MANUSCRIPT TITLE

CRIS: A Centralized Resource for High-Quality RNA Structure and Interaction Data in the AI Era

AUTHORS

Wilson H. Lee[1], Kongpan Li[1], Jianhui Bai[1], Minjie Zhang[3], Aman Sharma[1], Zhipeng Lu[1,2]

AFFILIATIONS

[1] Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences
[2] Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90033
[3] Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

ABSTRACT

The Crosslinking-based RNA Interactomes and Structurome (CRIS) database is a unified platform for RNA structure and interaction data, combining rigorously curated datasets, standardized workflows, and user-friendly tools. While advancements in experimental and computational methods expand our understanding of RNA structure and function, challenges in reproducibility, quality control, and data accessibility remain prevalent due to inconsistencies in data processing. As a result, many existing databases lack breadth in their data sources, are inconsistent in their processing pipelines, and limit accessibilities for end users. Designed for the AI-driven era, CRIS addresses critical gaps in reproducibility and accessibility by harmonizing crosslinking-based sequencing data (e.g., PARIS, SHARC) with built-in quality metrics, enabling robust comparative analyses and machine learning applications. Its structured workflow, built-in quality metrics, and detailed visualization guides ensure reproducibility and transparency while integrating seamlessly with existing experimental pipelines. By bridging complex RNA datasets with accessible computational tools, CRIS accelerates RNA biology and drug development, establishing a new standard for the field.

INTRODUCTION

Once considered only as a messenger of gene expression, RNA is now recognized as a key regulator of cellular processes across diverse organisms, from animals to viruses (1,2). Recent advancements in RNA biology have unveiled complex structure-dependent functions and networks of RNA-RNA and RNA-protein interactions that drive gene regulation, splicing, and epigenetic modifications, positioning RNA as a critical influencer of cellular function and disease (2–9).

RNA's therapeutic potential is now undeniable (10–12). RNA-based therapies, including mRNA vaccines against SARS-CoV-2 and antisense oligonucleotides (ASOs) for genetic disorders, underscore RNA's transformative therapeutic applications (10,13–16). Insights into viral RNA structures have also informed antiviral strategies, including targeting internal ribosome entry sites (IRES) in pathogens like Hepatitis C virus or poliovirus, and the 3'-UTR in West Nile virus to disrupt viral replication (17–19). However, these advances rely on precise RNA structural insights – for example, the SARS-CoV-2 vaccine efficacy depends on 5'-UTR stability and ASO targeting requires knowledge of splice-site accessibility.

Despite these advances, characterizing RNA structures in vivo remains a significant challenge (1,20–22). While traditional high-resolution methods like X-ray crystallography, cryo-electron microscopy, and NMR provide precise structural insights, they are limited by technical constraints, sample requirements, and potential artifacts (23–27). High-throughput chemical probing techniques such as PARIS and SHARC combine crosslink-ligation chemistry with next-generation sequencing to map RNA interactions in their native context (28,29). Thus, the rapid generation of data from such methods highlights the need for rigorous computational analysis to ensure reproducibility, quality control, and accurate biological interpretation (22,25,30,31).

Due to the complexity of RNA structural data across various methodologies, there is a need for standardized, high-quality pipelines that support experimental reproducibility and scalable computational modeling (32–35). Most current databases lack standardized data curation methods, built-in quality metrics, and integration capabilities for machine learning frameworks, limiting their utility in predictive modeling, large-scale data mining, and novel RNA structure discoveries (32,36).

CRIS integrates comprehensive datasets with robust analysis steps and visualization tools, providing a foundation for accelerated discoveries in RNA biology, therapeutic development, and the broader field of translational research. CRIS bridges the divide between complex RNA data and actionable insights. At the same time, it further serves as a training platform for machine learning models. With built-in quality metrics, user-friendly guides, and seamless integration with widely used bioinformatics tools, CRIS empowers researchers to explore RNA structure and function with unprecedented precision and scalability. In doing so, CRIS establishes a new standard for RNA informatics, fostering innovation and transformative breakthroughs across biological and medical sciences.

DATABASE CONTENT

CRIS Database Overview

The Crosslinking-based RNA Interactomes and Structurome (CRIS) database represents a significant advancement in RNA structure research by harmonizing crosslink-ligate RNA-sequencing data from diverse experimental pipelines, encompassing both animal and viral genomes (**Figure 1**). CRIS emphasizes rigorous quality control (QC) and consistent data processing, addressing a key gap in existing databases where inconsistent or unavailable QC hinders reliable cross-study comparisons. By providing processed data files (i.e., aligned, crosslink classified, duplex groups) ready for further analyses or comparisons, CRIS enables precise exploration of RNA structures and their functional implications.

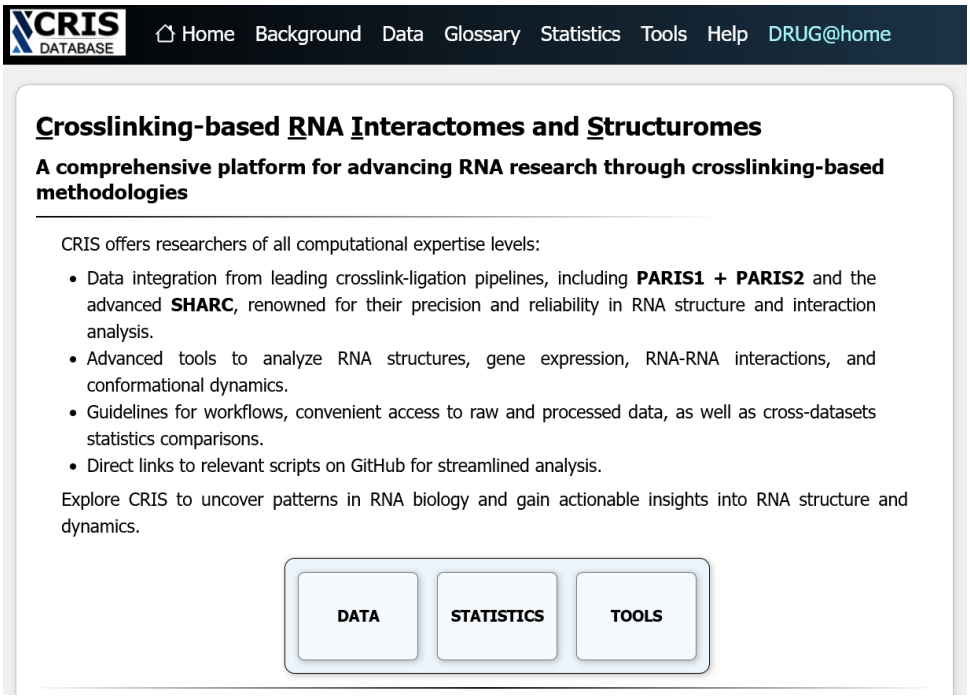

Figure 1. Landing page of the CRIS database. CRIS provides a user-friendly interface for accessing rigorously pre-processed RNA structure and interaction datasets. The navigation bar features intuitive access to data, statistical reports, and comprehensive tutorials for relevant computational tools, ensuring seamless integration and reproducibility in RNA research.

CRIS's web interface caters to users across experience levels, accommodating entry-level users who want to quickly browse dataset statistics, and providing example command-line scripts for advanced users. The landing page provides an overview of CRIS and its capabilities, while sections such as the data and tools pages allow efficient navigation of workflows. A glossary page standardizes terminology, and the quality control statistics page summarizes experimental data metrics, enhancing transparency. The tools page includes a quick-start guide and step-by-step instructions with command-line examples, enabling users to move from data download to publication-ready results with minimal complexity.

Data-rich intermediary analysis files detailing crosslinked RNA (XLRNA) configurations (e.g., gap1, gapm files) are provided, but CRIS does not store raw sequencing data. CRIS is not a service provider; currently it does not host source data or provide computational power for sequencing data analysis. Furthermore, the processed data is not intended to support global meta-analyses, as such approaches often lead to misleading comparisons. Rather, its purpose is harmonizing high-quality, rigorously curated datasets to empower researchers to conduct their own deeper analyses with confidence. CRIS prioritizes reproducibility by offering transparent workflows and detailed documentation. Its web interface includes intuitive download options, statistical summaries (e.g., read counts, segment lengths), and tutorials for data visualization through the IGV viewer. Users can follow the data from start-to-finish, guided stepwise from data collection to visualization without the need for a coding background.

Data Curation

CRIS includes a broad range of crosslink-ligation sequencing datasets, including PARIS1/2 (GSE74353, GSE149493), SPLASH (PRJNA318958), LIGR-seq (GSE80167), COMRADES (E-MTAB-6427), SHARC (GSE167812), hiCLIP (E-MTAB-2937, E-MTAB-2940, E-MTAB-2941), RIC/cRIC-seq (GSE127188,

GSE210583), and extensive miRNA data, selected for their variety in crosslinking technology and strengths in capturing RNA structures and interactions. These datasets cover diverse biological contexts, from various cell lines and model organisms to virus-host interactions, enhancing CRIS's utility for comparative analyses across species and experimental conditions. We reprocessed all public datasets (e.g., PARIS, RIC/cRIC-seq) from raw FASTQ files using a standardized pipeline. To ensure consistency, files were renamed with a structured convention: [Species]-*[CellLine]*-[Crosslinker]-*[LigationTime]*-[RNaseTreatment] (e.g., HS-HEK293T_Amo-0.5_T4-24h_exo-0h). This enables intuitive dataset selection and minimizes metadata ambiguity. For example, "HS-HEK293T_Amo-0.5_T4-24h_exo-0h" refers to PARIS data generated using samples of homo sapiens origin, HEK293T cell line, amotosalen crosslinked at 0.5 mg/mL, proximally ligated for 24 hours with T4 RNA ligase, and not trimmed with RNase R exonuclease.

Internally generated XLRNA data for PARIS and SHARC experiments were preprocessed for adapter removal, demultiplexing and PCR duplicate removal using a wrapper script that integrates custom scripts from the PARIS/SHARC computational pipeline for streamlined preprocessing (**Figure 2**) 5/5/25 1:59:00 PM. The processing tools (e.g., Trimmomatic, STAR) were selected for their flexibility in parameter customization, allowing for tailored analysis of crosslinking datasets, an area where commercially available software packages have processing limitations. All datasets available on CRIS were aligned to a modified "masked" reference genome using previously optimized STAR parameters prior to duplex group (DG) assembly using CRSSANT scripts. This approach ensures high-precision mapping and standardized classification of RNA interactions across species and experimental conditions.

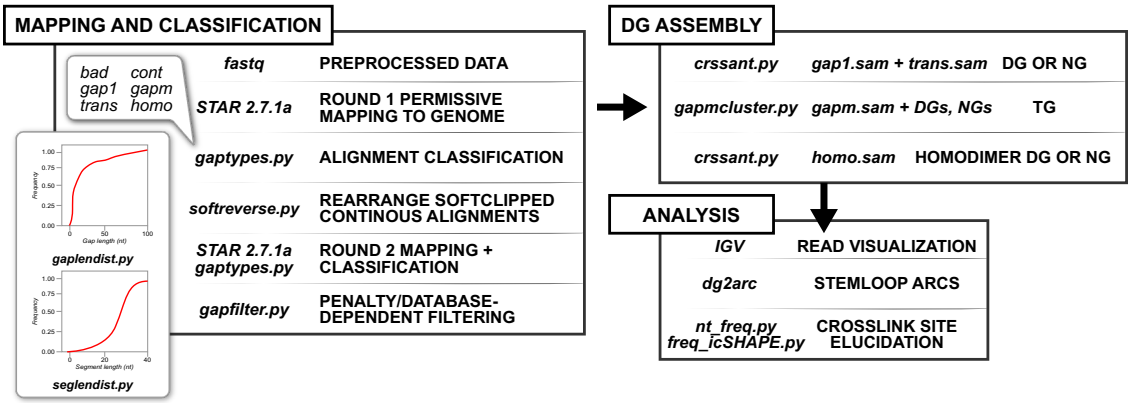

Figure 2. Overview of the data workflow – from raw sequencing files and pre-processing steps to prepared files available on demand from CRIS.

To reduce the storage footprint of XLRNA sequencing data, we developed bam2bedz, a specialized lossy compression tool tailored to extract essential information from RNA crosslinking alignments. This approach selectively extracts the read start position, CIGAR string, MD tag, read count, and strand information for simple gap-typed reads (prigap1, prigapm, prihomo) and stores the data in a compressed, modified BED format. For trans-gapped reads (pritrans), the two RNA arms are recorded in a modified BEDPE file that retains the chromosome and start position of each arm, CIGAR string, MD tag, read count, and strand information for both arms. Continuous reads (pricont), which resemble conventional RNA-seq coverage profiles, are processed and stored as compressed BEDgraph files that track per-base coverage. By selectively retaining only the most relevant features, bam2bedz significantly reduces data size while maintaining analytical utility.

Unlike reference-based formats like CRAM, bam2bedz eliminates redundancy from full sequences and removes reference reconstruction for visualization (37). Since the XLRNA approach for structural visualization do not rely on per-base sequence information, bam2bedz's design focuses on retaining positional and topological information. The resulting compressed files are lightweight, tab-delimited, and human-readable, allowing rapid inspection and manipulation using tools such as bedtools, awk, pandas, and tabix (for fast querying via bed.gz indexes). Critically, the format can readily be expanded using a reference genome to restore alignment-level fidelity for most structural analyses, providing flexibility for both storage and interpretation. To empirically benchmark compression efficiency, we applied bam2bedz to multiple datasets (e.g., PARIS, SHARC), sampling increasing numbers of mapped reads (e.g., 1 million to 100 million) using samtools and shuf, then compressing them using bam2bedz.

Across multiple datasets including PARIS and SHARC, bam2bedz achieved consistent and substantial reductions in file size across all XLRNA reads. In this case, compression efficiency is assessed by calculating the bytes per read (BPR) for both BAM and BED outputs, excluding BAM headers to ensure a direct comparison. Since pricont, prigap1, and pritrans files are the largest files in a typical XLRNA dataset, we benchmarked bam2bedz on subsets of these data.

For prigap1, 1000 randomized reads from PARIS data containing read name with experimental conditions, tags, CIGAR strings, and alignment metadata, average ~350 BPR. The equivalent BED lines contain ~30 BPR, resulting in an approximate 12-fold reduction in BPR where reads are unique and contain few or no overlaps. Most datasets in CRIS show the expected ~12-fold size reduction, since reads are short and only have mildly repetitive regions. For synthetic datasets with repetitive sequences or where sequencing coverage is near saturation such that every position in the genome is overlapped by multiple reads, we observe up to ~20-fold reduction in size. This can be visualized by plotting RPK (reads per kilobase) against compression ratio, where:

$$RPK = \frac{Number\ Reads \times Average\ Read\ Length}{Genome\ Size\ (in\ kb)}$$

High RPK values indicate denser coverage, which correlates with greater compression due to repeated alignments. In the example of SHARC data compression, 10 million ~50-nt reads mapped to the hg38 reference genome (~3.1 million kb) yields an RPK of approximately 155, suggesting low genomic saturation and high compressibility (**Figure 3a**). Decreasing the number of reads dramatically increases the uniqueness of each read, therefore reducing the compression ratio. The same logic can be applied to prigapm and prihomo reads, showing a decrease in file size that plateaus for higher read numbers.

When processing only ribosomal RNA (rRNA) reads, we note an increase in compression, up to ~13-fold BPR, due to the nature of repeated alignments in the densely covered rRNA regions (**Figure 3b**). In contrast, non-rRNA reads scale less favorably, with compression not exceeding 9-fold for up to 100 million reads (~1500 RPK) (**Figure 3c**). Compression scales logarithmically with read depth due to increased alignment redundancy but plateaus as data entropy declines and redundancy is exhausted. Compression of pricont reads follows a similar trend of logarithmic increase in compression, not exceeding ~3-fold BPR for combined reads (**Figure 3d, f**). This limited gain is expected, as reads mapping to distinct chromosomes retain separate coverage tracks in the BEDgraph format, preserving chromosome-specific detail. In contrast, rRNA-only reads compress more efficiently and scale linearly with read count, since all alignments are confined to a single "chromosome" in the custom STAR-masked genome used for mapping (**Figure 3e**).

Pritrans reads, which capture long-range or inter-molecular RNA interactions, require a distinct compression strategy due to their dual-arm structure. Unlike simple gapped or continuous reads that map to a single genomic locus, pritrans reads involve two separate alignments, each corresponding to one RNA arm on different chromosomes or distant loci within the same transcript. By compressing both arms into a single structured record, bam2bedz retains topological integrity essential for downstream structure inference, clustering, and visualization of long-range RNA contacts. The data exhibits a log decay pattern in file size compression, with ~1 million reads reducing ~8-fold BPR (**Figure 3g, h**).

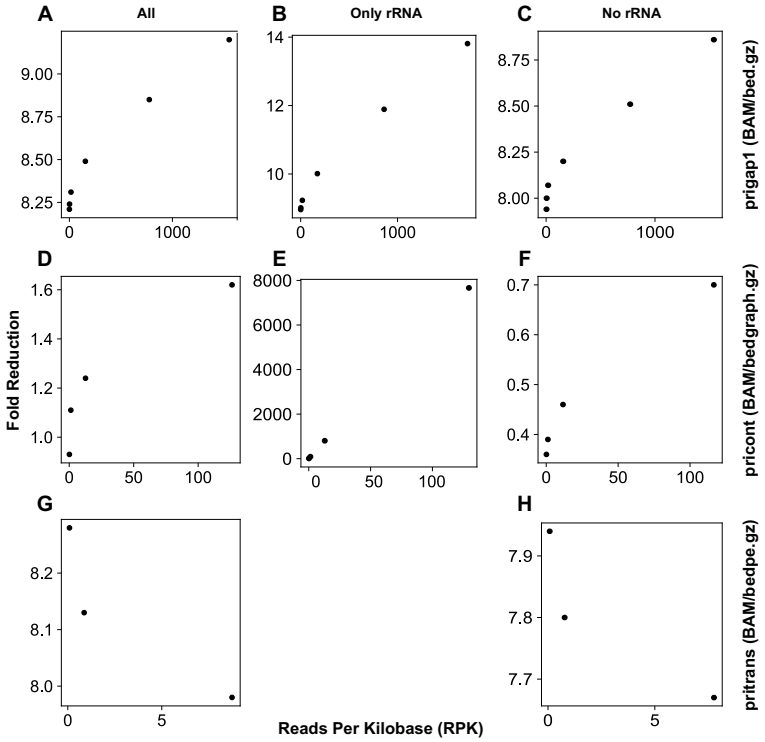

Figure 3. Compression performance of bam2bedz across read types and genomic content. Plots show the relationship between RPK (Reads Per Kilobase) and file size reduction for different classes of RNA-seq reads after bam2bedz compression. Each subpanel depicts trends across increasing read numbers for: (A) all prigap1 reads, (B) prigap1 reads mapping exclusively to rRNA regions, and (C) prigap1 reads excluding rRNA. Compression scales logarithmically with RPK, reflecting increasing alignment redundancy. (E, F) pricont reads show limited compression (~3-fold BPR) due to retention of position specific details in BEDgraph format. (E) pricont rRNA-only reads, by contrast, compress more efficiently and linearly with read count because all alignments localize to a single pseudo-chromosome in the STAR-masked genome. (G, H) pritrans reads represent dual-arm, long-range interactions and require structural pairing during compression. This category shows a log-decay compression profile, achieving up to ~8-fold BPR with ~1 million reads.

Genome References

Users are provided access to three categories of genome references to support accurate and consistent alignment of sequencing data across research applications.

1) “Standard”: Unmodified genomes human (hg38) and mouse (mm10) are linked, ensuring compatibility with common RNA research pipelines.

2) “Curated”: Repetitive-element masked genomes to improve mapping accuracy (30). This modified reference is particularly effective for studying noncoding and repetitive RNA sequences.
3) “Special”: Custom designed genome for RNA-RNA interaction studies. Manual preparation of concatenated interacting RNA genes into a single reference file, allowing mapping and analysis of complex RNA-RNA interactions within structured multi-gene contexts.

Data Quality Control

All downloaded datasets were first subjected to FastQC analysis to assess sequencing quality, overrepresented sequences, and adapter content. The resulting FastQC reports are accessible via the data tab, which also includes plots of gap and segment length distributions, hallmark measures of crosslinking data quality. A comprehensive statistics report generated using scripts to automate the extraction and aggregation of sequencing and alignment metrics from data processed through the CRSSANT pipeline. It parses read counts from raw FASTQ files, alignment metrics from STAR’s Log.final.out files, and gap analysis results from SLURM output logs generated during gap type processing. These metrics are consolidated into a structured CSV file, ensuring consistent formatting for downstream analysis and reporting. The sequencing statistics report enables reproducible, scalable assessment of sequencing quality, mapping performance, and gap-type distributions across datasets.

We also introduce the “bam2biotype” tool, which is used to classify sequencing reads into specific RNA biotypes, such as mRNA, rRNA, tRNA, snRNA, and viral RNA, based on the reference annotation. This classification is essential because different RNA classes have distinct structural and functional properties, and not all are relevant to every experimental context. By partitioning data, researchers can focus analyses on biologically meaningful targets, filtering out abundant rRNA to prioritize mRNA or miRNA interactions, thereby reducing background and improving interpretability. Moreover, biotype separation enhances quality control by enabling evaluation of sample composition across experiments and improves computational efficiency by streamlining downstream workflows. In essence, bam2biotype ensures that data analysis is precise, focused, and tailored to specific biological questions.

As a part of CRIS’s initiative to standardize the processing of crosslinked-RNA data for CRSSANT analysis, we integrate “genes2bed”, a script designed for filtering genes used in DG clustering by read count thresholds and extracting their corresponding genomic intervals in BED format. The data processing pipeline guarantees that structural and interactional insights are supported by robust experimental evidence, since the DGs are formed based on the already existing read information. This tool ultimately simplifies the process of clustering RNA duplexes and enables users to process data without needing to manually prepare the BED file necessary for CRSSANT.

Data Visualization

Integration of XLRNA data with genome browsers like IGV is typically not streamlined, lacking clear guidelines for simple data visualization, local feature extraction, and overlay with conservation tracks. The updated bedpe2arc12 script, introduced as part of the CRIS toolkit, enhances the visualization of RNA duplex groups (DGs) identified by CRSSANT by converting BEDPE records into BED12 format for arc-based genome browser display (**Figure 4a, b**). Unlike its predecessors (dg2arc.py and bedpetobed12.py), this version resolves prior alignment conversion issues and introduces a dynamic RGB coloring scheme to reflect DG score intensity (**Figure 4c**). Specifically, the script processes each DG group independently, cycling through a predefined set of base colors to assign a unique hue per group. Within each group, DGs are sorted by raw score, and their color intensity is modulated using a gamma-corrected log transformation of the score (**Figure 4d**). The formula applied is:

$$Intensity = (1 - \frac{(\log(1+S))}{\log(1+S_{max})})^{\gamma}$$

Where S is the DG score for the arc, S_{max} is the highest score in the group, and γ is the gamma correction factor (default = 1.5). Each RGB channel is then scaled by the computed intensity:

$$C_i^{adjusted} = \lfloor C_i \cdot Intensity\ Scale \rfloor\ for\ i\ \in \{R, G, B\}$$

This results in higher-scoring DGs appearing darker, drawing visual attention to biologically significant interactions. In addition to coloring, the script ensures strand concordance and precise block coordinate mapping, enabling accurate spatial representation of RNA-RNA interactions. Altogether, these improvements support more intuitive data interpretation and facilitate the prioritization of high-confidence DGs in large-scale structural studies.

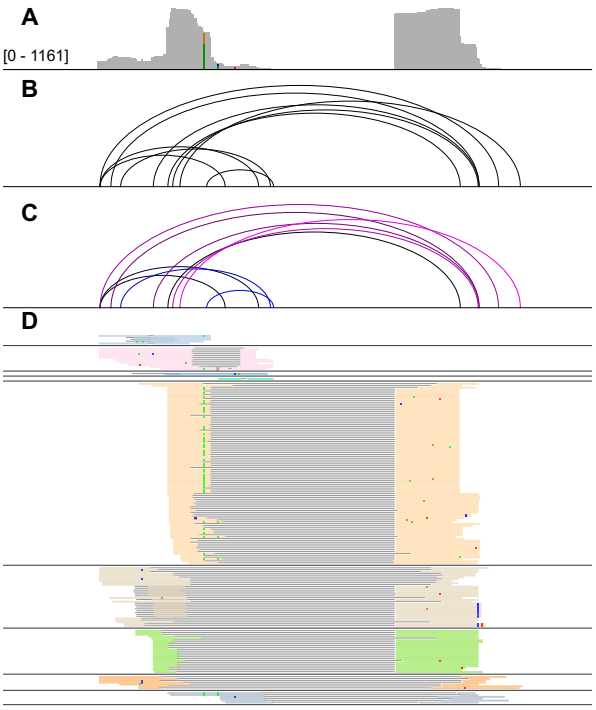

Figure 4. Visualization of RNA duplex groups (DGs) using the updated bedpe2arc12 script from the CRIS toolkit. (A) Read coverage track over a representative genomic region. (B) Arc visualization of DGs generated from BEDPE records using the legacy format. (C) Enhanced arc rendering using bedpe2arc12, showing group-wise coloring and score-modulated arc intensity. Darker arcs correspond to higher-scoring DGs, calculated using a gamma-corrected log transformation of the DG score. (D) Corresponding block-level alignments for DG pairs. Each color-coded group represents interactions processed independently, allowing comparison across subgroups.

Example Analysis from CRIS Data

A central challenge in transcriptome-wide RNA structure analysis lies in integrating data from distinct methodologies, each with its own biases, resolution, and structural assumptions. CRIS addresses this

challenge by transforming raw crosslinking and chemical probing data into a common coordinate framework, enabling rigorous comparison across methods. To illustrate this, we used *XIST*, a long noncoding RNA known for its complex regulatory roles and architectural modularity, as a model transcript for cross-platform structural analysis (**Figure 5**).

CRIS unifies PARIS (long-range interactions) and icSHAPE (local flexibility) data into a single genomic coordinate system, resolving a longstanding integration challenge. SHAPE-based methods such as produce nucleotide-resolution reactivity scores and indirectly infer secondary structure. However, these methods cannot resolve long-range or noncanonical pairings, nor do they directly report on duplex formation. In contrast, XLRNA techniques like PARIS physically capture base-pairing interactions, producing duplex maps that reflect actual RNA-RNA contacts across large sequence distances. By mapping both data types to the same transcript and representing them as aligned genome browser tracks, CRIS eliminates format disparities and allows positionally accurate overlay and comparison. This removes ambiguity in interpreting structural signals (target binding vs. base pairing or masking) and enables inference of local and global RNA architectures.

In the case of *XIST*, CRSSANT analysis of PARIS data revealed modular and highly structured intra-exonic duplexes (**Figure 5A**), many of which coincide with drops in icSHAPE reactivity (**Figure 5B**), validating their base-paired nature. Duplexes also map to peaks in PARIS read coverage, increasing confidence in RNA contact points (**Figure 5C**), further supporting their structural relevance. The structural domains identified by CRIS (**Figure 5D**) delineate assumptions of RNA folding, reflecting the likely hierarchical organization of XIST required for its function in X-chromosome silencing. While icSHAPE alone can suggest regions of low flexibility, only XLRNA data can determine whether these correspond to local helices or to distant interactions forming tertiary structures - critical distinctions for understanding RNA functionality.

More broadly, this analysis exemplifies a mode of structural inference made possible by CRIS: a multiscale, multi-modal analysis using both chemical probing and crosslinking data. Previous approaches typically relied on qualitative comparisons, overlaying reactivity scores and structural predictions without grounding them in empirically verified long-range interactions. CRIS allows researchers to identify such types of structurally insulated regions, cooperative domains, and potential tertiary motifs.

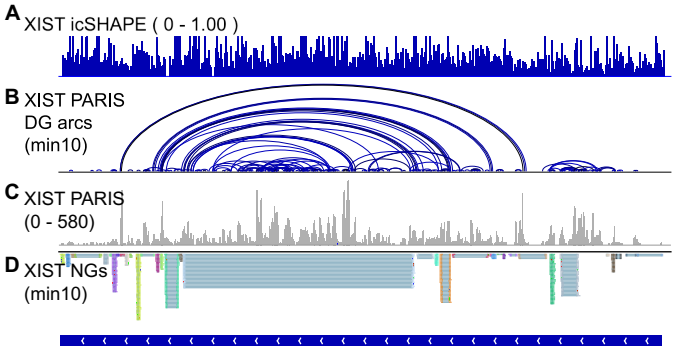

Figure 5. Multitrack representation of RNA structural features across an exon of the XIST locus. (A) Duplex group arcs identified by CRSSANT, representing predicted RNA-RNA base-pairing interactions along the XIST transcript; the span and density of arcs reflect the complexity of molecular interactions. (B) icSHAPE reactivity profile plotted as nucleotide-level scores (0 - 1.0), indicating single-strandedness and structural flexibility. (C) PARIS sequencing coverage showing read depth (0 - 580 reads), highlighting regions with high RNA-RNA contact frequency. (D) Collapsed, non-overlapping duplex groups from

PARIS, displaying independently folded structural modules. Genome browser-style track alignment facilitates cross-comparison of complementary structural data.

To further demonstrate CRIS's ability to resolve method-specific biases, we compared protein-agnostic and protein-centric datasets with the direct comparison of RNA duplexes identified by PARIS1 and STAU1 hiCLIP in the *SRSF1* exon 4 region, respectively **(Figure 6**). These data represent fundamentally different strategies for probing RNA structure: PARIS captures transcriptome-wide RNA duplexes without protein bias, while hiCLIP targets RNA structures bound specifically by STAU1, a double-stranded RNA-binding protein (dsRBP).

Such a comparison is inherently challenging due to differences in experimental design, resolution, and mapping assumptions. However, by processing both datasets through the CRIS pipeline, we normalize these differences, harmonizing data structure, clustering parameters, and genomic alignment, thus making the comparison fair. CRIS ensures that duplexes from both PARIS and hiCLIP are grouped, quantified, and visualized using the same computational framework, reducing confounding technical variability and emphasizing true biological differences.

dsRBPs such as STAU1 influence RNA metabolism by binding structured regions and occluding single-stranded regulatory elements. While CLIP-based methods allow one to study individual proteins, they do not provide a global picture of RNA structure or binding context. In contrast, PARIS-like approaches detect RNA duplexes genome-wide, enabling protein-agonistic detection of RNA structures. Here, we show that PARIS not only recapitulates all duplexes detected by STAU1 hiCLIP (**Figure 6A vs 6D**) but also reveals a broader set of additional duplexes in the same transcript region (**Figure 6B, 6C**), which likely represent either other dsRBP targets or functional RNA structures that evade detection in CLIP due to transient binding or low affinity. This complementary coverage demonstrates how CRIS enables data integration across methods with differing strengths, providing a richer and more accurate picture of the underlying RNA structure landscape.

The ability to unify PARIS and CLIP data at the level of RNA duplex groups is a significant improvement to the analysis pipeline. Historically, comparisons across structural and CLIP-based data were largely qualitative, often limited to overlaying raw read densities or peak annotations. CRIS enables a quantitative and interpretable structural comparison at the duplex level, showing not just whether a region is structured, but exactly how those structures align or diverge between methods. This reveals a broader structural context around dsRBP binding and enables hypothesis generation regarding cooperative or competitive binding by other factors. For example, the additional duplexes detected only by PARIS suggest possible alternative binding sites, structural switches, or regulatory elements masked in the protein-centric view of CLIP.

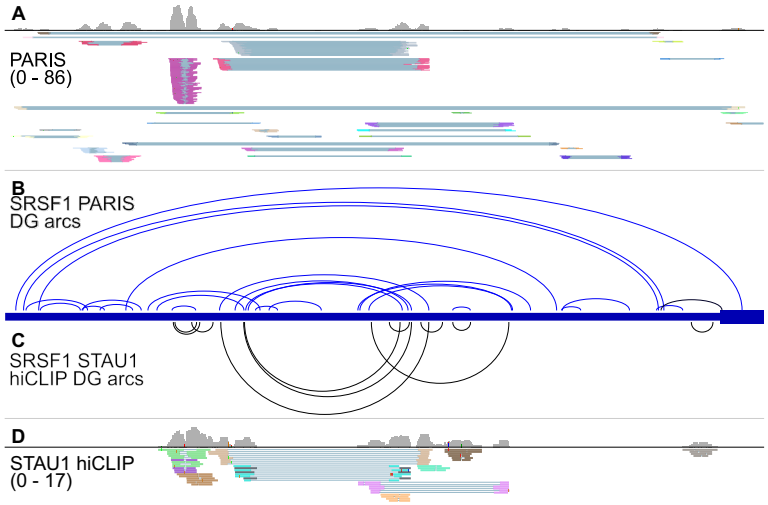

Figure 6. Comparative analysis of RNA duplexes identified by PARIS1 and STAU1 hiCLIP in the SRSF1 exon 4 region. (A) RNA duplex groups detected by PARIS1 in HeLa cells using CRSSANT, aligned to the transcriptome. (B) Arc diagram of PARIS1 duplex groups, highlighting long-range interactions and structural complexity. (C) STAU1 hiCLIP duplex groups, combined from three independent sequencing libraries and visualized as arcs. (D) Genomic alignment of STAU1 hiCLIP duplex groups, highlighting specific dsRBP-bound RNA structures. PARIS1 captures all previously known STAU1-bound duplexes and reveals additional RNA structures undetected by hiCLIP.

CRIS data also enables the discovery of novel RNA-RNA interaction networks by unifying data across distinct experimental inputs and sample contexts. To systematically map small nucleolar RNA (snoRNA) targets, we previously processed PARIS2 datasets derived from total RNA, chromatin-associated RNA, and antisense-oligo enriched snoRNA libraries using the CRSSANT pipeline (38). These data include interactions from diverse human cell lines and iPSC-derived cellular lineages, capturing a wide spectrum of cellular states and RNA populations. By integrating these three modalities, CRIS supports sensitive and comprehensive mapping of snoRNA-guided interactions with their targets. CRSSANT clustering identified >7,000 high-confidence RNA-RNA interactions involving hundreds of snoRNAs and their putative targets, revealing an unexpectedly expansive interaction network. Most interactions involved canonical targets such as rRNAs and snRNAs, but remarkably, CRIS analysis also uncovered extensive base-pairing between C/D box snoRNAs and nuclear-encoded tRNAs, previously underexplored as potential guide targets. Duplexes were detected consistently across both total and chromatin-associated RNA libraries and were further enriched in the antisense-capture datasets, underscoring their biological relevance. Visualization of the snoRNA-tRNA network reveals modular and cooperative targeting patterns, with specific snoRNAs forming interactions with distinct sets of tRNAs, often clustering by isotype or anticodon family. This suggests the possibility of snoRNA-mediated structural regulation of tRNA maturation or function, a hypothesis not easily inferred from sequence alone. Importantly, the identified interactions are not simply the result of sequence complementarity. This standardization allows for direct comparison of snoRNA interaction profiles across conditions and enables the identification of conserved, context-specific, or novel targets with high confidence.

While CRIS prioritizes experimental data integration, its high quality of standardized datasets and tools uniquely enable *de novo* RNA structure prediction. This is particularly useful for non-model organisms, engineered RNAs, or unannotated transcripts where crosslinking data are absent. As exemplified above, CRIS presents duplex groups (DGs) as arc-type graphs, where nodes correspond to RNA segments and edges reflect experimentally validated interactions. These graphs are preprocessed with biologically

meaningful features (e.g., segment length, conservation scores, DG confidence weights) and can readily be converted into machine-learning-ready formats (e.g., PyTorch Geometric files). Researchers can immediately train graph neural networks (GNNs) to predict novel RNA-RNA interactions or fine-tune transformer models using XLRNA-derived contact probabilities as training targets (**Figures 4-7**). For 3D structure modeling, CRIS provides scripts to convert DGs into pairwise distance constraints compatible with tools like AlphaFold-RNA (29). As AI-driven structural biology advances, CRIS datasets serve as experimentally grounded benchmarks to validate and iteratively improve predictive algorithms.

METHODS

Data pre-processing, read alignment, and statistical analysis were performed following the CRSSANT pipeline (30). Initially, all raw FASTQ files were inspected with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess sequencing quality and adapter content. Where appropriate, adapter sequences were trimmed from the 3' ends using Trimmomatic, and multiplexed libraries were demultiplexed and split by barcode using splitFastq.pl from the icSHAPE pipeline (https://github.com/qczhang/icSHAPE). PCR duplicates were removed using readCollapse, after which a second round of Trimmomatic trimming was completed to remove residual 5' adapter sequences.

Reads were aligned using STAR v2.7.1a (https://github.com/alexdobin/STAR) with permissive alignment parameters to maximize sensitivity for chimeric reads (38). Primary alignments were extracted and initially classified into gap types using gaptypes.py from the CRSSANT pipeline (https://github.com/zhipenglu/CRSSANT/). Reads identified as continuous (i.e., lacking internal gaps) were further examined for soft-clipped regions and potential reverse chimeras. These were processed with softreverse.py and remapped using STAR, followed by a second extraction and classification of primary alignments. Outputs from both rounds of mapping were merged and reads with splice junctions or short deletions (1-2 nt gaps) were filtered out using gapfilter.py to refine the prigap1 and prigapm datasets.

Finally, updated gaplendist.py and seglendist.py scripts from the rna2d3d repository (https://github.com/whl-usc/rna2d3d) were used to plot the gap-length and segment-length distributions, respectively. These distributions, along with summary statistics, were recorded for each dataset using seqstats to evaluate crosslinking quality and structural features of RNA interactions.

CONCLUSIONS

The Crosslinking-based RNA Interactomes and Structurome (CRIS) database makes a substantial contribution to the RNA research community by offering rigorously analyzed, pre-processed RNA structure and interaction datasets. Prioritizing data quality, transparency, and usability, CRIS addresses challenges like inconsistent quality control and the computational demands typically associated with raw data analysis. Its intuitive interface, detailed documentation, and ready-to-use datasets allow researchers to focus directly on biological interpretation and hypothesis generation without requiring prior coding knowledge or heavy computational resources.

By unifying disparate pipelines and XLRNA data, CRIS enables both quantitative and qualitative comparisons previously impractical, accelerating discoveries in RNA-based drug design and synthetic biology. Its emphasis on transparency and data integrity highlights its reliability, while promoting broader accessibility within the research community.

Designed as a dynamic and evolving platform, CRIS continually expands its features, datasets, and analytical capabilities to meet emerging research needs. Planned enhancements aim to improve user experience and data exploration through tools like interactive tables summarizing key RNA-RNA interactions. While the current focus highlights crosslink-ligation RNA sequencing datasets, future updates will integrate a broader diversity of data, such protein-RNA interactions. Expanding coverage to include more cell and tissue types, species, and RNA interactions will further support complex biological inquiries. CRIS is planned to incorporate information on RNA splicing, editing, and other processing events, offering a more comprehensive view of RNA biology. Continuous refinement of data curation standards and computational tools ensures the platform remains a relevant and valuable resource.

As RNA research evolves, addressing challenges like multi-omics integration, dynamic RNA structure analysis, and large-scale data handling will require adaptable tools. CRIS is committed to meeting these demands, serving as both a foundational database and a catalyst for future discoveries in RNA biology and therapeutics.

DATA AVAILABILITY

CRIS is accessible at https://whl-usc.github.io/cris/home. Published sequencing data can be found on NCBI GEO and the ENA Browser, while processed data are available for download via the links on the "Data" page of the CRIS website.

ACKNOWLEDGEMENTS

The Lu Lab is supported by multiple sources, including startup funds from USC, grants from the NHGRI (R00HG009662 and R01HG012928) and NIGMS (R35GM143068), and core support from several USC-affiliated centers and programs. These include the USC Research Center for Liver Disease (P30DK48522), the USC-Illumina Keck Genomics Platform Core Lab Partnership Program, the USC Research Center for Alcoholic Liver and Pancreatic Diseases and Cirrhosis (P50AA011999), the Norris Comprehensive Cancer Center (P30CA014089), and the USC Center for Advanced Research Computing.

REFERENCES

1. Das R. RNA structure: a renaissance begins? Nat Methods. 2021 May;18(5):439–439.
2. Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. Nat Rev Mol Cell Biol. 2019 Jan;20(1):21–37.
3. Bridges MC, Daulagala AC, Kourtidis A. LNCcation: lncRNA localization and function. J Cell Biol. 2021 Feb 1;220(2):e202009045.
4. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014 Jul;15(7):469–79.
5. Loda A, Heard E. Xist RNA in action: Past, present, and future. PLoS Genet. 2019 Sep;15(9):e1008333.
6. 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. 2011 Jun 14;108(24):10010–5.
7. Juźwik CA, S Drake S, Zhang Y, Paradis-Isler N, Sylvester A, Amar-Zifkin A, et al. microRNA dysregulation in neurodegenerative diseases: A systematic review. Prog Neurobiol. 2019 Nov;182:101664.
8. Bratkovič T, Božič J, Rogelj B. Functional diversity of small nucleolar RNAs. Nucleic Acids Res. 2020 Feb 28;48(4):1627–51.

9. Wang XW, Liu CX, Chen LL, Zhang QC. RNA structure probing uncovers RNA structure-dependent biological functions. Nat Chem Biol. 2021 Jul;17(7):755–66.
10. Zhou LY, Qin Z, Zhu YH, He ZY, Xu T. Current RNA-based Therapeutics in Clinical Trials. Curr Gene Ther. 19(3):172–96.
11. Zhu Y, Zhu L, Wang X, Jin H. RNA-based therapeutics: an overview and prospectus. Cell Death Dis. 2022 Jul 23;13(7):1–15.
12. Chen J, Lin J, Hu Y, Ye M, Yao L, Wu L, et al. RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. Nucleic Acids Res. 2023 Jan 6;51(D1):D1397–404.
13. Thomas SJ, Moreira ED, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine through 6 Months. N Engl J Med. 2021 Nov 3;385(19):1761–73.
14. Sahin U, Muik A, Vogler I, Derhovanessian E, Kranz LM, Vormehr M, et al. BNT162b2 vaccine induces neutralizing antibodies and poly-specific T cells in humans. Nature. 2021 Jul;595(7868):572–7.
15. Wang S, Li H, Lian Z, Deng S. The Role of RNA Modification in HIV-1 Infection. Int J Mol Sci. 2022 Jul 8;23(14):7571.
16. Yu H, Wen Y, Yu W, Lu L, Yang Y, Liu C, et al. Optimized circular RNA vaccines for superior cancer immunotherapy. Theranostics. 2025;15(4):1420–38.
17. Piñeiro D, Martinez-Salas E. RNA Structural Elements of Hepatitis C Virus Controlling Viral RNA Translation and the Implications for Viral Pathogenesis. Viruses. 2012 Oct 19;4(10):2233–50.
18. Burrill CP, Westesson O, Schulte MB, Strings VR, Segal M, Andino R. Global RNA Structure Analysis of Poliovirus Identifies a Conserved RNA Structure Involved in Viral Replication and Infectivity. J Virol. 2013 Nov;87(21):11670–83.
19. Zhu Y, Chaubey B, Olsen GL, Varani G. Structure of Essential RNA Regulatory Elements in the West Nile Virus 3′-Terminal Stem Loop. J Mol Biol. 2024 Nov;436(22):168767.
20. Lu Z, Chang HY. The RNA Base-Pairing Problem and Base-Pairing Solutions. Cold Spring Harb Perspect Biol. 2018 Dec 1;10(12):a034926.
21. Schroeder SJ. Challenges and approaches to predicting RNA with multiple functional structures. RNA N Y N. 2018 Dec;24(12):1615–24.
22. Bugnon LA, Edera AA, Prochetto S, Gerard M, Raad J, Fenoy E, et al. Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches. Brief Bioinform. 2022 Jul 18;23(4):bbac205.
23. Ma H, Jia X, Zhang K, Su Z. Cryo-EM advances in RNA structure determination. Signal Transduct Target Ther. 2022 Feb 23;7(1):1–6.
24. Jackson RW, Smathers CM, Robart AR. General Strategies for RNA X-ray Crystallography. Mol Basel Switz. 2023 Feb 23;28(5):2111.
25. Velema WA, Lu Z. Chemical RNA Cross-Linking: Mechanisms, Computational Analysis, and Biological Applications. JACS Au. 2023 Feb 27;3(2):316–32.
26. Marušič M, Toplishek M, Plavec J. NMR of RNA - Structure and interactions. Curr Opin Struct Biol. 2023 Apr 1;79:102532.
27. Kotar A, Foley HN, Baughman KM, Keane SC. Advanced approaches for elucidating structures of large RNAs using NMR spectroscopy and complementary methods. Methods. 2020 Nov 1;183:93–107.
28. Zhang M, Li K, Bai J, Velema WA, Yu C, van Damme R, et al. Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases. Nat Commun. 2021 Apr 20;12(1):2344.

29. Van Damme R, Li K, Zhang M, Bai J, Lee WH, Yesselman JD, et al. Chemical reversible crosslinking enables measurement of RNA 3D distances and alternative conformations in cells. Nat Commun. 2022 Feb 17;13(1):911.
30. Zhang M, Hwang IT, Li K, Bai J, Chen JF, Weissman T, et al. Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers. Genome Res. 2022 May;32(5):968–85.
31. Zhou J, Li P, Zeng W, Ma W, Lu Z, Jiang R, et al. IRIS: A method for predicting in vivo RNA secondary structures using PARIS data. Quant Biol. 2020;8(4):369–81.
32. Andrews RJ, Baber L, Moss WN. RNAStructuromeDB: A genome-wide database for RNA structural inference. Sci Rep. 2017 Dec 8;7(1):17269.
33. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003 Jul 1;31(13):3406–15.
34. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010 Mar 15;11(1):129.
35. Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, et al. Geometric deep learning of RNA structure. Science. 2021 Aug 27;373(6558):1047–51.
36. Gong J, Shao D, Xu K, Lu Z, Lu ZJ, Yang YT, et al. RISE: a database of RNA interactome from sequencing experiments. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D194–201.
37. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res. 2011 May;21(5):734–40.
38. Zhang M, Li K, Bai J, Van Damme R, Zhang W, Alba M, et al. A snoRNA–tRNA modification network governs codon-biased cellular states. Proc Natl Acad Sci U S A. 120(41):e2312126120.