Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

Nat Biotechnol. 2015 September ; 33(9): 980–984. doi:10.1038/nbt.3289.

High-throughput determination of RNA structure by proximity ligation

Vijay Ramani[1], **Ruolan Qiu**[1], and **Jay Shendure**[1]

[1] Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.

Abstract

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 [1]. Pioneering approaches for elucidating RNA structure—including crystallography[2], electron microscopy[3], and spectroscopy[4]—are technically complex and difficult to scale, motivating the development of computational algorithms for RNA structure prediction[5–7]. 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[8], less laborious experimental techniques have been developed for the global interrogation of RNA secondary structures. These include methods relying on structure-specific chemical modifications[9–11], such as DMS-seq and SHAPE-seq, as well as methods involving digestion with structure-specific RNases[12–14], 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, recent efforts have combined systematic mutagenesis and structure-specific probing to generate pairwise information for inferring

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RNA folds[15,16]. 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[17], to probe protein-protein interactions via antibody-bound oligonucleotides[18], and for targeted or global chromosome conformation capture (3C)[19,20]. Proximity ligation has also been applied in conjunction with crosslinking and either affinity purification or immunoprecipitation to characterize snoRNA-rRNA interactions[21] and Ago-mediated miRNA-target interactions[22]. However, these efforts have primarily focused on assessing specific *trans* interactions, rely on low-efficiency 254 nanometer UV crosslinking, 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 crosslinking, 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 non-denaturing conditions. These steps result in chimeric molecules formed from RNA strands intra-molecularly ligating across digested loops (**Fig. 1a**, inset). By deeply sequencing these resulting fragments and quantifying the relative abundance of specific intramolecular ligation junctions, we are 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 spheroplasted whole yeast cells for 1 h with zymolyase (dissolved in 1X PBS without DTT to allow endogenous RNases to remain active). 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 non-denaturing conditions. We then purified total RNA using acid guanidinium-phenol, and carried out a standard RNA-seq library preparation. Sequencing (Illumina) yielded 304 million (M) concatenated reads for a (+) ligase sample, and 342M concatenated reads for a (−) ligase control sample (**Methods**).

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 (RT) template switching, systematic mapping artifacts, PCR-mediated duplicates and non-specific ligation events. To reduce the impact of RT 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 (for example, gross deletions, un-annotated 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 nonspecific ligation, we performed an experiment in duplicate wherein human cells were taken through a modified version of the RPL protocol (**Methods**) 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 is comprised of 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 observe 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 is maintained even if we restrict analysis to candidate ligation junctions that span >100 bases in the linear sequence (**Supplementary Fig. 3**).

The observed signal is entirely dependent on the inclusion of ligase, and is not explained by sequencing errors, mapping artifacts or by proximity in sequence space (as opposed to structure space). As such, we conclude that it primarily derives from intramolecular ligation events between structurally proximal bases. Nonetheless, the signal shown in **Fig. 1b,c** is “noise-averaged” over all base pairs in these rRNA structures. Consistent with the stochastic nature of individual ligation events, we observe 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 observe are 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 observe that ligation junctions are enriched for single-stranded bases in the LSU and SSU rRNAs (Odds Ratio (OR) = 2.24; P < 2.2E-16,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Fisher's Exact Test). This bias, and the noisiness of the raw data, is evident when ligation junctions are 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 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[24] to account for other one-dimensional biases (for example, sequence biases of RNA ligase and PCR) (**Methods**). We then visualized these normalized RPL scores, calculated for pairwise windows, by directly overlaying them onto known secondary structures. RPL scores broadly mirror the secondary structures of the 5.8S/25S LSU rRNAs (**Fig. 1d**, **Fig. 2b**; **Supplementary Fig. 6a**) as well as the SSU 18S rRNA (**Supplementary Fig. 6b**). Furthermore, we observe signal corresponding to distal tertiary structures, including long-range “pseudo-knots” in the LSU rRNAs (**Fig. 1d**, right inset)[25].

We next sought to evaluate the correspondence between proximity ligation events and the structures of non-ribosomal RNA transcripts. Because we are 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 observe 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 INFERNAL[26] using covariance models taken from Rfam[27], our observations are 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 320-510—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 recognition particle (*SCR1*). Again, we observed broad agreement between RPL scores and regions containing paired bases (**Fig. 3c**), though we do find that certain expected long-range interactions (for example, folding between the molecule termini) are not seen. Further work will be needed to determine whether this is simply an artifact of insufficient depth-of-coverage, or is 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**) are consistent with previous work employing cross-linking, affinity-purification, and proximity ligation of RNA[21], which found ligation products supporting stem-formation between the two termini. In agreement

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with this cross-linking based approach, our data support the formation of both proximal (for example, stem formation at bases 1100 – 1150), 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 are 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 have 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 crosslinking may be used in tandem with anti-sense purification to capture *trans* RNA-RNA interactions[28]. In principle, RPL should be able to provide complementary information, as interacting RNAs may form ligation products at a higher rate than non-interacting 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 derives primarily from ligation products between the small and large ribosomal subunits (**Supplementary Fig. 8**). While no inter-subunit 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 (for example, 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 *C. elegans* curiously found that expected ligation products could form in the absence of exogenous T4 RNA Ligase I[29].

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 (for example, mRNAs)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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[9,15,16,30] of *in vivo* structure probing methods (for example, 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

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_2 in RPMI-1640 supplemented with 1X Anti-Anti (Gibco), 1X Plasmocin (Invivogen), and 15% FBS (Gibco).

RNA Proximity Ligation (RPL)

Individual yeast colonies were added directly to 0.5 U Zymolyase in 10 uL 1X phosphate buffered saline (PBS) (Gibco) w/ 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 uL SuperASE-In (Ambion), 2.5 uL T4 PNK (NEB), 5 uL 10X T4 DNA Ligase Buffer w/ 10 mM ATP (NEB), and 32 uL 1X PBS w/ 0.2% IGEPAL, after which the slurry was incubated at 37 °C for 30 min. Following end-repair, complexes were immediately transferred to 450 uL ligation reaction mix (50 uL 10X T4 DNA Ligase Buffer w/ 10 mM ATP (NEB); 5 uL SuperASE-In (Ambion), 12.5 uL T4 RNA Ligase I (NEB), 382.5 uL 1X PBS w/ 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 uL of RNace-IT (Agilent) diluted in 9.8 uL 1X PBS w/ 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[31]. To ensure that phosphatase activity was not limiting ligation efficiency, we also repeated our yeast RPL experiments using a low pH imidazole

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

buffer (50 mM imidazole-HCl, pH 6.0, 10 mM $MgCl_2$, 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.

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 contain 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 304M (for the treated sample) and 342M (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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

–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[32] and RNA structures were visualized through a modified version of VARNA. *H. sapiens* rRNA structures were inferred from a published cryo-EM structure[33], using 3DNA[34]. 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 sub-sampling 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_{ij} is the ligation count between base i and base j. To generate the RPL score matrix M^*, we compute the coverage at each base i and j (c_i; c_j) and generate a normalized matrix M_{norm} such that:

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

We then use 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 1 set of paired bases.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To generate secondary structures for *snR86* and *snR19*, we downloaded covariance models from Rfam (*snR86* Accession: RF01272; *snR19* Accession: RF00488), aligned respective yeast sequences to their covariance models using the cmalign method from INFERNAL v1.1.1, and converted the resulting Stockholm alignment files to BPSEQ format using VARNA.

Structures of the yeast ribosome (PDB Accession: 4V88) were visualized using PyMol (http://www.pymol.org/).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank members of the Shendure Lab (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. Tollervey 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 NIH Director's Pioneer Award (1DP1HG007811 to J.S.) and an NIH NGHRI Genome Training Grant (5T32HG000035 to V.R.).

References

1. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat. Rev. Genet. 2014; 15:469–479. CrossRef Medline. [PubMed: 24821474]
2. Cate JH, et al. Crystal Structure of a Group I Ribozyme Domain: Principles of RNA Packing. Science. 1996; 273:1678–1685. CrossRef Medline. [PubMed: 8781224]
3. Wang Y-H, Murphy FL, Cech TR, Griffith JD. Visualization of a Tertiary Structural Domain of the Tetrahymena Group I Intron by Electron Microscopy. J. Mol. Biol. 1994; 236:64–71. CrossRef Medline. [PubMed: 7508985]
4. Latham MP, Brown DJ, McCallum SA, Pardi A. NMR Methods for Studying the Structure and Dynamics of RNA. ChemBioChem. 2005; 6:1492–1505. CrossRef Medline. [PubMed: 16138301]
5. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31:3406–3415. CrossRef Medline. [PubMed: 12824337]
6. Reuter J, Mathews D. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010; 11:129. CrossRef Medline. [PubMed: 20230624]
7. Lorenz R, et al. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011; 6:26. CrossRef Medline. [PubMed: 22115189]
8. Shendure J, Aiden EL. The expanding scope of DNA sequencing. Nat. Biotechnol. 2012; 30:1084–1094. CrossRef Medline. [PubMed: 23138308]
9. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2014; 505:701–705. CrossRef Medline. [PubMed: 24336214]
10. Ding Y, et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014; 505:696–700. CrossRef Medline. [PubMed: 24270811]
11. Lucks JB, et al. Multiplexed RNA structure characterization with selective 2′- hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. USA. 2011; 108:11063–11068. CrossRef Medline. [PubMed: 21642531]
12. Kertesz M, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010; 467:103–107. CrossRef Medline. [PubMed: 20811459]
13. Wan Y, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014; 505:706–709. CrossRef Medline. [PubMed: 24476892]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

14. Underwood JG, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat. Methods. 2010; 7:995–1001. CrossRef Medline. [PubMed: 21057495]
15. Kladwang W, VanLang CC, Cordero P, Das R. A two-dimensional mutate- and-map strategy for non-coding RNA structure. Nat. Chem. 2011; 3:954–962. CrossRef Medline. [PubMed: 22109276]
16. Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Meth. 2014; 9:959–965.
17. Fredriksson S, et al. Protein detection using proximity-dependent DNA ligation assays. Nat. Biotechnol. 2002; 20:473–477. CrossRef Medline. [PubMed: 11981560]
18. Söderberg O, et al. Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nat. Methods. 2006; 3:995–1000. CrossRef Medline. [PubMed: 17072308]
19. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002; 295:1306–1311. CrossRef Medline. [PubMed: 11847345]
20. Lieberman-Aiden E, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009; 326:289–293. CrossRef Medline. [PubMed: 19815776]
21. Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. Proc. Natl. Acad. Sci. USA. 2011; 108:10010–10015. CrossRef Medline. [PubMed: 21610164]
22. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. Cell. 2013; 153:654–665. CrossRef Medline. [PubMed: 23622248]
23. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma. 2012 doi:10.1093/bioinformatics/bts635.
24. Rao SSP, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014; 159:1665–1680. Medline. [PubMed: 25497547]
25. Ben-Shem A, et al. The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. Science. 2011; 334:1524–1529. CrossRef Medline. [PubMed: 22096102]
26. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29:2933–2935. CrossRef Medline. [PubMed: 24008419]
27. Burge SW, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2013:D226–D232. Medline. [PubMed: 23125362]
28. Engreitz JM, et al. RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. Cell. 2014; 159:188–199. CrossRef Medline. [PubMed: 25259926]
29. Grosswendt S, et al. Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. Mol. Cell. 2014; 54:1042–1054. CrossRef Medline. [PubMed: 24857550]
30. Cordero P, Lucks JB, Das R. An RNA Mapping DataBase for curating RNA structure mapping experiments. Bioinformatics. 2012; 28:3006–3008. CrossRef Medline. [PubMed: 22976082]
31. Cameron V, Uhlenbeck OC. 3′-Phosphatase activity in T4 polynucleotide kinase. Biochemistry. 1977; 16:5120–5126. CrossRef Medline. [PubMed: 199248]
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. 2002; 3:2. CrossRef Medline. [PubMed: 11869452]
33. Anger AM, et al. Structures of the human and Drosophila 80S ribosome. Nature. 2013; 497:80–85. CrossRef Medline. [PubMed: 23636399]
34. Lu X, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003; 31:5108–5121. CrossRef Medline. [PubMed: 12930962]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

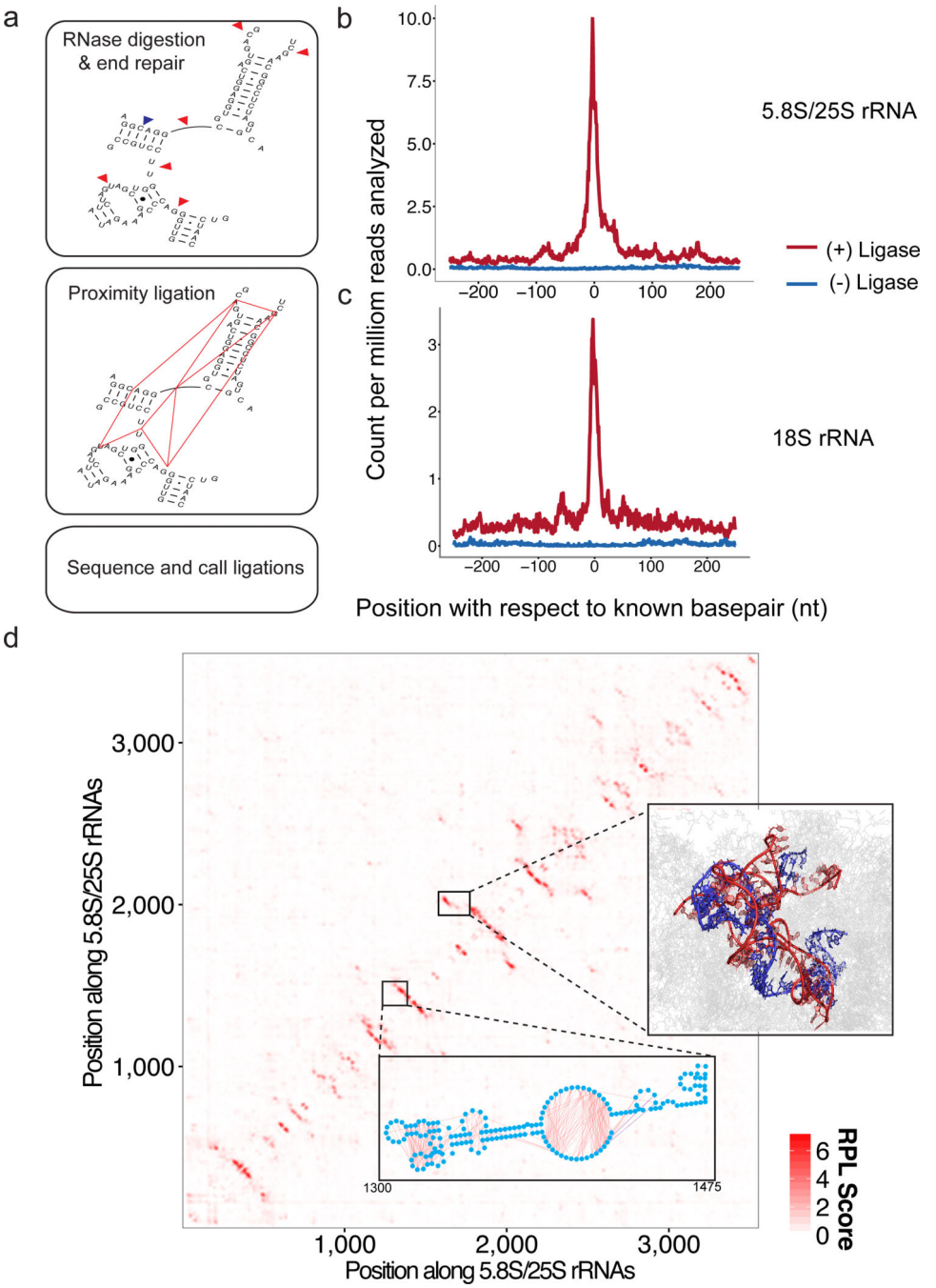

Figure 1. RNA Proximity Ligation identifies structurally proximate regions within the complex secondary structures of *S. cerevisiae* ribosomal RNAs. a.)

A schematic representation of the RPL method. Whole cells are spheroplasted with zymolyase 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.)** We examined the distribution of ligation junctions as a function of distance from known base-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pair partners in the 25S/5.8S rRNA and 18S rRNAs. Ligation products capture the structural proximity implied by base-pairing relationships, as evidenced by the enrichment for ligation junctions immediately near paired bases. Y-axes are shown as ligation counts per million reads analyzed. **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 (**Methods**). **Lower inset**: Ligation events, shown for bases 1300 to 1475 of the LSU rRNA in orange, primarily occur across digested single-stranded loops. RPL scores effectively smooth this noisy signal and are enriched for pairs of interacting regions. Plotted here are the 8,463 ligation events where both nucleotides fall within the displayed domain (compared to 17,029 ligation events where one nucleotide falls within the displayed domain and one does not, not shown). **Right inset:** RPL scores localize known pseudo-knots in the LSU rRNA structure, such as the interaction between bases 1727-1812 (shown in red) and bases 1941 – 2038 (shown in blue).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

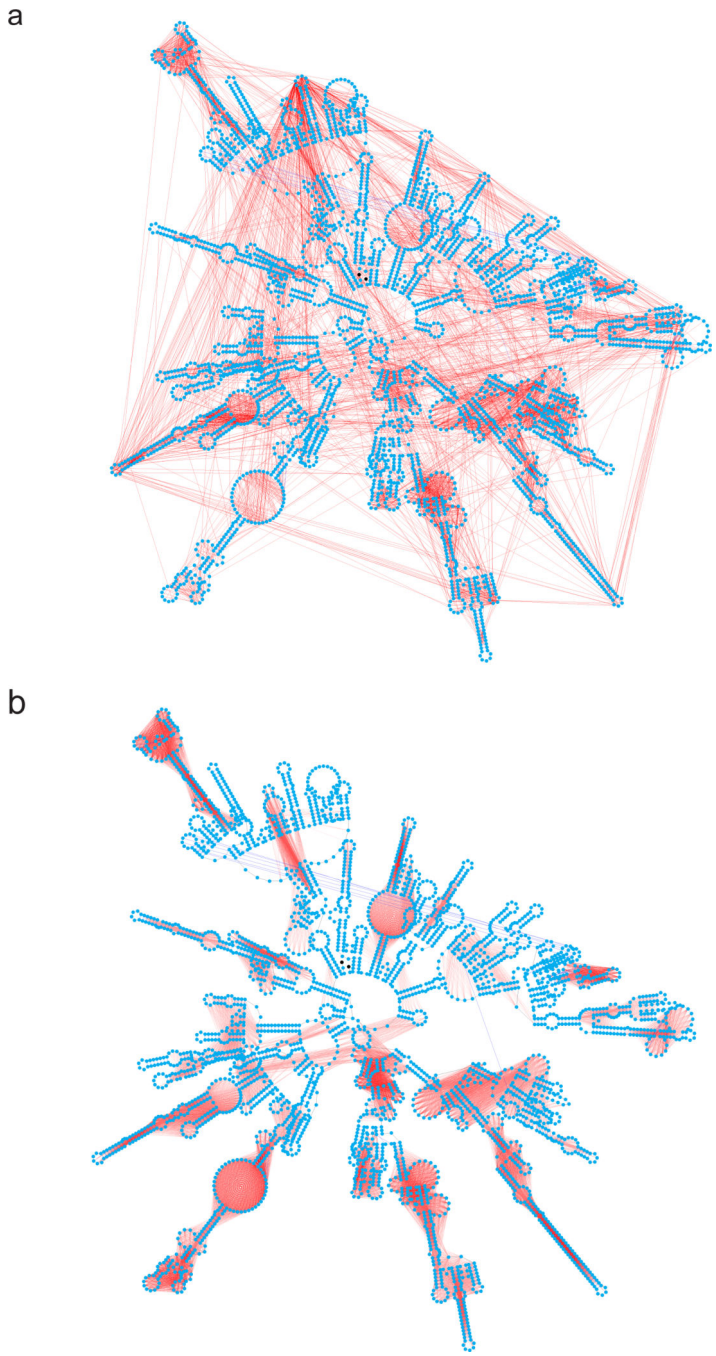

Figure 2.

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). While 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 (**Methods**), for the LSU rRNA in the (+) ligase sample (red), again overlaid onto the known

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

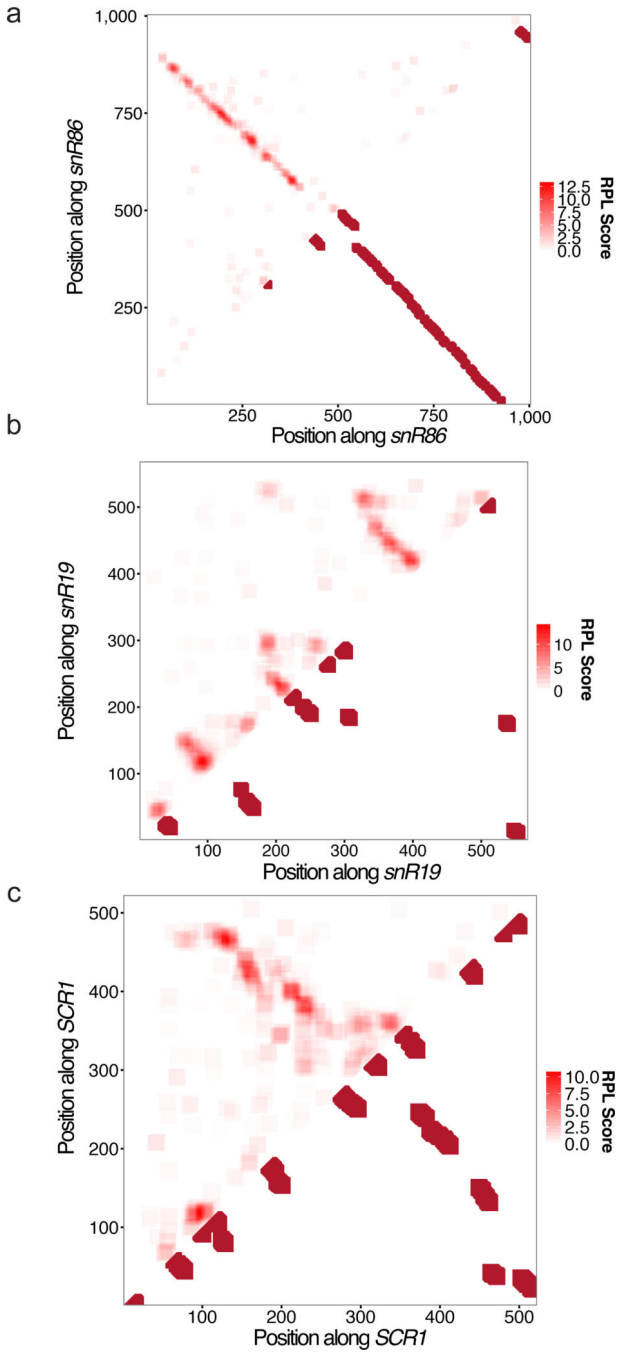

Figure 3.

2D RPL contact probability maps recapitulate known and predicted non-ribosomal 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bases, based on the known structure of *SCR1*. For all analyses shown here, RPL scores were calculating using a window size of 21 nt.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

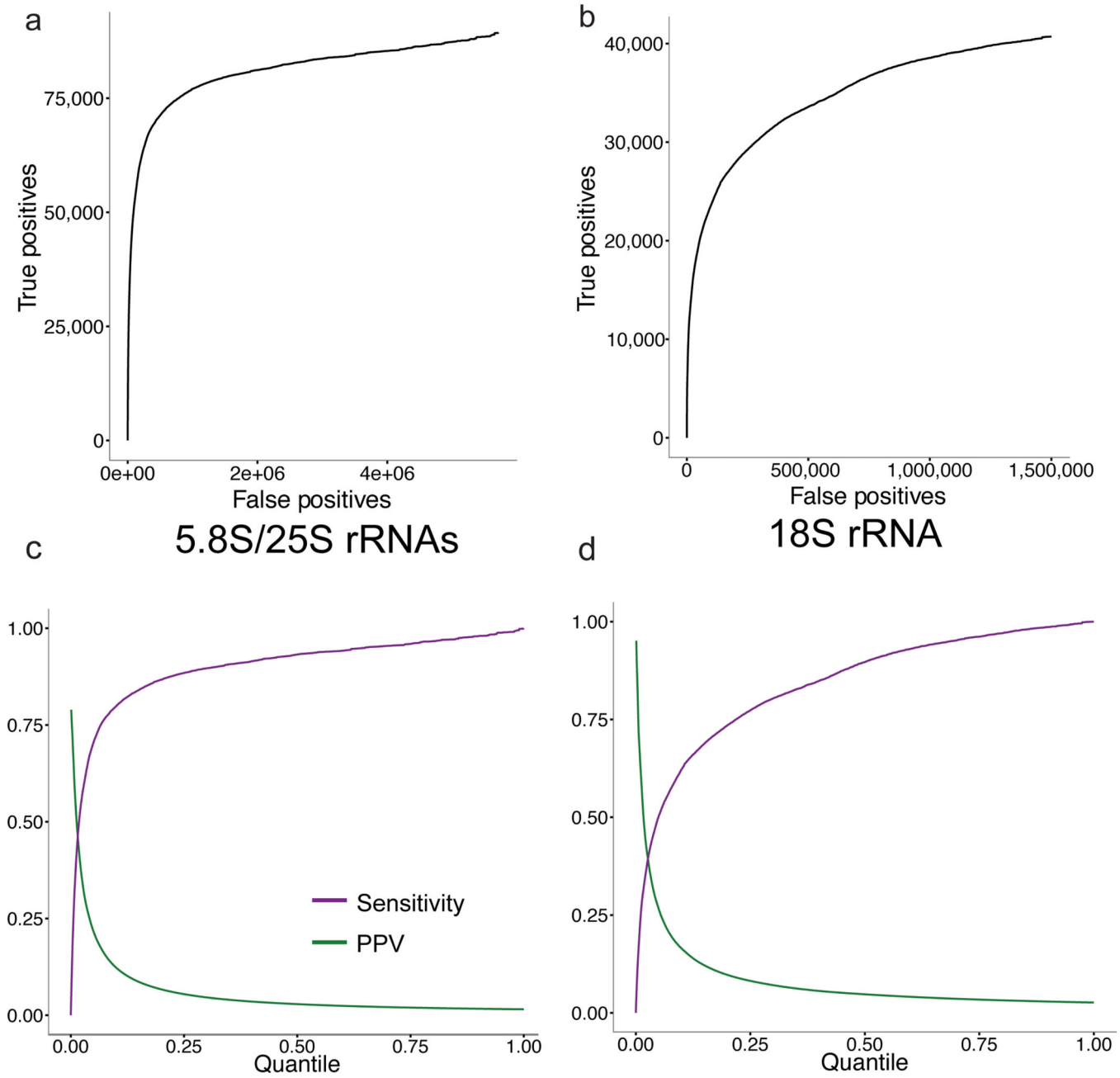

Figure 4. RPL scores demonstrate modest positive predictive value for pairs of interacting windows in RNA secondary structure. a-b.)

Plots of number of true positive interacting windows versus number of false positive interacting windows for the **(a)** 5.8SS/25S rRNAs and **(b)** 18S rRNA, 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 **(c)** 5.8S/25S and **(d)** 18S rRNAs. The quantile step size used for all analyses shown in this figure was 0.001.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript