Mapping RNA–RNA interactome and RNA structure in vivo by MARIO

Tri C. Nguyen1,*, Xiaoyi Cao1,*, Pengfei Yu1,*, Shu Xiao1, Jia Lu1, Fernando H. Biase1, Bharat Sridhar1, Norman Huang1, Kang Zhang2 & Sheng Zhong1

The pervasive transcription of our genome presents a possibility of revealing new genomic functions by investigating RNA interactions. Current methods for mapping RNA–RNA interactions have to rely on an ‘anchor’ protein or RNA and often require molecular perturbations. Here we present the MARIO (Mapping RNA interactome in vivo) technology to massively reveal RNA–RNA interactions from unperturbed cells. We mapped tens of thousands of endogenous RNA–RNA interactions from mouse embryonic stem cells and brain. We validated seven interactions by RNA antisense purification and one interaction using single-molecule RNA–FISH. The experimentally derived RNA interactome is a scale-free network, which is not expected from currently perceived promiscuity in RNA–RNA interactions. Base pairing is observed at the interacting regions between long RNAs, including transposon transcripts, suggesting a class of regulatory sequences acting in trans. In addition, MARIO data reveal thousands of intra-molecule interactions, providing in vivo data on high-order RNA structures.
More than 97% of the human genome are non-coding DNA, of which >98% do not appear to be cis-regulatory sequences\(^1\). Thus, the functions of the vast majority of the human genome remains unknown. We noted that >85% of the human genome is transcribed into RNA\(^2\). Important functions of several classes of RNA were discovered by studying RNA–RNA interactions. These discoveries include that transfer RNA interacts with messenger RNA to translate genetic code\(^3\), and that microRNA (miRNA) interacts with mRNA to promote their degradation\(^4\), as well as interactions related to RNA splicing, editing and ribosomal RNA maturation. Are there other unknown RNA–RNA interactions with important functions? The technical challenge lies in the astronomical number of possible RNA pairs. A genome-wide technology for identifying RNA interactions in vivo is much needed.

Interactions between RNA molecules are often mediated by RNA-binding proteins\(^5\) such as ARGONAUTE proteins\(^6\), PUM2, QKI\(^7\) and small nuclear RNA proteins\(^8\). However, it is difficult to directly observe RNA–RNA interactions facilitated by single proteins in normal cellular conditions. CLASH\(^9\)–\(^11\) and hiCLIP\(^12\) use transformed cell lines that overexpress the facilitating protein. It is unclear to what extent that ectopic expression or genome-insertion-based cell transformation would influence RNA–RNA interactions. PAR-CLIP\(^14\) and HITS-CLIP\(^15\) assay RNAs attached to an RNA-binding protein, which do not directly assay RNA–RNA interactions. Most importantly, all the methods above trace the interactions ‘anchored’ at a known protein or RNA. It is infeasible to map the entire RNA–RNA interactome by extensions of these one-RNA-at-a-time or one-protein-at-a-time methods.

As the previous technologies relied on an ‘anchor’ RNA or protein, the topology of the entire RNA–RNA interactome remains unknown. Inferring from the notion that regulatory RNAs ‘promiscuously’ interact with 300–1,000 target RNAs\(^11\)–\(^20\), one would probably guess that the RNA–RNA interactome has a flat topology, as opposed to a hierarchical topology\(^21\)–\(^22\) that is shared by many other biological networks\(^21\)–\(^22\).

The MARIO technology maps RNA–RNA interactions in a massive scale. MARIO can identify protein-assisted between-molecule and within-molecule RNA interactions. The MARIO identified RNA–RNA interactome is composed of tens of thousands of interactions, which involve mRNA, long intergenic noncoding RNA (lincRNA), small nuclear RNA (snoRNA), small nuclear RNA, tRNA, miRNA, transposon RNA, pseudogene RNA, antisense RNA and novel transcripts. The MARIO identified RNA–RNA interactome is a scale-free network. Long non-coding RNA including lincRNA, transposon RNA and pseudogene RNA are observed to interact with mRNA. Sequence complementation is observed in interactions between transposon (LINE and LTR) RNA and mRNA, as well as in mRNA–mRNA, mRNA–pseudogeneRNA, lincRNA–mRNA, miRNA–mRNA and LINE–mRNA interactions. MARIO data also provide spatial-proximity information related to RNA folding in three-dimensional space.

**Results**

**The MARIO technology.** We developed the MARIO technology to detect RNA–RNA interactions facilitated by any single protein in vivo. In this procedure, RNA molecules are cross-linked with their bound proteins and then ligated to a biotinylated RNA linker such that proximal RNA molecules co-bound by the same protein form a chimeric RNA in the form of RNA1–Linker–RNA2. These linker-containing chimeric RNAs are isolated using streptavidin-coated magnetic beads and subjected to paired-end sequencing (Fig. 1a and Supplementary Fig. 1). Thus, each non-redundant paired-end read reflects a molecular interaction.

We carried out two independent MARIO assays on mouse embryonic stem (ES) cells with minor technical differences (Supplementary Table 1 and Supplementary Figs 2–5), which we designated as ES-1 and ES-2. A library for indirect RNA interactions was produced using two cross-linking agents (formaldehyde and EthylGlycol bis (SuccinimidylSuccinate))\(^23\)–\(^25\), which ‘effectively captures RNAs linked indirectly through multiple protein intermediates\(^26\) (ES indirect). Two other unique libraries were produced from mouse embryonic fibroblasts (MEFs) and mouse brain, offering two additional data sets for bioinformatic quality assessment (Supplementary Fig. 6). We confirmed that each library contained RNA constructs of the desired form (RNA1–Linker–RNA2) and lengths (Fig. 1b). We sequenced each library to yield, on average, 47.3 million paired-end reads, among

![Figure 1](image-url)
which ~ 15.1 million non-redundant paired-end reads represented the desired chimeric form (Supplementary Fig. 7). In addition, we carried out three control experiments. The first and the second control experiments excluded the cross-linking step (non-cross-linking control) and the protein biotinylation step (non-biotinylation control), respectively (Supplementary Note 1). The third control experiment used Drosophila S2 cells and mouse ES cells to test the extent of random ligation of RNAs (cross-species control). After cross-linking and cell lysis, the lysates from the two cell lines were immediately mixed before any subsequent steps. The mixture was subjected to the rest of the experimental procedure and resulted in a sequenced library (Fly-Mm). The proportion of RNA pairs mapped to two species is in the range of 2.5–6.8%, depending on whether the Drosophila genome and the mouse genome were assembled into a pan genome16,27 before mapping (Supplementary Note 1). We chose the more conservative estimate (derived from mapping to the pan genome) that 6.8% of the ligation products were generated from random ligations. This estimate is comparable to that (7.0%) derived from in silico simulations (Supplementary Note 2).

A suite of bioinformatics tools was created (MARIO tools) to analyse and visualize MARIO data. MARIO tools automated the analysis steps, including removing PCR duplicates, splitting multiplexed samples, identifying the linker sequence, splitting junction reads, calling interacting RNAs, performing statistical assessments, categorizing RNA interaction types, calling interacting sites and analysing RNA structure (http://mariotools.ucsd.edu). It also provides visualization tools for both the RNA–RNA interactome and the proximal sites within each RNA (Supplementary Fig. 8).

Figure 2 | MARIO-identified RNA–RNA interactions. (a) Distribution of MARIO-identified RNA–RNA interactions among different types of RNAs in ES cells. RNAs were experimentally and computationally removed from the analysis. (b) Degree distribution of the ES cell RNA–RNA interactome composed of mRNA, lincRNA, miRNA, pseudogene transcripts and antisense transcripts. The number of RNAs (as a proportion of all RNAs in the network, y axis) is inversely correlated with the number of interactions they participated (degree, x axis) in log scale, characteristic of scale-free networks. (c) The numbers of lincRNAs and miRNAs (x axis) categorized by the number of their interacting mRNAs (y axis).

RNA–RNA interactome in ES cells. We compared the five MARIO libraries. ES-1 and ES-2 were most similar as judged by correlations of FPKMs (fragments per kilobase of transcript per million mapped reads; separately calculated for the read fragments on the left and the right sides of the linker), followed by ES indirect, MEF, and then brain tissue (Supplementary Fig. 6). The interacting RNA pairs identified from ES-1 and those from ES-2 exhibited strong overlaps (P-value < 10^-35, permutation test) (Supplementary Table 2). The interactions identified in MEF did not exhibit significant overlaps with those in either of the ES samples (P-value for each overlap = 1, permutation tests). For example, an interaction between the 3’ untranslated region of Trim25 mRNA and Snoa1 snoRNA was supported by multiple paired-end reads in ES-1 and ES-2 samples, but was not detected in MEF (Supplementary Fig. 7). We did not expect many interactions identified from ES-1 and ES-2 to show up in ES indirect data, because a cross-linked protein complex can bury an RNA molecule, limiting the RNA’s accessibility to RNA ligase, which is required to form the chimeric RNA product. Among the snoRNAs identified as having interactions with mRNAs in our data sets, 172 of them, including Snoa1, were detected both as enzymatically processed small RNAs28 (red lane, Supplementary Figs 7 and 9–11) and in ARGONAUTE HTS-CLIP data (Supplementary Figs 9–11). This supports the proposition that transcripts from snoRNA genes could be enzymatically processed into miRNA-like small RNAs and interact with mRNAs in RISC complex29,30 (Supplementary Note 3).

The ES-1 and ES-2 libraries were merged to infer the RNA–RNA interactome in ES cells. This dataset included 4.54 million non-duplicated paired-end reads that were unambiguously split into two RNA fragments with both fragments uniquely mapping to the genome (mm9). We identified tens of thousands of inter-RNA interactions (false discovery rate < 0.05, Fisher’s exact test with Benjamin–Hochberg correction) (Supplementary Fig. 12). As expected, the RNA expression level (FPKM) is weakly correlated with the number of MARIO reads on each RNA, but FPKM is not correlated with the statistical significance (false discovery rate) of the interactions (Supplementary Fig. 12C,D). mRNA–snoRNA interactions were the most abundant type, although thousands of mRNA–mRNA and hundreds of lincRNA–mRNA, pseudogeneRNA–mRNA and miRNA–mRNA interactions were also detected (Fig. 2a). Our simulation suggested ~66% sensitivity and ~93% specificity for the entire experimental and analysis procedure (Supplementary Note 2).

Validation of selected interactions. We used two methods to validate selected interactions in the MARIO identified interactome. These two methods were selected because they do not perturb the cells or change RNA expression levels. First, we examined co-localization between Malat1 lincRNA and Slc2a3 mRNA in vivo by two-colour single-molecule RNA fluorescence in situ hybridization (smRNA–FISH)31. Quantum dots (qDots)
were used instead of organic dyes for increased fluorescence signal intensities and narrower ranges of emission wavelengths. When designed to target the same transcript, the qDot labeled (five 25–30-nt probes) and the organic dye-based (forty-three 20-nt hybridization probes) smRNA–FISH identified the same Actb mRNAs in the cytoplasm (Supplementary Figs 13 and 14). We designed probes for Malat1 and Slc2a3 (Fig. 3a) and labelled them with 605 and 525 nm qDots, respectively, for an unequivocal distinction of signal from each qDot (Supplementary Fig. 15). We note that qDots cannot penetrate into the nuclei unless specific delivery methods are used, and in mouse ES cells approximately a quarter of Malat1 RNA is in the cytoplasm. Cytoplasmic Malat1 and Slc2a3 RNAs were detected in 27 ES cells, with an average of 7.6 and 4.5 copies per cell, respectively (Supplementary Fig. 16). We called co-localization of two RNAs with a threshold on the distance between the estimated centres of two FISH spots, which corresponds to a physical distance of 0.2 μm (ref. 36). Sixteen pairs of co-localized Malat1 and Slc2a3 RNA molecules were detected from a total of 80 copies of Malat1 and 50 copies of Slc2a3 (Fig. 3b,c). In the first control experiment, we detected 1 co-localization of Malat1 and Actb RNAs from 124 copies of Malat1 and 34 copies of Actb. In the second control experiment, we detected 1 co-localization of Slc2a3 and Actb RNAs from 545 copies of Slc2a3 and 298 copies of Actb (odds ratio between experiment (Malat1-Slc2a3) and controls (Malat1-Actb and Slc2a3-Actb) = 446, P-value < 10^-20, χ²-test).

To test interactions at a larger scale, we carried out RNA interactome analysis and sequencing (RIA-seq). We chose RIA-seq, because it does not require genetic perturbation. First, we did Malat1 RIA-seq and Actb RIA-seq (control) to test the interactions involving Malat1 (Supplementary Note 4). Malat1 RNA itself exhibited a 5.81-fold increase in Malat1 RIA-seq over Actb RIA-seq, confirming the validity of the RIA purification. Malat1-interacting RNAs reported by MARIO showed 14.6 (O610007P14Rik), 4.53 (Slc2a3), 3.88 (Eif4a2) and 2.39 (Tfrc)-fold increase in Malat1 RIA-seq over Actb RIA-seq (P-value < 0.0003, χ²-test). This suggests a strong overlap of Malat1 targets in MARIO and Malat1 RIA-seq. Next, we asked whether Tfrc RIA could reversely identify Malat1 by Tfrc RIA-seq (Supplementary Note 4). The Tfrc RNA itself showed 2.87-fold of increase in Tfrc RIA-seq compared with Actb RIA-seq. Malat1 exhibited 3.84-fold increase (P-value < 2.2 × 10^-16, derived from testing the null hypothesis fold change = 1), suggesting that antisense purification of Tfrc could reversely pull down Malat1. In addition, three out of four other Tfrc-interacting RNAs identified by MARIO exhibited 1.4- to 13.6-fold increases (P-value < 0.00002, χ²-test). Taken together, seven additional MARIO-identified interactions were validated by RIA-seq.

Figure 3 | Testing co-localization of Malat1 lincRNA and Slc2a3 mRNA with smRNA-FISH. (a) Positions of hybridization probes labelled with 605-nm (red) and 525-nm (blue) qDots. (b) Quantification of Malat1 and Slc2a3 RNA molecules. Green: number of co-localized molecules. (c) Representative images of two-colour smRNA-FISH showing four co-localized molecules (arrows). Scale bar, 10 μm.
**MARIO-derived RNA–RNA interactomes are scale free.** The experimentally derived RNA interactome offered an opportunity to test a fundamental physical property. Most known biological networks are hierarchical networks\(^{21,22}\), where the number of nodes is reversely correlated with the number of interactions (edges) they participate. This reverse correlation takes a linear form in log scale, which is called the power law, also referred to as the scale-free property\(^{39}\). However, RNA–RNA interactions have been reported as ‘surprisingly promiscuous’\(^{20}\). It was suggested that each miRNA interacts with 300–1,000 mRNAs in one cell\(^{19}\) and a similar picture was proposed for lincRNAs\(^{39}\). In addition, pseudouridines\(^{42}\) RNAs (Supplementary Table 3), indicative of their frequent interactions with other RNAs\(^{40,41}\). Extrapolating from these information, one would not expect the RNA interactome to be scale free. For example, an artificial network with 100 miRNAs and each miRNA randomly connected to 300–1,000 mRNAs is not a hierarchical network (Supplementary Fig. 17A).

The MARIO-derived RNA–RNA interactome is scale free (Fig. 2b and Supplementary Fig. 17B). In other words, the quantity of RNAs with a given number of interaction partners (the number of RNAs at the same level of promiscuity) decreases exponentially as that number of interaction partners (promiscuity) increases. In theory, the scale-free property of the experimentally derived network should not be affected by the bioinformatic threshold used for calling the interactions. This is because uniformly adding or withdrawing edges at random should not affect the log linearity in the network’s degree distribution. We tested this empirically. Indeed, changing the threshold of calling interactions did not affect the log linearity between the number of RNAs and their number of interaction partners (Supplementary Fig. 17D). Furthermore, the scale-free property of the observed RNA interactome does not change if we remove all rRNA and snoRNA from the network (Fig. 2b). In addition, the RNA interactome derived from mouse brain is scale free (Supplementary Fig. 17C), suggesting this global property is not cell-type specific. In each cell type, there are exponentially more miRNAs and lincRNAs that had relatively specific (fewer) mRNAs than those with greater number of targets (Fig. 2c).

In summary, regardless of the thresholds and cell types, MARIO-derived RNA interactomes were hierarchical networks. Therefore, MARIO-derived RNA interactome shares the most essential network property (scale free) as other types of known biological networks. We speculate that this fundamental topological property has not been reported, because previous methods requiring ectopic expression or genetic perturbation may affect the network topology by altering the concentrations of critical molecules.

**Frequently used RNA segments for interactions.** A number of the interacting RNAs exhibited overlapping MARIO reads (Fig. 4a), suggesting interactions were often concentrated at specific segments of an RNA. ‘Peaks’ of overlapping read fragments were identified and termed ‘interaction sites’ (Fig. 4b). Interaction sites appeared not only on miRNAs (the entire mature miRNA), miRNAs and lincRNAs, but also on pseudogene and transposon RNAs (Fig. 4c). Over 2,000 interaction sites were produced from L1, SINE, ERVK, MaLR and ERV1 transposon RNAs (Supplementary Table 3), indicative of their frequent interactions with other RNAs\(^{10,41}\). In addition, pseudouridines\(^{12}\) were enriched in the RNA interactions sites of snoRNA–mRNA interactions, corroborating the idea that some RNA segments were favoured in certain types of RNA interactions (Supplementary Note 5).

**Sequence complementation on RNA interaction regions.** We asked whether base complementation is used by different types of RNA–RNA interactions. We estimated the hybridization energy of a pair of interacting RNAs by the average hybridization energy of all pairs of ligated fragments (RNA1 and RNA2)\(^{43}\) that were mapped to this RNA pair, and compared it with the hybridization energy of control fragment pairs generated by random shuffling of the bases. Complementary bases were preferred in nearly all types of RNA–RNA interactions and were most pronounced in transposonRNA–mRNA, mRNA–mRNA, pseudogeneRNA–mRNA, lincRNA–mRNA and miRNA–mRNA interactions (P-values < 2.4 \(\times\) 18, test by random shuffling), but was not observed in LTR-pseudogeneRNA interactions (Fig. 4d and Supplementary Fig. 18). This data led us to speculate that base pairing facilitates sequence-specific posttranscriptional regulation in long RNAs.

**Evolutionary conservation of interaction sites.** If these RNA–RNA interactions are sequence specific\(^{44}\), the RNA interaction sites should be under selective pressure. We found that the interspecies conservation levels\(^{45}\) are strongly increased at the interaction sites and the peak of conservation precisely pinpointed the junction of the two RNA fragments (Fig. 4e). When interacting with lincRNAs, pseudogene RNAs, transposon RNAs or other mRNAs, the interaction sites on mRNAs were more conserved than the rest of the transcripts (Supplementary Fig. 19). The interactions sites on lincRNAs and pseudogene RNAs exhibited increased conservation in lincRNA–mRNA, pseudogeneRNA–mRNA and pseudogeneRNA–transposonRNA interactions (Supplementary Fig. 19). The increased conservation at interaction sites was not due to exon-intron boundaries (Supplementary Fig. 20). Taken together, base complementarity is widespread in the interactions of long RNAs. The complementary regions are evolutionarily conserved.

**Three-dimensional RNA structure.** Although we originally designed MARIO for mapping inter-molecule interactions, we also found that MARIO revealed RNA secondary and tertiary structures. All the analyses above were based on intermolecular reads. By looking at intramolecular reads, we learned two characteristics of RNA structure. First, the footprint of single-stranded regions of an RNA were identified by the density of RNase I digestion sites (RNase I digestion was applied before ligation, see Step 2 in Fig. 1a and Supplementary Fig. 21). Second, the spatially proximal sites of each RNA were captured by proximity ligation (Step 5 in Fig. 1a). Over 60,000 linker-containing read pairs were mapped to individual genes and thus were determined as intramolecule cutting and ligation (Supplementary Fig. 22A). Each cut-and-ligated sequence can be unambiguously assigned to one of two structural classes by comparing the orientations of RNA1 and RNA2 in the sequencing read with their orientations in the genome (Fig. 5a). These reads provided spatial proximity information for 2,374 RNAs, including those from 1,696 known genes and 678 novel genes. For example, 277 cut-and-ligated sequences were produced from Snorat73 transcripts (Fig. 5b). The density of RNase I digestion sites (Fig. 5c) was strongly predictive of the single-stranded regions of the RNA (heatmap; Fig. 5e). Six pairs of proximal sites were detected (circles; Fig. 5d). Each pair was supported by three or more cut-and-ligated sequences with overlapping ligation positions (black spots; Fig. 5b). Five out of the six proximal site pairs were physically close in the generally accepted secondary structure (arrows of the same colour; Fig. 5e). On Snorat14, a pair of inferred proximal sites appeared distant, according to sequenced inferred secondary structure (green arrows; Supplementary Fig. 23). However, ribonucleoprotein DYSKERIN bent Snorat14 transcript in vivo\(^{46,47}\), making the two pseudouridylation loops close to each other, as predicted by the cut-and-ligated sequence (green arrows; Fig. 5f). Structural...
Discussion
The MARIO method offers several advantages for mapping RNA–RNA interactions. First, MARIO directly analyses the endogenous cellular features without introducing any exogenous nucleotides or protein-coding genes before cross-linking. This eliminates the uncertainty of reporting spurious interactions produced by changing the RNA or protein expression levels. Moreover, it makes MARIO well-suited for assaying tissue samples. Second, the introduction of a selectable linker enables an unbiased selection of interacting RNAs, making it possible to globally map an RNA–RNA interactome. This method circumvents the requirement for a protein-specific antibody or the need to express a tagged protein. It also removes the limit of working with one RNA-binding protein at a time. Third, this method only captures the RNA molecules co-bound with a single protein molecule, avoiding capture of RNA molecules that are independently bound to different copies of a protein, which would potentially lead to reporting spurious interactions. Fourth, false positives that result from RNAs ligating randomly to other nearby RNAs are minimized by performing the RNA ligation step on streptavidin beads in extremely dilute conditions. Fifth, the RNA linker provides a clear boundary delineating the position of ligation site in the sequencing reads, thus avoiding ambiguities in mapping the ligated chimeric RNA. Sixth, potential PCR amplification biases are removed by attaching a random four- or six-nucleotide barcode to each chimeric RNA before PCR amplification and subsequently counting completely overlapping sequencing reads with identical barcodes only.
MARIO should facilitate future investigations of RNA functions and regulatory roles.

**Methods**

**Cell culture.** Undifferentiated mouse E14 ES cells (gift from Huck-Hui Ng) were cultured under feeder-free conditions. ES cells were seeded on gelatin-coated dishes and were cultured in DMEM medium (GIBCO) supplemented with 15% fetal bovine serum (FBS; Gemini Gemcell), 0.055 mM 2-mercaptoethanol (Sigma), 2 mM Glutax (GIBCO), 0.1 mM MEM non-essential amino acid (GIBCO), 5,000 U ml⁻¹ penicillin/streptomycin (GIBCO) and 1,000 U ml⁻¹ of LIF (Millipore). The cells were maintained in an incubator at 37°C and 5% CO₂. MEFs (C57BL/6, GlobalStem) were cultivated in 15-cm dishes in DMEM (GIBCO) supplemented with 15% FBS (Gemini Gemcell), 0.055 mM 2-mercaptoethanol (Sigma), 2 mM Glutax (GIBCO), 0.1 mM MEM non-essential amino acid (GIBCO), 5,000 U ml⁻¹ penicillin/streptomycin (GIBCO). MEFs were also maintained in an incubator at 37°C and 5% CO₂.

**Tissue dissection and preparation.** Mice handling was approved by the Institutional Animal Care and Use Committee of the University of California, San Diego. Adult female (C57BL/6) background) was killed by cervical dislocation and the whole brain was immediately collected, rinsed with ice-cold PBS three times and snap frozen. Frozen, whole mouse brain tissue was ground into fine powder in liquid nitrogen using a mortar and pestle. The tissue powder was quickly transferred into a Petri dish on a bed of dry ice and irradiated on dry ice three times at 400 mJ cm⁻² in an ultraviolet cross-linker (254 nm) with gentle swirling between each irradiation. Cross-linked, powdered tissue was immediately lysed and subjected to MARIO procedure as described.

**Overview of the MARIO method.** MARIO was designed to: (i) capture interacting RNAs *in vivo* in an unbiased manner without genetically or transiently introducing exogenous molecules; (ii) allow stringent removal of non-physiologic associations that form after cell lysis53; (iii) select the proximity-ligated chimeric RNAs; and (iv) allow unambiguous bioinformatic identification of interacting RNAs.
We achieved these objectives by: (i) cross-linking and immobilization of all RNA–protein complexes in streptavidin beads and removal of nonspecific binding (ii) capturing RNA–protein complexes as well. This was to capture the RNA that were brought together in vivo, particularly in the context of protein–protein interactions. We applied an in vivo dual cross-linking method with previously validated parameters.55,57 Briefly, cells were first rinsed with room temperature PBS and treated with 1.5 mM Ethyl(iGlycol (Succinimidy)iGlycate (Pierce Protein Research Products, Rockford, Illinois) freshly prepared in PBS for 45 min at room temperature on a shaker. Cells were further treated with formaldehyde (Pierce Protein Research Products) to a final concentration of 1% and incubated for 20 min at room temperature with rocking. Glycine was added to a final concentration of 250 mM and incubated for 10 min at room temperature, to quench the cross-linking reaction. Cells were then washed once with cold PBS, 5 min at room temperature, scraped off, pelleted at 1,000 g, 10 min at 4°C. The pellets were snap-frozen in liquid nitrogen and stored at −80°C.

Cell lysis, RNA fragmentation, and protein biotinylation. Approximately 6 × 10^6 cross-linked cells were cultured for 20 minutes on ice and resuspended in 3 volumes of ice-cold PBS with 0.1% Tween-20, 100 mM NaCl, 0.1% SDS, 1% IGEPA CA-630, 0.5% sodium deoxycholate, 1 mM EDTA supplemented with 1.20 volume of EDTA-free complete protease inhibitor cocktail (Roche). Lysis was performed on ice for 20 min. Cell debris and insoluble chromatin were removed by centrifugation at 20,000 × g for 10 min at 4°C. The supernatant was collected and used as the lysate solution.

To fragment the proteins, 500 mM NaCl final concentration was added and the solution was incubated at 4°C for 10 min with rotation. To further dissociate protein complexes and non-cross-linked RNAs, and halt the activities of T4 RNA ligase and prevent bead aggregation. Polyethylene glycol (PEG) was used to enhance intermolecular ligation by increasing the concentrations of the donor and the acceptor ends.

Proximity ligation. Next, the beads were washed twice with ice-cold wash buffer II, once with ice-cold wash buffer III and PNK wash buffer. To prepare the proximity ligation, we first dephosphorylated the 3′-end using the 3′-phosphate activities of T4 PNK, a 3′-hydroxyl group62. After discarding wash buffer, the beads were mixed with 73 l of RNase-free water, 20 ml of 5 x PNK buffer pH 6.5 (350 mM Tris-HCl pH 6.5, 50 mM MgCl2, 10 mM DTT) and 5 l of 10 U ml−1 T4 PNK (3′-phosphate minus) (NEB), 2 ml of RNAsin Plus (Promega) and incubating for 1 h at 37°C with intermittent shaking, for 5 x every 2 min. The mixture was washed once with wash buffer I, II, III and PNK, two times with each buffer, with rotation for 5 min at 4°C in the second wash. The ice-cold washes were used to eliminate any leftover PNK, which may phosphorylate the RNA linker, inducing it to be potentially ligated to the 3′-end of RNAs. After wash buffer was removed, the T4 phosphatase-linked 3′ end was added by 160 µl RNA ligation reaction mixture, which contained 2 µl RNAsin Plus (Promega), 16 l of 10 mM ATP, 16 l of 10 RNA ligase buffer, 16 l of 1 mg ml−1 BSA, 30 µl of 20 mM biotin-labelled linker, 64 l of 50% PEG8000 (NEB), 16 l of 10 µl T4 RNA ligase I (NEB). Ligation was carried out at 16°C overnight with intermittent shaking at 1,200 r.p.m. for 15 s every 2 min. BSA was added to enhance the activities of T4 RNA ligase and prevent bead aggregation. Polyethylene glycol (PEG) was used to enhance intermolecular ligation by increasing the concentration of the donor and the acceptor ends.

RNA purification and reverse transcription. The following day, ligation was performed under extremely diluted conditions in a 15-ml total volume reaction (8.9 ml of RNase-free water, 1.5 ml of 10 mM ATP, 1.5 ml of 10 µl T4 RNA ligase I (NEB)), for 1 h at 37°C with intermittent shaking. Following phosphosylation, the beads were twice in PNK wash buffer and proximity ligation was then performed. After adding EDTA to a final concentration of 25 mM and 15 min at 4°C, to prevent inter-molecular ligation from happening as the beads were collected on the wall of the tube. The beads were washed once in PBST. We next eluted protein–RNA complexes from streptavidin beads twice in 100 µl of Elution Buffer (100 mM Tris-HCl pH 7.5, 30 mM NaCl, 10 mM EDTA, 1% SDS, 10 mM DTT, 2.5 mM D-biotin (Invitrogen)) by heating to 95°C for 5 min. The resulting solutions were combined, mixed with 50 l of 800 µl Proteinase K stock solution (100 mg ml−1), incubated for 10 min at 65°C, and snap-frozen in liquid nitrogen. The RNA solution was extracted with two volumes of chloroform, followed by washing with two volumes of chloroform to remove any residual proteins. The RNA solutions were resuspended in 400 µl of RNase-free water and extracted with phenol/chloroform/isoamyl alcohol (25:24:1, pH 4.5) (Ambion) and incubating at 37°C for 20 min with shaking at 1,200 r.p.m. The mixture was transferred into a 2-ml MaxExtract high-density phase lock gel tube (Qagen) and centrifuged at 16,000 g for 20 min at 4°C. The upper aqueous phase was removed, replaced with 400 µl of chloroform to the same MaxExtract tube and centrifuged at 16,000g for 5 min at room temperature. Following centrifugation, the aqueous phase was transferred...
into a new tube and RNAs were precipitated by adding 1.9 volume of 3 M sodium acetate pH 5.2, 1.5 µl of glycocholate (Ambion) together with 1 M of 1:1 ethanol (propanol-2) to precipitating at –20°C overnight. We treated the precipitated RNA by centrifugation at 21,000 g for 30 min at 4°C. After discarding the supernatant, the pellet was washed twice with 80% ethanol and air dried until ethanol completely evaporated. The purified RNAs at this stage were a mixture of RNAs without linkers (RNA1 or RNA2), RNAs ligated with linkers but not proceeding RT, RNAs ligation with linkers but not proceeding RT, and RNAs proceeding RT. The specific RNAs in the mixture were digested with restriction enzymes in different cut sites. The RNAs in the mixture were then digested with T7 exonuclease, which not only removes 5’-mononucleotides from duplex DNA but also exerts exonuclease activity on the RNA strand from a RNA–DNA hybrid. A complementary DNA oligonucleotide (5’-TGCGGATCCTGAGGAGGATGCTG-3’) was annealed to the RNA linker, creating a double-stranded DNA–RNA hybrid between the RNA linker and the cDNA strand. The cDNA strand was designed so that after annealing the 5’-end of the RNA linker was recessed, while the 3’-end of the DNA strand was protruding. The annealed products were then treated with T7 exonuclease.

Removal of rRNAs by antibody-based depletion of RNA–DNA hybrid (GeneRead rRNA Depletion Kit (Qiagen)) in ES-2 MEF samples. rRNA was removed according to the manufacturer’s instructions with the following modifications. Instead of cleaning up depleted RNA by RNeasy MinElute spin column which will remove RNAs shorter than 200 nucleotides, we removed excess rRNA capture probes by rigorous DNase treatment. DNase-treated RNA was also purified by phenolchloroform extraction and ethanol precipitation as described above.

Next, cDNA was eluted from SPRISelect beads (Beckman Coulter Genomics) and eluted in RNase-free water. The following reaction was described as above.

Reverse transcription. Following ligation, RNA was fragmented in size range of 150–400 bp, optimal for sequencing by Illumina HiSeq, by using the RNase III fragmentation kit according to the manufacturer’s protocol. Fragmented RNA was purified by 2.2 × SPRISelect beads (Beckman Coulter Genomics) and ethanol precipitated as described above.

Ligation with reverse transcription (RT) adapter. Next, the RNAs were ligated with a 3’-RT adapter (5’/5’-App/AGATCGGAGGCGGGCCACCTAG3/3’-Ad) that served as a primer for RT reaction. Following ethanol precipitation, the RNA pellet was resuspended in 20 µl of ligation reaction mixture: 1 µl RNAin Plus (Promega), 2 µl of 10 µM lRNA ligase buffer, 7 µl of 20 µM pre-adenylated L3-App adapter, 8 µl of 50% PEG8000 (NEB), 2 µl of 200 µM T4 RNA ligase 2, truncated KQ (NEB). The reaction was over night at 16°C.

Reverse transcription. Following ligation, RNA was purified by 2 × SPRISelect beads (Beckman Coulter Genomics) and eluted in RNase-free water. The following reaction is described for 2 µl of RNA and was scaled up accordingly for higher amount of RNAs. For each experiment or replicate, a different RT primer containing individual experimental barcode sequence was used. Each RT primer has the form of 5’/5’-Phos/NXNNNNNNNAGATCGGAGGAGGCGGTGCG/C/TGAAAGCCTCCTGAGGT-3’. According to this scheme, the first read of every sequencing read pairs contains a barcode that takes the configuration of 3’-end of the RNA–linker or linker–RNA junctions, in one end of the read pair. A substantial fraction of the sequencing read pairs cover to the RNA–linker or linker–RNA junctions, in one end of the read pair.

Biotin pull-down of chimeric RNA–DNA hybrids. Streptavidin–biotin affinity purification was used to enrich for chimeric RNA–DNA hybrids. This pull-down was carried out after the second RNA fragmentation and RT, to allow

Construction of sequencing library. Considering the ultraviolet-induced cross-link site sometimes stalls RT, resulting in truncated cDNAs that lack the 5’-adapter, we adopted a circularization strategy that allows for constructing sequencing libraries even from truncated cDNAs (Supplementary Fig. 1). The RT mixture was ligated to sequences of the PCR amplification by Illumina PE PCR Forward Primer 1.0 (5’-AATGATACGGCGACCAGAGATCTTCTCTTCCTCAGACGCTCTCTGATCT-3’) and PE PCR Reverse Primer 2.0 (5’-CAACGAAGGAAACGCACTAGGTGCGTGCGTCTCGAGTCCGAGTCTCTTCACT-3’), flanking a BamHI restriction site and a sequencing barcode.

Circularization. cDNA was circularized by CircLigase II (Epicentre). Briefly, cDNA was eluted from SPRISelect beads in 20 µl CircLigase reaction mixture (12 µl of sterile water, 2 µl of CircLigase II 10 × reaction buffer, 1 µl of 50 mM MnCl2, 4 µl of 3 M Betaine, 1 µl of 100 µM 5’-linker–RNA2, PE PCR Reverse Primer 2.0 (5’-GTTCAGGATCCAGGCGGTCCTCAGTCTGCAACGTTAGAGCTGCGTCTCAGATCT-3’)), and incubated for 2 h at 60°C. CircLigase II was inactivated by incubating the reaction at 80°C for 10 min.

Linearization. A DNA oligo was annealed to the RT primer, generating a short double-stranded region suitable for BamHI restriction. This strategy also prevents BamHI activities on other endogenous BamHI restriction sites. Next, BamHI was implied, creating linear cDNAs with adapters at both 5’- and 3’-ends to prime subsequent PCR amplification. Next, oligo annealing mixing (43 µl water, 6 µl 10 × FastDigest Buffer (Fermentas), 5 µl 200 µM Cut-oligo (5’-GTTCAGGATCCAGGCGGTCCTCAGTCTGCAACGTTAGAGC/3’/3’)), was added into the CircLigase II reaction. Annealing was carried out by heating to 95°C for 2 min, followed by 71 cycles of 20 s each, starting from 95°C and decreasing the temperature by 1°C every cycle down to 25°C and holding at 25°C for 1 min.

Six microliters of FastDigest BamHI (Fermentas) was added and incubated at 37°C for 30 min. Re-linearized cDNA was purified by 2 × SPRISelect beads (Beckman Coulter Genomics) (v/v) and eluted in nuclease-free water.

PCR pre-amplification and size selection. Single-stranded cDNA was first pre-amplified by PCR using a truncated version of PCR primers (forward primer D5, 5’-CAGGCGCTTCCGAGGCTCTGAGGCTCCTGAGGCTCTGAGTGAGTCGCTGCGGT/3’-linker–RNA2). This was based on the idea that larger fragments (primer–primer mismatches) might contain only the barcode and/or RNA linker by doing size selection at this stage.

Six cycles of PCR were performed in a 40 µl reaction, which contained 20 µl of NEBNext High-Fidelity 2 × PCR Master Mix (NEB), 0.625 µM of each DPS/DPS primer using the following temperatures: 1 cycle of initial denaturation at 98°C for 30 s; 6 cycles of amplification with 98°C for 10 s, 65°C for 30 s, 72°C for 30 s; followed by final extension at 72°C for 5 min; and hold at 4°C. The PCR product was purified by 1.8 × SPRISelect beads (v/v) and size-selected using E-gel 2% Agarose gels (Invitrogen). The DNA fragments between 150 bp and 350 were excised from the gel and purified using MinElute gel extraction kit (Qiagen) (gel-extracted).

rRNA removal by duplex-specific nuclease (DSN) approach (ES-1, ES-indirect). To reduce rRNA cDNAs from ES-1 and ES-indirect library, we also pre-amplified ss-cDNA using the truncated PCR primer DPS/DPS. However, the PCR cycle number was increased until we could obtain 80–100 ng of cDNA after purification by 1.8 × SPRISelect beads (Beckman Coulter Genomics) (v/v). We skipped the size selection by agarose gel, as this would largely reduce the amount of DNA. The eluted DNA from SPRISelect beads was mixed with 4.5 µl hybridization buffer (2 M NaCl, 200 mM HEPES pH 8.0) and sterile water (if necessary) to a final concentration of 10 µg/ml rRNA cDNA. The resulting mixture was denatured at 98°C for 2 min and re-annealed at 68°C for 5 h on an agarose gel slab. Although the rRNA was still in the thermal cycle, we added 20 µl of 68°C-preheated 2 × DSN buffer (Axxora) to the reaction mix, mixed well by pipetting up and down ten times and incubated the reaction for 10 min at 68°C. Two microliters of 1 U/µl 5’-DSN enzyme (Axxora) was added, mixed and incubated at 68°C for 25 more minutes. We stopped the reaction by adding 40 µl of 2 × DSN stop solution (Axxora) to the reaction mix, mixed well and incubated the tube to ice. The reaction mixture was then purified using 1.8 × SPRISelect beads.
Final PCR amplification. We performed PCR amplification of DNA produced from previous steps using full-length PCR primer PE 1.0 and 2.0 (Illumina). The number of PCR cycles was carefully titrated by running pilot PCRs with small aliquots of DNA to avoid overamplification. We purified the PCR products by 1.8 × 5PRISSelect beads (Invitrogen) and size-selected fragments between 250 and 550 bp. The combined length of Illumina PE 1.0/2.0. Final libraries were quantified by Qubit (Invitrogen) and quantitative PCR, quality-checked by Bioanalyzer (Agilent Technologies) and submitted for paired-end sequencing on Illumina HiSeq platform.

Oligonucleotide sequences used in MARIO. The custom-designed RNA and DNA oligonucleotides used in the procedure are as follows: Biotinylated RNA linker (RNase-free HPLC-purified from IDT): 5′-rC+rA-rG+iBd+iL/rA-rG+rC-rG+rA-rG+cDNA strand with a RNA linker (RNase-free HPLC-purified from Sigma): 5′-CCTGCGATTCGAGCTGATCATGATC-3′ Pre-adenylated RT adapter (RNase-free HPLC-purified from IDT): 5′-5′RApp/AGATCGGAAGACGGTTCAGA3′dC/dC-3′

RT primers (adapted from ref. 62) (RNase-free HPLC-purified from Sigma): RT Primer for the ES-1 sample: 5′-5′Phos/NNAGTTINNAGATCCGGAGCTGTCGATCTGAAGACCGCTCCTCGATCT-3′ RT Primer for the ES-2 and MEF samples (sequences on different lanes): 5′-5′Phos/NNGCCTCNNNNAGATCCGGAGCTGTCGATCTGAAGACCGCTCCTCGATCT-3′ RT Primer for the ES-indirect sample: 5′-5′Phos/NNCATTTNNNAGATCCGGAGCTGTCGATCTGAAGACCGCTCCTCGATCT-3′ BamHI restriction site is highlighted: Truncated PCR Forward Primer DP5 (HPLC-purified from IDT): 5′-CAGCAGGCTCTTCCGATCT-3′ Truncated PCR Reverse Primer DP5 (HPLC-purified from IDT): 5′-GCAAGCCGCTCTTCCGATCT-3′ Illumina PE PCR Forward Primer 1.0 (Purge from Illumina); 5′-AATGATACGGCGACCCAGAATGTACATCACTTCTCCCTGCACTACAGCCGCTCTCCGATCT-3′ Illumina PE PCR Reverse Primer 2.0 (Purge from Illumina): 5′-GCAAGCCGCTCTTCCGATCT-3′

Single-molecule RNA-FISH. Oligonucleotides were synthesized with a biotin attached to their 5′-end (IDT). Labelling was achieved by incubation of oligonucleotides and dyes (Alexa 555 or qDots, Invitrogen) coated with poly-D-lysine (5 mg/ml) and laminin (0.01 mg/ml) on tissue culture dishes (C. smRNA-FISH experiments were conducted on mouse cells) and allowed to permeabilize with methanol at −20°C. The smRNA-FISH experiments were conducted using a modified version of an established protocol11. Hybridizations were carried out as follows: oligonucleotides in hybridization buffer for 30 min at 40°C. Excess of probes and dyes was removed by two washes (2× SSC, 2× formamide, 10%) at 55°C for 30 min. The cells were then imaged in SSC 2× buffer (pH 7.5).

Wide-field fluorescence imaging was conducted in an Olympus IX83 inverted microscope equipped with appropriate cubes for the dyes used (Supplementary Table 1, Chroma) and xen 60 oil-immersion objective (numerical aperture = 1.4, Olympus). Images were captured with ORCA-R2 charge-coupled device camera (Hamamatsu) at intervals of 0.2 μm on the z axis. Following imaging, raw image stacks were processed in ImageJ, first by applying the Laplacian of Gaussian filter32, second by counting the three-dimensional-rendered signal spots at incrementing intervals of 500 fluorescence arbitrary units. The number of spots in each image stack was obtained at the threshold interval of three consecutive equal counts. We observed that by selecting fluorescence intensities beyond this plateau one would have reduced the number of spots identified in the order of one unit, with the potential increase in false-negative spots. In addition to the counts, we collected data for the spot’s respective x-y-z centre of mass.

The computational pipeline (MARIO tools). MARIO tools is a package of command-line tools for analyses of MARIO data. It is written in Python and R, and is version controlled by Github. The full documentation of the MARIO tools software is available at http://mariotools.ucsd.edu. The pipeline takes pair-end sequencing reads as input (Supplementary Fig. 26A). The main outputs include: (1) a parsed cDNA library, including the list of chimeric cDNAs in the form of RNA1–Linker–RNA2 (see the final products in Supplementary Figs 1 and 26C); (2) the genomic locations of RNA1 and RNA2 of every chimeric cDNA (Supplementary Fig. 26D); and (3) interacting RNA pairs inferred from statistical enrichment of chimeric cDNAs (Supplementary Fig. 26E). The major analysis steps of MARIO tools are as follows: (1) removing PCR duplicates; (2) assigning multiplexed sequencing reads into corresponding experimental samples; (3) recovering the cDNAs in the sequencing library; (4) parsing the chimeric cDNAs; (5) mapping to the genome; (6) identifying interacting RNA pairs; and (7) identifying RNA interaction sites. Detailed documentation of MARIO tools is available at http://mariotools.ucsd.edu

Binding energies between RNA interaction sites. The binding energies between two RNA interaction sites were determined by the DuplexFold programme from RNAstructure version 5.6 (ref. 43). The base pairing between two interaction sites was determined by MiRanda version 3.3a.

Conservation levels of RNA interaction sites. For every read pair in the RNA1–Linker–RNA2 category (output of Step 4 from MARIO tools), we obtained the Phylp0 conservation scores42 of two 1,000-bp genomic regions, one centred at the ligation junction of RNA1–Linker and the other centred at the ligation junction of Linker–RNA2. The average Phylp0 scores of all the RNA1–Linker–RNA2-type read pairs were plotted. As a control, we obtained average Phylp0 scores from the same number of random genomic regions of the same lengths.

Detecting read pairs. Starting from the RNA1–Linker–RNA2 type of read pairs (output of ‘Step 6. Selection and extraction of desired RNA–RNA interactions and reverse transcript’ from MARIO tools), we applied the following filters to identify the pair-end reads generated from self-interacting RNAs. We removed those read pairs mapped to different genes. If a read pair mapped to the same gene, we also removed those pairs that: (1) did not contain any fraction of the linker sequence; (2) the forward and the reverse reads mapped to opposite strands within 2,000 bp; (3) the read mapped to plus strand has smaller coordinates than the read mapped to minus strand in the genome within the pair. This step minimizes the inclusion of any intact (continuous) RNA fragment in the structural analysis.

RNA folding and secondary structure prediction. Structural information of the RNAs with known or generally accepted structures was downloaded from RNAdb database v3.4 in DOT format (graph description language). We drew figures from the DOT files using the command line version of VARNA Applet version 3.9. For the RNAs without structural information in RNAdb, we predicted their secondary structures based on the sequence using the ‘Fold’ programme in RNAstructure version 5.646.

Data availability. Data that support the findings of this study have been deposited in Gene Expression Omnibus with the accession number GSE61489.

References
1. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–100 (2012).
2. Suyyn, A., Lajoie, B. R., Johnson, J. C. G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109–113 (2012).
3. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).
4. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83–90 (2012).
5. Djebari, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
6. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 9, e1003569 (2013).
7. Holley, R. W., Everett, G. A., Madison, J. T. & Zamir, A. Nucleotide sequences in the yeast alanine transfer ribonucleic acid. J. Biol. Chem. 240, 2112–2128 (1965).
8. Holley, R. W. et al. Structure of a ribonucleic acid. Science 147, 1462–1465 (1965).
9. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391, 806–811 (1998).
10. Ambros, V. The functions of animal microRNAs. Nature 431, 350–355 (2004).
11. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297 (2004).
12. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172–177 (2013).
13. Meister, G. Argonaute proteins: functional insights and emerging roles. Nat. Rev. Genet. 14, 447–459 (2013).
14. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141, 129–141 (2010).
