



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Global mapping of RNA homodimers in living cells

Citation for published version:

Gabryelska, MM, Badrock, AP, Lau, JY, O'Keefe, RT, Crow, YJ & Kudla, G 2022, 'Global mapping of RNA homodimers in living cells', *Genome Research*, vol. 32, pp. 956-967. <https://doi.org/10.1101/gr.275900.121>

Digital Object Identifier (DOI):

[10.1101/gr.275900.121](https://doi.org/10.1101/gr.275900.121)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Genome Research

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Method

Global mapping of RNA homodimers in living cells

Marta M. Gabryelska,^{1,2} Andrew P. Badrock,³ Jian You Lau,¹ Raymond T. O’Keefe,⁴ Yanick J. Crow,³ and Grzegorz Kudla¹

¹MRC Human Genetics Unit, University of Edinburgh, Edinburgh, EH4 2XU, United Kingdom; ²Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Bedford Park, South Australia 5042, Australia; ³Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, United Kingdom; ⁴Division of Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, School of Biological Sciences, University of Manchester, Manchester, M13 9PL, United Kingdom

RNA homodimerization is important for various physiological processes, including the assembly of membraneless organelles, RNA subcellular localization, and packaging of viral genomes. However, understanding RNA dimerization has been hampered by the lack of systematic *in vivo* detection methods. Here, we show that CLASH, PARIS, and other RNA proximity ligation methods detect RNA homodimers transcriptome-wide as “overlapping” chimeric reads that contain more than one copy of the same sequence. Analyzing published proximity ligation data sets, we show that RNA:RNA homodimers mediated by direct base-pairing are rare across the human transcriptome, but highly enriched in specific transcripts, including *U8* snoRNA, *U2* snRNA, and a subset of tRNAs. Mutations in the homodimerization domain of *U8* snoRNA impede dimerization *in vitro* and disrupt zebrafish development *in vivo*, suggesting an evolutionarily conserved role of this domain. Analysis of virus-infected cells reveals homodimerization of SARS-CoV-2 and Zika genomes, mediated by specific palindromic sequences located within protein-coding regions of *N* gene in SARS-CoV-2 and *NS2A* gene in Zika. We speculate that regions of viral genomes involved in homodimerization may constitute effective targets for antiviral therapies.

[Supplemental material is available for this article.]

The biological functions of RNA molecules depend on their ability to form intra- and intermolecular interactions, often mediated by Watson-Crick base-pairing. Intramolecular base-pairing determines the structure and function of RNA, including rRNA and tRNA; it regulates viral replication; and it influences the efficiency of mRNA translation into proteins. Intermolecular RNA–RNA base-pairing underlies codon–anticodon recognition, splicing, and regulation of gene expression by miRNA and siRNAs. Intramolecular and intermolecular interactions are interdependent, and according to the competing endogenous RNA (ceRNA) hypothesis (Salmena et al. 2011; Gardiner et al. 2015), intermolecular RNA interactions have the potential to rewire regulatory networks and expand the information encoded in a genome.

An intermolecular interaction between two identical molecules is known as homodimerization. Although homodimers are common in proteins (Bergendahl and Marsh 2017), relatively few homodimers of RNA molecules have been described *in vivo* (for review, see Bou-Nader and Zhang 2020). Perhaps the best studied are dimers of the HIV genome, which are initiated by an interaction between two copies of the palindromic sequence known as DIS (Berkhout and van Wamel 1996). This interaction leads to the formation of an extended double helix that joins together two copies of the genome, launching a series of events that leads to the packaging of the pair of genomes into one capsid (Paillart et al. 2004). Homodimerization events have been described in retroviruses, hepatitis C virus, SARS coronavirus, and in bacteriophages (Clever et al. 2002; Shetty et al. 2010; Ishimaru et al. 2013; Dubois et al. 2018).

RNA oligomerization also plays a role in the process of phase separation, which leads to the formation of membraneless RNA-containing organelles, such as P-bodies, stress granules, nucleoli, Cajal bodies, and others (Jain and Vale 2017; Khong et al. 2017; Nguyen et al. 2018; Van Treec et al. 2018; Van Treec and Parker 2018). There is growing evidence that such granules are formed via transient protein–RNA and RNA–RNA interactions. As an example, homo- and heterodimerization of mRNA induces the formation of distinct types of phase-separated droplets in a filamentous fungus (Langdon et al. 2018). Homodimerization also influences the localization of *oskar* and *bicoid* mRNAs in *Drosophila* embryos (Ferrandon et al. 1997; Wagner et al. 2001, 2004; Jambor et al. 2011; Masliyah et al. 2013). Strong interaction between mRNA and pre-mRNA of *CUP1* gene leads to RNA miscompartmentalization and localization to cytoplasmic foci, possibly including P-bodies and stress granules (Qu et al. 2014).

An example of pathogenic homodimerization has been observed in a mutated variant of a human mitochondrial tRNA (Wittenhagen and Kelley 2002; Roy et al. 2005). Additionally, tRNA fragments (tRFs) were shown to form homodimers (Tosar et al. 2018) and tetramers (Lyons et al. 2017). CAG and other repeats underlying RNA expansion disorders form hairpin structures, with a stem composed of periodically occurring standard C-G and G-C base pairs (Ciesiolka et al. 2017). Repeat expansion, correlated with the severity of disorders, increases the possibility of homodimer formation. Sufficiently long trinucleotide repeats can form foci *in vivo* through phase separation (Jain and Vale 2017). Homodimers are also formed by various ribozymes and riboswitches (Bou-Nader and Zhang 2020). Dimerization of RNAs is used in nanobiotechnology for the design and

Corresponding author: gkudla@gmail.com

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.275900.121>. Freely available online through the *Genome Research* Open Access option.

© 2022 Gabryelska et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

construction of RNA architectures through controlled self-assembly of modular RNA units (tectoRNAs) (Chworos et al. 2004; Guo 2010; Ishikawa et al. 2013; Geary et al. 2014; Grabow and Jaeger 2014; Tanaka et al. 2016). These observations suggest that transient and stable RNA homodimers play a role in a variety of physiological and pathological processes.

The last few years have seen the development of RNA proximity ligation methods to map cellular RNA–RNA interactions (Kudla et al. 2020). CLASH (Kudla et al. 2011), miR-CLIP (Imig et al. 2015), and hiCLIP (Sugimoto et al. 2015) use a protein bait to detect protein-associated RNA duplexes, whereas PARIS (Lu et al. 2016), LIGR-seq (Sharma et al. 2016), SPLASH (Aw et al. 2016), and COMRADES (Ziv et al. 2018) use a small molecule, psoralen, to cross-link interacting RNA strands. Proximity ligation methods have been commonly used to identify heterotypic interactions, such as interactions between snoRNA, miRNA, piRNA, or sRNA, and their respective targets (Kudla et al. 2011; Helwak et al. 2013; Grosswendt et al. 2014; Ramani et al. 2015). However, these methods also uncover many homotypic interactions, in which the two partners can be mapped to the same gene. Homotypic interactions, usually assumed to originate from the same RNA molecule, have been used to reveal the secondary structures of cellular RNAs (Kudla et al. 2011; Aw et al. 2016; Lu et al. 2016; Sharma et al. 2016) and structural dynamics of viral genomes (Ziv et al. 2018, 2020; Huber et al. 2019; Zhang et al. 2021b). However, homotypic interactions could also in principle represent binding between a pair of identical molecules forming an RNA homodimer. Here, we aimed

to establish methods for the identification of homodimers in RNA proximity ligation data, benchmark experimental and computational protocols for mapping of homodimers, and globally profile the homodimerization of RNA in yeast and human cells and in Zika and SARS-CoV-2 viruses.

Results

Overlapping chimeras indicate intermolecular interactions

CLASH, PARIS, and other RNA proximity ligation methods rely on the ligation of interacting fragments of RNA, which are then detected as chimeric reads by high-throughput sequencing. We reasoned that chimeras that represent intra- and intermolecular interactions can be distinguished from each other by an analysis of sequence overlap between the arms of each chimera.

When an RNA molecule that comprises an intramolecular interaction is subjected to proximity ligation, the RNA is fragmented into smaller pieces, which (by definition) originate from distinct parts of the RNA. When these fragments are ligated, sequenced, and mapped back to the reference, they should never be mapped to the same region of the RNA. The possible arrangements of the two fragments on the source RNA are shown in Figure 1A. Of these arrangements, the ungapped 5'-3' chimera is indistinguishable from the source RNA sequence and cannot be identified as a chimera by a simple mapping approach. The other arrangements—gapped 5'-3', ungapped 3'-5', and gapped

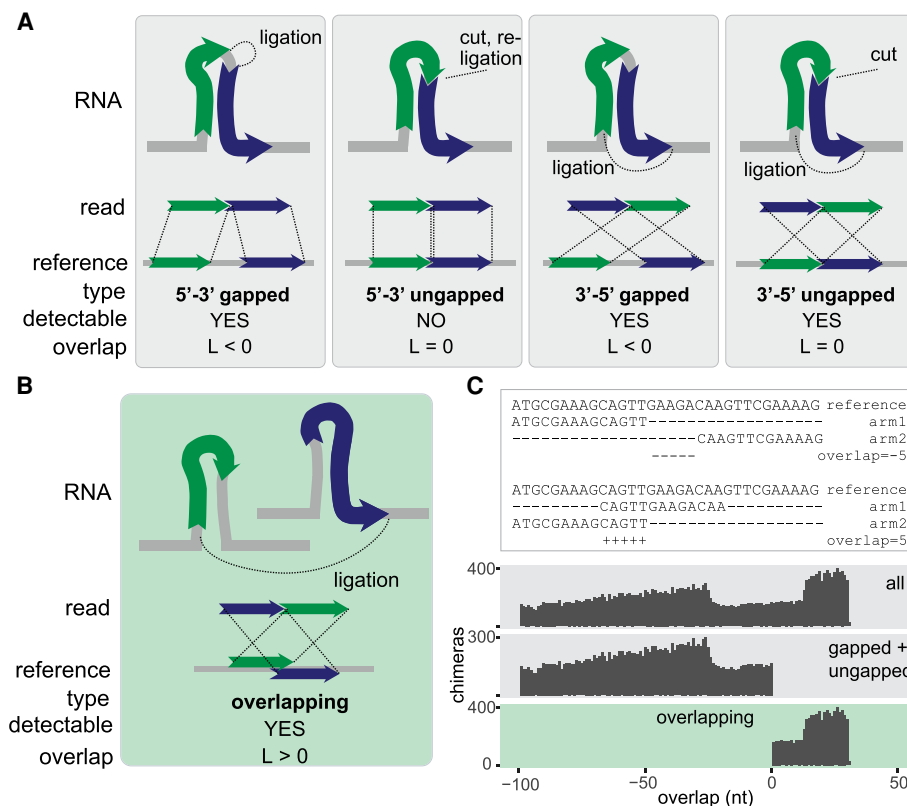


Figure 1. Classification of chimeric reads. (A) Types of nonoverlapping chimeras that may be formed by proximity ligation. The RNA fragments, shown as green and blue arrows, can originate from the same transcript (as shown in the figure) or from distinct transcripts (not shown). The dotted lines in the upper panel indicate ligation sites. (B) Diagram of an overlapping chimera. The RNA fragments originate from two distinct copies of a transcript but are mapped to the same region of the reference gene. (C, top) Examples of nonoverlapping and overlapping chimeric reads mapped to a reference gene. (Bottom) Distribution of the calculated overlap score (L) in all simulated chimeras, simulated nonoverlapping chimeras, and simulated overlapping chimeras.

3'-5'—are all feasible and are commonly detected in proximity ligation experiments.

In contrast, when an intermolecular interaction exists between two copies of the same RNA molecule, the two interacting fragments may or may not originate from the same part of the RNA. When these fragments are mapped to the source RNA sequence, they can be found in any of the arrangements shown in Figure 1A and in an additional overlapping arrangement (Fig. 1B). Thus, gapped and ungapped chimeras can result from intra- and intermolecular interactions, but overlapping chimeras are diagnostic of intermolecular interactions.

In the following sections, we discuss the suitability of bioinformatic methods to detect the relevant types of chimeric reads; the types of chimeras we identify in various RNA proximity ligation experiments; and the possible origin and interpretation of the interactions we detect.

Detection of overlapping chimeras in simulated sequencing data

To identify chimeras, we used the *hyb* pipeline (Travis et al. 2014). *Hyb* maps reads against a reference sequence database with one of several tools (BLAST, Bowtie 2, or BLAT) (Altschul et al. 1990; Kent 2002; Langmead and Salzberg 2012) and detects chimeric reads with two separate local matches in the database. To test whether *hyb* is suitable for detection of gapped, ungapped, and overlapping chimeras, we assembled a test data set with simulated chimeras by concatenating all possible pairs of 30-nt substrings from an arbitrary 228-nt RNA sequence. Using either BLAST or Bowtie 2 as the mapping engine, *hyb* correctly identified the majority of sequences as chimeric. A subset of 5'-3' gapped chimeras and overlapping chimeras were not called by the algorithm (Supplemental Fig. S1). Inspection of the BLAST and Bowtie 2 outputs showed that these chimeras were interpreted by the mapping programs as nonchimeric reads with internal deletions or insertions. We also called chimeras with STAR, a general-purpose mapping tool that has been used in some RNA proximity ligation studies (Dobin et al. 2013). Although the results were comparable with *hyb*, STAR missed most 5'-3' chimeras and a subset of 3'-5' chimeras (Supplemental Fig. S1). Both *hyb* and STAR commonly misidentified the position of the ligation junction between the two arms of the chimera by 1–3 nt, but this did not affect the identification of overlapping chimeras, which typically relies on the mapped position of nonligated ends of reads (Fig. 1B). Using an alternative test data set with more than 1 million simulated nonchimeric reads and more than 1 million chimeric reads (Methods), we found 99.9% specificity and 95% sensitivity in the detection of nonoverlapping chimeras, and 100.0% specificity and 69% sensitivity in the detection of overlapping chimeras, where most false negatives were chimeras with very short overlaps.

We then quantified the degree of overlap between arms of chimeric reads using an overlap metric *L* defined for chimeras where both arms are mapped to the same reference transcript as follows:

$$L = 1 + \min(e_1, e_2) - \max(s_1, s_2),$$

in which s_1 and s_2 are start mapping coordinates of arms 1 and 2 of the chimera on the reference transcript, and e_1 and e_2 are end coordinates of arms 1 and 2. *L* is positive for overlapping chimeras, null for ungapped 3'-5' chimeras, and negative for gapped 5'-3' or 3'-5' chimeras. In the test data set, *L* was positive for simulated overlapping reads and negative for simulated nonoverlapping reads (Fig. 1C), as expected. These results show that

our methods are appropriate for the identification of overlapping chimeras in RNA proximity ligation data.

Overlapping chimeras in RNA proximity ligation data

We analyzed representative RNA proximity ligation data sets generated by several experimental protocols (Helwak et al. 2013; Ramani et al. 2015; Aw et al. 2016; Lu et al. 2016; Sharma et al. 2016; Waters et al. 2016; Li et al. 2018; Ziv et al. 2018, 2020; Huber et al. 2019; Cai et al. 2020; Methods). The protocols differ, among other ways, in the method used to stabilize RNA–RNA interactions. CLASH uses UV-protein cross-linking, with only one RNA strand expected to be covalently linked to a protein and the other bound by complementarity. SPLASH, PARIS, and COMRADES use psoralen cross-linking, whereas RIC-Seq is based on protein-dependent formaldehyde cross-linking, and RPL omits the cross-linking step altogether.

We focused on homotypic chimeras, that is, those in which both arms are mapped to the same transcript. Among homotypic chimeras, we detected gapped, ungapped, and overlapping chimeras in all data sets, but the relative proportions of these three types varied greatly between data sets (Fig. 2). Methods that use UV and psoralen cross-linking to recover direct RNA:RNA interactions yielded large numbers of gapped and ungapped, but few overlapping chimeras. For example, out of 4.1 million chimeras we detected in the PARIS HEK293 data set, 3.2 million were homotypic chimeras, but only 42,000 were overlapping homotypic chimeras, indicative of RNA homodimerization (Supplemental Fig. S2). Although gapped chimeras could originate from inter- or intramolecular interactions, the near absence of overlapping chimeras suggests that homomeric intermolecular interactions are rare in these data sets. In contrast, RPL and RIC-Seq recovered large numbers of overlapping chimeras. Both RPL and RIC-Seq can plausibly recover indirect interactions: RIC-Seq was specifically designed to detect indirect contacts through protein formaldehyde cross-linking, and RPL might allow for reassociation of RNA:RNA complexes during chemical processing in situ, owing to the absence of a covalent linkage between RNA strands. These results suggest that RNA homodimerization mediated by direct RNA–RNA base-pairing is uncommon in vivo. The results also show that RNA duplexes are generally stable during library preparation, at least in the CLASH, SPLASH, PARIS, and COMRADES methods, because random reassociation of duplexes would lead to the formation of similar proportions of gapped and overlapping chimeras.

An intriguing pattern is the peak at overlap = 0 in Figure 2, indicating the preferential recovery of 3'-5' ungapped chimeras relative to gapped and overlapping chimeras (as discussed above, 5'-3' ungapped chimeras cannot be detected with our methods). We propose that ungapped chimeras typically arise from local RNA stem-loop structures, which are subject to three endonucleolytic cleavages, followed by ligation of the distal ends to each other, whereas gapped chimeras could be created either by four independent endonucleolytic events, or by a combination of three endonucleolytic cuts combined with exonucleolytic trimming of RNA ends. Although enrichment of ungapped chimeras can be readily explained for intramolecular interactions, it is difficult to imagine a mechanism that could enrich ungapped chimeras for intermolecular interactions. These results reinforce our conclusion that stable RNA homodimers are rarely formed in vivo.

We also observed an overrepresentation of overlapping chimeras with very short overlaps ($0 < L < 5$) in all UV and psoralen data sets (Fig. 2; Supplemental Fig. S2). We speculate that most

such chimeras are derived from the same type of interaction that gives rise to 3'-5' ungapped chimeras, but the apparent overlap is caused by mapping errors. Although we could not detect similar artifacts in our simulated benchmarking data, the artifacts could arise in experimental data because of sequencing errors or adapter mutations. We thus conservatively exclude chimeras with $L < 5$ from consideration in the calling of homodimers.

Homodimerization of human and yeast RNAs

Although few RNA homodimers were found in UV and psoralen cross-linking experiments, we hypothesized that homodimers might be limited to a specific subset of RNAs. To investigate this possibility, we analyzed chimeras detected in individual genes in transcriptome-wide PARIS data from HEK293 cells. To increase the stringency of our analysis, we filtered the data to remove likely mapping errors, chimeras with thermodynamically unstable interactions, homopolymers, and chimeras with very short overlaps (< 5 nt) (Supplemental Fig. S2). After filtering, gapped chimeras were more common overall, but 84 genes contained overlapping chimeras, including 17 genes that were significantly enriched in overlapping chimeras (Fisher's exact test with Benjamini-Hochberg correction, $P < 0.05$) (Supplemental Fig. S3; Supplemental Data Set S1). The most highly enriched transcript was mRNA *TMEM107*, which contained 100 times more overlapping than gapped chimeras (Fig. 3).

TMEM107 contains a small nucleolar RNA (snRNA), *U8*, in its 3' untranslated region, and almost all *TMEM107*:*TMEM107* chimeras mapped to that region, suggesting that these chimeras represent *U8*:*U8* interactions. The chimeras were concentrated around the 5' end of *U8* (Fig. 4A,B), and RNA folding prediction showed extended self-complementarity in this part of the transcript, consistent with homodimerization with a predicted free energy of -21 kcal/mol. The same *U8*:*U8* interaction was identified in CLASH data, and in an independent analysis of a new PARIS2 data set (Zhang et al. 2021a, 2022). Previous studies showed that the 5' region of *U8* may base pair with pre-ribosomal RNA (Peculis 1997; Zhang et al. 2021a) and with the 3' end of a 3'-extended precursor of *U8* (Badrock et al. 2020). Because homodimerization seems incompatible with these interactions, it might represent an immature form of *U8* or play a role in the regulation of *U8* function. This is potentially important for the pathogenesis of LCC, a neurological dis-

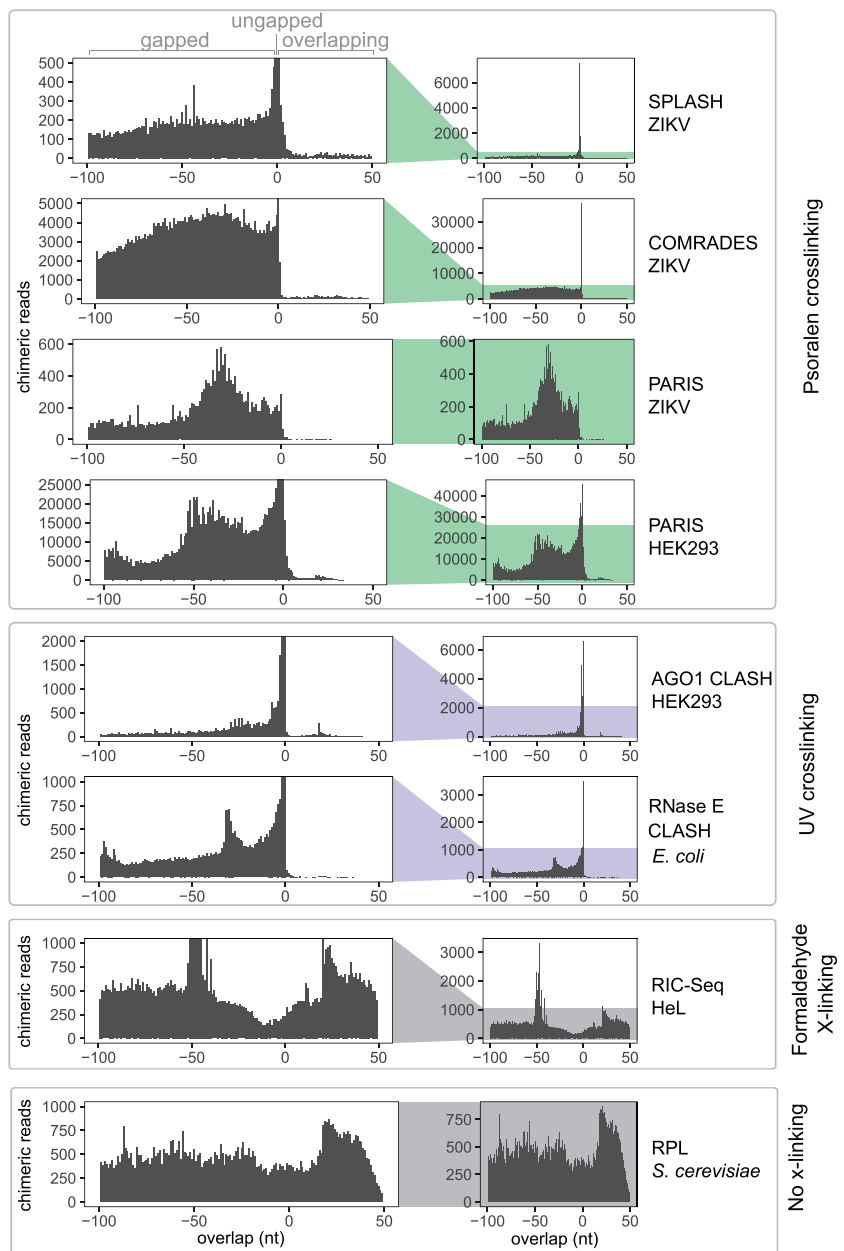


Figure 2. UV and psoralen cross-linking predominantly generate gapped and ungapped chimeras. The distribution of overlap scores (L) across proximity ligation data sets: SPLASH ZIKV (Zika virus) (Huber et al. 2019), COMRADES ZIKV (Ziv et al. 2018), PARIS ZIKV (Li et al. 2018), PARIS HEK293 (human) (Lu et al. 2016), AGO1 CLASH HEK293 (Helwak et al. 2013), RNase E CLASH (*Escherichia coli*) (Waters et al. 2016), RIC-Seq HeLa (human) (Cai et al. 2020), and RPL (*Saccharomyces cerevisiae*) (Ramani et al. 2015; Methods).

ease caused by loss-of-function mutations in *U8* (Jenkinson et al. 2016; Badrock et al. 2020). Some overlapping chimeras comprised a 5'-extended form of *U8*, indicating that the homodimers may be formed during snRNA maturation. We have not found any interactions involving other regions of the *TMEM107* transcript (Supplemental Fig. S3), nor have we found homodimer enrichment in other mRNAs from the *TMEM* family.

In addition to *U8*:*U8* interactions, analysis of PARIS data showed enrichment of homodimers in *U1* and *U2* snRNA

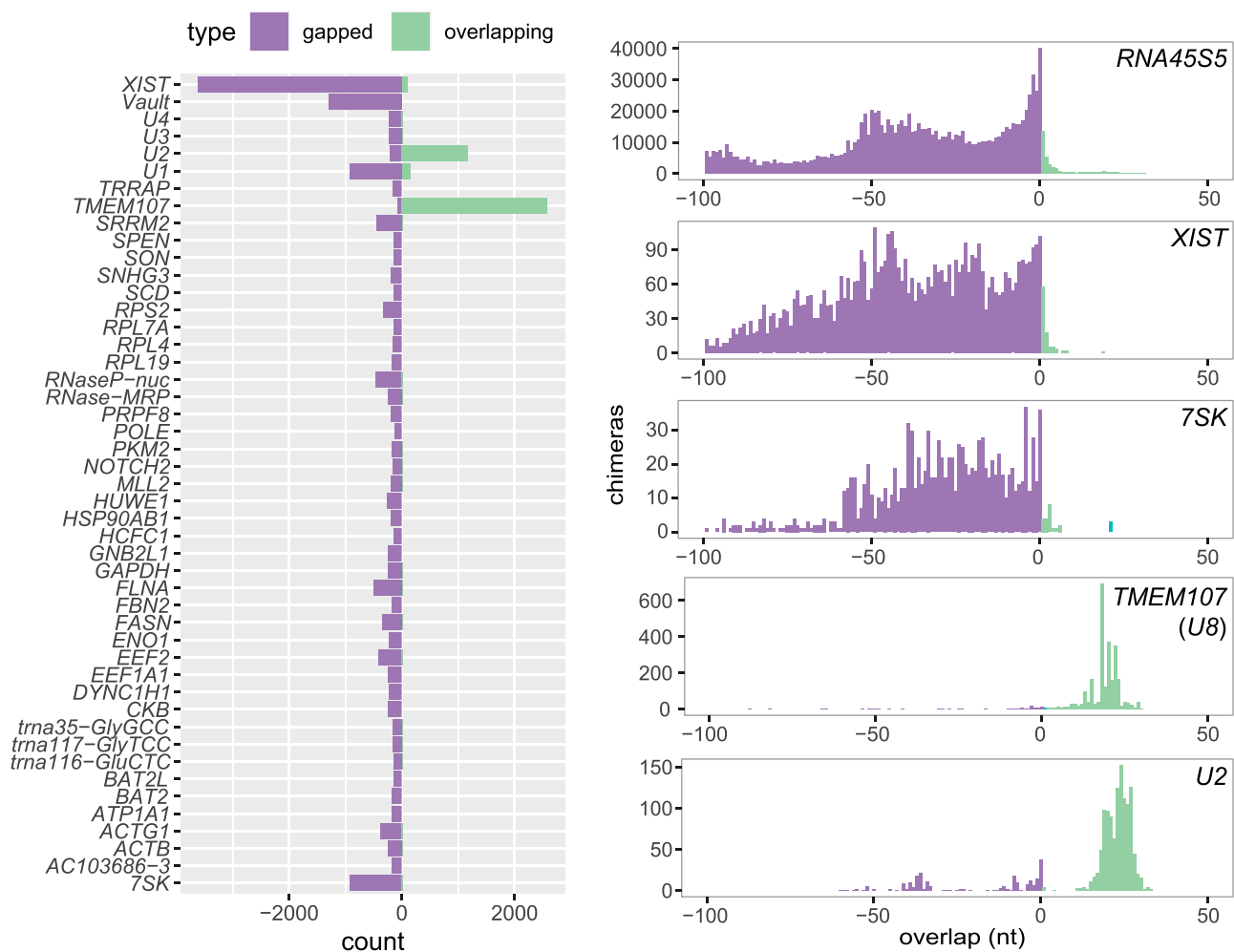


Figure 3. Distribution of overlap scores in individual genes in PARIS experiment. Counts of gapped and ungapped chimeras (purple) and overlapping chimeras (green) in individual genes in PARIS data from human HEK293 cells (Lu et al. 2016) (*left*), and distribution of the overlap value L in selected genes.

(Figs. 3,4C; Supplemental Fig. S4). *U2* snRNA contained 20 times as many overlapping as gapped chimeras. Regions involved in homodimeric interactions in *U1* and *U2* are limited to a particular fragment of the RNA, whereas other types of interactions can be found along the transcript (Supplemental Fig. S3). Most overlapping chimeras in *U2* included the sequence of stem-loop III, downstream from the Sm binding site, suggesting that in a fraction of *U2* molecules found in the cell, stem-loop III is unfolded and forms homomeric intermolecular interactions (Fig. 4C). Out of the two major isoforms of *U2*, *U2-1* shows more efficient deposition of the Sm ring and incorporation into snRNP complexes than *U2-2* (Kosmyna et al. 2020), and we hypothesized that failure of assembly into an snRNP complex might be associated with *U2:U2* dimerization. However, analysis of the exact sequences of *U2:U2* overlapping chimeras showed that out of 223 reads that could be assigned to specific *U2* isoforms, 162 were of the *U2-1:U2-1* type, and 59 were *U2-1:U2-2* chimeras, suggesting that both *U2* isoforms may form homo- and heteromeric intermolecular interactions. *U2* homodimers were also found in CLASH, SPLASH (this study), and PARIS2 (Zhang et al. 2022).

Altogether, across the five studies we analyzed (AGO1 CLASH, PARIS, Zika COMRADES, Zika SPLASH, and human

SPLASH), we found 50 transcripts with homodimers found across two or more studies (Supplemental Fig. S3). These transcripts include ribosomal RNA, *U1* and *U2* snRNA, *U3* and *U8* snoRNA, 4 tRNAs, and 36 mRNAs. We then compared specific homodimerization events detected by different proximity ligation methods. PARIS and COMRADES showed the largest fractions of homotypic chimeras, most of which were nonoverlapping and likely represented intramolecular interactions. Across all RNA biotypes, rRNAs formed most homodimers, particularly in PARIS and SPLASH, but such homodimers were not statistically enriched, when compared to intramolecular interactions. tRNAs were enriched for homodimers in AGO1 CLASH and SPLASH (Supplemental Figs. S3, S5). tRNA-derived small RNAs (tsRNAs), including tRNA-derived fragments (tRFs) and tRNA halves (tiRNAs), are small regulatory RNAs processed from mature tRNAs or precursor tRNAs (Xie et al. 2020). *tX(XXX)D*, a yeast tRNA similar to serine tRNAs (Chan and Lowe 2009) formed a homodimer through a 12 base pair long stem in SPLASH data (Supplemental Fig. S6). The tRNA homodimers detected by AGO1 CLASH in human cells (Supplemental Fig. S6) may indicate a miRNA-tRNA network resulting in competition for binding sites and availability for gene silencing, as reported previously (Shigematsu and Kirino 2015).

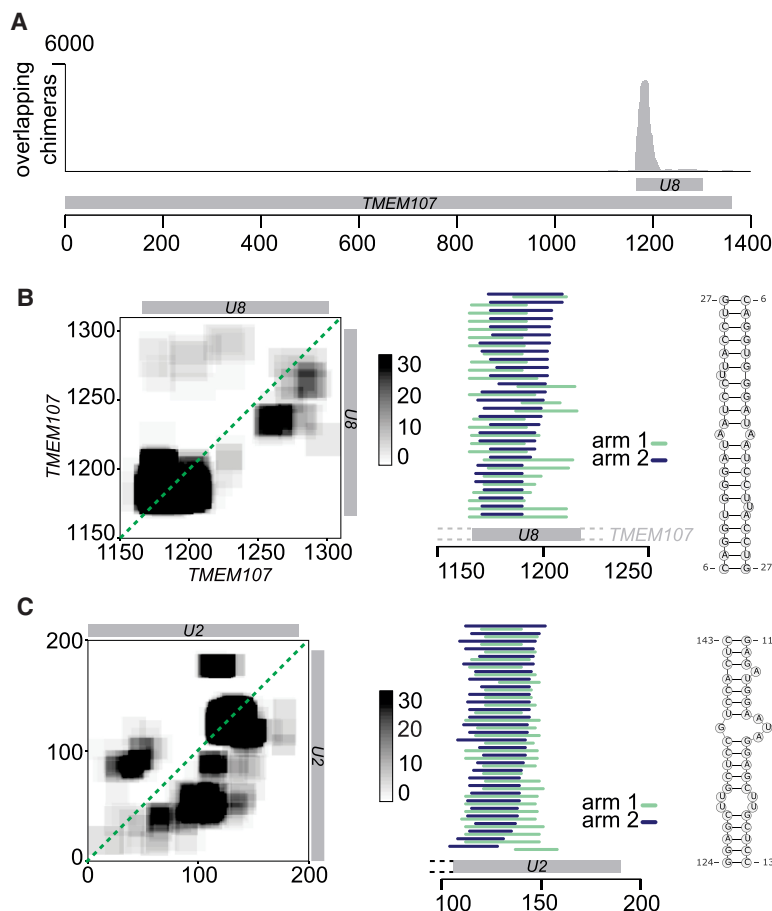


Figure 4. Homodimerization of *U8* snoRNA and *U2* snoRNA. (A) Distribution of overlapping chimeras along the *TMEM107* gene in PARIS (HEK293 cells) (Lu et al. 2016). All overlapping chimeras coincide with the position of *U8* snoRNA in the 3' UTR of *TMEM107* transcript. (B, left) Coverage map of *U8:U8* chimeras. Numbers on the x- and y-axes indicate positions in the *TMEM107* transcript. The position of mature *U8* snoRNA is indicated by the gray bars. (Middle) Mapping positions of both arms of chimeras along *TMEM107*. Some chimeras extend beyond the 5' end of mature *U8* (gray bar), indicating that they originate from the *U8* precursor molecule. (Right) Predicted base-pairing of *U8:U8* homodimer. (C) As above, analysis of *U2:U2* chimeras. Coordinates indicate positions along the *U2* transcript.

The largest ratio of overlapping to homotypic chimeras was recovered by RIC-Seq (7%). RIC-Seq also recovered the highest number of genes with overlapping chimeras (more than a thousand), 17 of which were significantly enriched for overlapping chimeras (Supplemental Fig. S3). As discussed above, the overlapping chimeras found by RIC-Seq probably represent indirect interactions rather than RNA homodimers. RIC-Seq showed significant enrichment of overlapping chimeras in some mitochondrial mRNAs, with *CO1*, *ND2*, and *ND4* containing the highest numbers of overlaps (Supplemental Figs. S3, S7). Bidirectional transcription of mitochondrial RNA is known to result in hybridization of complementary strands (Dhir et al. 2018; Kim et al. 2018), but in the RIC-Seq data, both partners come from the same strand, suggesting that they represent a distinct type of interaction. The mitochondrial mRNA:mRNA chimeras showed low thermodynamic stability and short regions of complementarity (2–8 base pairs), suggesting that these chimeras represent indirect interactions facilitated by the high local concentrations of transcripts in mitochondria. COMRADES, PARIS, and SPLASH also detected homodimers among mitochondrial transcripts (Supplemental Fig. S4).

YLR154W-E, a possible ncRNA from yeast with a strong enrichment in overlapping chimeras in the RPL data, can be predicted to dimerize through an extended stem structure (Supplemental Fig. S7).

The *U8* homodimerization domain plays an important role in vivo

To study the function of homodimers in more detail, we focused on the *U8:U8* interaction, the most abundant homodimer in our analysis of the PARIS data. We took advantage of an experimental system in which *U8-3^{-/-}* zebrafish embryos are injected with human *U8* precursor RNA (pre-*U8*) to analyze the functional consequences of human *U8* mutations (Badrock et al. 2020). Previous experiments showed abnormal yolk sac and brain development in *U8-3^{-/-}* embryos, and that these phenotypes were complemented by injection of wild-type human pre-*U8*, but not by injection of known disease mutants of *U8* (Badrock et al. 2020). Thus, the zebrafish model can identify loss-of-function mutations in human *U8*.

We first asked if mutations predicted to affect homodimer formation disrupt *U8* function in zebrafish. We selected three candidate mutations: 19C > G, 20C > G, and 24C > G, which are expected to strongly disrupt *U8* homodimerization, but have little or no effect on the predicted interactions of *U8* with the preribosome or on any other known domain or function of *U8* (Fig. 5A; Supplemental Fig. S8). We found that all three mutants fail to complement developmental phenotypes observed in the *U8-3^{-/-}* embryos (Fig. 5B,C). We also tested candidate mutations 20C > T and 24C > T, which are predicted to disrupt the *U8* homodimer, and that had been found in patients suffering from LCC, a neurodegenerative disease caused by the loss of *U8* function. Again, these mutants fail to complement the zebrafish phenotype (Supplemental Fig. S8). Furthermore, mutations in the homodimerization domain disrupt the formation of slowly migrating conformers by in vitro transcribed *U8* RNA in native polyacrylamide gel electrophoresis (Supplemental Fig. S9). Taken together, these results suggest that the *U8* homodimerization domain we discovered plays an important biological role that is conserved across vertebrates.

Next, we attempted to rescue homodimerization mutants by compensatory mutations. None of the five double mutants we tested rescued the phenotype of zebrafish embryos (Fig. 5B,C; Supplemental Fig. S8). We postulate that compensatory mutations disrupt other important aspects of *U8* function, and thus they do not complement function, although they may rescue dimerization. This interpretation is supported by the predicted effect of our compensatory mutations on the *U8:28S* pre-rRNA interaction

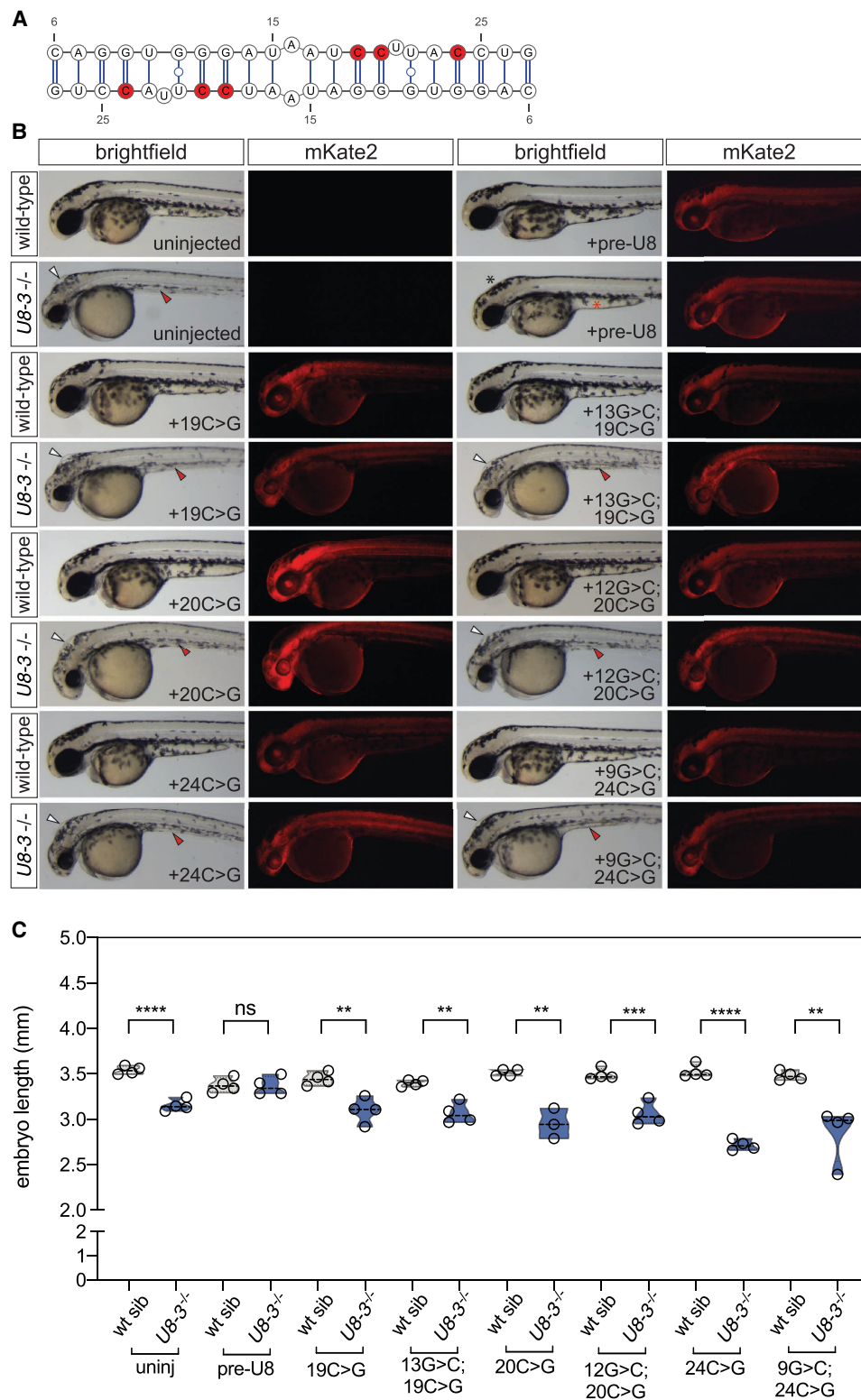


Figure 5. Functional analysis of *U8* homodimerization domain. (A) *U8:U8* homodimer with positions of selected destabilizing mutations indicated in red. (B) Representative brightfield and fluorescent images of the indicated genotype and exogenous precursor *U8* snoRNA variants taken at 48 h post-fertilization. mRNA encoding the mKate2 fluorescent protein was coinjected to show successful uptake of the injected solution into the embryo. White arrowheads denote hydrocephaly, red arrowheads denote aberrant yolk extension, whereas a black asterisk denotes an absence of hydrocephaly and a red asterisk denotes rescued yolk morphology. (C) Quantitation of embryo length for the genotypes and introduced precursor *U8* snoRNA variants indicated. (uninj) Uninjected. $n = 3-4$ embryos per genotype. Black dashed line indicates median value; all data points are shown.

(Supplemental Fig. S8). Further biochemical studies are required to dissect the functions of the *U8* homodimerization domain and its mutated variants. Altogether, our experiments are consistent with the hypothesis that *U8:U8* and *U8:28S* pre-rRNA interactions are essential for *U8* function in vivo.

Homodimerization of virus RNA

We next turned to COMRADES data from cells that have been infected with SARS-CoV-2 and Zika viruses, to detect possible homodimers of virus RNA. Although Zika RNA is not known to homodimerize, dimerization is an essential step in the packaging of some viruses, including HIV, whereas dimerization of SARS-CoV RNA was suggested to play a role in translational frameshifting (Ishimaru et al. 2013). To detect dimers of virus RNA, we analyzed the coverage of overlapping chimeras along viral genomes. Unlike gapped chimeras, which covered the Zika genome relatively evenly, overlapping chimeras were strongly enriched in several positions within the *NS2A*, *NS2B*, and *NS5* coding sequences of the Zika virus, indicating possible dimerization sites (Fig. 6A). RNA folding prediction showed regions of self-complementarity in the interaction sites, including a pair of uninterrupted 11-bp duplexes in the (3578–3656):(3578–3656) region in the *NS2A* gene. However, folding energy alone was not enough to predict dimerization sites, as evidenced by the weak negative correlation between the count of overlapping chimeras in a genomic window and the predicted strength of homodimeric interaction in that window (Pearson $R = -0.17$, $P = 3 \times 10^{-8}$).

We also detected dimerization events in the SARS-CoV-2 genome, with the largest peak in the nucleoprotein (*N*) gene, and additional peaks in the *N*, *Orf6*, and *Orf1a* coding sequences (Fig. 6B). The region with the largest coverage of overlapping chimeras was 200 nt long (coordinates 28,610–28,810) and the resolution was insufficient to indicate the exact base-pairing, but RNA folding analysis showed several high-scoring interactions, including a 10-nt duplex formed by the palindromic sequence, GGTTG CAACT. Although a previous NMR study detected a functionally important dimerization site near the frameshifting element of the SARS-CoV virus (Ishimaru et al. 2013), our analysis shows no obvious enrichment of overlapping chimeras in the homologous region of SARS-CoV-2.

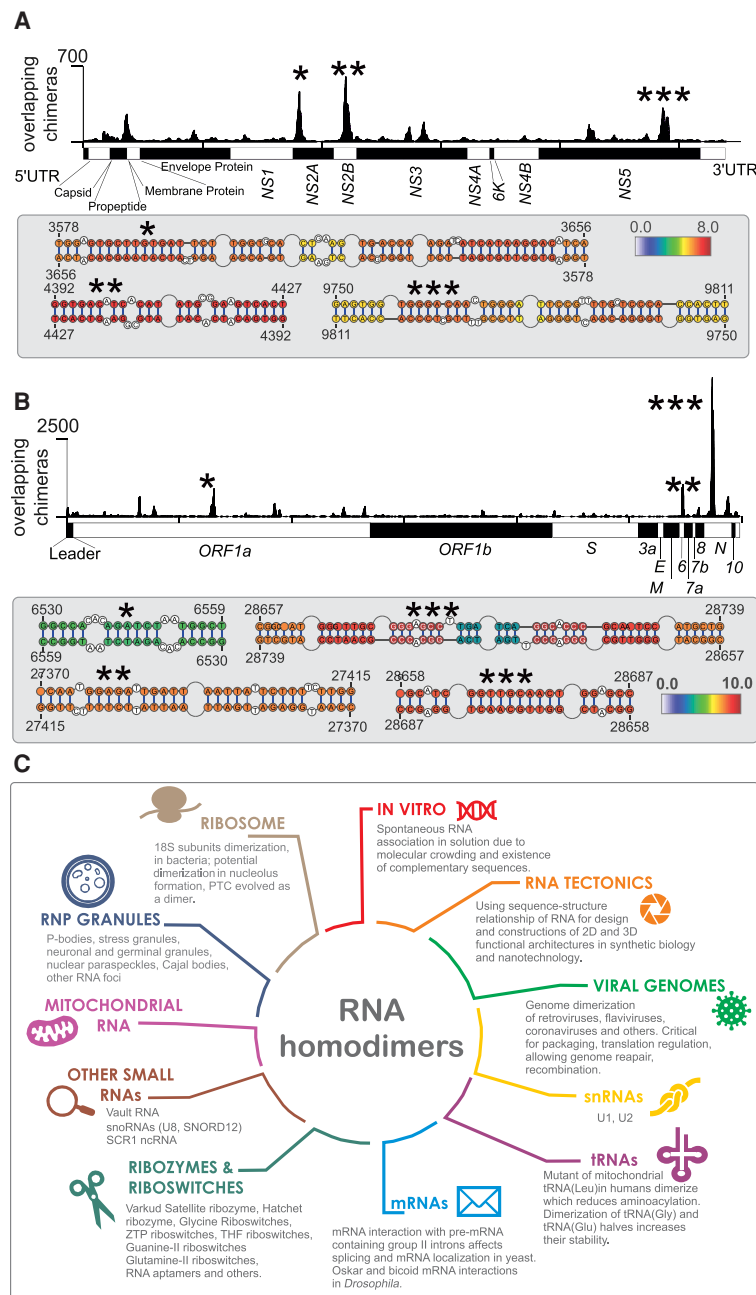


Figure 6. Homodimers of viral genomes. (A, top) Coverage of overlapping chimeras in Zika virus identified by COMRADES (Ziv et al. 2018). The regions with the largest coverage of homodimers are indicated by asterisks. (Bottom) Predicted secondary structures of regions with high coverage of overlapping chimeras. The colors indicate the COMRADES score, that is, the \log_2 of the number of chimeric reads supporting each base pair. (B) Coverage of overlapping chimeras along the SARS-CoV-2 genome identified by COMRADES (Ziv et al. 2020). The top homodimeric regions and their predicted structures are shown, as described above. (C) Known types of RNA homodimers.

Discussion

Although homo-oligomerization is common in proteins, few RNA homo-oligomers have been described in vivo. This is somewhat surprising, given that RNA molecules readily homodimerize in vitro, to the point that special procedures have to be used to isolate

monomeric forms of certain RNAs for structural studies (Zhang and Ferre-D'Amare 2014; Bou-Nader and Zhang 2020). The paucity of *in vivo* homo-oligomers might be explained by the folding of RNAs and by their association with protein complexes, which reduce the propensity for *trans*-RNA–RNA interactions. Alternatively, the apparent lack of *in vivo* homodimers might simply reflect the lack of systematic studies of dimerization. Here, by analyzing the relative proportions of gapped, ungapped, and overlapping chimeric reads in RNA proximity ligation experiments, we find that homodimerization mediated by direct RNA base-pairing is indeed rare *in vivo*. However, we find that certain human RNAs and some regions of the RNA genomes of the Zika and SARS-CoV-2 viruses are enriched for *in vivo* homodimers.

Out of thousands of RNAs we examined, only a handful show clear evidence of dimerization. The propensity to dimerize is necessarily influenced by the primary sequence of the RNA: for example, palindromic sequences or CAG repeats might be prone to form intermolecular interactions. A recent review of RNA homodimer structures detected in viruses, ribozymes, and riboswitches identified preferences for certain sequence and structural arrangements, such as palindromes, complementary strand swapping, and kissing-loop interactions (Bou-Nader and Zhang 2020). Indeed, in the present study, palindromic sequences were found in several RNA homodimers. Homodimerization is also likely to be influenced by folding kinetics, association with proteins and other RNAs, subcellular localization, and local concentration of RNA and metal ions. RNA molecules that fold cotranscriptionally into stable secondary structures are unlikely to form extended duplexes with other RNAs (Yu et al. 2021), whereas molecules that are unfolded by helicases, or located in granules with high local concentrations of a given RNA, might be more likely to form transient or stable oligomers. Copies of RNA molecules located in close proximity may initially interact with a few nucleotides, followed by destabilization of local structure and nucleation of longer interactions (Ganser et al. 2019).

Are the RNA homodimers detected by proximity ligation biologically relevant, or are they experimental artifacts? We argue that nonspecific dimerization and oligomerization of RNA during library preparation, if present, should lead to the formation of many overlapping chimeras, distributed across a large variety of RNAs. Indeed, this is what we observe in the RPL and RIC-Seq data sets. RPL is performed without cross-linking, whereas RIC-Seq involves formaldehyde cross-linking. As a result, overlapping chimeras detected by these methods likely indicate local transcript proximity rather than direct base-pairing, although it is also possible that a fraction of overlapping chimeras arises during the library preparation step.

In contrast, techniques that rely on UV or psoralen cross-linking—CLASH, SPLASH, PARIS, and COMRADES—are expected to detect RNA–RNA contacts mediated by direct base-pairing. We observed that these methods generate few overlapping chimeras, but these chimeras are strongly enriched in a small subset of RNAs, suggestive of bona fide interactions. Alternatively, overlapping reads might theoretically arise through reverse transcription of an endogenous circular RNA (circRNA), or of an artificial circRNA created *in vitro* by ligation, producing a concatemeric cDNA. However, the low abundance of circRNAs and low efficiency of RNA ligases makes such events unlikely. We also note that proximity ligation can only identify a subset of possible RNA homodimers, namely, those where both RNAs interact via the same part of their sequence, or via two regions that are close enough in the primary sequence to detect overlaps in chimeric

reads. Although many known RNA homodimers are of this type (e.g., the DIS kissing loop interaction in HIV, the SL2-SL2 interaction in Moloney murine sarcoma virus [MoMuSV], or the dimerization of *Oskar* RNA via its 3' UTR in *Drosophila* embryos) (Berkhout and van Wamel 1996; Kim and Tinoco 2000; Jambor et al. 2011), interactions mediated via distant fragments of RNA would not be detectable by proximity ligation.

Homodimerization has now been reported for most major biotypes of RNA, and known roles of homodimers include the packaging of viral genomes, assembly of membraneless organelles, regulation of RNA localization (Fig. 6C). Given its dependence on the local concentration of RNA, dimerization might play a role in RNA quorum sensing—a process analogous to that used by bacteria and viruses to coordinate their behavior in response to the local population density. Nevertheless, many RNA homodimers do not have a known biological function, and indeed might be detrimental. Stretches of dsRNA are known to trigger antiviral immunity through PKR and other cellular factors (Hull and Bevilacqua 2016), and some types of homodimers might be misidentified as foreign RNA. RNA multimerization has also been associated with general cellular stress (Van Treeck and Parker 2018; Van Treeck et al. 2018). In RNA proximity ligation methods, the use of psoralen, formaldehyde, and UV light is a stress factor that might contribute to RNA multimerization. In any case, further functional studies are required to elucidate the roles of the wide variety of RNA homodimers that can be detected in our cells.

Methods

Benchmarking chimera detection on test data

To benchmark methods for detection of overlapping chimeras, we assembled a test data set using an arbitrary RNA sequence (nucleotides 1–228 of *S. cerevisiae* *RDN37* gene: NCBI GenBank database (<https://www.ncbi.nlm.nih.gov/genbank/>) Sequence ID: CP026300.1, range 448,071–448,298, minus strand). We generated all 30-nt substrings of the reference sequence and concatenated all possible pairs of substrings, which yielded 10,871 overlapping chimeras and 28,730 nonoverlapping chimeras.

We then called chimeras in the overlapping and nonoverlapping data sets using *hyb* and STAR, using the following commands:

```
hyb (Bowtie 2 mapping):
hyb analyse in = input.fasta db = RDN37 format = comp eval = 0.001
```

```
hyb (BLAST mapping):
hyb analyse in = input.fasta db = RDN37 format = comp align = blastall eval = 0.001
```

```
STAR:
STAR --genomeDir . --readFilesIn input.fasta --outFileNamePrefix 06 --outReadsUnmapped Fastx --outFilterMismatchNoverLmax 0.05 --outFilterMatchNmin 16 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --clip3pAdapterMMp 0.1 --chimSegmentMin 15 --scoreGapNoncan -4 --scoreGapATAC -4 --chimJunctionOverhangMin 15
```

We used the *ua.hyb* files from *hyb* and *Chimeric.out.junc* files from STAR for downstream analysis. To generate the coverage heatmaps of chimeras detected in the test data set, we extracted the coordinates of chimera junctions and plotted them using Java TreeView (Saldanha 2004).

As a second test data set, we used more than 1 million simulated chimeric reads and more than 1 million nonchimeric reads from the *S. cerevisiae* preribosomal RNA (*RDN37*; 6858 nt in length). The lengths of chimeric arms ranged from 20 to 40 nt, to approximate

the sizes of chimeric fragments found by CLASH, PARIS, and related methods. We mapped these simulated reads against the entire yeast transcriptome, using hyb with default settings. To estimate sensitivity and specificity for chimera detection, we used $L \geq 5$ (see next section) as the cutoff for calling overlapping chimeras.

Calculation of chimera overlaps

To quantify overlap between arms of chimeric reads, we defined the overlap metric, L , as

$$L = 1 + \min(e_1, e_2) - \max(s_1, s_2),$$

where L is defined for chimeras in which both arms are mapped to the same reference transcript, or same chromosome in case of mapping to a genome reference. e_1 represents the end mapping coordinate of the left arm of the chimera (arm 1); e_2 represents the end mapping coordinate of the right arm of the chimera (arm 2); s_1 and s_2 represent the start mapping coordinates of the respective arms. Calculation of L was implemented as a custom awk script, taking the `ua.hyb` files produced by the `hyb` pipeline as inputs.

RNA proximity ligation data

We downloaded the data from the following and including NCBI (<https://www.ncbi.nlm.nih.gov/>) and ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>) accession numbers: *Escherichia coli* RNase E CLASH: GSE77463 (Waters et al. 2016); human AGO1 CLASH: GSE50452 (Helwak et al. 2013); human SPLASH: SRR3404931 (Aw et al. 2016); Zika SPLASH: SRR6252011 (Huber et al. 2019); Zika COMRADES: E-MTAB-6427 (Ziv et al. 2018); human LIGR-Seq: SRR3361013 (Sharma et al. 2016); human PARIS: SRR2814765 (Lu et al. 2016); Zika PARIS: PRJEB28648 (Li et al. 2018); human RIC-Seq: SRR8632820 (Cai et al. 2020); *Saccharomyces cerevisiae* RPL: SRR2048219 (Ramani et al. 2015); and SARS-CoV-2 COMRADES: GSM4676632 (Ziv et al. 2020).

Sequencing data were downloaded in FASTQ format (except for the SARS-CoV-2 data set, from which `hyb` output files were downloaded). Chimeric reads were called and annotated with the `hyb` package (Travis et al. 2014) with default settings, using the appropriate transcriptome database (Helwak et al. 2013; Waters et al. 2016; Ziv et al. 2018), as described in Supplemental Data Set S2.

Overlap statistics across experimental data sets were visualized in R (R Core Team 2017) using the `ggplot2` and `ggforce` libraries (`facet_zoom` function). To identify genes enriched in overlapping chimeras, we filtered `hyb` outputs to remove possible mapping errors (any reads with nucleotide repeats of length 15 or more, and chimeras where either arm had a mapping e -value greater than 0.001); we also removed chimeras with predicted interaction energy weaker than -5 kcal/mol. Because very short overlaps might represent mapping or sequencing errors, we conservatively called chimeras with overlap score $L \geq 5$ as overlapping, and chimeras with $-50 < L < 0$ as nonoverlapping. We then assembled a contingency table with counts of overlapping and nonoverlapping chimeras for the focal gene and for all other genes, and we used a Fisher's exact test with Benjamini-Hochberg multiple testing correction to identify genes with significant enrichment of overlapping chimeras.

Zebrafish *U8-3* mutant rescue experiments

Human *U8* RNA variants were in vitro transcribed from DNA templates containing a T7 consensus sequence as described (Badrock et al. 2020). `pCS2+-mKate2` was linearized with `NotI`, and mRNA transcribed from the DNA template using the `mMESSAGE mMACHINE SP6` transcription kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Microinjections of 2

nL of solution containing 500 pg of a *U8* variant and 100 pg *mKate2* mRNA were microinjected into the yolk of one-cell stage zebrafish embryos through use of the PicoSpritzer III (Parker Instruments) apparatus. Where *U8* variants were found to rescue the morphology of *U8-3* mutant zebrafish, genotyping was performed as described in Badrock et al. (2020) to confirm the genotype of the assessed embryos.

Imaging and embryo measurement

Zebrafish embryos were anesthetized using MS-222 (Sigma-Aldrich), embedded in 3% Methyl cellulose (M0387), and imaged on an MZFLIII fluorescent stereomicroscope (Leica) with a MicroPublisher 3.3 RTV camera, using Micro-Manager 1.4.23 software. Embryo length was quantified using images taken at 1 \times magnification with a 1-mm scale bar as a reference point. Embryo length was measured in Microsoft Powerpoint (Microsoft) by drawing a line from head to tail.

Statistics and reproducibility

Statistical analyses were performed using GraphPad Prism 8.0. Results are presented as violin plots, with bold dashed line representing the median value. All data points are shown. For all analyses, $P < 0.05$ was considered statistically significant (using a student t -test). Statistical methods were not used to predetermine sample size. Experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Primers unique to this study were mutated nucleotide lowercase, T7 sequence red, bolded Gs to ensure accurate transcription of entire *U8* snoRNA. Common reverse primer and primers used to transcribe the wild-type precursor form of *U8*, and primers used to genotype *U8-3* mutant zebrafish are reported in Badrock et al. (2020).

8G>A sense: TAATACGACTCACTATA**GGGG**ATCGTCAaGTGG
GATAATCCTTACCTG
19G>C sense: TAATACGACTCACTATA**GGGG**ATCGTCAGGT
GGGATAATgCTTACCTGTTCTCCTC
13C_19G sense: TAATACGACTCACTATA**GGGG**ATCGTCAGG
TGGcATAATgCTTACCTGTTCTCCTC
20C>G sense: TAATACGACTCACTATA**GGGG**ATCGTCAGGT
GGGATAATCgTTACCTGTTCTCCTC
12G>C_20C>G sense: TAATACGACTCACTATA**GGGG**ATCGT
CAGGTGcGATAATCgTTACCTGTTCTCCTC
20C>T sense: TAATACGACTCACTATA**GGGG**ATCGTCAGGTG
GGATAATCtTTACCTGTTCTCCTC
12G>A_20C>T sense: TAATACGACTCACTATA**GGGG**ATCGT
CAGGTGaGATAATCtTTACCTGTTCTCCTC
24C>G sense: TAATACGACTCACTATA**GGGG**ATCGTCAGGT
GGGATAATCCTTAgCTGTTCTCCTC
9G>C_24C>G sense: TAATACGACTCACTATA**GGGG**ATCGTC
AGcTGGGATAATCCTTAgCTGTTCTCCTC
24C>T sense: TAATACGACTCACTATA**GGGG**ATCGTCAGGTG
GGATAATCCTTAtCTGTTCTCCTC
9A_24T sense: TAATACGACTCACTATA**GGGG**ATCGTCAGaTG
GGATAATCCTTAtCTGTTCTCCTC

Native gel electrophoresis

In vitro transcription templates for the mature *U8* snoRNA (136 nt), *U8* snoRNA variants (19G, 20T, and homodimer mutant) and a truncated *U8* snoRNA (nt 1–26 removed) were produced by PCR from the plasmid `pRNA-hU8-GFP` (Badrock et al. 2020) using the primers listed in the next section. PCR products were purified with the PureLink PCR Purification Kit (Invitrogen). In vitro

transcription was performed using the T7 RiboMAX Express Large Scale RNA Production System (Promega) in a 20- μ L reaction using 800 ng PCR product as template. PCR product was removed by RQ1 DNase treatment, and the reaction phenol extracted was then precipitated with ammonium acetate. U8 snoRNAs were resuspended in water at a concentration of 0.4 μ g/ μ L, then purity and size checked by denaturing polyacrylamide gel electrophoresis using UreaGel-6 (National Diagnostics). RNA was visualized using SYBR Green II RNA Gel Stain and a LI-COR Odyssey FC imager using the 600 channel. For native gel electrophoresis, U8 snoRNAs were put in 50 mM Tris-Cl at pH 8, 100 mM NaCl, heated to 95°C, and cooled slowly to room temperature. U8 snoRNAs were run at 4°C on a 6% native gel (37.5:1 Acrylamide:Bis-acrylamide) with TBM buffer (47.5 mM Tris-base, 47.5 mM Boric Acid, 5 mM MgCl₂) at 120 V. RNA was visualized using SYBR Green II RNA Gel Stain and a LI-COR Odyssey FC imager using the 600 channel.

PCR primers

hU8_T7_F
TAATACGACTCACTATAGGGGATCGTCAGGTGGGATAATCC
mat_hU8_R
AATCAGACAGGAGCAATCAGGGTGTGCAAG
phU8T7_19G_F
TAATACGACTCACTATAGGGGATCGTCAGGTGGGATAATgCT
TACCTGTTCTCCTCCTC
phU8T7_20T_F
TAATACGACTCACTATAGGGGATCGTCAGGTGGGATAATCtT
TACCTGTTCTCCTCCTC
phU8T7_delta-homo_F
TAATACGACTCACTATAGGGGATCGTCAccTcccATAATggTTA
CCTGTTCTCCTCCCG
phU8T7_trunc_F
TAATACGACTCACTATAGGGGTTCTCCTCCCGAGGGCAG

Software availability

The software used to perform these analyses can be downloaded from GitHub (<https://github.com/gkudla/hyb>).

The script used to calculate L (the overlap between arms of chimeras) is available at GitHub (https://github.com/gkudla/hyb/blob/master/bin/hyb_overlaps.awk) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Aleksandra Helwak and members of the Kudla laboratory for helpful discussions. This work was supported by the Wellcome Trust (Fellowship 207507 to G.K.) and the Medical Research Council (grants MC_UU_00007/12 to G.K. and MR/V009273/1 to Y.J.C. and R.T.O.).

Author contributions: G.K. and M.M.G. conceived the work, designed the analysis methods, and analyzed the data. J.Y.L. analyzed the data. A.P.B. performed experiments in zebrafish under Y.J.C.'s supervision. R.T.O. performed EMSA experiments. G.K. wrote the paper.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410. doi:10.1016/S0022-2836(05)80360-2

Aw JG, Shen Y, Wilm A, Sun M, Lim XN, Boon KL, Tapsin S, Chan YS, Tan CP, Sim AY, et al. 2016. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Mol Cell* **62**: 603–617. doi:10.1016/j.molcel.2016.04.028

Badrock AP, Ugenti C, Wacheul L, Crilly S, Jenkinson EM, Rice GI, Kasher PR, Lafontaine DJ, Crow YJ, O'Keefe RT. 2020. Analysis of U8 snoRNA variants in zebrafish reveals how bi-allelic variants cause leukoencephalopathy with calcifications and cysts. *Am J Hum Genet* **106**: 694–706. doi:10.1016/j.ajhg.2020.04.003

Bergendahl LT, Marsh JA. 2017. Functional determinants of protein assembly into homomeric complexes. *Sci Rep* **7**: 4932. doi:10.1038/s41598-017-05084-8

Berkhout B, van Wamel JL. 1996. Role of the DIS hairpin in replication of human immunodeficiency virus type 1. *J Virol* **70**: 6723–6732. doi:10.1128/jvi.70.10.6723-6732.1996

Bou-Nader C, Zhang J. 2020. Structural insights into RNA dimerization: motifs, interfaces and functions. *Molecules* **25**: 2881. doi:10.3390/molecules25122881

Cai Z, Cao C, Ji L, Ye R, Wang D, Xia C, Wang S, Du Z, Hu N, Yu X, et al. 2020. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* **582**: 432–437. doi:10.1038/s41586-020-2249-1

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**: D93–D97. doi:10.1093/nar/gkn787

Chworos A, Severcan I, Koefman AY, Weinkam P, Oroudjev E, Hansma HG, Jaeger L. 2004. Building programmable jigsaw puzzles with RNA. *Science* **306**: 2068–2072. doi:10.1126/science.1104686

Ciesiolka A, Jazurek M, Drzalkowska K, Krzyzosiak WJ. 2017. Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front Cell Neurosci* **11**: 97. doi:10.3389/fncel.2017.00097

Clever JL, Miranda D Jr, Parslow TG. 2002. RNA structure and packaging signals in the 5' leader region of the human immunodeficiency virus type 1 genome. *J Virol* **76**: 12381–12387. doi:10.1128/JVI.76.23.12381-12387.2002

Dhir A, Dhir S, Borowski LS, Jimenez L, Teitell M, Rötig A, Crow YJ, Rice GI, Duffy D, Tamby C, et al. 2018. Mitochondrial double-stranded RNA triggers antiviral signalling in humans. *Nature* **560**: 238–242. doi:10.1038/s41586-018-0363-0

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. doi:10.1093/bioinformatics/bts635

Dubois N, Marquet R, Paillart JC, Bernacchi S. 2018. Retroviral RNA dimerization: from structure to functions. *Front Microbiol* **9**: 527. doi:10.3389/fmicb.2018.00527

Ferrandon D, Koch I, Westhof E, Nusslein-Volhard C. 1997. RNA–RNA interaction is required for the formation of specific bicoid mRNA 3' UTR–STAUEN ribonucleoprotein particles. *EMBO J* **16**: 1751–1758. doi:10.1093/emboj/16.7.1751

Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. 2019. The roles of structural dynamics in the cellular functions of RNAs. *Nat Rev Mol Cell Biol* **20**: 474–489. doi:10.1038/s41580-019-0136-0

Gardiner AS, Twiss JL, Perrone-Bizzozero NI. 2015. Competing interactions of RNA-binding proteins, microRNAs, and their targets control neuronal development and function. *Biomolecules* **5**: 2903–2918. doi:10.3390/biom5042903

Geary C, Rothmund PW, Andersen ES. 2014. A single-stranded architecture for cotranscriptional folding of RNA nanostructures. *Science* **345**: 799–804. doi:10.1126/science.1253920

Grabow WW, Jaeger L. 2014. RNA self-assembly and RNA nanotechnology. *Acc Chem Res* **47**: 1871–1880. doi:10.1021/ar500076k

Grosswendt S, Filipchuk A, Manzano M, Klironomos F, Schilling M, Herzog M, Gottwein E, Rajewsky N. 2014. Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol Cell* **54**: 1042–1054. doi:10.1016/j.molcel.2014.03.049

Guo P. 2010. The emerging field of RNA nanotechnology. *Nat Nanotechnol* **5**: 833–842. doi:10.1038/nnano.2010.231

Helwak A, Kudla G, Dudnakova T, Tollervey D. 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**: 654–665. doi:10.1016/j.cell.2013.03.043

Huber RG, Lim XN, Ng WC, Sim AYL, Poh HX, Shen Y, Lim SY, Sundstrom KB, Sun X, Aw JG, et al. 2019. Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nat Commun* **10**: 1408. doi:10.1038/s41467-019-09391-8

Hull CM, Bevilacqua PC. 2016. Discriminating self and non-self by RNA: roles for RNA structure, misfolding, and modification in regulating the innate immune sensor PKR. *Acc Chem Res* **49**: 1242–1249. doi:10.1021/acs.accounts.6b00151

Imig J, Brunschweiler A, Brümmer A, Guennewig B, Mittal N, Kishore S, Tsikrika P, Gerber AP, Zavolan M, Hall J. 2015. miR-CLIP capture of a

- miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nat Chem Biol* **11**: 107–114. doi:10.1038/nchembio.1713
- Ishikawa J, Furuta H, Ikawa Y. 2013. RNA tectonics (tectoRNA) for RNA nanostructure design and its application in synthetic biology. *Wiley Interdiscip Rev RNA* **4**: 651–664. doi:10.1002/wrna.1185
- Ishimaru D, Plant EP, Sims AC, Yount BL Jr, Roth BM, Eldho NV, Pérez-Alvarado GC, Armbruster DW, Baric RS, Dinman JD, et al. 2013. RNA dimerization plays a role in ribosomal frameshifting of the SARS coronavirus. *Nucleic Acids Res* **41**: 2594–2608. doi:10.1093/nar/gks1361
- Jain A, Vale RD. 2017. RNA phase transitions in repeat expansion disorders. *Nature* **546**: 243–247. doi:10.1038/nature22386
- Jambor H, Brunel C, Ephrussi A. 2011. Dimerization of *oskar* 3' UTRs promotes hitchhiking for RNA localization in the *Drosophila* oocyte. *RNA* **17**: 2049–2057. doi:10.1261/rna.2686411
- Jenkinson EM, Rodero MP, Kashner PR, Uggenti C, Oojageer A, Goosey LC, Rose Y, Kershaw CJ, Urquhart JE, Williams SG, et al. 2016. Mutations in *SNORD118* cause the cerebral microangiopathy leukoencephalopathy with calcifications and cysts. *Nat Genet* **48**: 1185–1192. doi:10.1038/ng.3661
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12**: 656–664. doi:10.1101/gr.229202
- Khong A, Matheny T, Jain S, Mitchell SF, Wheeler JR, Parker R. 2017. The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. *Mol Cell* **68**: 808–820.e5. doi:10.1016/j.molcel.2017.10.015
- Kim CH, Tinoco I Jr. 2000. A retroviral RNA kissing complex containing only two G-C base pairs. *Proc Natl Acad Sci* **97**: 9396–9401. doi:10.1073/pnas.170283697
- Kim Y, Park J, Kim S, Kim M, Kang MG, Kwak C, Kang M, Kim B, Rhee HW, Kim VN. 2018. PKR senses nuclear and mitochondrial signals by interacting with endogenous double-stranded RNAs. *Mol Cell* **71**: 1051–1063.e6. doi:10.1016/j.molcel.2018.07.029
- Kosmyna B, Gupta V, Query C. 2020. Transcriptional analysis supports the expression of human snRNA variants and reveals U2 snRNA homeostasis by an abundant U2 variant. bioRxiv doi:10.1101/2020.01.24.917260
- Kudla G, Granneman S, Hahn D, Beggs JD, Tollervey D. 2011. Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc Natl Acad Sci* **108**: 10010–10015. doi:10.1073/pnas.1017386108
- Kudla G, Wan Y, Helwak A. 2020. RNA conformation capture by proximity ligation. *Annu Rev Genomics Hum Genet* **21**: 81–100. doi:10.1146/annurev-genom-120219-073766
- Langdon EM, Qiu Y, Ghanbari Niaki A, McLaughlin GA, Weidmann CA, Gerbich TM, Smith JA, Crutchley JM, Termini CM, Weeks KM, et al. 2018. mRNA structure determines specificity of a polyQ-driven phase separation. *Science* **360**: 922–927. doi:10.1126/science.aar7432
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. doi:10.1038/nmeth.1923
- Li P, Wei Y, Mei M, Tang L, Sun L, Huang W, Zhou J, Zou C, Zhang S, Qin CF, et al. 2018. Integrative analysis of Zika virus genome RNA structure reveals critical determinants of viral infectivity. *Cell Host Microbe* **24**: 875–886.e5. doi:10.1016/j.chom.2018.10.011
- Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS, et al. 2016. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**: 1267–1279. doi:10.1016/j.cell.2016.04.028
- Lyons SM, Gudanis D, Coyne SM, Gdaniec Z, Ivanov P. 2017. Identification of functional tetramolecular RNA G-quadruplexes derived from transfer RNAs. *Nat Commun* **8**: 1127. doi:10.1038/s41467-017-01278-w
- Masliyah G, Barraud P, Allain FH. 2013. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell Mol Life Sci* **70**: 1875–1895. doi:10.1007/s00018-012-1119-x
- Nguyen TC, Zaleta-Rivera K, Huang X, Dai X, Zhong S. 2018. RNA, action through interactions. *Trends Genet* **34**: 867–882. doi:10.1016/j.tig.2018.08.001
- Paillart JC, Shehu-Xhilaga M, Marquet R, Mak J. 2004. Dimerization of retroviral RNA genomes: an inseparable pair. *Nat Rev Microbiol* **2**: 461–472. doi:10.1038/nrmicro903
- Peculis BA. 1997. The sequence of the 5' end of the U8 small nucleolar RNA is critical for 5.8S and 28S rRNA maturation. *Mol Cell Biol* **17**: 3702–3713. doi:10.1128/MCB.17.7.3702
- Qu G, Dong X, Piazza CL, Chalamcharla VR, Lutz S, Curcio MJ, Belfort M. 2014. RNA–RNA interactions and pre-mRNA mislocalization as drivers of group II intron loss from nuclear genomes. *Proc Natl Acad Sci* **111**: 6612–6617. doi:10.1073/pnas.1404276111
- Ramani V, Qiu R, Shendure J. 2015. High-throughput determination of RNA structure by proximity ligation. *Nat Biotechnol* **33**: 980–984. doi:10.1038/nbt.3289
- R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>.
- Roy MD, Wittenhagen LM, Kelley SO. 2005. Structural probing of a pathogenic tRNA dimer. *RNA* **11**: 254–260. doi:10.1261/rna.7143305
- Saldanha AJ. 2004. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**: 3246–3248. doi:10.1093/bioinformatics/bth349
- Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. 2011. A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? *Cell* **146**: 353–358. doi:10.1016/j.cell.2011.07.014
- Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ. 2016. Global mapping of human RNA–RNA interactions. *Mol Cell* **62**: 618–626. doi:10.1016/j.molcel.2016.04.030
- Shetty S, Kim S, Shimakami T, Lemon SM, Mihailescu MR. 2010. Hepatitis C virus genomic RNA dimerization is mediated via a kissing complex intermediate. *RNA* **16**: 913–925. doi:10.1261/rna.1960410
- Shigematsu M, Kirino Y. 2015. tRNA-derived short non-coding RNA as interacting partners of argonaute proteins. *Gene Regul Syst Bio* **9**: 27–33. doi:10.4137/GRSB.S29411
- Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, Luscombe NM, Ule J. 2015. hiCLIP reveals the *in vivo* atlas of mRNA secondary structures recognized by Staufen 1. *Nature* **519**: 491–494. doi:10.1038/nature14280
- Tanaka T, Matsumura S, Furuta H, Ikawa Y. 2016. Tecto-GIRZ: engineered group I ribozyme the catalytic ability of which can be controlled by self-dimerization. *ChemBioChem* **17**: 1448–1455. doi:10.1002/cbic.201600190
- Tosar JP, Gámbaro F, Darré L, Pantano S, Westhof E, Cayota A. 2018. Dimerization confers increased stability to nucleases in 5' halves from glycine and glutamic acid tRNAs. *Nucleic Acids Res* **46**: 9081–9093. doi:10.1093/nar/gky495
- Travis AJ, Moody J, Helwak A, Tollervey D, Kudla G. 2014. Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods* **65**: 263–273. doi:10.1016/j.ymeth.2013.10.015
- Van Treeck B, Parker R. 2018. Emerging roles for intermolecular RNA–RNA interactions in RNP assemblies. *Cell* **174**: 791–802. doi:10.1016/j.cell.2018.07.023
- Van Treeck B, Protter DSW, Matheny T, Khong A, Link CD, Parker R. 2018. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proc Natl Acad Sci* **115**: 2734–2739. doi:10.1073/pnas.1800038115
- Wagner C, Palacios I, Jaeger L, St Johnston D, Ehresmann B, Ehresmann C, Brunel C. 2001. Dimerization of the 3' UTR of *bicoid* mRNA involves a two-step mechanism. *J Mol Biol* **313**: 511–524. doi:10.1006/jmbi.2001.5057
- Wagner C, Ehresmann C, Ehresmann B, Brunel C. 2004. Mechanism of dimerization of *bicoid* mRNA: initiation and stabilization. *J Biol Chem* **279**: 4560–4569. doi:10.1074/jbc.M306511200
- Waters SA, McAteer SP, Kudla G, Pang I, Deshpande NP, Amos TG, Leong KW, Wilkins MR, Strugnell R, Gally DL, et al. 2016. Small RNA interactome of pathogenic *E. coli* revealed through crosslinking of rNase E. *EMBO J* **36**: 374–387. doi:10.15252/embj.201694639
- Wittenhagen LM, Kelley SO. 2002. Dimerization of a pathogenic human mitochondrial tRNA. *Nat Struct Biol* **9**: 586–590. doi:10.1038/nsb820
- Xie Y, Yao L, Yu X, Ruan Y, Li Z, Guo J. 2020. Action mechanisms and research methods of tRNA-derived small RNAs. *Signal Transduct Target Ther* **5**: 109. doi:10.1038/s41392-020-00217-4
- Yu AM, Gasper PM, Cheng L, Lai LB, Kaur S, Gopalan V, Chen AA, Lucks JB. 2021. Computationally reconstructing cotranscriptional RNA folding from experimental data reveals rearrangement of non-native folding intermediates. *Mol Cell* **81**: 870–883.e10. doi:10.1016/j.molcel.2020.12.017
- Zhang J, Ferré-D'Amaré AR. 2014. New molecular engineering approaches for crystallographic studies of large RNAs. *Curr Opin Struct Biol* **26**: 9–15. doi:10.1016/j.sbi.2014.02.001
- Zhang M, Li K, Bai J, Velema WA, Yu C, van Damme R, Lee WH, Corpuz ML, Chen JF, Lu Z. 2021a. Optimized photochemistry enables efficient analysis of dynamic RNA structures and interactomes in genetic and infectious diseases. *Nat Commun* **12**: 2344. doi:10.1038/s41467-021-22552-y
- Zhang Y, Huang K, Xie D, Lau JY, Shen W, Li P, Wang D, Zou Z, Shi S, Ren H, et al. 2021b. *In vivo* structure and dynamics of the SARS-CoV-2 RNA genome. *Nat Commun* **12**: 5695. doi:10.1038/s41467-021-25999-1
- Zhang M, Hwang IT, Li K, Bai J, Chen JF, Weissman T, Zou JY, Lu Z. 2022. Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers. *Genome Res* (this issue) **32**: 968–985. doi:10.1101/gr.275979.121
- Ziv O, Gabryelska MM, Lun ATL, Gebert LFR, Sheu-Gruttadauria J, Meredith LW, Liu ZY, Kwok CK, Qin CF, MacRae IJ, et al. 2018. COMRADES determines *in vivo* RNA structures and interactions. *Nat Methods* **15**: 785–788. doi:10.1038/s41592-018-0121-0
- Ziv O, Price J, Shalamova L, Kamenova T, Goodfellow I, Weber F, Miska EA. 2020. The short- and long-range RNA–RNA interactome of SARS-CoV-2. *Mol Cell* **80**: 1067–1077.e5. doi:10.1016/j.molcel.2020.11.004

Received June 18, 2021; accepted in revised form March 18, 2022.



Global mapping of RNA homodimers in living cells

Marta M. Gabryelska, Andrew P. Badrock, Jian You Lau, et al.

Genome Res. 2022 32: 956-967 originally published online March 24, 2022

Access the most recent version at doi:[10.1101/gr.275900.121](https://doi.org/10.1101/gr.275900.121)

Supplemental Material <http://genome.cshlp.org/content/suppl/2022/05/02/gr.275900.121.DC1>

Related Content **Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers**
Minjie Zhang, Irena T. Hwang, Kongpan Li, et al.
Genome Res. May , 2022 32: 968-985

References This article cites 72 articles, 18 of which can be accessed free at:
<http://genome.cshlp.org/content/32/5/956.full.html#ref-list-1>

Articles cited in:
<http://genome.cshlp.org/content/32/5/956.full.html#related-urls>

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Affordable, Accurate
Sequencing.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>