RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure

Highlights

- PARIS yields in vivo RNA duplex maps of human and mouse cells.
- In vivo maps discover extensive long-range and alternative RNA structures.
- PARIS guides evolution analysis and validation of duplex function.
- Unique duplex fold of XIST A-repeat nucleates XIST-SPEN IncRNP.

Authors

Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, ..., Jill P. Mesirov, Thomas R. Cech, Howard Y. Chang

Correspondence

howchang@stanford.edu

In Brief

A method for global mapping of RNA duplexes in living cells with near base-pair resolution, called PARIS, reveals extensive long-range structures, higher-order architectures, alternative structures, and RNA-RNA interactions across the transcriptome and uncovers a unique architecture of Xist and its specific interaction with a key silencing factor.

Accession Numbers

GSE74353

Lu et al., 2016, Cell 165, 1–13
May 19, 2016 © 2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2016.04.028
RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure

Zhipeng Lu,1,9 Qiangfeng Cliff Zhang,1,2,9 Byron Lee,1 Ryan A. Flynn,1 Martin A. Smith,3,4 James T. Robinson,6 Chen Davidovich,5,6,7,8 Anne R. Gooding,6 Karen J. Goodrich,6 John S. Mattick,3,4 Jill P. Mesirov,5 Thomas R. Cech,6 and Howard Y. Chang1,6

1Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA
2MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China
3RNA Biology and Plasticity Group, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
4St Vincent’s Clinical School, UNSW Medicine, NSW 2052, Australia
5Department of Medicine and Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA
6HHMI and Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80303, USA
7Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Victoria 3800, Australia
8EMBL Australia and the ARC Centre of Excellence in Advanced Molecular Imaging, Clayton, VIC 3800, Australia
9Co-first author

Correspondence: howchang@stanford.edu
http://dx.doi.org/10.1016/j.cell.2016.04.028

SUMMARY

RNA has the intrinsic property to base pair, forming complex structures fundamental to its diverse functions. Here, we develop PARIS, a method based on reversible psoralen crosslinking for global mapping of RNA duplexes with near base-pair resolution in living cells. PARIS analysis in three human and mouse cell types reveals frequent long-range structures, higher-order architectures, and RNA-RNA interactions in trans across the transcriptome. PARIS determines base-pairing interactions on an individual-molecule level, revealing pervasive alternative conformations. We used PARIS-determined helices to guide phylogenetic analysis of RNA structures and discovered conserved long-range and alternative structures. XIST, a long noncoding RNA (lncRNA) essential for X chromosome inactivation, folds into evolutionarily conserved RNA structural domains that span many kilobases. XIST A-repeat forms complex inter-repeat duplexes that nucleate higher-order assembly of the key epigenetic silencing protein SPEN. PARIS is a generally applicable and versatile method that provides novel insights into the RNA structurome and interactome.

INTRODUCTION

RNA structure and intermolecular interactions are essential in nearly every step of the gene expression program. Structured RNAs are critical components of key molecular machines in the cell, such as the spliceosome, ribosome, and telomerase, and RNA structures play important roles in the control of mRNA and noncoding RNA functions (Cech and Steitz, 2014). Base pairing dominates the energetics of both RNA folding and RNA-RNA interactions. Despite recent advances in measuring RNA structures in living cells (Ding et al., 2014; Rouskin et al., 2014; Smola et al., 2015; Spitale et al., 2015), current methods largely provide one-dimensional information. That is, these methods identify which bases are single or double stranded but do not directly reveal the counter-parties in each base pair (Lu and Chang, 2016). Inferring transcriptome structure in living cells is especially challenging, due to the presence of long-range structures, pseudoknots, alternative structures, repetitive sequences, and RNA-RNA interactions. One example illustrating these difficulties is XIST, a long noncoding RNA (lncRNA) required for X chromosome inactivation in female cells of eutherian animals (Penny et al., 1996). The key region for XIST-mediated epigenetic silencing, termed the A-repeat, is comprised of 7.5 or 8.5 near-identical copies of a sequence, and multiple structural models have been proposed (Fang et al., 2015; Maenner et al., 2010; Wutz et al., 2002). The structural basis for XIST interaction with key protein partners like SPEN is also not known (Chu et al., 2015; McHugh et al., 2015; Moindrot et al., 2015; Monfort et al., 2015). These challenges highlight the need for further advances to address the structures of the vast majority of coding and noncoding RNAs in the cell.

RNA affinity capture and proximity ligation may offer the next generation of solutions (Engreitz et al., 2014; Helvak et al., 2013; Ramani et al., 2015; Sugimoto et al., 2015). While these methods can identify RNA base pairs, current methods are limited by specific protein or RNA baits, which are performed one at a time and have limited resolution especially for longer RNAs (Engreitz et al., 2014). Here, we describe a general method that directly identifies base-pairing interactions in living cells and, by doing so, determines both RNA structures and RNA-RNA interactions. PARIS (psoralen analysis of RNA interactions and structures) combines several critical steps (in vivo crosslinking, 2D purification of RNA duplexes, and proximity ligation) that yield excellent sensitivity and specificity, as validated by...
numerous known structures and evolutionary conservation. We discovered a large number of long-range and alternative structures. PARIS-determined structures contain many targets of double-stranded RNA (dsRNA)-binding proteins (STAU1, DICER1, DGCR8). Furthermore, the high confidence structures guide two new approaches for phylogenetic analysis of RNA structures, revealing conserved architectures in housekeeping gene mRNAs. The combination of PARIS, icSHAPE (in vivo click selective 2'-hydroxyl acylation and profiling experiment), phylogenetic analysis, and iCLIP (individual-nucleotide resolution UV crosslinking and immunoprecipitation) reveals the overall architecture of the XIST lncRNA and the mechanism of SPEN binding to XIST A-repeat.

RESULTS

The PARIS Method and Validation

Current methods for in vivo probing generate averaged reactivity profiles and fail to capture the complexity of RNA structures that include long-range structures, pseudo-knots, and alternative conformations. To address these challenges, we developed PARIS to directly identify base-paired helices, the most basic elements in RNA structures and RNA-RNA interactions (Figures 1 and S1; Table S1; Experimental Procedures). The PARIS method employs the highly specific and reversible nucleic acid cross-linker psoralen-derivative 4'-aminomethyltrioxsalen (AMT) to fix base pairs in living cells (Calvet and Pederson, 1979). AMT
intercalates in RNA helices and, upon photo-activation, cross-links the two strands, with a preference for staggered uridines (Climo et al., 1985). Partial RNase and complete proteinase digestion during RNA purification ensures that the identified crosslinks are limited to small and directly base-paired RNA fragments (Figures S1A–S1C). Two-dimension electrophoresis of the RNase-digested fragments enables purification of only crosslinked fragments (above the main diagonal, Figures 1A, 1B, S1D, and S1E). The 2D purification consistently recovers 0.2%–0.5% of input RNA as double-stranded (above the diagonal), demonstrating that 2D purification is essential for enriching dsRNA fragments. Proximity ligation of duplex RNA fragments, photo-reversal of crosslinks, and high throughput sequencing reveal the direct base pairing between fragments. Each PARIS read is an individual-molecule evidence of a duplex between two RNA fragments (arms). The multiplicity of PARIS reads can thus reveal a single common structure, multiple alternative structures, or interactions between two RNAs in trans (Figures 1D–1G). The combination of these important features allows us to model RNA structures and interactions with high specificity and sensitivity.

We performed PARIS on human HeLa, HEK293T, and mouse embryonic stem (mES) cells and generated a total of 350 million reads after removing duplicates. The gapped reads, arising from RNase digestion of single-stranded loops in RNA structure, constitute 2.5%–6% of all mappable reads (Figures 1B and S1; Table S1). Given the absence of any background above the diagonal in the –AMT controls (Figure 1B), the non-gapped reads may come from failed ligations of duplexes due to steric hindrance (Sugimoto et al., 2015). PARIS is highly reproducible across biological replicates in each of the three cell types (R = 0.94–0.98 between replicates; Figures 1C, S1F, and S1G).

We assembled gapped reads into duplex groups (DG), each corresponding to an RNA stem loop, with the two arms from the stem and the gap from the RNase-cleaved loop, or an RNA-RNA interaction, with the two arms from the two interacting RNAs (Figures 1E–1G). DGs are filtered to retain only the ones with high confidence supported by multiple reads. To visualize this new type of RNA structurome data and associated structure models, we developed new features in the Integrative Genome Viewer (Robinson et al., 2011) (Figure 1D). Gapped reads are displayed in groups by DG, and structure models are visualized as arcs connecting the two arms of each DG (see Supplemental Experimental Procedures for the detailed analysis methods, directions, and links to visualization of PARIS data).

We validated the sensitivity and specificity of PARIS using a number of well-studied RNAs, such as rRNA, small nuclear RNA (snRNA), and microRNA (Figures 1H, 1I, and S2). Complex RNA structures are currently difficult to detect using one-dimensional chemical probing or computational prediction. Among the most difficult structures to predict are pseudoknots, comprised of interlocked helices, in which the loop of one stem loop participates in base pairing with an outside region. We were able to detect well-known pseudoknots in telomerase RNA (TERC; Figure 1J), RMRP, and RPPH1 (the RNA components of RNase MR and RNase P, data not shown) in both human and mouse PARIS data.

### Global Properties of the RNA Structurome Revealed by PARIS

Having established the PARIS method, we investigated the global properties of the RNA structurome. Most previous experimental and computational methods can only identify short-range structures (i.e., the span from the beginning of the left arm of the duplex to the end of the right arm), typically focusing on <200 nt windows. We found that a large number of RNA duplexes (29%–40%) span >200 nt in the three cell types and 4%–11% of duplexes span greater than 1,000 nt (Figure 2A).

We next investigated the extent to which RNA duplexes are organized into higher-level architectures. Many genomic studies categorize mRNAs into 5′ UTR, coding sequence (CDS), and 3′ UTR and perform analyses on these units assuming they are separate entities. We observed extensive RNA duplexes that cross these artificial boundaries. To illustrate the long-range structures, we plotted the number of DGs connecting among the first three and last three exons (Figure 2B). Even though most structures are local, as shown in Figure 2A, we observed many structures that span multiple exons (Figure 2B). For example in the RPS4X mRNA and other mRNAs, we observe multiple independent loops between the 5′ UTR and CDS, the CDS and 3′ UTR, the 5′ and 3′ UTRs (Figures 2C, 2D, S3A, and S3B), and the structures that cover the start and stop codons (Figures S3C and S3D). In addition, we also identified structures formed by repetitive elements like Alu elements (Figures S3E and S3F). The RNA structural features that dictate the specific recognition of dsRNA binding proteins (dsRBPs) to their cognate targets are not known, and PARIS identified the RNA structures associated with the dsRBP binding sites (Figure S4).

### PARIS-Guided Analysis of RNA Structure Conservation and Covariation

The large number and diversity of RNA duplexes identified by PARIS poses a challenge to distinguish the subset of structures with important biological functions. Evolutionary conservation of RNA secondary structure across several species is a strong indicator of function (Smith et al., 2013). Conserved RNA duplexes are supported by conservation of base pairs between the two arms or, more convincingly, by covariation in evolution (e.g., swapping Watson-Crick base pairs across the helix, i.e., less conservation). Genomic screens of conserved structures usually employ sliding window analysis of multiple sequence alignments and therefore are limited by the window size, sliding step, and general lack of experimental validation. Whereas typical covariation analysis uses sliding windows of 200 nucleotides to achieve reasonable runtimes (Smith et al., 2013), PARIS data reveal that a substantial fraction of the RNA duplexes spans more than 200 nt. This observation suggests that a large number of the structures (at least 23%–46%; Figure 2A) have been missed and are in fact incorrectly assigned to nearby neighbors by current methods. We reasoned that PARIS data can focus evolutionary analyses to the biologically relevant helix arms and overcome length limitation imposed by current methods and that evolutionary conservation can globally validate and highlight functional RNA duplexes.

RNA duplex determination by PARIS in human and mouse cells enables direct analysis of global structure conservation in two ways. First, direct determination of RNA duplexes by PARIS...
enabled us to precisely position the two arms of RNA helices in whole-genome alignments and guide covariation analyses regardless of their linear distance (Figure 3A). We measured the significance of base-pair covariation and structure conservation by shuffling sequences within each duplex, and calculating a Z score based on the distribution of structure energies in 100 shuffled alignments for each DG (Gesell and von Haeseler, 2006). This guided analysis revealed 25% of the well-aligned helices in amniotes’ genomes are highly conserved (Z score < 2.326, corresponding to a p value of 0.01). Many of these conserved structures also show strong covariation (46% conserved DGs with less than 10 kcal/mol covariation energy contribution; Figure 3B; Table S2). Among these conserved structures, we found that 43% of them span long distances (>200nt; Figure 3C). This analysis further validates the PARIS method by showing that a significant fraction of the experimentally derived structures are potentially functional (Figures S3D and S5).

Prior computational genomic screens have identified large numbers of conserved elements, yet little is known about their function. Bejerano et al. (2004) reported the identification of a 481 ultra-conserved elements (UCEs) in human, mouse, and rat, and 95 of them are located in mature RNA transcripts. We intersected the 95 UCEs with the PARIS-defined structures and found 14 overlapping with mES cell PARIS DGs and 34 overlapping with human PARIS DGs, and 12 of them overlap with both human and mouse PARIS DGs (Figures S5B–S5D; Table S3). This analysis suggests that at least some of the UCEs encode structural elements.

Second, the PARIS-determined structures in two distantly related species—human and mouse—allowed us to directly compare the structures on homologous sequences. We lifted the coordinates of mouse RNA structures to the human genome based on human-mouse pairwise genome alignments and intersected the helices between the two species (Figure 3A; Table S4). Despite the limited coverage of homologous RNAs between the two cell types, different cell type origins, and the dramatic difference of noncoding regions, we identified 10% of the structures to be shared between human and mouse. Among these 3000 structures shared between human and mouse, 22% of them span regions longer than 200 nt (Figure 3C). In addition, 29% of the direct-comparison-discovered (approach II) conserved helices are also found by structure-based phylogenetic analysis (approach I; Figure 3A).

Direct comparison of PARIS data in human and mouse validated conserved long-range structures in mRNAs and lncRNAs (Figures 3D, 3E, and S5A). In the RPL8 mRNA, of the 44 DGs identified in human cells and of the 46 in mouse cells, 23 are shared (Figure 3D). Many of these conserved structures span...
different exons, revealing conserved architecture of the RPL8 mRNA (p < 0.001 with 1,000 shuffles). The conserved long-range structures that connect exon3 to exon6 are also supported by icSHAPE data (low SHAPE reactivity in the base-paired region) in both species and phylogenetic analysis of vertebrates (Figure 3D). In addition, analysis of five mRNAs and the well-known lncRNA MALAT1 with similar numbers of PARIS-detected DGs in human and mouse showed that architectures are conserved for all of them (Figures 3E and S5E–S5H).

PARIS Reveals Pervasive Alternative RNA Structures

Dynamic RNA structures play important roles in regulating gene expression and catalyzing enzymatic reactions (Dethoff et al., 2012). Previous methods for identifying dynamic or alternative structures typically use McCaskill’s partition functions, with or without flexibility measurements as soft constraints (McCaskill, 1990; Ritz et al., 2013). These methods are often limited by sequence length and lack experimental validation. Since PARIS detects individual RNA duplexes in cells, alternative structures are directly detected as conflicting duplexes (Figure 4A). As a positive control, we detected the important U4:U6 alternative structures in the U4:U6 dimer in addition to their individual structures (Figures 4B and 4C).

We also identified new alternative structures, for example in the 3’ UTR of TUBB mRNA (Figures 4D and 4E) and lncRNAs MALAT1 and XIST (Figures S5E and S7B). The TUBB cluster of
alternative structures consists of five helices (DG1–DG5). Among these structures, DG1, 2, 4, and 5 appear to be mutually exclusive (Figure 4E). DG2 and 3 also have strong conflicts with each other, and thus cannot simultaneously take place on the same molecule. We analyzed the top 50 mRNAs with the highest numbers of detected helices in the three cell types and found that about 20%–50% of them are involved in at least one pair of alternative structures, suggesting that alternative structures are pervasive (Figure 4F; Table S5). Interestingly, a substantial number of the helices are involved in more than three pairs of alternative structures, suggesting highly complex networks of structures in living cells. These results are consistent with recent in vitro studies showing mRNAs sampling multiple structures (Kutchko et al., 2015).

Alternative RNA structures could be simply a result of the degeneracy of base pairing or, in contrast, be important for the RNA’s function. The latter scenario predicts that some of the alternative structures should be evolutionarily conserved. To
test this, we integrated PARIS, icSHAPE, and phylogenetic analysis to examine both high-level architecture and high-resolution structures in a functional context. The matched PARIS and icSHAPE datasets in HEK293 and mES cells showed that the alternative structures are evolutionarily conserved. Out of the 44 DGs for human RPL8, 32 of them form 42 alternative structure pairs, and 19 of the 42 pairs of alternative structures are conserved between human and mouse. An example alternative structure is shown in the coding region of RPL8 mRNA (Figure 4G). Both human and mouse PARIS and icSHAPE in the same cell types support this pair of alternative structures. Approximately 5% of the alternative structures examined have both structures supported by sequence conservation or covariation in evolution. Thus, some alternative structures in mRNAs are evolutionarily conserved and therefore likely functional.

PARIS Identifies RNA-RNA Interactions in trans with High Precision

RNA-RNA interactions are used by many ncRNAs to build macromolecular complexes and regulate gene expression (Lee et al., 2015). Current methods to identify RNA-RNA interactions require a “bait” protein or RNA; thus, it can be limited in scope (Helwak et al., 2013; Sugimoto et al., 2015). In contrast, PARIS is a general method that can detect RNA-RNA interactions in a protein/RNA-agnostic fashion. SnoRNAs (small nucleolar RNAs) and scaRNAs (small Cajal body RNAs) guide modification and processing of rRNAs and snRNAs (Figure 5A) (Kiss, 2001). We compared all known snoRNA: RNA interactions with the PARIS data from HEK293 cells. All the arms mapped to the rRNAs are centered on the modification sites, with a very narrow distribution (~20 nt at half height; Figure 5B, 5D, and S6). Given that snoRNA:rRNA interactions are around 10–20 base pairs, PARIS determines the interaction with near base-pair resolution. Furthermore, because rRNA and snoRNAs are among the most abundant RNAs, the precise mapping of their interaction sites confirms the high specificity of PARIS. The availability of both human and mouse PARIS data and the identical location of the interaction sites provides even stronger evidence to the authenticity of the interactions (Figures 5C, 5D, and S6).

We highlight two applications of PARIS to understand RNA-RNA interactions. First, PARIS can identify new RNA interactions, such as between snoRNAs and rRNAs. U8 snoRNA is essential for the processing of 5.8S and 28S rRNAs (Peculis and Steitz, 1993). U8 depletion leads to accumulation of pre-rRNA intermediates in Xenopus. Previous studies suggested that the 5’ end of U8 snoRNA base pairs with the 5’ end of 28S rRNA based on accessibility measurement (Peculis, 1997). Phylogenetic analysis revealed high conservation of ~15 nt at U8 snoRNA 5’ end (Peculis, 1997), suggesting that this region is essential. We found that, in both human and mouse cells, the primary U8 snoRNA interaction sequence is located on the 5’ end, consistent with previous studies (Figure 5E). However, the 28S interaction site is near the 3’ end in both human and mouse cells (Figure 5F, blue shaded area). No crosslinking is observed on the previously proposed binding site, even though uridine crosslinking sites are present (Figure 5F, gray shaded area). Phylogenetic analysis using the Rfam database provided independent support and showed that the highly conserved nucleotides correspond to the base-paired nucleotides in the new model (Figures 5G and 5H). The new model is more energetically favorable, with a minimum free energy of ~19.9 kcal/mol versus ~2.5 kcal/mol for the current model (Peculis model). Thus, PARIS can nominate new RNA interactions that derive further support from comparisons of human and mouse PARIS data, evolutionary conservation, and computational modeling.

Second, PARIS can refine the resolution of RNA-RNA interaction sites. U1 snRNA has been shown to bind 5’ splice sites and other cognate sequences throughout the transcriptome (Almada et al., 2013; Lu et al., 2014; Ntini et al., 2013, Engreitz et al. (2014) used RAP-RNA to enrich for U1-associated RNAs and identify U1 binding sites across transcripts. However, this purification approach recovers broad regions (Figure 5J). In contrast, PARIS determines high-resolution binding sites for U1. In both human and mouse, the first ~20 nucleotides of U1 are involved in trans interactions, consistent with the accessibility of the first 12 nt (Figure 5I), and the interaction with target RNAs is focal (Figure 5J). For instance, Engreitz et al. (2014) reported strong interactions between U1 and Malat1 in mES cells. We find precise PARIS interactions between U1 and Malat1 within the broad RAP peaks. The U1:Malat1 PARIS interactions are conserved between human and mouse (p = 2.3 × 10^-160). Fisher’s exact test; Figure 5J). These results are consistent with strong predictive power of complementary U1 motifs in target RNAs for U1-dependent RNA stabilization (Almada et al., 2013), indicative of precise sequence-dependent interactions.

XIST Structure Informs Higher-Order Assembly of XIST-Spen Complex

XIST is a 19kb IncRNA essential for X chromosome inactivation in placental mammals (Brown et al., 1991; Penny et al., 1996). However, one-dimensional methods have produced conflicting models of its structure. We used a combination of three orthogonal methods—PARIS, icSHAPE, and phylogenetic conservation—to determine the structure of the XIST IncRNA in living cells (Figures 6A and 6B). Global analysis of the PARIS data reveals both local helices and multiple long-range structures that span up to 7 kb (Figure 6B). The long-range helices organize regions of the RNA into four major domains. To determine if the identified secondary structures are biologically meaningful, we used the PARIS-determined helices to guide phylogenetic analysis. Our analysis reveals that 10% of the PARIS-determined helices are conserved, and the domain structures for domains 1, 2, and 4 are conserved (Figures 6C and S7A; Table S6). A conserved long-range structure over 7 kb that anchors domain 2 is shown in Figure 6D. A large number of the helices in XIST are involved in alternative structures, suggesting that this IncRNA is highly dynamic (Figures S7B and S7C). Interestingly, another IncRNA MALAT1 also contains many long-range structures, yet NEAT1 does not (Figures S5B and S7D).

The A-repeat, located at the 5’ end of the XIST RNA, contains up to 8.5 copies of a highly conserved sequence separated by uridine-rich variable spacers (8.5 repeats in human and 7.5 in mouse, ~400 nt; Figure 6E). A mouse Xist mutant lacking the A-repeat is unable to silence genes but is still capable of coating the X chromosome (Wutz et al., 2002). The A-repeat is thus a critical link in RNA-mediated epigenetic silencing. The A-repeat was
(A) Models of H/ACA box sno/scaRNA-guided RNA pseudouridylation and C/D box sno/scaRNA-guided 2’-O-methylation. Ψ, pseudouridine. 2’-O-Me,
2’-O-methyl.

(B) Specificity and resolution of the snoRNA-guided modification of human rRNAs. For each known snoRNA:rRNA interaction, the number of reads were normalized so that the maximum is one. All identified snoRNA:rRNA interactions from HEK293 cells were averaged.

(C–D) Base-pairing model from snoRNABase (C) and PARIS data (D) were shown for the SNORD95:28S interaction. The asterisk indicates the known modification site.

(E–F) PARIS in human and mouse cells reveals the interaction site on U8 snoRNA (E) and 28S rRNA (F). PARIS-determined interaction sites were marked by the blue box, while the previously reported binding site is shaded gray (Peculis 1997).

(G–H) The original U8:rRNA interaction was not supported by phylogenetic conservation and hybridization energy (G), whereas the newly identified U8:rRNA interaction is (H). The consensus sequences were from Rfam.

(I) Meta-analysis of the U1 target site. The U1:MALAT1 interactions use the 5’ end of the U1 snRNA in both human and mouse cells. (J) U1 snRNA interacts with MALAT1 RNA in human and mouse cells. PARIS achieves higher resolution than RAP (RNA antisense purification) (Engreitz et al., 2014). The blue-shaded peaks are shared between human/mouse PARIS and RAP data. The red-shaded peaks are shared between one of the PARIS datasets and RAP data. Fisher’s exact test was used to show the significant overlap between human and mouse PARIS-determined U1 sites.

(K) Example gapped reads for a conserved U1:MALAT1 interaction. The 5’ end of the U1 snRNA interacts with MALAT1 (at nt position ~5400).

See also Figure S6.
recently found to be required for Xist to interact with a small number of proteins (Chu et al., 2015), and among these, Spen emerged as a factor linking Xist to histone deacetylase complexes and gene silencing (Chu et al., 2015; McHugh et al., 2015; Moindrot et al., 2015; Monfort et al., 2015). Despite its importance, the repetitive nature of the A-repeat has complicated structural studies. Indeed, several contradictory models have been proposed, suggesting that each repeat base pairs within itself (“intra-repeat”) (Wutz et al., 2002), base pairs with other repeats (“inter-repeat”) (Maenner et al., 2010), or a combination of both (Fang et al., 2015). Prior studies were limited by the use of one-dimensional RNA structure data and computational models that arbitrarily precluded long-range RNA interactions (Fang et al., 2015; Maenner et al., 2010; Wutz et al., 2002).

PARIS highlighted several key structural features of the ~400 nt A-repeat region in vivo. First, the A-repeat does not form duplexes with any sequence far from the region, suggesting that this region mostly folds as an isolated domain (Figure 6B). Second, the repeats form extensive duplexes. All the detected RNA duplexes are between repeats (Figure 6E). While we cannot rule out the possibility that intra-repeat structures can form, our data suggest inter-repeat structures are more likely to occur in vivo, consistent with the higher stability of inter-repeat helices (ΔG = −15.2 kcal/mol for inter-repeat versus −5.8 kcal/mol for intra-repeat duplex). Each repeat tends to contact the closest repeats, but long-range contacts (bigger arcs) are also observed, suggesting 3D folding of the A-repeat region. In addition to the inter-repeat structures, we also observed structures between spacer 4 and several repeats. Repeat 4 and spacer 4 are not conserved rodents (Elisaphenko et al., 2008): these spacer-repeat structures may have species-specific function. Notably, the inter-repeat helices form between the first halves of the

---

Figure 6. Integrated Structure Analysis of the Human XIST RNA

(A) Overview of XIST IncRNA. Xist exons and repeat, phylogenetic conservation (PhyloP), icSHAPE, and PARIS data in HEK293 cells are shown.

(B) Architecture of the XIST RNA. Each point in the triangular heatmap shows the PARIS connection between the two regions indicated by the feet of the triangle. Data are plotted in 100 nt x 100 nt bins. Each RNA duplex detected by PARIS is plotted below. The duplex loops are clustered into four major RNA structure domains. The repeat A region is a small domain before the domain 1.

(C) Conservation of RNA duplexes determined by phylogenetic analysis of eutherian XIST homologs. Conserved helices (p value < 0.01) are plotted.

(D) An example long range (~7 kb gap) structure with PARIS, icSHAPE, and phylogenetic support (9.4% of all base pairs are one- or two-sided covariants).

(E) Integrated structure analysis of the conserved repeats in the A-repeat region. Conservation track, phyloP score for the eutherian alignments. PARIS coverage is shown in log scale. All detected inter-repeat and are illustrated in the arcs of structure models. A1–A8, repeats. Numbers in parentheses are the numbers of reads in each DG. The non-conserved repeat-spacer DGs (lower) were shown separately from the conserved ones (upper).

(F) Consensus model of the A-repeat inter-repeat structure. The consensus model depicts two repeats base-paired to each other. The red highlighted regions indicate the conserved repeats, while the non-highlighted regions indicate the spacers. Non-canonical: non-Watson-Crick base pairs with intermediate icSHAPE reactivity (constrained by the surrounding base pairs). Conservation and icSHAPE: average for all eight repeats. Mouse Xist in vitro SHAPE is similar to the HEK293 Xist icSHAPE. SPEN is crosslinked to 3–5 nt upstream of the inter-repeat duplex (see Figure 7 for SPEN iCLIP).

See also Figure S7 and Table S6
two repeats, flanked by single-stranded U-rich sequences on the 5' and the second half of the repeat on the 3' side (Figure 6F). Each inter-repeat unit has nearly identical structure, which is also supported by icSHAPE data that delineate precisely the complementarity (Figure 6F). Since each instance of the A-repeat can contact one of several other repeats, our data imply that the A-repeat exists as a family of multiple complex structures in living cells.

The presence of at least 7.5 copies of repeats in XIST and the unique structural unit raised the hypothesis that its higher-order structure may be important for the interaction with the key silencing factor SPEN. Previous studies of SPEN RRM domains suggest that they bind many RNA species, without preference for single copies of the A-repeat motif (Monfort et al., 2015). To address this issue, we performed individual nucleotide crosslinking and immunoprecipitation (iCLIP) with recombinant SPEN RRM domains (RRM2-4) and an ~1.6 kb region of mouse Xist RNA containing the A-repeat in vitro. We used a GFP mRNA matched in length as a negative control (Figure 7A and S7E).

iCLIP on both GFP mRNA and A-repeat RNA generated a radioactive SPEN band, but A-repeat RNA also generated a higher molecular weight band the size of a dimer SPEN RRM2-4 crosslinked to RNA (Figures 7A and S7E).

We sequenced RNA from the monomer and dimer bands separately and found that SPEN interacted nearly exclusively with the single-stranded nucleotides 3–5 nt upstream of the inter-repeat duplex. See also Figure S7.

Figure 7. The A-Repeat Structure Promotes SPEN Binding and Higher-Order RNP Formation

(A) In vitro iCLIP with human SPEN RRM2-4 and IRES-GFP or mouse repA RNA. The diagram shows the domain organization of SPEN. The autoradiograph shows one iCLIP experiment. The entire A-repeat region is 1630 nt. The IRES-GFP RNA is 1,533 nt. The dimer band relative intensity is 1 for the repA RNA and 0.61 for the GFP RNA control. See Figure S7 for another replicate of the iCLIP experiment.

(B) All the six iCLIP tracks are normalized by total read count and scaled to 0–2,300.

(C) For each of the four SPEN+repA iCLIP tracks, the crosslinking frequency for top 5% of crosslinked nucleotides was extracted from the repeats region and the outside region. This analysis shows that SPEN binds the repeats region more than the outside region.

(D) Nucleotides with the top 5% and bottom 5% of iCLIP signal were extracted from each of the four tracks, and then the icSHAPE signals were compared. This analysis shows that SPEN RRM2-4 is preferentially crosslinked to single-stranded regions (high icSHAPE signal).

(E) Model of SPEN-repA association. The base pairing among the repeats are stochastic and only one specific conformation is shown here. SPEN binding requires both single-stranded and double-stranded regions but is only crosslinked to the single-stranded nucleotides 3–5 nt upstream of the inter-repeat duplex. See also Figure S7.
Second, these targeted approaches are more appropriate if the structure is at near base-pair resolution, independent of the RNA of interest. The relationship of PARIS to HiCLIP and CLIP is also quite analogous to the structural organization of Droso phila roX RNA (Ilik et al., 2013). The advent of PARIS and related methods should catalyze discoveries of higher-order lncRNA structures in the future.

EXPERIMENTAL PROCEDURES

PARIS Experimental Method

HeLa, HEK293T, and mES cells were treated with or without AMT and crosslinked with 365 nm UV. Cell lysates were digested with S1 nuclease and RNA purified using TRIzol. Purified RNA was further digested with ShortCut RNase Ill to smaller fragments. RNA was separated by 12% native polyacrylamide gel and then the first dimension gel slices were further electrophoresed in a second dimension 20% urea-denatured gel. Crosslinked RNA above the main diagonal was eluted, proximity ligated with T4 RNA ligase I and photo-reversed with 254 nm UV. The proximity-ligated RNA molecules were then ligated to barcoded adapters and converted to libraries for Illumina sequencing. See Supplemental Experimental Procedures for details.

Determination of RNA Structure and Interactions

Sequencing reads were mapped to the human, mouse, or artificial genomes (such as the rDNA unit or the snRNAs) using STAR (spliced transcripts alignment to a reference) (Dobin et al., 2013), allowing chimeric mapping.
(in a chiasmic manner). Mapped reads were filtered to retain only gapped reads and the gapped reads were assembled into DGs and visualized together with the predicted or known secondary structures using newly implemented features in IGV (Integrative Genomics Viewer). To analyze RNA-RNA interactions, reads were mapped to the Rfam database and chimeric reads mapped to two RNA molecules were assembled into DGs. See Supplemental Experimental Procedures for details.

**Analysis of Structure Conservation/Covariation and Alternative Structures**

For structure-based analysis (approach I), DG coordinates in hg38 were used to extract alignment blocks from the amniote23 or other multiple genome alignments. The extracted alignments were scored for structure conservation and covariation. For direct comparison (approach II), DGs in mm10 were lifted to hg38 coordinates and conserved structures were defined as human and mouse DGs with both arms overlapped. Alternative structures were extracted such that, for each pair of DGs, one arm should overlap while the other should not. See Supplemental Experimental Procedures for details.

**In Vitro SPEN iCLIP**

SPEN RRMs-4 was mixed with the repA RNA or control GFP mRNA, cross-linked with 254 nm UV, digested with RNase, and labeled with radioactivity. The monomer and dimer bands were purified separately for iCLIP library construction.

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.04.028.


**AUTHOR CONTRIBUTIONS**

Z.L. conceived this project and designed the PARIS method and overall analysis strategy. Q.C.Z. and Z.L. implemented the PARIS analysis programs. Z.L. and B.L. performed all the PARIS experiments. R.A.F. performed the icSHAPE experiments. Z.L. and Q.C.Z. performed the analysis on most of the data. Z.L., B.L., C.D., A.R.G., K.J.G., and T.R.C. performed the in vitro studies on XIST and SPEN. Z.L., M.A.S., and J.S.M. performed conservation and covariation analysis. Z.L., J.T.R., and J.P.M. implemented the new features for structure visualization in IGV. H.Y.C. supervised the project. Z.L. and H.Y.C. wrote the manuscript with input from all authors.

**ACKNOWLEDGMENTS**

We thank Anton Petrov (Georgia Tech) for the compiled rRNA structures, Alexey Amurts (Stockholm University) for the mitochondrial rRNA structure, and Sebastian Will (MIT) for help with LocARNA software. We also want to thank Yoon-Jae Cho, Sekyung Oh, Y. Grace Chen for reagents. H.Y.C. is supported by NIH R01-HS004361 and P50-HG007735. Z.L. is a Layton Family Fellow of the Damon Runyon-Sloan Foundation Pediatric Cancer Fellowship Award (DRSG-14-15). M.A.S. and J.S.M. are supported by a Cancer Council Fellow of the Damon Runyon-Sohn Foundation Pediatric Cancer Fellowship and Sebastian Will (MIT) for help with LocARNA software. We also want to thank Anton Petrov (Georgia Tech) for the compiled rRNA structures, Alexey Amurts (Stockholm University) for the mitochondrial rRNA structure, and Sebastian Will (MIT) for help with LocARNA software. We also want to thank Yoon-Jae Cho, Sekyung Oh, Y. Grace Chen for reagents. H.Y.C. is supported by NIH R01-HS004361 and P50-HG007735. Z.L. is a Layton Family Fellow of the Damon Runyon-Sloan Foundation Pediatric Cancer Fellowship Award (DRSG-14-15). M.A.S. and J.S.M. are supported by a Cancer Council NSW project grant (RG 14-18). T.R.C. is an investigator of the Howard Hughes Medical Institute.

Received: December 18, 2015
Revised: March 3, 2016
Accepted: April 5, 2016
Published: May 12, 2016

**REFERENCES**


Development and Optimization of the PARIS Method and PARIS Experiments on HEK293T Cells and Mouse ES Cells, Related to Figure 1

(A) RNase digestion is required for efficient extraction of RNA from AMT-crosslinked cells. Cells crosslinked with AMT contain TRIzol-insoluble material that retains a large fraction of the cellular RNA. RNA recovery can be improved by S1 nuclease digestion of lysate (lysed in urea/SDS buffer). Experiments were performed in duplicates and the recovery of RNA is quantified. The non-AMT crosslinked without S1 sample is set to 1. Without S1: RNA recovery when RNA is extracted with TRIzol without S1 nuclease digestion. With S1: RNA recovery when S1 nuclease digestion of lysate is performed prior to TRIzol extraction.

(B) Bioanalyzer analysis of the S1/PK purified total RNA (without or with AMT crosslinking). The broad peak between 1000 and 4000 nt indicates successful crosslinking and purification of crosslinked RNA.

(C) ShortCut RNase III digestion reduces the size of the dsRNA for better separation by 2D gels. Bioanalyzer results are shown here.

(D) The 2D gel for high RNase digestion sample from HeLa cells.

(E) Second dimension denatured gel separation of the RNA fragments. RNA above the diagonal, as indicated by the blue box, was purified for library preparation. Only one blue box is drawn here for each gel as an example. Only AMT-crosslinked RNA samples are shown here. The recovery rates for crosslinked RNA are indicated under the gels.

(F and G) Reproducibility of the HEK and mES cell PARIS libraries. The numbers of total mapped reads for each transcript were plotted for each library.

(H) Overall analysis strategies used in this study. For the “STAR map to genomes” step, the “genome” can be a normal genome like hg38, or an artificial one that contains selected sequences (such as an rDNA unit, or an rDNA unit plus spliceosomal snRNAs, as shown in Figure S2, or the Rfam database of human RNA sequences) to facilitate special subsequent analysis.

(I) Despite the previously reported preference of psoralen for uridines, we noticed a higher amount of GC in the helices we detected using PARIS. The matrix represents the occurrences of all 16 dinucleotides in the HEK293_1 dataset.

(J) Violin plot of the size distribution of PARIS-determined stems, i.e., the base paired regions between the two arms (in base pairs). An example is shown for the HEK293_1 dataset.
A

Human 45S pre-rRNA (13357nt)

PARIS DGs at 1pm

PARIS DGs at 5pm

B

187 human ribosome crystal helices

human 1pm reads, score=0.01, 424 DGs; identified 88 of the helices in crystal structure

Blue arcs: only in crystal structure
Red arcs: only identified by PARIS
Black arcs: common between the two

E

human 1pm reads, score=0.01, noncrystal helices 336

mouse 1pm reads, score=0.01, noncrystal helices 77

28 common between human and mouse PARIS data.

F

Blue arcs: PARIS helices only in human non-crystal
Red arcs: only in human hiCLIP
Black arcs: common between the two

G

All STAU1 hiCLIP 1pm reads, score=0.01, 260 helices

52 common between PARIS and hiCLIP (black arcs)

K

U1 U2 U4 U6 U5 U11 U12 U4atac U6atac

(legend on next page)
Figure S2. PARIS Is Highly Specific and Sensitive, Related to Figure 1

(A) Specificity of the PARIS method is examined by mapping PARIS reads to a combination of the human 45S pre-rRNA and the 9 spliceosomal snRNAs (U1, U2, U4, U6, U5, U11, U12, U4atac and U6atac). The red line separates the two groups of RNAs. Both of these two classes of RNAs are highly abundant and reside in different subcellular compartments. No interactions were observed between the rRNAs and snRNAs when either 1 RPN (read per nucleotide, e.g., 1000 reads for a 1000nt long RNA) or 5 RPN reads were used in the analysis.

(B) Sensitivity of the PARIS method is determined by comparison to the helices in the crystal structure of the human ribosome with 1 RPN reads for human PARIS data. We were able to directly detect 88 out of the 187 helices (47%) in 18S, 5.8S and 28S rRNAs reported in the human ribosome crystal structure, including the 5.8S:28S intermolecular interactions (Anger et al., 2013). This is in contrast to flexibility measurements, where the detection of helices is based on prediction. Given that we measured rRNA structures in whole cells, which contain large amounts of ribosome assembly intermediates (Kim et al., 2014; McGinnis and Weeks, 2014), it is not surprising that we also identified many helices not present in the ribosome crystal structure (336 non-crystal helices at 1 RPN coverage).

(C and D) Number of reads per helix (C) and scores (D) for helices in the crystal structure versus those not in the crystal structure are plotted. In general the structures from the mature ribosome have higher coverage than the potential assembly intermediates (27 versus 10 median gapped reads respectively, p = 7 x10⁻¹⁰; and 0.057 versus 0.027 in connection score, p = 5 x10⁻⁸).

(E–G) To demonstrate that the non-crystal helices are real, we performed two types of comparisons. First of all, among the helices not present in the ribosome crystal structure (336), 29 of them (38%) are shared between mouse (77 non-crystal helices) and human PARIS data, despite that the two cell types are not of similar origin (E). Second, we compared PARIS-defined non-crystal helices with the STAU1 hiCLIP in HeLa cells (Sugimoto et al., 2015), and found that 52 of them are detected by both methods (F). (G) zooms in for (H) to show the local structures that are identified by both PARIS and hiCLIP. These analyses suggest that these helices are real and conserved; they likely represent rRNA assembly and processing intermediates.

(H) To illustrate the RNA helices identified by both PARIS and hiCLIP, two examples are shown that are present in (DG1) or absent from (DG2) the human ribosome crystal structure model of 18S rRNA 5’ domain. The red lines highlighted regions identified by hiCLIP and PARIS.

(I) Zoom-in view of the snRNA structure and interactions. In addition to the U4:U6 interactions described in Figure 1D-G, we also detected the known U2:U6 and U4atac:U6atac interactions, and recapitulated the structures for each individual snRNA. These interactions have previously been mapped using psoralen crosslinking methods (Hausner et al., 1990; Rinke et al., 1985; Tarn and Steitz, 1996). The U12:U6atac interaction was not detected due to their lower abundance. In summary, these results suggest that PARIS achieves high specificity and precision but with moderate sensitivity. The moderate sensitivity may arise from sequence bias of psoralen, dsRNA recovery efficiency, sequencing depth, and from occlusion of psoralen intercalation by proteins bound to the RNA.
Figure S3. PARIS Reveals Global Properties of RNA Structures, Related to Figure 2

(A and B) Example architectures of mRNAs. Human HEK293 PARIS DGs for two mRNAs are shown.

(C) Distribution of helices among different classes of RNA features.

(D) Transcriptome-wide structure probing experiments (e.g., icSHAPE and DMS-seq) revealed a general lack of structures covering the start and stop codons (Ding et al., 2014; Rouskin et al., 2014; Spitale et al., 2015). Despite the overall lack of structures near start and stop codons, we were able to detect structures that directly block the start and stop codons in a subset of mRNAs, suggesting that some of them may have consequences on mRNA translation. Here shows an example of conserved structure covering the stop codon and 3’UTR for the CD9 mRNA. The two blue tracks represent phyloP for 20 primates and 100 vertebrate
alignments (Primates Multiz and multiz100 from UCSC). The alignments are from the multiz100 multiple genome alignments and major clades are labeled on the species tree. The stop codon and anti-stop codon (i.e., base-paired with the stop codon) are labeled below the multiple alignments. For this structure, 8.8% of all potential base pairs are one- or two-sided covariants. 

(E and F) Nearly half of the human genome is comprised of repetitive elements, such as from transposons (relics of ancient viruses) and pseudogenes. Many repetitive sequences are transcribed on their own as IncRNAs, or embedded in other transcripts (Kim et al., 2004). Repeat transcripts are now known to have important roles in early human development, control of gene expression, and innate immunity (Gong and Maquat, 2011; Grow et al., 2015; Hung et al., 2015). Their repetitive nature makes them intrinsically prone to forming RNA duplexes and provides a mechanism for long-range structures and interactions. We identified a large number of structures in the repeat-containing RNAs, such as Alu elements (using uniquely mapped reads). Two examples of helices formed by Alu elements are shown from HeLa PARIS data. Our data thus provide a potentially useful resource for the investigation of repetitive transcripts in diverse biological investigations.
Figure S4. PARIS Identifies Binding Sites of dsRBPs, Related to Figure 2

RNA duplexes can create binding sites for dedicated dsRNA-binding proteins (dsRBPs) as well as occlude single-stranded regulatory RNA elements. Eukaryotic genomes encode a large number of dedicated dsRBPs, and several play important roles in regulating gene expression. Important dsRBPs include DGCR8 and DICER that are essential for microRNA and siRNA biogenesis, a distinct role for DICER in binding dsRNA stem-loops without processing (Macias et al., 2012; Rybak-Wolf et al., 2014), and STAU1 in regulated RNA decay (Kim et al., 2005). The targets of most dsRBPs studied so far are only known on the primary sequence level based on CLIP (presumably intra-molecular); a notable exception is STAU1, whose structure targets have been defined using crosslinking and protein pull-down followed by proximity ligation termed hiCLIP (Sugimoto et al., 2015). In contrast to hiCLIP, which analyzes dsRBPs one by one, PARIS provides a comprehensive overview of all dsRBP binding sites. The vast majority of RNA duplexes we observe were not previously known from individual dsRBPs, and we expand the universe of experimentally confirmed RNA duplexes by ~20-fold. Collectively, these results extend the concept of extensive long-range RNA structures in the transcriptome (Sugimoto et al., 2015), and indicate that this is a general feature rather than a private feature of individual RBPs. Given that the human genome encodes many dsRBPs, investigation of individual binding sites is laborious, PARIS can identify and validate targets for many dsRBPs at once (of course subject to the limitation of coverage). The incomplete overlap between PARIS and individual dsRBP targets is most likely due to current sequence

(legend continued on next page)
The sequence coverage of PARIS is spread across all RNA duplexes rather than just the targets of one RBP. The PARIS data alone does not reveal the identity of the proteins associated with the RNA target due to the lack of sequence specificity of most dsRBPs, instead, the integration of PARIS with CLIP data determines the structural targets of dsRBPs. Although likely, RBP binding does not necessarily exclude crosslinking since the helices can be longer than the RBP footprint and the RBP binding does not necessarily wrap the entire RBP footprint. It is possible that some RBP binding will inhibit the psoralen crosslinking, and this could contribute to the RBP binding sites not identified by PARIS.

(A) First of all, we intersected the STAU1-associated structures with the PARIS-defined structures in HeLa cells and found that 21% of the STAU1-associated structures were identified by PARIS with nucleotide resolution overlap, despite that completely different experimental methods were used.

(B and C) Comparison of the STAU1 hiCLIP (all three libraries combined) duplex groups to PARIS duplex groups (HeLa_LowRNase) in the exon4 region of SRSF1 (chr17: 58000919-58007346). All 8 of the STAU1 hiCLIP duplex groups were detected in the PARIS data. In addition, there are many other duplex groups in PARIS not present in the STAU1 hiCLIP data. Black arcs: structures detected by both PARIS and STAU1 hiCLIP. Blue arcs: structures only detected by PARIS. DG1 (detected by both methods) and DG2 (detected by only PARIS) are shown in details in (C).

(D) As an independent validation in HEK293 cells, we compared the PARIS data with the binding sites of dsRBP DGCR8 and DICER1, which are essential for miRNA biogenesis and also play miRNA-independent roles in gene expression. PARIS identifies 38% and 32% of the DICER1 and DGCR8 binding sites as defined by CLIP, respectively.

(E) PARIS identifies MIRLET7F1 miRNA precursor, consistent with the DGCR8 and DICER1 iCLIP data. HeLa cells PARIS data were used here, since the HEK293 data were not deep enough for this RNA.

(F) PARIS identifies non-canonical binding sites of DGCR8 and DICER1. Whereas CLIP data only showed the binding sites of the dsRBPs, PARIS revealed the structures that are associated with the proteins, and the structures may span a very long linear distance. Here shows the comparison between PARIS data and DGCR8 (Macias 2014), DICER1 (Rybak-Wolfe et al., 2013) CLIP data (HEK293T cells) in exon6 of ACTG1 gene (chr17: 81510479-81510833). PARIS: show arcs, coverage and DG alignments. DICER1 and DGCR8: show bed tracks and coverage tracks for both duplicates. Blue boxes: PARIS compared to DGCR8 binding sites. Red boxes: PARIS compared to DICER1 binding sites.

The relationship of PARIS to HiCLIP and CLIP is analogous to Hi-C versus ChIA-PET (Fullwood et al., 2009). PARIS has the advantage of extensive view of higher-order transcriptome structure, while hiCLIP can provide more in depth coverage of a single target dsRBP of interest.
Figure S5. PARIS Guides Phylogenetic Analysis of RNA Structures, Related to Figure 3

(A) An example of conserved structure in the 3'UTR of KRT5 mRNA based on phylogenetic analysis of vertebrate homologs. This stable helix spans 250nt, and is conserved down to metatherians (marsupials). For this structure, 9.0% of all potential base pairs are one- or two-sided covariants. The multiple alignments presented here were directly retrieved from the Vertebrate 100 genome alignments that included species beyond mammals. In this database, however, no alignments could be detected for the 3'UTR of KRT5 in species beyond metatherian. This could be due to the low alignment quality for distantly related species. To verify this independently, we used the Infernal program (Nawrocki et al., 2009) to build a covariance model based on the structure alignments that include eutherian and metatherian (Tasmanian devil) KRT5 sequences. Then we used this covariance model to search against the chicken, Xenopus and zebrafish KRT5 3'UTRs. This search did not yield positive structural alignments beyond what have been included in the multiz100 database. Note, no KRT5 homolog is known in platypus, a prototherian.

(B) Comparison of exonic UCEs to PARIS determined RNA structures.

(C) Notable examples of structured UCEs include HOXC4 5'UTR, a member of a class of RNAs now known to have unique translational control (Xue et al., 2015). Two high confidence structures were detected here, one covering the UCE uc.345.

(D) Since our PARIS experiments were performed on total RNA, intronic coverage is too low for systematic comparison with the intronic UCEs. However, we did find a few intronic structures overlapping UCEs. An example UCE (uc.28) in the third intron of human SRSF11 gene overlapped with human PARIS DGs.

(E) Phylogenetic analysis of MALAT1 based on human and mouse PARIS data. Conserved helices identified in both human (HEK293 cells, 3 replicates combined) and mouse (mouse ES cells, 3 replicates combined). The shown tracks are human and mouse PARIS DG represented by arcs. Mouse Malat1 DGs were lifted to human hg38 coordinates. Blue: PARIS DG only in human. Red: PARIS DG only in mouse. Black: common structures. Multiz alignments for 100 vertebrate species. Three conserved domains are discernable (Domain 1-3).

(F) Example conserved structure in human and mouse MALAT1. The size of the stem is around 20bp.

(G) Venn diagram of the structures directly identified by PARIS in human and mouse. Out of the 85 retrievable multiple alignments (from the 95 human DGs), 21 of them have strong conservation support ($p < 0.01, z < -2.326$). Out of these 21 DGs, 6 overlap with the 25 human/mouse common ones.

(H) Conserved alternative structures in the domain 3 of MALAT1 RNA. The highlighted region: H1H2 and M1M2 are conserved alternative structures.
Figure S6. PARIS Identifies Known and New snoRNA/rRNA and scaRNA/snRNA Interactions, Related to Figure 5

(A–E) Examples of known RNA-RNA interactions identified in human and mouse PARIS data. The base pairing models were retrieved from snoRNAbase (https://www-snorna.biotoul.fr). Up to ten reads per group are used to make each gapped read alignment example. The specificity of the interactions were demonstrated by the histograms of PARIS gapped reads for snoRNA/scaRNA binding sites on the target RNAs (18S, 28S and U6). Both human and mouse PARIS were available for the SNORD95:28S, SNORD10:U6 and SNOU13:18S interactions. The SNOU13:18S interaction involves two separated regions. Another example, snoU83B was identified by Duga et al. in 2000, and the RNA is predicted to bind 18S (Duga et al., 2000). Interestingly a modification site at 18S:468 was detected by Maden 1986, but the guide was not known (Maden, 1986). Here we show that snoU83 binds 18S:468 in living cells.
Figure S7. IncRNA XIST Architecture and Interaction with SPEN, Related to Figures 6 and 7

(A) Conserved architecture of the XIST RNA. Only conserved helices are shown in the structure. This is an alternative way to plot the structure in Figure 6C.

(B) Helices involved in alternative structures are plotted as a fraction of total helices for XIST. In total, 1252 out of the 1386 DGs are involved in alternative structures.

(C) Among all the pairs of alternative structures in XIST, 91 of them are supported by conservation/covariation (both considered together). Here shows an example pair of alternative structures in XIST RNA supported by conservation/covariation. For this pair of alternative structures, 9.1% and 11.7% of all potential base pairs are one- or two-sided covariants, respectively. Among the three regions, the last two came from the same repeat unit and diverged in evolution.

(D) Comparison of architectures for 3 lncRNAs. The RNA sizes and numbers of DGs are indicated on the left side. The NEAT1 (human) / Neat1 (mouse) lncRNA is dominated by short helices, while the MALAT1 (human) / Malat1 (mouse) and XIST (human) lncRNAs have longer helices in general. The length distribution is plotted either as a fraction of the RNA length (left) or their real size (right). The p values were calculated using two-sided Mann-Whitney U test. The MALAT1/Malat1 and NEAT1/Neat1 lncRNAs do not have introns. The DG size for XIST RNA was calculated on the mature RNA without the introns. Boxplot widths are proportional to the square root of the numbers of DGs.

(E) Repeating the in vitro SPEN iCLIP, as shown in Figure 7.

(F) Quantifying the SPEN iCLIP data within versus outside of the repeats region for Xist using Gini index.

(G) In vitro gel shift assay for the RRM2-4 peptide and Xist repA RNA fragment (504nt), compared to a size-matched bacterial beta-lactamase RNA control. Radiolabeled RNA was heated to 95°C for 1 min and snap cooled for 2 min, then allowed to fold for 30 min at 37°C in binding buffer (50 mM Tris-HCl, pH [7.5] at 25°C, 100 mM KCl, 2.5 mM MgCl2, 0.1 mM ZnCl2, 2 mM 2-mercaptoethanol, 0.1 mg/ml BSA, 0.1 mg/ml fragmented yeast tRNA (Sigma R5636), 5% v/v glycerol). Trace RNA was incubated with various concentrations of protein for 30 min at 30°C in the above binding buffer. Samples were cooled to 4°C before being loaded on a 0.7 % agarose gel (SeaKem GTG Agarose, Fisher Scientific cat # BMA 50070), buffered with TBE. Gels (15 x 10 cm) ran 90 min under 66 V at 4°C and were vacuum dried on a positively charged membrane before being exposed to a storage phosphor screen. Empty arrows indicate the lowest protein concentration, for a given RNA, that was sufficient for the formation of substantial protein-RNA complex. P-RNA complex: protein-RNA complex. Right, quantification of initial binding events, n = 3, and error bars represent SDs.
Supplemental Information

RNA Duplex Map in Living Cells Reveals
Higher-Order Transcriptome Structure

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Cell culture. HeLa-S3 cells were cultured in DMEM+ 10% FBS + Pen/Strep at 37°C with 5% CO₂. Flp-In HEK293T-REx cells express doxycycline inducible 3XFLAG-DDX3X at a level comparable to endogenous DDX3X (from YJ Cho lab). The HEK293T cells were cultured with DMEM + 10% Tetracycline-free FBS (Clontech, Tet System Approved FBS, 631106) + Pen/Strep + 2mM Glutamax (Gibco). Cells were induced for 8 hours with 1 µg/ml Doxycycline before AMT crosslinking. J1 mouse ES cells (male) were cultured on gelatin coated plates at 37°C with mES media: 500ml Knockout DMEM (Gibco), 90ml FBS, 6ml non-essential amino acid (MEM NEAA, 100x, Gibco), 6ml glutamine or glutamax (200mM stock solution), 6ml Pen/Strep, 1ml BME and 60µLIF (Millipore, ESG1106).

icSHAPE in HEK293 cells. The icSHAPE experiments were performed as previously described (Flynn et al., 2016). The sequencing data were deposited in GEO with the accession number: GSE74353.

PARIS (Step 1. in vivo crosslinking and RNA purification): AMT (Sigma-Aldrich A4330) was dissolved in pure water at a concentration of 1mg/ml. Cells cultured to 70% confluency in 10cm or 15cm plates were treated with the normal culture media plus AMT for 30min in 37°C incubator (40% v/v media and 0.5mg/ml AMT, adjusting osmolarity to physiological with 10x PBS). After the 30min incubation, the media was replaced with 0.5mg/ml AMT in 1x PBS. The control cells were incubated without AMT in either the incubation phase or the crosslinking phase. The plates in crosslinking solution were placed on ice bed in a Stratalink 2400 UV crosslinker and crosslinked for 30 min under UV365nm bulbs (Calvet and Pederson, 1981; Thompson and Hearst, 1983). The distance between cells and the bulbs is about 15cm. After crosslinking, cells were collected and frozen for further use. In general, the AMT-treated cell pellets have a darker color than the non-treated cells. This color difference can be used as an indicator of successful crosslinking. To purify RNA from the crosslinked cells, cell pellets were dissolved in 3 volumes of urea/SDS buffer (4M urea with 0.1% SDS), and pipetted vigorously. To each tube of dissolved cells (~ 20 million), add S1 nuclease buffer to a final concentration of 1x and add 2µl S1 nuclease (ThermoFisher EN0321), incubate at room temperature for 10min, with frequent pipetting to break viscous material. The S1 nuclease digestion is necessary to extract RNA from the crosslinked cells, since the crosslinking generates insoluble complexes in TRIzol (See Figure S1A-C for details of the digestion results). Both the S1 nuclease and subsequent ShortCut RNase III (NEB M0245) digestion produces 5’-phosphate and 3’-hydroxyl groups that are directly usable for subsequent proximity ligation and library preparation. After S1 digestion, add 10% SDS to a final concentration of 1% and then add Proteinase K (PK) to a final concentration of 10µg/ml. Perform the PK digestion at 50C for 30min to remove all proteins crosslinked to RNA. Even though psoralen crosslinking is selective for nucleic acids, proteins are also crosslinked to a lower degree (Sastry et al., 1997; Sastry et al., 1993). Then RNA is purified by TRIzol (Life Technologies) and resuspended in water. The S1/PK purified RNA was run on a Bioanalyzer to confirm that the crosslinking and purification worked. Successfully crosslinked and extracted RNA has a major broad peak between 1000 and 4000nt not present in the controls (Figure S1B).

PARIS (Step 2. Digestion and 2D purification of crosslinked RNA). The S1/PK purified RNA was further digested with ShortCut RNase III (dsRNase) at 37°C for 20min to reduce the size of the RNA fragments. This digestion is critical for proper 2D
separation of crosslinked RNA fragments (See Figure S1C for representative results). Each digestion reaction in 50µl included 20µg S1/PK purified RNA and 2µl (low RNase experiment in Figure 1) or 4µl (high RNase in Figure 1 and all other experiments) ShortCut RNase III. MnCl₂ and ShortCut buffer were used as described in manufacturer’s manual. After digestion, the RNA was usually quantified by Bioanalyzer, and 10-15µg RNA was usually recovered from the digestion of 20µg RNA. Purified RNA was separated by 2D gel electrophoresis as follows. For first dimension electrophoresis, RNA samples were loaded on 12% native polyacrylamide gel (1.5mm x 10cm x 10cm) and run at 150V for 100 min, at room temperature. Higher voltage would heat up the gel and should be avoided. After electrophoresis, the gel was stained with SYBR Gold (ThermoFisher S11494) and imaged under 312nm UV light. The 254nm UV reverses the crosslinking and should be avoided. Gel slices of each lane were embedded and polymerized into the top of the second dimension 20% urea-TBE denatured polyacrylamide gel (1.5mm x 10cm x 10cm). Two to three slices of gels can be multiplexed in one second-dimension gel. The second dimension is typically run at 55W for 50min. The second dimension gel has higher temperature due to the high power, and the high temperature facilitates denaturation of the dsRNA fragments. After the second dimension electrophoresis, the gel is stained with SYBR Gold and imaged. Destaining the gel with TBE increases sensitivity and is recommended since the crosslinked RNA is not abundant. Gel containing crosslinked RNA above the diagonal is cut out and crushed for RNA extraction.

PARIS (Step 3. proximity ligation, reverse crosslinking and library preparation). Purified dsRNA fragments were quantified by Bioanalyzer and the yield is about 50-100 ng for each 2D gel. The proximity ligation was performed under the following conditions. Each ligation reaction contains 10µl RNA, 2µl 10x ligation Buffer, 5µl T4 RNA Ligase 1 (NEB M0437M), 1µl SuperaseIn (Life Technologies), 1µl 0.3mM ATP and 1µl water. The ligation mixture is incubated at room temperature overnight (16-20h). After ligation, the samples were boiled for 2 minutes to terminate the reaction. After heat denaturation, the samples turned murky. These samples were centrifuged to remove the precipitate and then precipitated by ethanol and glycoblue (Ambion) (no need for Trizol extraction). Samples were resuspended in 15µl water. To reverse crosslinking, samples were transferred to the lid of a multi-well cell culture plate and directly irradiated with UV254nm for 15min on ice bed.

Two different methods were used for library preparation, one for HeLa, and the other for HEK293 and mES cells. For the HeLa PARIS library preparation, the photo-reversed RNA samples were reverse transcribed using a set of barcoded primers to generate cDNA. The primers are designed as follows (ordered from IDTDNA): /5phos/DDDNNXXXXNNNNNTACCCTTCGCTTCACACAAG/iSp18/GGATCC/iSp18/TACTGAACCGCNNNNNN, where the DDD indicates non-cytosine bases, XXXX indicates the barcode, N indicates any of the 4 bases, iSp18 is spacer. The order of the barcodes are the same as described before (Spitale et al., 2015). The first random hexamer region is used to check for PCR duplication, whereas the second random hexamer is the primer region. The subsequent library preparation for the HEK293 and mES cells is essentially the same as described before, starting from step 26 in the Flynn et al. protocol paper (Flynn et al., 2016). Briefly, RNA was first ligated with barcoded adapters and then reverse transcribed using primers annealed to the adapters. In each of the gel purification steps in library preparation, fragments with >40nt inserts are selected. Libraries are multiplexed and sequenced on the Illumina NextSeq. We noticed that the direct ligation of RNA to adapters was a better approach for the library
preparation because of higher reverse transcription efficiency, and therefore used the
direct ligation method for all subsequent libraries.

**Mapping and filtering of PARIS data.** See Figure S1H for an overview of the analysis
strategy. The custom scripts for data analysis are deposited in Github:
https://github.com/qczhang/ and https://github.com/zhipenglu/. After sequencing of
PARIS libraries, adapters from the 3’ end were trimmed off using Trimmomatic (Bolger
2014). We then collapsed these reads to remove PCR duplicates with identical
sequences including barcode region (using the readCollapse script from the icSHAPE
pipeline: https://github.com/qczhang/icSHAPE). The inclusion of random hexamer in the
middle of the primer/adapter allows us to remove PCR duplicates efficiently to achieve
the goal that each read from the sequencing represent an independent single molecule
measurement of an RNA structure. Then the libraries were split based on barcodes
(using splitFastqLibrary from icSHAPE pipeline, https://github.com/qczhang/icshape)
and 5’ adapters were removed using Trimmomatic.

After primary preprocessing, reads were mapped to hg38 or mm10 genome indices
using the STAR program (Dobin 2013). The parameters used are as follows. STAR --
runMode alignReads --genomeDir STAR_index --readFilesIn fastq_file --
outFileNamePrefix name_prefix --outReadsUnmapped Fastq --outFilterMultimapNmax
100 --outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4 --scoreGapATAC -4
--chimSegmentMin 15 --chimJunctionOverhangMin 15. The parameters chosen here
reduce the penalty for gapped reads and allow mapping of chiasmatic reads. For the
detailed analysis of a select subset of repetitive or intron-containing RNA genes, such as
45S rRNAs, mitochondrial rRNAs, snRNAs and XIST, STAR indices were made for each
one individually. The human snRNA index contains the complete set of 9 snRNAs (U1,
U2, U4, U6, U5, U11, U12, U4atac and U6atac). The human XIST index contains the
XIST RNA without the introns. To make these mini-genome STAR indices, it is important
to note that the STAR index generation step requires a custom value for the --
genomeSAindexNbases option (see STAR manual for details). To map reads to the
mini-genomes, the STAR parameters were adjusted so that --outFilterMultimapNmax is
1 and --alignSJoverhangMin is 15.

The STAR mapping produces normally mapped reads (xxxAligned_out.sam) and
chiasmatically mapped reads (xxxChimeric_out.junction and xxxChimeric_out.sam). The
gapped reads were extracted from the normally mapped reads and combined with
chiasmatically mapped reads for making duplex groups. These gapped reads (including
normal and chiasmatic reads) are the useful data for determining RNA structures and
RNA:RNA interactions. We obtained 2.5%-6% gapped reads in all mapped reads (Table
S1). The higher percentage of gapped reads in PARIS [compared to 0.28% in RPL
(Ramani et al., 2015) and 2% in hiCLIP (Sugimoto et al., 2015)], demonstrates the
efficiency of the experimental strategy. The ligation could happen on both ends of the
crosslinked helices, and these two ligation events would generate reads in normal
“splicing-like” configuration or in a chiasmatic manner (e.g. normal: LLLLLL-RRRRRR
or chiasmatic: RRRRRR-LLLLLL, L for bases from the left arm, while R for the right arm). The
presence of reads in a DG with both configurations suggest that these reads not only
come from distinct molecules but also ligated in different ways, thus lending stronger
support that these gapped reads come from structures.

**Processing of gapped and chiasmatic reads to read groups.** We implemented a
pipeline that automatically processes the mapped and filtered reads into duplex groups
(DG tag in the SAM file). The processing first removed gapped reads that are gapped as a result of splicing and further removed PCR duplicates. This step of analysis was implemented in the samPairingCalling.pl script in the paris pipeline (https://github.com/qczhang/paris). Furthermore, we implemented an additional tag NG (non-overlapping group) in the SAM file to maximally pack DGs. The XG (chiastic group, XG:i:0 for normal gapped, XG:i:1 for chiastic on the same strand of the same chromosome, and XG:i:2 for all others) tag was implemented to differentiate normal gapped reads and chiastic reads, where the two arms are swapped in relative position. Chiastic reads also include additional reads that are mapped to different strands or different chromosomes.

The DG assembly uses a two-step greedy algorithm to maximize speed while maintaining the reasonable clustering of reads into distinct groups that support individual RNA helices. Before DG assembly, gapped reads are first sorted by coordinates and examined in one round. In the first step of DG assembly, we generate intermediate DGs by grouping gapped together. Each read is either added to an existing DG or used to establish a new DG based on this criterion: all reads in a DG must share at least 5nt in both arms. Because every intermediate DG (and also the final DG) is represented using the coordinates of the core regions where all reads share (this is to ensure that all reads in one DG overlap with each other), some intermediate DGs that are not overlapped by core regions actually represent the same duplex structure. So in the second step, we assemble final DGs by merging those intermediate DGs as long as the maximum gap of both arms between two intermediate DGs is less than 10nt and the maximum length of both arms of the final DG less than 30nt.

To identify DGs with high confidence, we used two criteria. First each DG must have at least two unique gapped reads that have different termini, which should come from structures in two individual RNA molecules. The ligation in solution ensures that similar gapped reads are unlikely to form multiple times in solution by chance alone. Second, we computed a connection score for the two arms of a DG, as follows: number of reads connecting the two arms divided by the coverage of gapped reads at the two arms (connection_A_B/sqrt(coverage_A * coverage_B), A and B representing the two arms). The coverage of the two arms are different from the gapped reads connecting them because each region could be covered by multiple DGs, and some of them are likely to be alternative structures, which are pervasive in the transcriptome (Figure 4). We used a cutoff of 0.01 to remove low score DGs. This second criterion in effect normalizes for transcript abundance and ensures that our RNA duplex map is not dominated by low frequency duplexes in abundant transcripts.

The NG assembly algorithm is summarized as follows (https://github.com/zhipenglu/). All DGs are first ranked by read numbers (more reads, higher rank). Then going through all the DGs, each DG is either considered a new NG if it overlaps with all existing NGs, or added to the first existing non-overlapping NG. The end result is that for any NG, none of the DGs overlap with each other.

PARIS determines RNA helices with two arms of ~20-30nt each. At this size, the inferred base pairs are nearly always unique, which justifies the term “near base pair resolution”. This is in contrast to 1D flexibility measurements of RNA structures (e.g. DMS-seq, SHAPE etc.), where the data are nucleotide resolution but the precise base pairs are not determined. Although PARIS is performed on cells rather than single molecules, the
information we obtain is for single molecules (one-arm to one-arm Watson-Crick basepairs for each structured RNA fragment).

**Visualization of RNA structure models and PARIS data on IGV genome browser** 
(Robinson et al., 2011). As described above, gapped reads were organized into DGs based on newly implemented tags, each providing strong support for a potential RNA duplex. At the same time, DGs can be visualized as arcs connecting the two arms in zoomed-out view of transcripts, highlighting the architecture of entire transcripts. PARIS-constrained predicted base pairs are also displayed as arcs in zoomed-in views. This arrangement allows easy integration with flexibility measurements of RNA structures (like icSHAPE and DMS-seq), protein-binding sites (as determined by various CLIP methods), together with phylogenetic information, DNA sequence variants, copy number variation, and additional annotations of functional motifs in the transcriptome. This integration provides a holistic view of the structural basis of RNA functions.

Specifically, RNA structure models were prepared in bed format where each pair of second and third column coordinates constitutes a base pair. This format is similar to the connect format described by the mfold program (Zuker, 2003). This method allows simultaneous presentation of alternative and complex structures, and easy comparison with other types of data on genome browser tracks. The PARIS gapped reads alignment data were visualized with the following options in IGV. Gapped reads can be grouped by tag XG (chiastic group), DG (duplex group) or NG (non-overlapping group), colored by tag XG or DG as needed (see IGV website for detailed updates). Visualization of entire bam files is not recommended since long-range and especially inter-molecular interactions would extremely compress the DGs in arcs or read alignments. Instead, individual RNAs should be extracted from the bam files (aligned gapped reads) and bed files (arcs representing DGs) using the SAMtools and BEDtools programs. Examples and instructions can be downloaded from the following link: https://www.dropbox.com/s/1oqkcfzlfafdhq/PARIS_visdata.tgz?dl=0

**Analysis of RNA:RNA interactions.** PARIS determines RNA:RNA interactions in a “all-to-all” fashion. This is the strength of the method, yet it makes the analysis very complicated. The majority of the human genome is duplicated sequences such as repetitive DNA, genes with multiple copies and/or pseudogenes. This property makes unambiguous identification of RNA:RNA interactions very difficult on a genomic scale. To identify true RNA-RNA interactions from PARIS data, the reads were mapped to selected subsets of RNAs, each one as a small “chromosome”. The non-redundant sets of RNA families from Rfam, most of which are noncoding, were used to make the human and mouse reference “genomes”. One sequence from each family was randomly chosen for each species. The 12S, 16S, 18S and 28S rRNAs were added since they are not extracted well from the Rfam. STAR mapping was performed with these specific parameters: --outFilterMultimapNmax 10 --alignSJoverhangMin 15 --outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4 --scoreGapATAC -4 --chimSegmentMin 15 --chimJunctionOverhangMin 15. Mapped reads were assembled into duplex groups and filtered to remove interactions within RNAs and ones with identical break points. To identify the interactions between ncRNAs and the 45S pre-rRNA, we mapped the reads to the Rfam non-redundant ncRNAs plus one copy of the 45S rRNA. To identify the interactions between ncRNAs and mRNAs/IncRNAs, we constructed a reference containing the Rfam ncRNAs described above, all the miRNAs from miRBase and all RNAs longer than 300nt from UCSC annotations.
Although we cannot give an accurate estimate of the proportion of reads from RNA:RNA interactions in all gapped reads for the reasons described above, we counted the number of gapped reads for several abundant RNA:RNA interactions, such as snRNAs (U4:U6, U2:U6, U4atac:U6atac etc.) and rRNAs (5.8S:28S) in the HEK293_1 dataset, to provide a glimpse of the RNA:RNA interaction abundance. In the 45S rRNA mini-genome, 5.8S:28S interaction is supported by 34862 reads for the 3 helices. U8:28S, 87 reads (shown in Figure 6), 376 reads in the mES dataset. U4:U6, 1734 reads. U2:U6, 1105 reads. U4atac:U6atac, 365 reads. U12:U6atac, 4 reads (filtered out due to the high stringency of the analysis).

MicroRNAs are detected in our PARIS data sets but generally toward the lower end of read coverage. For example, the Mir-295 cluster of microRNAs, which make up more than 50% of total miRNA in mouse ES cells, generated less than 20 gapped reads of in our mES dataset. We can detect the precursor structure of these miRNAs but not interactions with the known targets such as Casp2 or Ei24 mRNAs. Some possible explanation for the low coverage may be that the miRNA-mRNA duplex is typically short (~5 nt of the miRNA seed), and the Argonaute proteins entirely envelop miRNAs, which may blocks the psoralen crosslinking.

To examine the evolutionary conservation and covariation of the RNA:RNA interactions, the relevant sequences were downloaded from Rfam, and the interacting regions were aligned structurally using LocARNA (Will et al., 2007).

**Comparison of PARIS data with STAU1 hiCLIP data.** The STAU1 hiCLIP raw data from Sugimoto et al. were mapped to hg38 using STAR with the same parameters described above (Sugimoto et al., 2015). This mapping strategy yields even better results than the original pipeline since STAR deals with gapped mapping natively and chaotic reads are also mapped properly at the same time. The adapters from the hiCLIP were simply ignored by the STAR mapper. The mapped reads were then processed in the same way as the PARIS data. To compare the PARIS and STAU1 hiCLIP data, the duplex groups (DG) in bed12 format were intersected using the bedtools package using the following parameters.

**Analysis of dsRBP CLIP data (DICER1 and DGCR8).** The HEK293 cell DICER1 PAR-CLIP raw data rep1 and rep2 (Rybak-Wolf et al., 2014) were mapped to hg38 genome using STAR (Dobin et al., 2013) and the following parameters: --outFilterMultimapNmax 10 --outSAMattributes All. Significant target sites were called using the PARalyzer program with default parameters (Corcoran et al., 2011). Numbers of DICER1 binding sites identified are 8899 and 3496, respectively. The HEK293 cell DGCR8 HITS-CLIP mapped data D8_2 and T7_2 in bedgraph format were lifted to hg38 coordinates. Significant DGCR8 binding sites were called using the Piranha program with the following parameters: -s -a 0.7 -b 40 -u 0. Numbers of DGCR8 binding sites identified are 57110 and 7290, respectively. To compare CLIP data with PARIS defined structures, binding sites that are common in both replicates were extracted. A total of 2383 DICER1 and 6415 DGCR8 target sites were used in the comparison.

**Phylogenetic analysis of RNA structure.** The two arm intervals of each DG were used to extract multiple alignments from whole-genome alignments of 23 amniote vertebrate species (Ensembl, hg38 version) with the python script maf_extract_ranges_indexed.py (bxpython package, https://github.com/bxlab/bx-python). RNAalifold was used to predict
a consensus structure from the alignments for each DG with or without inter-arm base-pairing constraints (Lorenz et al., 2011). The significance of each conserved structure was assessed using SISSIz shuffling with the RIBOSUM matrix (Gesell and von Haeseler, 2006).

For the direct comparison between human and mouse PARIS determined structures, the mouse DGs were lifted from mm10 to hg38 coordinates using the liftOver utility and the mm10ToHg38.over.chain file (UCSC). The liftOver program was run with the following parameters: liftOver -minMatch=0.2 -minBlocks=0.2 -fudgeThick. The minMatch was reduced from the default so that most regions can be properly aligned between species. In order to visualize the mouse PARIS reads on the human genome in IGV, the mouse PARIS reads were first converted to bed format using bedtools, lifted to hg38 coordinates, and then converted back to bam format using bedtools. It is noted that this strategy is limited by the quality of the available genome alignments, and improvement of these alignments is beyond the scope of the current study. To visualize the multiple alignments of RNA structures, the aligned sequences were plotted using the R-package using the default settings, except the choice of color codes (Lai et al., 2012).

**Analysis of alternative structures.** We defined alternative structures as helices that overlap on one arm by more than 50% (see the scripts in https://github.com/zhipenglu/duplex). These alternative structures are critical for the dynamic remodeling of the spliceosome during splicing. DGs were intersected with each other to identify pairs of DGs that have one pair of overlapped arms (left-left, left-right or right-right), but not two pairs at the same time. Inter-arm structures were predicted and significant overlapping of base pairs were used as another filter for alternative structures (at least 50% overlap). For alternative structures in the HEK293T and mES cells, we also confirmed them using the icSHAPE data from matched cell types (Figure 4G) and conservation/covariation data (Figure S7C). We note that interlocked RNA structures are also detected, likely representing alternative structures or pseudoknots. Although we can not definitively show that the interlocked structures are indeed pseudoknots, the results represent a stepforward in the identification of pseudoknots, which were previously only identified by crystallography or pure prediction.

**Determination of XIST structure.** Among the three cell types we performed PARIS on, only HEK293 cells express XIST and undergo X inactivation even though their karyotype is hypotriploid. HEK293 cells have been used to study human X inactivation. In 293 cells, human Xist silences linked genes in an A-repeat-dependent manner; 293 cells are thus a reasonable model for our Xist structural analysis (Chow et al., 2007). HeLa cells are female, but do not undergo X inactivation, while the J1 mES cells are male. To facilitate the analysis and visualization of the intron-containing XIST RNA, PARIS reads were mapped to a single mature XIST RNA reference without the introns. Since the XIST RNA contains multiple repetitive regions, only uniquely mapped reads were used. Gapped reads with mismatches were removed from both the Aligned.out.sam and Chimeric.out.sam files before duplex group assembly. For the repA region, although we cannot definitively determine whether the inter-repeat structures are intramolecular or between several identical XIST RNA molecules, the tendency for proximal interactions suggests that the inter-repeat contacts are within the same molecule, as would arise during co-transcriptional folding. Trans interactions would not follow such distance constraints (Figure 6).

**In vitro SPEN iCLIP**
In iCLIP experiments, we use UV radiation to crosslink direct RNA-protein interactions, and then use RNase to digest the RNA, followed by end-labeling the RNA with $^{32}$P to visualize the RNA-protein complex. A brief summary of the protocol is as follows.

Cloning and Purification of Recombinant SPEN RRMs as previously described (Arieti et al., 2014). RepA and control RNA were in vitro transcribed with T7 polymerase (MEGAscript T7 Transcription Kit), DNase Treated (TURBO DNase), and then purified with TRIzol. Purified RNA was resuspended and folded as described (Maenner et al., 2010). Until completion of UV crosslinking, samples were kept at room temperature. In a 1:20 RNA:protein molar ratio, 400ng of control or RepA RNA was incubated with the recombinant RRMs in the Maenner buffer for 20 minutes. The mixture was crosslinked twice with 2500 x 100 uJ using a UV Stratalinker 2400 (254nm bulbs), with mixing by pipette in between. SPEN-RNA crosslinked samples were treated with RNaseA, and stopped with RNase inhibitor. RNaseA was removed by spin filtration using a 30kDa Amicon column (Millipore) three times at 6000xg for 5 minutes, using RNase-free water for each dilution. Samples were then radiolabeled by 5’ kinasing with gamma-32P ATP (Optikinase). The completed kinase reactions were heated in denaturing, reducing sample buffer. The denatured samples were run on SDS Bis-Tris PAGE, transferred to nitrocellulose, and visualized by radioblot. The membrane in the region >8 kDa above the apparent protein size was cut out for RNA isolation. The RNA was isolated by Proteinase K treatment at 55C followed by acid phenol chloroform extraction and alcohol precipitation.

**SPEN iCLIP library preparation and analysis**

Preadenylated 3’ biotinylated adapters were ligated to the isolated RNA through 3’ end repair by a T4 PNK (NEB) dephosphorylation reaction, followed by RNA Ligase 1 (NEB) ligation reaction without ATP. Free adapter was removed by digestion with RecJ exonuclease. Subsequent library preparation (cDNA synthesis, circularization, library amplification and purification) were all completed as described by the FAST-iCLIP method (Flynn et al., 2015). Libraries were sequenced on an Illumina MiSeq using a 150-cycle v3 kit for single-end 75 base reads. Reads were mapped to the XIST or GFP RNA reference and RT stop sites were extracted.
Supplemental References


