massively reveal native RNA-chromatin interactions from unperturbed cells. The gist of this technology is to ligate chromatin associated RNAs (caRNAs) with their target genomic sequences by proximity ligation, forming RNA-DNA chimeric sequences, which are converted to sequencing library for paired-end sequencing. Using MARGI, we produced RNA-genome interaction maps for human embryonic stem (ES) cells and HEK cells. MARGI revealed hundreds of caRNAs including previously known *XIST*, *SNHG1, NEAT1, MALAT1,* as well as each caRNA's genomic interaction loci. Using a cross-species experiment, we estimated that approximately 2.2% of MARGI identified interactions were false positives. In ES and HEK cells, the RNA ends of more than 5% of MARGI read pairs were mapped to distal or inter-chromosomal locations as compared to the locations of their corresponding DNA ends. The majority of transcription start sites are associated with distal or inter-chromosomal caRNAs. ChIP-seq reported H3K27ac and H3K4me3 levels are positively while H3K9me3 is negatively correlated with MARGI reported RNA attachment levels. The MARGI technology should facilitate revealing novel RNA functions and their genomic target regions.

Graphical abstract

Sridhar et al. develop a technology to map global RNA-chromatin interactions in unperturbed cells. They discover hundreds of chromatin associated RNAs. They find that the majority of

Correspondence to: Sheng Zhong.

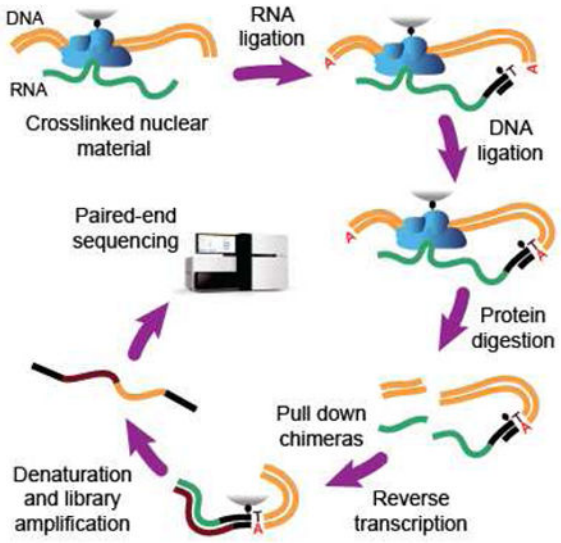
[3]Co-first author

[4]Lead Contact

Supplemental Information: Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online. All sequencing data are available at Gene Expression Omnibus with access number GSE92345.

Author Contributions: B.S., T.C.N., and S.Z. designed the experiments. B.S., T.C.N., and L.H performed the experiments. All authors analyzed and interpreted the data.

known transcription start sites are associated with chromatin associated RNAs, suggesting that this phenomenon is more widespread than previously thought.



Results and Discussion

Development of the MARGI technology

We developed MARGI (Mapping RNA-genome interactions), a technology to massively reveal RNA-chromatin interactions from unperturbed cells. MARGI simultaneously identifies all caRNAs and the respective genomic target loci of each caRNA. This changes the paradigm of analyzing one-RNA-at-a-time, and enables the mapping of the native RNA-chromatin interaction network in a single experiment. The major innovation of the MARGI technology is to ligate caRNAs with their target genomic sequences by proximity ligation, forming RNA-DNA chimeric sequences, which are subsequently converted into a sequencing library and subjected to paired-end sequencing. We have developed two variations of the MARGI technology. The first, which we refer to as proximity MARGI (pxMARGI), does not differentiate passive and direct interactions. The second, which we refer to as direct MARGI (diMARGI), is designed to reveal protein or RNA tethered interactions. Throughout this paper, we call pxMARGI- and diMARGI-identified caRNAs pxRNA and diRNA, respectively.

In the MARGI procedure, chromatin is crosslinked, fragmented, and subsequently biotinylated and stabilized on streptavidin beads (Figure 1). A specially designed linker sequence is introduced to first ligate with the 3′ end of RNA (RNA ligation) and then this RNA ligated linker is ligated to DNA (DNA ligation). These ligation steps are controlled by the configuration of the linker and by sequential applications of different end modifications and ligation enzymes. Successfully ligated products, in the form of RNA – linker – DNA, are selected and converted to cDNA. This cDNA is circularized and then cut in the middle of the linker, producing cDNA with the configuration: left.half.Linker – DNA – RNA – right.half.Linker, which are subjected to paired-end sequencing. The left.half.Linker and the

right.half.Linker are the two halves of the linker separated by a restriction site in the middle. These two halves serve to differentiate which end originated from RNA and which end was from DNA, respectively.

The linker was designed to achieve three goals. The first is to maximize the ligation efficiency of the linker to the RNA and to the DNA. The second is to ensure that RNA and DNA are only ligated to the designated strand of the linker in the desired orientation. Third, when made into sequencing library, the cDNA originated from RNA is next to SP1 sequencing primer and the DNA side is next to SP2 sequencing primer. This makes it straightforward to associate Read 1 and Read 2 from a paired-end sequencing readout with the RNA and the DNA side, respectively.

To optimize the ligation efficiency with RNA, the 5′ end of the top strand of the linker starts with an 5′-Adenylation (5′App) followed by 10 single-stranded DNA bases (Figure 1A). In addition, the first two single-stranded bases are random DNA bases (“NN” in the top strand, Figure 1A), designed to equalize the ligation efficiency with all four bases at the 3′ end of the RNA. The RNA-linker ligation is catalyzed by a mutated version of T4 RNA Ligase 2 (T4 RNA Ligase 2, truncated KQ), which selectively ligates single-stranded RNA with 3′ hydroxyl (3′-OH) to single-stranded RNA or DNA with 5′App without using ATP [1], therefore minimizing spurious RNA-RNA ligation (Figure 1B). To maximize the ligation efficiency of linker and DNA, sticky-end ligation is used where the 3′-T overhang of the linker is joined with the 3′-A overhang produced by A-tailing of the DNA (Figure 1C).

In order to arrange the RNA side (cDNA) next to the SP1 and the DNA side next to the SP2 sequencing primers, we designed a “phasing” strategy that involves circularization and re-linearization (Figure 1D). The center of the linker sequence is a BamHI restriction site. The right and left flanking sequences of the linker are designed to be complementary to the SP1 and SP2 sequencing primers, respectively. After circularization and restriction enzyme digestion, the linker is split and the two halves are rearranged to the two ends of the product sequence, respectively. This enables the product sequences originated from the RNA and the DNA to be sequenced as Read 1 and Read 2, respectively, from the two distinct sequencing primers by paired-end sequencing.

To test the phasing strategy, we generated two MARGI libraries from ES cells, where the genome was fragmented, respectively, by sonication (sonication library), and by HaeIII restriction enzyme (HaeIII library), which recognizes “GGCC” sequence and leaves “CC” at the 5′ end of the cut. If the circularization strategy can phase the RNA and the DNA ends into Read 1 and Read 2, we would expect to see the first two bases of Read 2 (DNA end) in the HaeIII library to be enriched with “CC”, the signature of HaeIII cut. Indeed, Read 1 (RNA end) exhibited nearly equal frequencies of A, C, G, and T, whereas more than 97% of the Read 2 sequences started with “CC” (Figure 2A). In contrast, both Read 1 and Read 2 in the sonication library exhibited similar amounts of the four nucleotides in their first two bases (Figure 2B). Hereafter, we call the two ends of a paired-end read the RNA mate and the DNA mate, respectively.

To assess the specificity of MARGI, we carried out two control experiments. In the first control experiment, we generated from HEK293T cells three MARGI libraries in parallel with identical steps, except that in Library 2 we left out the RNA-Linker ligation step (RNA ligation -, Figure S1A) and in Library 3 we excluded the Linker-DNA ligation step (DNA ligation -, Figure S1A). The MARGI procedure without either the RNA or the DNA ligation step could not produce a detectable amount of DNA in the final sequencing library. In the second control experiment, we used a mixture of *Drosophila* S2 cells and human HEK293T cells to estimate the extent of unspecific ligations (cross-species control)[2, 3]. After cross-linking and cell lysis, the lysates from the two species were mixed before any subsequent steps. The mixture was subjected to the rest of the MARGI procedure, and resulting in a sequenced library (Fly-Hs). A total of 7,066,395 paired-end reads were mapped to either of the two genomes (both ends were uniquely mapped), among which 969,034 (13.71%) had both ends mapped to the fly genome (dm6) and 5,942,976 (84.1%) had both ends mapped to the human genome (hg38), whereas 154,385 (2.18%) had one end mapped to fly genome and the other end mapped to human genome. This suggests that random ligations are uncommon in the MARGI procedure.

Genome-wide analyses of short- and long-range RNA-chromatin interactions

We designed pxMARGI to reveal caRNA in spatial proximity of any genomic region. This was achieved by a combination of formaldehyde crosslinking [4] and complete genome fragmentation achieved by overnight HaeIII digestion to ensure all genomic regions including heterochromatin were fragmented (Figure S1B-C) before the subsequent DNA ligation step. We generated two pxMARGI libraries from human HEK (HEK293T, biological replicates, Samples 1-2 in Table S1) and two pxMARGI libraries from H9 human ES cells (biological replicates, Samples 4-5, Table S1). Sequencing and mapping these libraries yielded approximately 105 million and 65 million non-redundant and mappable read pairs (both ends are uniquely mapped) for HEK and H9 cells, respectively. We separated the mapped read pairs into three categories, namely proximal (mapped within 2,000 bp on the same chromosome), distal (mapped to the same chromosome with a distance larger than 2,000 bp), and inter-chromosomal. The two cell types yielded similar distributions of the three types of read pairs, where approximately 80% were proximal, less than 4% were distal, and 15 – 20 % were inter-chromosomal, suggesting a non-trivial amount of long-range interactions (Table S1).

To prioritize potentially “direct” interactions, we developed diMARGI. In this procedure, after crosslinking with formaldehyde and DSG [5], the chromatin was fragmented by sonication, and only the soluble fraction was passed onto the subsequent ligation steps. Non-crosslinked RNA was removed prior to the ligation steps by protein denaturation with stringent washing and binding buffers [6]. We generated one diMARGI library from HEK and one library from ES cells. Compared to pxMARGI libraries, the proximal read pairs increased to 94% and 95% in HEK and human ES cells, respectively, distal pairs dropped to 1% in both cell types, inter-chromosomal pairs dropped to 5% and 4%, respectively (Table S1). In the rest of this paper, all the analyses are based on long-range (distal and inter-chromosomal) interactions.

Identification of chromatin-associated pseudogene RNAs, antisense RNAs, and lincRNAs

To identify the caRNAs, we calculated RPKM for every gene based on the RNA-end (Read 1) from mapped distal and inter-chromosomal read pairs, and tested whether the RPKM equals 0. This led to identification of 2,864 and 1,933 non-coding pxRNAs and 747 and 467 non-coding diRNAs, from HEK and ES cells, respectively (FDR < 0.0001) (Figure 2C-J). The largest components of non-coding pxRNAs were pseudogene, antisense, and lincRNA, which was the same for both cell types (Figure 2G-H), whereas the largest component of diRNAs in both cell types was snoRNA (Figure 2I-J), consistent with its role in maintaining accessible structures of euchromatin [7] and attachment to nascent target transcripts [8, 9] (reviewed by [10]).

The pair of MARGI technologies provided an opportunity to evaluate the chances for each caRNA to interact passively or in the “direct model”, and whether the model of interaction is cell type specific. We compared pxMARGI and diMARGI data of previously known caRNAs including *MALAT1, NEAT1, SNHG1*, and *XIST* [11, 12]. All these lincRNAs were identified as pxRNAs in both HEK and ES cells (Figure 2C-D). *MALAT1, SNHG1, NEAT1* became even more significant in diMARGI, and *MALAT1* and *SNHG1* rose into the few most significant diRNAs in both cell types (blue and red circles in Figure 2E-F, Figure 3A-B). The identification of *XIST* as pxRNA is consistent with the reported *XIST* activity in H9 ES cells [13]. Interestingly, *XIST* did not exhibit diRNA activity in H9 ES cells (Figure 2F), but it was one of the most significant diRNA in HEK cells (black circle in Figure 2E, Figure 3C). This may suggest an underappreciated difference between the modes of *XIST*-X chromosome interactions in H9 ES cells and differentiated cells.

Overview of pxRNA and diRNA associated genomic regions

We compared the genomic regions identified by pxMARGI and diMARGI. We called peaks from the DNA ends of distal and inter-chromosomal read pairs using MACS v1.4.2 [14], which we call pxPeaks and diPeaks. As anticipated, pxPeaks were greater in numbers than diPeaks (Figure S2A), and larger in sizes (Figure 3A), whereas the size distributions were similar between the two cell types for either pxPeaks or diPeaks (Figure 4A). HEK cells yielded 120,872 pxPeaks, which was more than twice of that from ES cells (57,154). In comparison, the amount of diPeaks from HEK (7,212) remains larger than that from ES (5,247), but the difference was not as great as that in pxPeaks. This is reminiscent of the idea that pxRNA may be trapped in closed chromatin, because stem cell differentiation is usually coupled with chromatin condensation [15, 16].

The majority of known promoters is associated with pxRNA and diRNA

We asked whether pxPeaks were correlated to any known genomic features. We analyzed the overlaps with known genomic features including promoters, 5′ UTR, 3′ UTR, exons, introns, downstream sequence (3 kb), and intergenic sequence, accounting for overlaps to multiple genomic features by the recently developed “Upset” tool [17]. In HEK cells, approximately 37% of pxPeaks overlapped with intergenic regions and 17% overlapped with promoters (Figure 4B, Figure S3A). Adjusting for the sizes of these genomic features, pxPeaks were enriched in promoters (Odds ratio = 1.7, p-value < 2×10^{-16}, Chi-squared test)

(Figure S2B). ES cells exhibited very similar proportions, and an enrichment in promoters (p-value < 2×10^{-16}) (Figure 3B, Figure S2B).

This observation led us to directly assess the degree of association between promoters and pxRNA. We plotted the density of pxRNA across the 20,000 bp flanking regions of every transcription start site (Figure 4C). Out of 34,475 human genes (GRCh38) with non-redundant transcription start sites, 23,838 (69.1%) exhibited increased pxRNA intensities at their transcription start sites in HEK cells. Even more transcription start sites (25,392, 73.7%) exhibited increased pxRNA intensity in ES cells (Figure 4C).

The overlaps of diPeaks to promoters increased to 61% (4,391) in HEK and 63% (3,306) in ES cells (Figure 4B), and the odds ratios for these overlaps increased to 16.8 (HEK, p-value < 2×10^{-16}, Chi-squared test) and 17.7 (ES, p-value < 2×10^{-16}, Chi-squared test) (Figure S2B, Figure S3B). Similar to pxRNA, diRNA intensities increased in promoters, but became more concentrated; forming sharp peaks centered at transcription start sites (Figure 4D). A total of 18,135 transcription start sites in HEK and 6,551 transcription start sites in ES exhibited clear increases of diRNA attachments.

Correlations of caRNA intensity and histone modification levels

We asked whether caRNA intensity is correlated with ChIP-seq defined histone modification levels. To this end, we calculated RAL (RNA attachment level) for each genomic segment as the average read count of the DNA-end of distal and inter-chromosomal read pairs. Proximal read pairs were excluded from RAL calculation. Across all transcription start sites, RAL exhibited positive correlations with H3K4me3, H3K27ac, and negative correlation with H3K9me3 (Figure 4C-D). These correlations were preserved in the datasets generated by pxMARGI and diMARGI.

We proceeded to analyze the entire genome by scanning the genome with 1,000 bp windows. diRNA RAL exhibited positive correlations with H3K4me3 and H3K27ac, and a negative correlation with H3K9me3 (Figure 4E). In comparison, pxRNA RAL did not exhibit clear genome-wide correlations to H3K4me3 and H3K27ac (Figure 4E), possibly attributable to lack of H3K4me3 and H3K27ac in condensed chromatin. Interestingly, pxRNA RAL retained genome-wide anti-correlation with H3K9me3 (pxMARGI, Figure 4E). Moreover, H3K9me3 was depleted in nearly all diPeaks and all pxPeaks (Figure 4F), suggesting a competition between RNA attachment and H3K9me3 events in closed chromatin.

H9 ES cells may represent an incomplete state of X chromosome inactivation

XIST exhibited pxRNA activity but not diRNA activity in H9 cells (Figure 2D,F), which led to a possibility that H9 ES cells represent an incomplete state of X chromosome inactivation (XCI). More specifically, H9 and HEK share *XIST* involvement in heterochromatin, but *XIST* also attaches to less compact parts of the X chromosome through protein bridges specifically in HEK cells. If this was the case, we would anticipate to see differences in diRNA activities on *XIST* associated lincRNAs *TSIX* [18, 19] and *FTX* [20] between HEK and H9 cells. The two lincRNAs exhibited similar degrees of pxRNA activities between HEK and H9 (Figure S4A-B), consistent with the observed *XIST* pxRNA activities and

reported XCI in H9 cells [13]. However, *TSIX* and *FTX* exhibited much reduced diRNA activities in H9 as compared to HEK (Figure S4C-D), consistent with the minimal diRNA activity of *XIST* in H9.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by NIH 1DP1HD087990.

References

1. Yin S, Ho CK, Shuman S. Structure-function analysis of T4 RNA ligase 2. The Journal of biological chemistry. 2003; 278:17601–17608. [PubMed: 12611899]
2. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013; 153:654–665. [PubMed: 23622248]
3. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502:59–64. [PubMed: 24067610]
4. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
5. Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science. 2013; 341:1237973. [PubMed: 23828888]
6. Tagwerker C, Flick K, Cui M, Guerrero C, Dou Y, Auer B, Baldi P, Huang L, Kaiser P. A tandem affinity tag for two-step purification under fully denaturing conditions: application in ubiquitin profiling and protein complex identification combined with in vivocross-linking. Molecular & cellular proteomics : MCP. 2006; 5:737–748. [PubMed: 16432255]
7. Schubert T, Pusch MC, Diermeier S, Benes V, Kremmer E, Imhof A, Langst G. Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. Molecular cell. 2012; 48:434–444. [PubMed: 23022379]
8. Cavaille J, Nicoloso M, Bachellerie JP. Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. Nature. 1996; 383:732–735. [PubMed: 8878486]
9. Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, Leon-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell. 2014; 159:148–162. [PubMed: 25219674]
10. Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. Cell. 2011; 144:16–26. [PubMed: 21215366]
11. West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, Tolstorukov MY, Kingston RE. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Molecular cell. 2014; 55:791–802. [PubMed: 25155612]
12. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. Xist RNA and the mechanism of X chromosome inactivation. Annual review of genetics. 2002; 36:233–278.
13. Shen Y, Matsuno Y, Fouse SD, Rao N, Root S, Xu R, Pellegrini M, Riggs AD, Fan G. X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:4709–4714. [PubMed: 18339804]

14. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008; 9:R137. [PubMed: 18798982]

15. Chalut KJ, Hopfler M, Lautenschlager F, Boyde L, Chan CJ, Ekpenyong A, Martinez-Arias A, Guck J. Chromatin decondensation and nuclear softening accompany Nanog downregulation in embryonic stem cells. Biophysical journal. 2012; 103:2060–2070. [PubMed: 23200040]

16. Ugarte F, Sousae R, Cinquin B, Martin EW, Krietsch J, Sanchez G, Inman M, Tsang H, Warr M, Passegue E, et al. Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells. Stem cell reports. 2015; 5:728–740. [PubMed: 26489895]

17. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. IEEE transactions on visualization and computer graphics. 2014; 20:1983–1992. [PubMed: 26356912]

18. Lee JT, Davidow LS, Warshawsky D. Tsix, a gene antisense to Xist at the X-inactivation centre. Nat Genet. 1999; 21:400–404. [PubMed: 10192391]

19. Migeon BR, Chowdhury AK, Dunston JA, McIntosh I. Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. American journal of human genetics. 2001; 69:951–960. [PubMed: 11555794]

20. Chureau C, Chantalat S, Romito A, Galvani A, Duret L, Avner P, Rougeulle C. Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region. Human molecular genetics. 2011; 20:705–718. [PubMed: 21118898]

Highlights

- MARGI globally profiles native RNA-chromatin interactions.
- Two variations of MARGI interrogate proximity-based and direct interactions.
- The majority of transcription start sites are associated with interacting RNAs.
- RNA attachment is positively correlated with active histone marks in promoters.

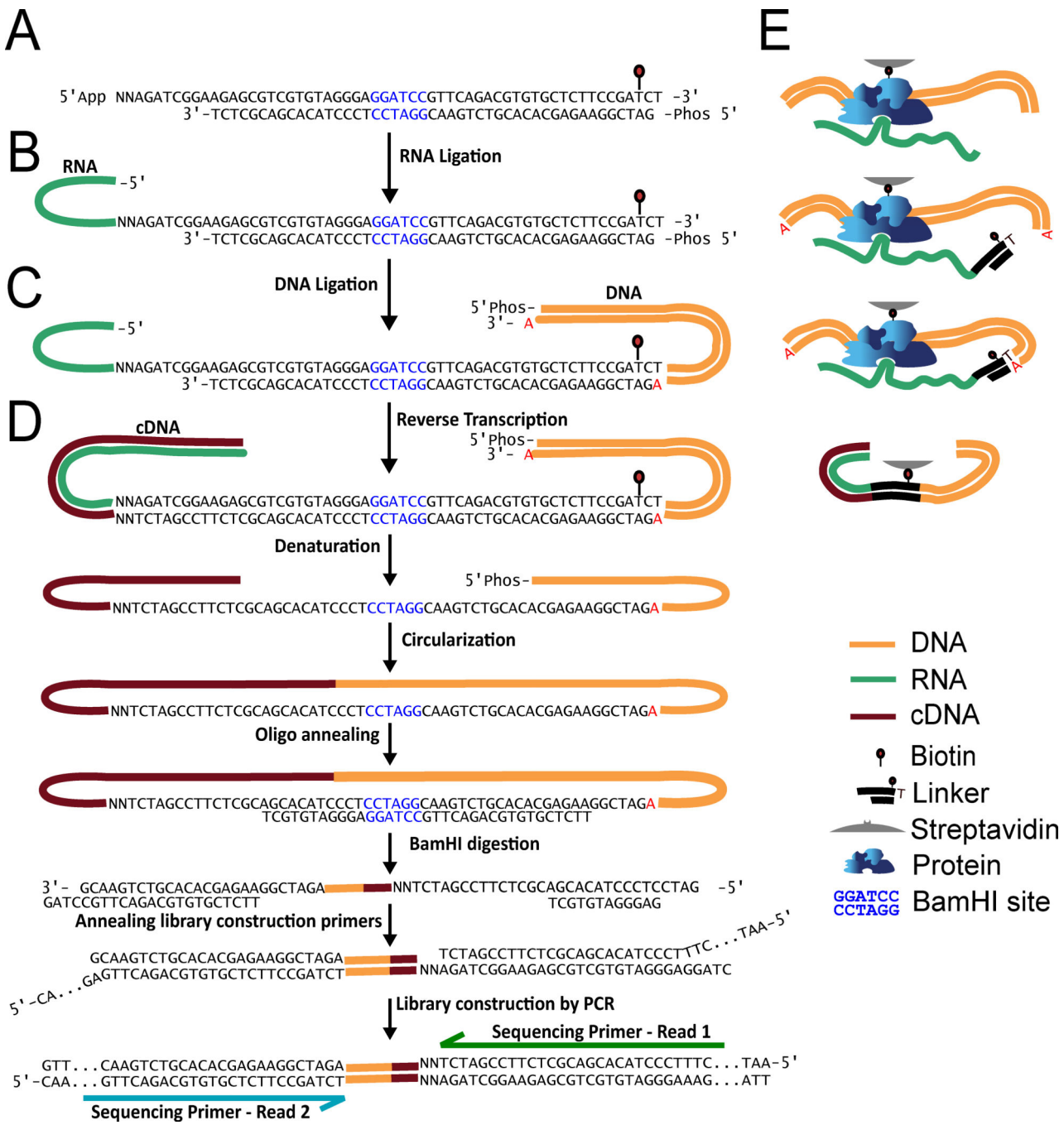


Figure 1. Schema for the MARGI technology

(A) Linker sequence. The linker is composed of double-stranded DNA in the center and single-stranded DNA on the two ends. The top strand of the linker is 11 bases longer than the bottom strand, leaving 10 bases of single-stranded DNA at the 5′ end and 1 base of single-stranded DNA at the 3′ end. The 5′ end of the top strand is adenylated (5′ App) and the 5′ end of the bottom strand is phosphorylated (Phos). N: random base. Letters in blue: BamHI restriction site. (B) RNA-Linker ligation. RNA with 3′-OH was produced by T4 PNK treatment. (C) Linker-DNA ligation. A single base “A” tail (in red) is added to the 3′

end of DNA, which enables a sticky-end ligation to the linker. (D) Circularization and re-linearization. After BamH1 digestion, the linker sequence is split and re-allocated to the two ends, which were by design complementary to the library construction primers. (E) Another perspective of the ligation and reverse transcription steps as shown in A-D. See also Figure S1.

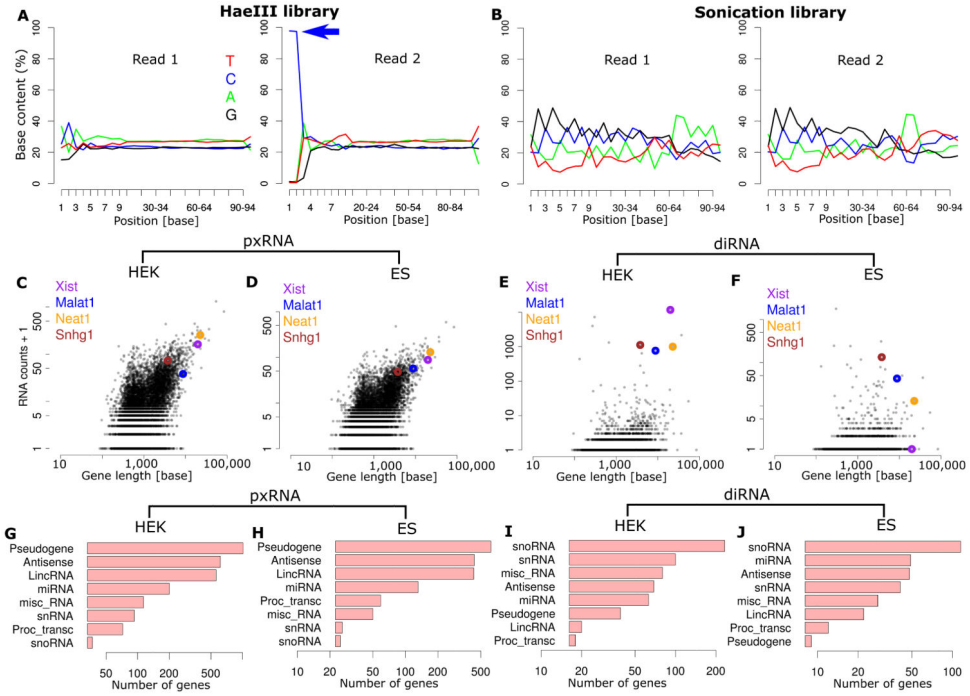


Figure 2. MARGI identified caRNAs

(A-B) Composition of the four nucleotide bases on every position of paired-end reads in a HaeIII library (A) and in a sonication library (B). Arrow: increased proportions of Cytosine (C) in the first two positions, which is specific to Read 2 (DNA end) of the HaeIII library. (C-F) Scatter plots of lincRNAs with MARGI RNA-end read counts per gene (y axis) plotted against gene length (x axis). (G-J) Numbers of non-coding pxRNAs (G-H) and diRNAs (I-J) categorized by RNA type. Proc_transcripts: processed transcripts, misc_RNA: miscellaneous other RNA. See also Figure S4.

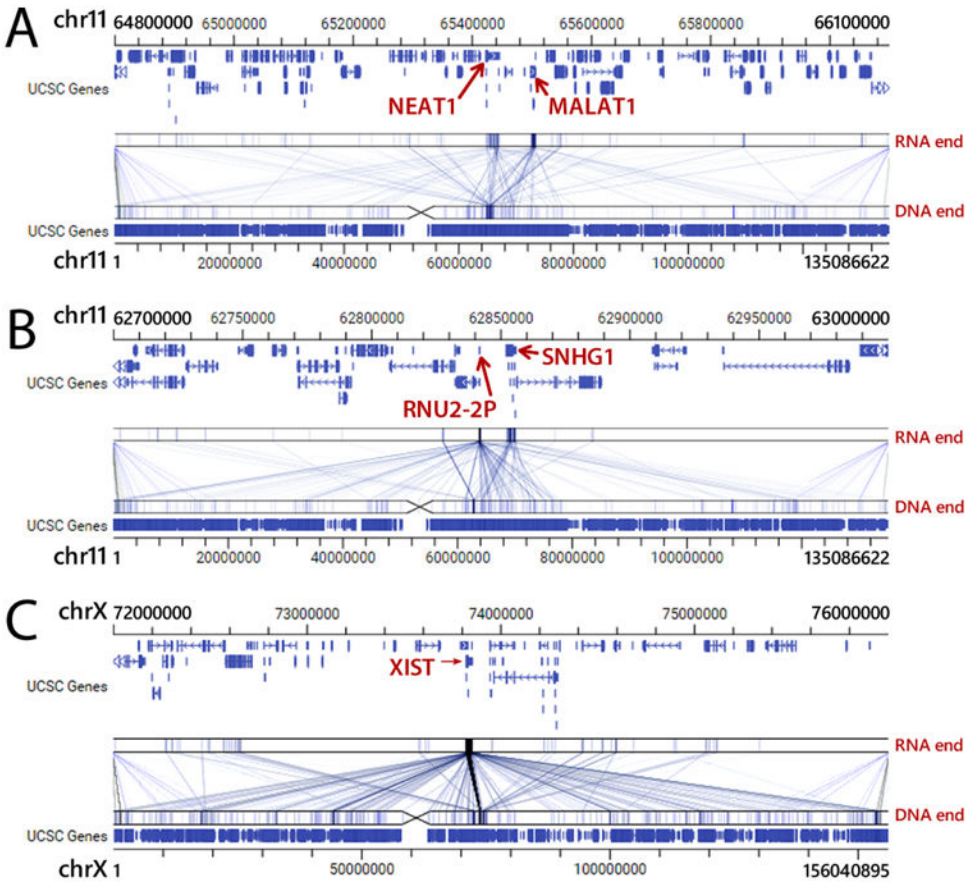


Figure 3. Genome view of mapped MARGI read pairs

The mapped MARGI reads are plotted with Genome Interaction Visualizer (GIVe, http://give.genemo.org/?hg38), where the reference genome is plotted horizontally, twice (top and bottom bars). The top and the bottom bars can be zoomed in or out independent of each other. The mapped RNA ends are shown on the top bar (genome), and the DNA ends are shown on the bottom bar (genome). Each MARGI read pair is represented as a line linking the locations of RNA end (top) and the DNA end (bottom). The HEK diMARGI data at are shown with *MALAT1* locus (top) versus the entire Chromosome 11 (bottom) (A), the *SNHG1* locus versus the entire Chromosome 11 (B), and the *XIST* locus versus the entire Chromosome X (C). Red arrows point to known caRNAs.

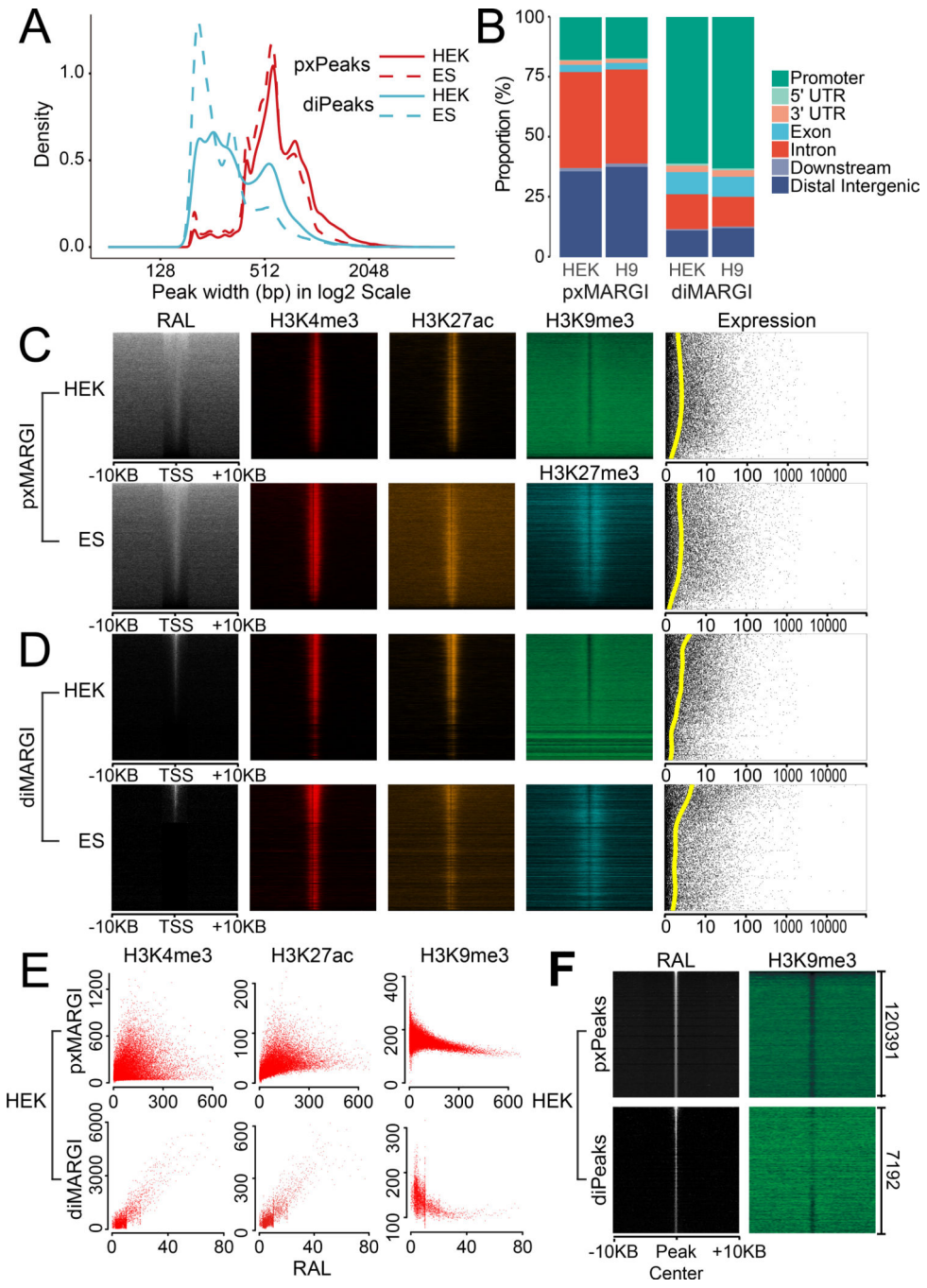


Figure 4. caRNA associated genomic regions

(A) Size distributions of pxPeaks (red) and diPeaks (blue) from HEK (solid) and ES cells (dashed curve). (B) Proportions of pxPeaks and diPeaks in promoters, UTRs, exons, introns, downstream regions, and distal intergenic regions. (C-D) RNA attachment levels (RAL) (white) in 20,000 bp flanking regions of all (34,475) non-redundant TSSs (GRCh38), as derived from pxMARGI (C), diMARGI (D) in descending orders. Shown in parallel are H3K4me3 (red), H3K27ac (orange), H3K9me3 (green), and H3K27me3 (blue, ES only, no available data in HEK) intensities, and gene expression levels. Yellow curve: smoothed

average of gene expression levels. TSS: transcription start site. (E) Scatter plots of 1,000 bp genomic windows with histone modification levels (y axis) versus RALs (x axis). Each data point represents the average of 100 windows. (F) RAL were plotted for all identified pxPeaks and diPeaks and their 20,000 flanking regions (white). Shown in parallel are H3K9me3 levels on the same regions (green). See also Figures S2, S3.