

# High-throughput determination of RNA structure by proximity ligation

Vijay Ramani, Ruolan Qiu & Jay Shendure

**We present an unbiased method to globally resolve RNA structures through pairwise contact measurements between interacting regions. RNA proximity ligation (RPL) uses proximity ligation of native RNA followed by deep sequencing to yield chimeric reads with ligation junctions in the vicinity of structurally proximate bases. We apply RPL in both baker's yeast (*Saccharomyces cerevisiae*) and human cells and generate contact probability maps for ribosomal and other abundant RNAs, including yeast snoRNAs, the RNA subunit of the signal recognition particle and the yeast U2 spliceosomal RNA homolog. RPL measurements correlate with established secondary structures for these RNA molecules, including stem-loop structures and long-range pseudoknots. We anticipate that RPL will complement the current repertoire of computational and experimental approaches in enabling the high-throughput determination of secondary and tertiary RNA structures.**

The folding of RNA species into complex secondary and tertiary structures is central to RNA's catalytic, regulatory and information-carrying roles<sup>1</sup>. Pioneering approaches for elucidating RNA structure—including crystallography<sup>2</sup>, electron microscopy<sup>3</sup> and spectroscopy<sup>4</sup>—are technically complex and difficult to scale, motivating the development of computational algorithms for RNA structure prediction<sup>5–7</sup>. Current algorithms have limited predictive power, particularly for long-range interactions such as pseudoknots (secondary structures involving intercalated stem loops). With the advent of massively parallel sequencing<sup>8</sup>, less laborious experimental techniques have been developed for the global interrogation of RNA secondary structures. These include methods relying on structure-specific chemical modifications<sup>9–11</sup>, such as DMS-seq and SHAPE-seq, as well as methods involving digestion with structure-specific RNases<sup>12–14</sup>, like PARS-seq and Frag-seq. Although these methods probe the extent to which individual bases participate in secondary structures, they do not directly query which specific pairs of bases or regions interact to form these structures. To address this, researchers have combined systematic mutagenesis and structure-specific probing to generate pairwise information for inferring RNA folds<sup>15,16</sup>. However, despite considerable progress, the high-throughput determination of RNA secondary and tertiary structures remains a challenging problem.

Here we show that proximity ligation is a straightforward means of generating global pairwise data about RNA secondary and tertiary structure. Proximity ligation records the physical proximity

of two nucleic acid termini through their ligation and has been applied to detect DNA aptamer-bound proteins<sup>17</sup>, to probe protein-protein interactions through antibody-bound oligonucleotides<sup>18</sup>, and to achieve targeted or global chromosome conformation capture (3C)<sup>19,20</sup>. Proximity ligation has also been applied in conjunction with cross-linking and either affinity purification or immunoprecipitation to characterize small nucleolar RNA (snoRNA)-rRNA interactions<sup>21</sup> and Argonaute-mediated microRNA (miRNA)-target interactions<sup>22</sup>. However, these efforts have primarily focused on assessing specific *trans* interactions, rely on low-efficiency, 254-nanometer UV cross-linking and require time-consuming purification steps.

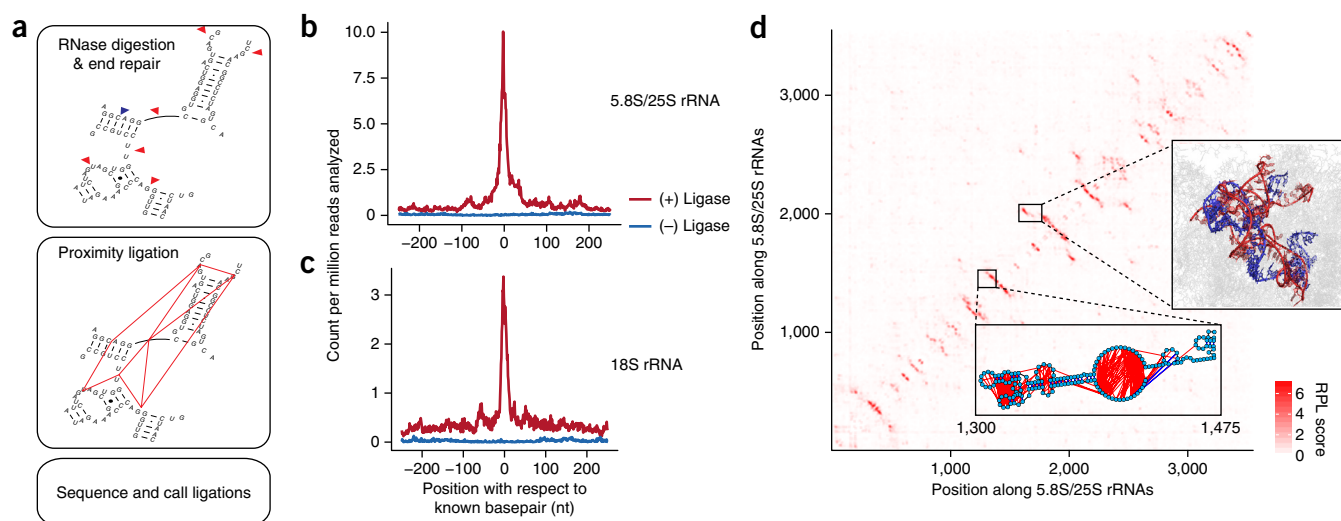
RPL ('ripple') globally assesses which pairs of regions are interacting to form intramolecular RNA structure (Fig. 1). Similar to 3C methods for DNA conformation, RPL uses digestion and re-ligation of RNA, but omits cross-linking, relying instead on the inherent spatial proximity of RNA nucleobases in secondary structural features (i.e., stem loops). To generate RPL libraries, we performed RNase digestion *in situ* (or, for yeast, took advantage of endogenous single-stranded RNases), followed by treatment with exogenous T4 RNA ligase I under nondenaturing conditions. These steps resulted in chimeric molecules formed from RNA strands intramolecularly ligating across digested loops (Fig. 1a, inset). By deeply sequencing these resulting fragments and quantifying the relative abundance of specific intramolecular ligation junctions, we were able to create pairwise contact maps that reflect the short- and long-range stem-loop and pseudoknot interactions of intramolecular RNA secondary structures.

First, we tested RPL in the budding yeast *S. cerevisiae*. To create libraries, we prepared spheroplasts from whole yeast cells. We then treated the resulting slurries with T4 polynucleotide kinase (PNK), to convert 5'-hydroxyl to 5'-phosphate termini, and diluted and incubated these mixtures overnight in the presence of a single-stranded RNA ligase (T4 RNA ligase I) under nondenaturing conditions. We then purified total RNA using acid guanidinium-phenol and carried out a standard RNA-seq library preparation. Sequencing (Illumina) yielded 304 million concatenated reads for a (+) ligase sample, and 342 million concatenated reads for a (–) ligase control sample (Online Methods).

To identify candidate ligation junctions in these sequencing reads, we adapted an algorithm for identifying novel RNA isoforms from RNA-seq data<sup>23</sup>, relaxing constraints on splice-site composition to more generally recognize intramolecular chimeric reads that map discontinuously to a single RNA sequence. To quantify the enrichment of

Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.S. (shendure@uw.edu).

Received 8 April; accepted 15 June; published online 3 August 2015; doi:10.1038/nbt.3289



**Figure 1** RNA proximity ligation identifies structurally proximate regions within the complex secondary structures of *S. cerevisiae* ribosomal RNAs. (a) A schematic representation of the RPL method. Spheroplasts are obtained from whole cells and RNA is allowed to react with endogenous RNases. RNA ends are repaired *in situ* via T4 PNK to yield 5'-phosphate termini. Complexes are ligated overnight in the presence of T4 RNA Ligase I. Ligation products are cleaned up via acid guanidinium-phenol and subsequent DNase treatment, and subjected to Illumina TruSeq RNA-seq library preparation. These libraries are sequenced to map and count ligation junctions. (b,c) Distribution of ligation junctions as a function of distance from known base-pair partners in the 25S/5.8S rRNA (b) and 18S rRNAs (c). Ligation products capture the structural proximity implied by base-pairing relationships, as evidenced by the enrichment for ligation junctions immediately near paired bases. (d) Contact probability map for the eukaryotic 5.8S/25S rRNA based on RPL scores, which are calculated from the frequencies of ligation events between pairs of 21-nt windows. Lower inset: ligation events, shown for bases 1,300 to 1,475 of the LSU rRNA in red, primarily occur across digested single-stranded loops. Plotted here are the 8,463 ligation events where both nucleotides fall within the displayed domain. Right inset: RPL scores localize known pseudoknots in the LSU rRNA structure, such as the interaction between bases 1,727–1,812 (red) and bases 1,941–2,038 (blue).

candidate ligations in our samples, we first examined the distribution of spanned distances of intramolecular chimeric reads (i.e., gap sizes), per million reads, in both (+) and (–) ligase samples. Although the overall fraction of reads corresponding to candidate intramolecular ligation junctions is low, the (+) ligase sample is enriched for these across a broad range of spanned distances (0.28% in (+) ligase sample vs. 0.011% in (–) ligase sample; **Supplementary Fig. 1**).

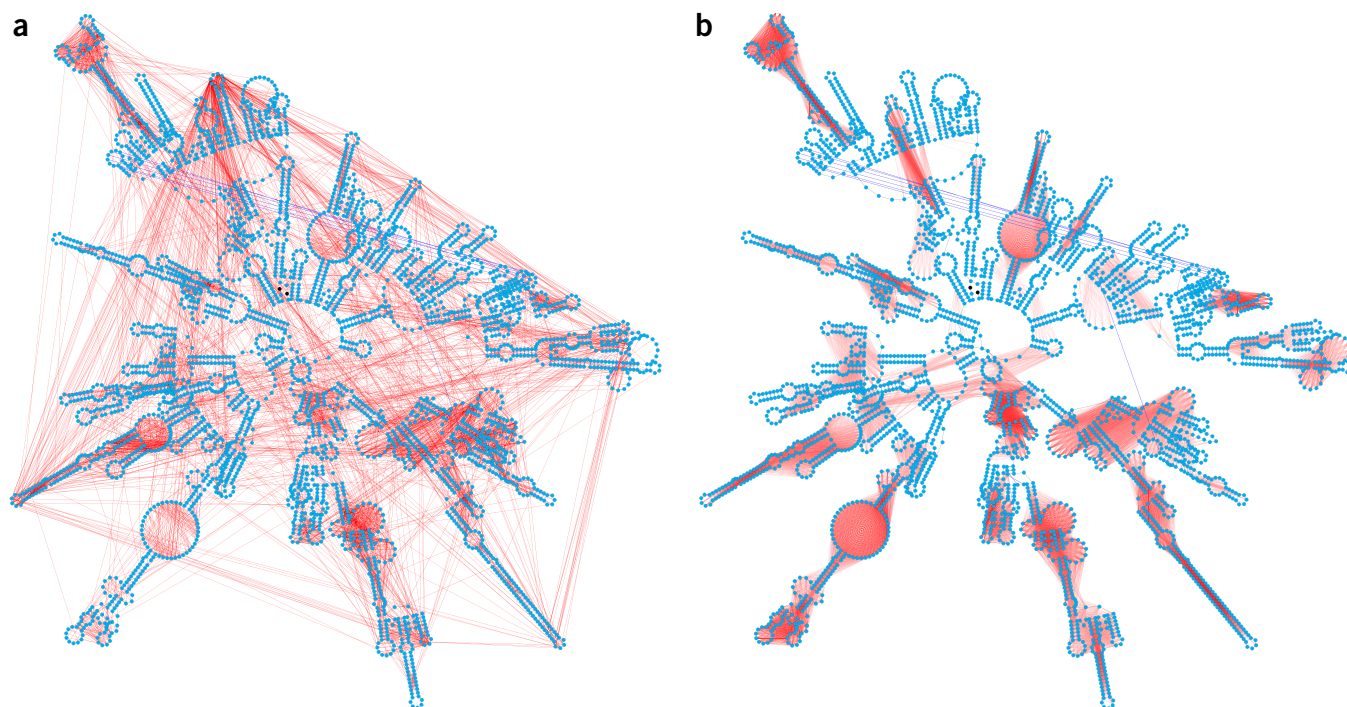
Potential sources of technical artifacts in these data include the formation of chimeric molecules by reverse transcriptase template switching, systematic mapping artifacts, PCR-mediated duplicates and nonspecific ligation events. To reduce the impact of reverse transcriptase template switching, we discarded candidate ligation junctions with >5 nucleotides (nt) microhomology, as well as those mapping to opposite strands. To remove PCR-mediated duplicates, we collapsed all reads with identical mapping coordinates and CIGAR alignment strings. To reduce the impact of systematic mapping artifacts caused by errors within our reference transcriptome (e.g., gross deletions, unannotated splice junctions), we conservatively discarded candidate ligation junctions containing the highest 1% of ligation counts, for each RNA species analyzed. Finally, to quantify the extent of non-specific ligation, we performed an experiment in duplicate, wherein human cells were taken through a modified version of the RPL protocol and spiked into yeast slurries immediately before proximity ligation. The resulting data demonstrate marked enrichment for intraspecies, intramolecular chimeric reads (**Supplementary Fig. 2**).

We first analyzed RPL data in the context of the complex but extensively validated secondary structures of the yeast ribosomal RNAs (rRNAs). The yeast ribosome comprises the 60S large subunit (LSU), which includes the 3.4 kb 25S rRNA and short 5.8S and 5S rRNAs, and the 40S small subunit (SSU), which includes the 1.8 kb 18S rRNA. To assess whether RPL captures the proximity implied by secondary structure base-pairing, we tallied candidate ligation

junctions in a 500 base-pair window centered on known base pairs of the established rRNA structures, effectively quantifying ligation probability as a function of distance (in linear sequence) from known base pairs (in secondary structure). We observed an enrichment of candidate ligation junctions immediately proximal (i.e., within 10 nt) to known base pairs in both the 5.8S/25S rRNAs (~9-fold; **Fig. 1b**) and 18S rRNA (~6-fold; **Fig. 1c**). Furthermore, in the case of the 5.8S/25S rRNAs, which contain many long-range, base-pairing interactions, this enrichment was maintained even if we restricted analysis to candidate ligation junctions that spanned >100 bases in the linear sequence (**Supplementary Fig. 3**).

The observed signal was entirely dependent on the inclusion of ligase and was not explained by sequencing errors, mapping artifacts or by proximity in sequence space (as opposed to structure space). We therefore conclude that it derived primarily from intramolecular ligation events between structurally proximal bases. Nonetheless, this signal is 'noise-averaged' over all base pairs in these rRNA structures (**Fig. 1b,c**). Consistent with the stochastic nature of individual ligation events, we observed weaker enrichment when repeating our analysis with a randomly selected subset of 10, 25 or 50 paired bases in either the LSU or SSU rRNAs (**Supplementary Fig. 4**). The ligation junctions that we observed were also clearly affected by other biases, including the bias against G/C extremes routinely seen with Illumina sequencing, as well as more subtle base-composition preferences at the ligation junction (**Supplementary Fig. 5**). We also observed that ligation junctions were enriched for single-stranded bases in the LSU and SSU rRNAs (odds ratio = 2.24;  $P < 2.2 \times 10^{-16}$ , Fisher's exact test). This bias, and the noisiness of the raw data, was evident when ligation junctions were overlaid onto a known secondary structure (**Fig. 2a**).

Given these observations, we concluded that the signal of RPL likely arises from the combinatorial digestion and ligation of predominantly unpaired ribonucleotides across broken loop structures. Considering



**Figure 2** Smoothing of ligation junction data results in ligase-dependent signal around known stem-loop formations. **(a)** The 10,000 most abundant ligation pairs for the LSU rRNA (red) overlaid onto the known secondary structure (blue). Although signal across stem-loops is evident, there is considerable noise. **(b)** Top 25,000 interacting windows based on RPL scores, which are calculated from the frequencies of ligation between pairs of 21-nt windows, for the LSU rRNA in the (+) ligase sample (red), again overlaid onto the known secondary structure (blue). Lines are drawn between the central bases of two interacting 21-nt windows. For **b**, the shading of the red lines is proportional to the ligation frequency.

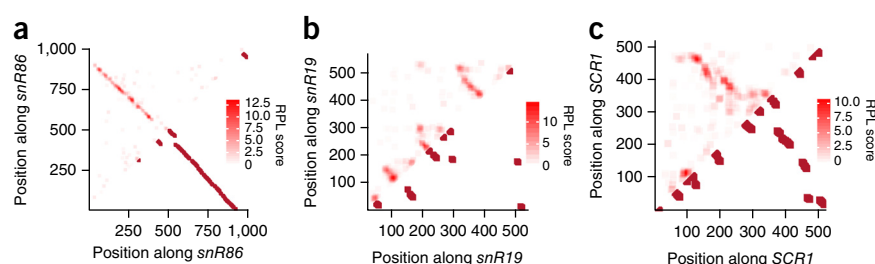
this, along with the stochastic, biased nature of individual ligation events, we speculated that our ability to resolve secondary structure would improve by calculating the frequency of ligation events between pairs of sliding windows (21 nt each), effectively capturing a combinatorial diversity of ligation events surrounding secondary structural elements. Concurrent with this, we adapted normalization methods developed for Hi-C matrices<sup>24</sup> to account for other one-dimensional biases (e.g., sequence biases of RNA ligase and PCR). We then visualized these normalized RPL scores, calculated for pairwise windows, by directly overlaying them onto known secondary structures. RPL scores broadly mirrored the secondary structures of the 5.8S/25S LSU rRNAs (Figs. 1d, 2b and Supplementary Fig. 6a) as well as the SSU 18S rRNA (Supplementary Fig. 6b). Furthermore, we observed signal corresponding to distal tertiary structures, including long-range pseudoknots in the LSU rRNAs (Fig. 1d, right inset)<sup>25</sup>.

We next sought to evaluate the correspondence between proximity ligation events and the structures of nonribosomal RNA transcripts. Because we were limited by sampling depth, we focused on

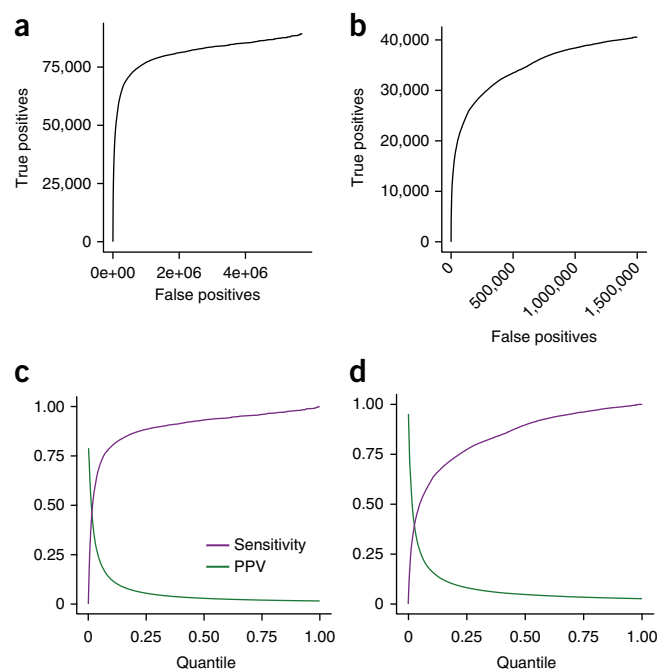
well-characterized, abundant RNAs; specifically, the snoRNA *snR86* (Fig. 3a), which guides uridylation of the LSU rRNA, the U1 spliceosomal RNA (*snR19*) (Fig. 3b), the RNA component of the signal recognition particle (*SCR1*) (Fig. 3c) and the U2 spliceosomal RNA homolog (*LSR1*) (Supplementary Fig. 7). In ‘contact probability maps’ for these RNAs (based on the normalized RPL scores described above), we observed a striking anti-diagonal pattern, reminiscent of signal observed at known stems in the 5.8S/25S and 18S rRNAs. When comparing our contact probability maps to secondary structure predictions generated with INFERNA<sup>26</sup>, using covariance models taken from Rfam<sup>27</sup>, our observations were consistent with conserved stems in both *snR86* and *snR19* (Fig. 3a,b). In RPL measurements for *snR19*, we also observed signal indicative of stem formation in the region comprising bases 320 to 510—minimum free energy (MFE) predictions suggest that this region can form a helix, raising the possibility that this structure is present endogenously.

We also analyzed RPL measurements in the context of a non-ribosomal RNA with a solved structure, the RNA subunit of the signal

**Figure 3** 2D RPL contact probability maps recapitulate known and predicted nonribosomal RNA structures. **(a)** Contact probability map for *snR86* mirrored against interacting windows containing paired bases, based on conserved secondary structure. **(b)** Contact probability map for *snR19* mirrored against interacting windows containing paired bases, based on conserved secondary structure. RPL signal indicating the formation of a stem-loop in bases 320–510 within the molecule is supported by MFE predictions, but not conservation. **(c)** Contact probability map for *SCR1* mirrored against interacting windows containing paired bases, based on the known structure of *SCR1*. For all analyses shown here, RPL scores were calculated using a window size of 21 nt.



**Figure 4** RPL scores demonstrate modest positive predictive value for pairs of interacting windows in RNA secondary structure. **(a,b)** Plots of number of true-positive interacting windows versus number of false-positive interacting windows for the 5.8S/25S rRNAs **(a)** and 18S rRNA **(b)**, at various quantile thresholds on RPL scores. This analysis shows that RPL scores have predictive value in classifying interacting regions containing at least one set of paired bases within RNA secondary structure. **(c,d)** Plots of the positive predictive value (green) and sensitivity (purple) of RPL-based classification of interacting regions, as a function of quantile threshold used for 5.8S/25S **(c)** and 18S rRNAs **(d)**. The quantile step size used for all analyses was 0.001.



recognition particle (*SCR1*). Again, we observed broad agreement between RPL scores and regions containing paired bases (Fig. 3c), though we did find that certain expected long-range interactions (e.g., folding between the molecule termini) were not seen. Further work will be needed to determine whether this was simply an artifact of insufficient depth of coverage or was symptomatic of some other bias with respect to the classes of structural elements that proximity ligation can resolve.

Finally, our observations for *LSR1* (Supplementary Fig. 7) were consistent with previous work employing cross-linking, affinity purification and proximity ligation of RNA<sup>21</sup>, which found ligation products supporting stem formation between the two termini. In agreement with this cross-linking-based approach, our data support the formation of both proximal (e.g., stem formation at bases 1,100–1,150) and distal folds.

We next explored the value of RPL scores as a predictive tool for classifying pairs of interacting regions within a structured RNA. To show that RPL scores can be used in this manner, we examined their positive predictive value (PPV) at varying quantile thresholds for the gold-standard 5.8S/25S and 18S rRNAs (Fig. 4a,b). This is a challenging classification problem (92,392 true-positive interacting windows out of 6,317,235 possible interacting windows for the LSU rRNAs (1.5%); 41,981 true-positive interacting windows out of 1,620,900 possible interacting windows for the SSU rRNA (2.6%)). The highest RPL scores were strongly enriched for true-positive interacting windows (LSU rRNA: PPV of 54% using the top 1% of RPL scores; SSU rRNA: PPV of 61% using the top 1% of RPL scores). Plotting PPV as a function of threshold illustrates the tradeoff with sensitivity (Fig. 4c,d). For example, at a sensitivity of 50%, RPL scores had a PPV of 43% for the LSU rRNA and 27% for the SSU rRNA, for predicting structurally interacting pairs of regions.

The high-throughput, unbiased identification of intermolecular RNA-RNA interactions is of strong interest in the RNA biology field. Recent work has shown that psoralen-mediated cross-linking may be used in tandem with antisense purification to capture *trans* RNA-RNA interactions<sup>28</sup>. In principle, RPL should be able to provide complementary information, as interacting RNAs may form ligation products at a higher rate than noninteracting RNAs. Although we observed a modest enrichment for intermolecular yeast ligation junctions in the species mixing experiment (Supplementary Fig. 2), this enrichment in our yeast RPL experiment derived primarily from ligation products between the small and large ribosomal subunits (Supplementary Fig. 8). Although no intersubunit RPL scores approached those of strongly interacting intramolecular windows, it remains possible that a combination of methodological improvements to reduce background and deeper sequencing of RPL libraries may enable global surveys of *trans* RNA-RNA interactions (e.g., the signal recognition particle–ribosome interaction; subunit interactions in the translating ribosome).

We next sought to adapt RPL to generate secondary structure information corresponding to RNAs in human cells. Most notably,

we replaced the zymolyase treatment with a limited *in situ* digestion with exogenous single-stranded RNases A and T1. In analyzing the resulting data in the context of the well-studied human ribosomal RNAs, we again observed correlation of high RPL scores with known interacting regions (Supplementary Fig. 9). However, an (–) RNase, (–) ligase control also demonstrated signal that correlated with secondary structure, albeit much more weakly and possibly reflecting endogenous nuclease and ligase activity (Supplementary Fig. 10). The possibility that endogenous enzymatic activity may contribute to the formation of chimeric RNAs is not novel; recent work using a cross-linking approach to characterize the miRNA interactome of *Caenorhabditis elegans* curiously found that expected ligation products could form in the absence of exogenous T4 RNA ligase I<sup>29</sup>.

We anticipate several directions for improving RPL. First, RPL libraries require deep sequencing to reliably map interacting regions, even for highly abundant RNA species. The sufficient sampling of lower-abundance RNA species of interest (e.g., mRNAs) might be achieved by optimizing the enzymatic steps of the protocol, by adopting hybrid capture enrichment or subtraction, or simply by brute force deep sequencing.

Second, given the high predictive value<sup>9,15,16,30</sup> of *in vivo* structure-probing methods (e.g., DMS-seq, SHAPE-seq) in determining the pairedness of individual bases in secondary structures, a framework that integrates two-dimensional, lower-resolution RPL data with one-dimensional, higher-resolution, structure-probing data seems highly attractive. Ideally, computational predictions would be integrated at the same time, thereby taking advantage of three largely orthogonal approaches to maximize the accuracy of RNA structural predictions.

The current repertoire of high-throughput empirical assays for RNA secondary structure provides us with a deep, but ultimately one-dimensional window into the structural landscape of RNA molecules. In contrast, RPL globally captures information with respect to pairwise interactions within RNA secondary structures. Through its integration with complementary computational and experimental approaches, we anticipate that RPL will facilitate the high-throughput elucidation of RNA secondary structures in diverse organisms.

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

**Accession codes.** GEO: [GSE69472](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank members of the Shendure laboratory (particularly D. Cusanovich, M. Kircher, A. McKenna and M. Snyder), D. Fowler, C. Trapnell and J. Underwood for helpful discussions and comments on the manuscript. We thank G. Kudla, A. Helwak and D. Tollervy for answering questions pertaining to the CLASH protocol. We would also like to acknowledge A. Dobin for making auxiliary scripts for processing STAR alignments publicly available. This work was funded by National Institutes of Health (NIH) Director's Pioneer Award (1DP1HG007811 to J.S.) and an NIH National Human Genome Research Institute (NHGRI) Genome Training Grant (5T32HG000035 to V.R.).

## AUTHOR CONTRIBUTIONS

V.R. and J.S. conceived of the project and devised experiments. V.R. and R.Q. carried out the experiments. V.R. performed computational analyses. V.R. and J.S. wrote the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Mortimer, S.A., Kidwell, M.A. & Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469–479 (2014).
- Cate, J.H. *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273**, 1678–1685 (1996).
- Wang, Y.-H., Murphy, F.L., Cech, T.R. & Griffith, J.D. Visualization of a tertiary structural domain of the tetrahymena group I intron by electron microscopy. *J. Mol. Biol.* **236**, 64–71 (1994).
- Latham, M.P., Brown, D.J., McCallum, S.A. & Pardi, A. NMR methods for studying the structure and dynamics of RNA. *ChemBioChem* **6**, 1492–1505 (2005).
- Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
- Reuter, J. & Mathews, D. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
- Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
- Shendure, J. & Aiden, E.L. The expanding scope of DNA sequencing. *Nat. Biotechnol.* **30**, 1084–1094 (2012).
- Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
- Ding, Y. *et al.* *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
- Lucks, J.B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. USA* **108**, 11063–11068 (2011).
- Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
- Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
- Underwood, J.G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7**, 995–1001 (2010).
- Kladwang, W., VanLang, C.C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* **3**, 954–962 (2011).
- Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **9**, 959–965 (2014).
- Fredriksson, S. *et al.* Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.* **20**, 473–477 (2002).
- Söderberg, O. *et al.* Direct observation of individual endogenous protein complexes *in situ* by proximity ligation. *Nat. Methods* **3**, 995–1000 (2006).
- Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
- Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- Kudla, G., Granneman, S., Hahn, D., Beggs, J.D. & Tollervy, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc. Natl. Acad. Sci. USA* **108**, 10010–10015 (2011).
- Helwak, A., Kudla, G., Dudnakova, T. & Tollervy, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654–665 (2013).
- Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Rao, S.S.P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529 (2011).
- Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
- Burge, S.W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **D226–D232** (2013).
- Engreitz, J.M. *et al.* RNA–RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
- Grosswendt, S. *et al.* Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell* **54**, 1042–1054 (2014).
- Cordero, P., Lucks, J.B. & Das, R. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics* **28**, 3006–3008 (2012).



## ONLINE METHODS

**Cell culture.** *S. cerevisiae* strain FY3 was struck out on YPD plates and grown at 30 °C. Mammalian cells (lymphoblastoid cell line GM12878; Coriell) were cultured at 37 °C, 5% CO<sub>2</sub> in RPMI-1640 supplemented with 1× Anti-Anti (Gibco), 1× Plasmocin, a mycoplasma prophylactic, (Invivogen) and 15% FBS (Gibco).

**RPL.** Individual yeast colonies were added directly to 0.5 U zymolyase in 10 µl 1× phosphate buffered saline (PBS) (Gibco) with 0.2% IGEPAL (Sigma) and incubated at 37 °C for 60 min to spheroplast while maintaining endogenous RNase activity. Spheroplasted yeast were immediately transferred to ice, and mixed with 0.5 µl SuperASE-In (Ambion), 2.5 µl T4 PNK (New England Biolabs), 5 µl 10× T4 DNA Ligase Buffer with 10 mM ATP (NEB), and 32 µl 1× PBS with 0.2% IGEPAL, after which the slurry was incubated at 37 °C for 30 min. Following end-repair, complexes were immediately transferred to 450 µl ligation reaction mix (50 µl 10× T4 DNA Ligase Buffer with 10 mM ATP (NEB); 5 µl SuperASE-In (Ambion), 12.5 µl T4 RNA Ligase I (NEB), 382.5 µl 1× PBS with 0.2% IGEPAL), and incubated overnight in a 16 °C water bath, after which complexes were added to 1.5 ml TriZOL (Ambion). Samples were then purified using Direct-ZOL spin columns (Zymo) according to manufacturer's protocols. For mammalian experiments a modified version of RPL was performed wherein 2E6 whole human lymphoblastoid cells (GM12878, Coriell) were treated *in situ* with 0.2 µl of RNase-IT (Agilent) diluted in 9.8 µl 1× PBS with 0.2% IGEPAL for 10 min at 22 °C, after which the RPL protocol was followed, beginning with PNK treatment.

T4 PNK is known to have minimal 3' phosphatase activity under the buffer conditions we use during our end-repair step<sup>31</sup>. To ensure that phosphatase activity was not limiting ligation efficiency, we also repeated our yeast RPL experiments using a low pH imidazole buffer (50 mM imidazole-HCl, pH 6.0, 10 mM MgCl<sub>2</sub>, 1 mM ATP, and 10 mM DTT) for our PNK reactions. We observed comparable ligation efficiencies independent of the use of low pH buffer (0.28% of analyzed reads in our sample compared to 0.21% and 0.14% in imidazole experiments performed in duplicate).

For spike-in experiments, an individual yeast colony and 5E5 human lymphoblastoid cells were treated with respective RPL treatments described above. Following PNK treatment, the two slurries were mixed and treated with T4 RNA Ligase I overnight, after which complexes were purified as described above.

To quantify the extent of RNA degradation during the yeast RPL protocol, we repeated the yeast RPL experiment, isolating RNA after PNK treatment, as well as after overnight incubations both in the presence and absence of T4 RNA Ligase I. We then analyzed the integrity of these RNA products using an RNA 6000 Nano Lab-on-Chip (Agilent), finding our products were mildly degraded following PNK treatment (RIN Score of ~7), though this degradation appears to have been halted before ligation (**Supplementary Fig. 11**).

**Library preparation.** Libraries were prepared according to standard Illumina TruSeq RNA guidelines, with minor changes. Notably, polyA-selection steps were skipped, RNA fragmentation (Elute, Prime, Fragment) was carried out for 2.5 min, and PCR amplification of the final library was carried out using qPCR for 8–12 cycles on a BioRad OpticonMini to prevent library overamplification. Two biological replicate libraries were generated and sequenced for (+) ligase yeast experiments, one of which was selected for deep sequencing and analyzed further in this paper. Two biological replicate libraries each were generated for imidazole and species-mixing experiments, for both (+) and (–) ligase samples.

**Sequencing and sequence alignment.** Sequencing of libraries was carried out using the Illumina MiSeq, NextSeq 500 and HiSeq 2000 instruments, generating paired-end 80 bp and 101 bp reads. All raw sequencing data and processed data files are accessible at GEO Accession [GSE69472](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69472).

**FASTQ Post-processing.** Raw paired-end FASTQ files were adaptor-trimmed and merged with SeqPrep (<https://github.com/jstjohn/SeqPrep>) to account for all read pairs that contained redundant information (i.e., sequence) content. We then took the resulting “singleton” forward and reverse reads (i.e., those that did not contain sufficient overlap to be fused) and concatenated them along with fused reads to yield 304 million (for the treated sample) and 342 million (for the negative control) concatenated reads, which were then analyzed.

**Alignment.** These resulting FASTQ files were aligned to references generated from either a manually curated list of yeast transcripts with duplicated transcripts removed, taken from the Saccharomyces Genome Database (<http://yeastgenome.org/>), or a selected list of deduplicated RefSeq human transcripts, using the STAR aligner with the following parameters:

```

-outSJfilterOverhangMin 6 6 6 6
-outSJfilterCountTotalMin 1 1 1 1
-outSJfilterDistToOtherSJmin 0 0 0 0
-alignIntronMin 10
-chimSegmentMin 15
-chimScoreJunctionNonGTAG 0
-chimJunctionOverhangMin 6

```

**Bioinformatic analyses.** Secondary structures in BPSEQ format for *S. cerevisiae* were downloaded from the Comparative RNA Website<sup>32</sup> and RNA structures were visualized through a modified version of VARNA. *Homo sapiens* rRNA structures were inferred from a published cryo-EM structure<sup>33</sup>, using 3DNA<sup>34</sup>. STAR-generated output was analyzed with custom Python and R scripts to generate contact probability maps (all custom scripts used to analyze aligned data are provided in **Supplementary Scripts**). First, STAR alignments were deduplicated by collapsing all alignments with identical start coordinates and CIGAR strings. These deduplicated alignments were then converted to “splice junction” and “chimer” files using awk, and ligation junctions were parsed from these files. For specific species of interest, these ligation counts were then filtered further to remove the highest 1% of counts between individual pairs of bases. To calculate the distribution of ligations around known base-pairs, we looked at all pairs of bases (*i, j*) in our secondary structure BPSEQ files and calculated the abundance of ligation events between (*i, j* – 250) to (*i, j* + 250) for each base. For subsampling experiments, we randomly sampled 10, 25 or 50 paired-bases and repeated these calculations.

To compute RPL scores, which measure the extent of ligation between two regions of a molecule, we first considered the sparse matrix *M* where *M<sub>ij</sub>* is the ligation count between base *i* and base *j*. To generate the RPL score matrix *M\**, we computed the coverage at each base *i* and *j* (*c<sub>i</sub>*; *c<sub>j</sub>*) and generated a normalized matrix *M<sub>norm</sub>* such that

$$M_{ij}^{norm} = \frac{M_{ij}}{\sqrt{c_i c_j}}$$

We then used this normalized matrix to generate *M\** by binning all normalized scores

$$M_{ij}^* = \sum_{a=i-10}^{i+10} \sum_{b=j-10}^{j+10} M_{ab}^{norm}$$

Classification analyses were performed as follows: we thresholded the RPL scores resulting from the above smoothing by quantiles, with a quantile step size of 0.001, and classified true positive interacting windows as those interacting 21 nt windows with RPL scores greater than our specified threshold, that also contain at least one set of paired bases.

To generate secondary structures for *snR86* and *snR19*, we downloaded covariance models from Rfam (*snR86* Accession: [RF01272](https://www.ebi.ac.uk/EMBL/GenBank/CCDS/entry/view/CCDS1272); *snR19* Accession: [RF00488](https://www.ebi.ac.uk/EMBL/GenBank/CCDS/entry/view/CCDS488)), aligned respective yeast sequences to their covariance models using the cmalign method from INFERNA v1.1.1 and converted the resulting Stockholm alignment files to BPSEQ format using VARNA.

Structures of the yeast ribosome (PDB Accession: [4V88](https://www.rcsb.org/structure/4V88)) were visualized using PyMOL (<http://www.pymol.org/>).

31. Cameron, V. & Uhlenbeck, O.C. 3'-Phosphatase activity in T4 polynucleotide kinase. *Biochemistry* **16**, 5120–5126 (1977).

32. Cannone, J. *et al.* The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2 (2002).

33. Anger, A.M. *et al.* Structures of the human and *Drosophila* 80S ribosome. *Nature* **497**, 80–85 (2013).

34. Lu, X. & Olson, W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).