PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Identifying proximal RNA interactions from cDNA-encoded crosslinks with ShapeJumper

Thomas W. Christy[1,2], **Catherine A. Giannetti**[1], **Alain Laederach**[3], **Kevin M. Weeks**[1*]

[1] Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina, United States of America, [2] Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, North Carolina, United States of America, [3] Department of Biology, University of North Carolina, Chapel Hill, North Carolina, United States of America

* weeks@unc.edu

Abstract

SHAPE-JuMP is a concise strategy for identifying close-in-space interactions in RNA molecules. Nucleotides in close three-dimensional proximity are crosslinked with a bi-reactive reagent that covalently links the 2'-hydroxyl groups of the ribose moieties. The identities of crosslinked nucleotides are determined using an engineered reverse transcriptase that jumps across crosslinked sites, resulting in a deletion in the cDNA that is detected using massively parallel sequencing. Here we introduce ShapeJumper, a bioinformatics pipeline to process SHAPE-JuMP sequencing data and to accurately identify through-space interactions, as observed in complex JuMP datasets. ShapeJumper identifies proximal interactions with near-nucleotide resolution using an alignment strategy that is optimized to tolerate the unique non-templated reverse-transcription profile of the engineered crosslink-traversing reverse-transcriptase. JuMP-inspired strategies are now poised to replace adapter-ligation for detecting RNA-RNA interactions in most crosslinking experiments.

Author summary

In principle, crosslinking represents a simple and elegant way to measure important features of RNA structure. Crosslinking-derived, close-in-space structural information can be highly useful for modeling complex higher-order RNA structure and for generating hypotheses regarding how an RNA functions. In practice, extracting the information from an RNA crosslinking experiment, rigorously and at nucleotide resolution, has been difficult and imprecise. This work outlines the development and optimization of an analysis pipeline, called ShapeJumper, that substantially facilitates analysis of RNA crosslinking experiments, based on easily implemented JuMP technology. Both the crosslinking experiment and the analysis software described here are readily implemented by non-expert users.

This is a *PLOS Computational Biology* Software paper.

Citation: Christy TW, Giannetti CA, Laederach A, Weeks KM (2021) Identifying proximal RNA interactions from cDNA-encoded crosslinks with ShapeJumper. PLoS Comput Biol 17(12): e1009632. https://doi.org/10.1371/journal.pcbi.1009632

Editor: Mihaela Pertea, Johns Hopkins University, UNITED STATES

Received: January 22, 2021

Accepted: November 11, 2021

Published: December 14, 2021

Data Availability Statement: Alignments and raw and processed deletions are available at https://weekslab.com/ and in S1 Data. The ShapeJumper pipeline is available at https://github.com/Weeks-UNC/ShapeJumper_V1. Structure analysis and visualization scripts available at https://github.com/Weeks-UNC/StructureAnalysis. Raw sequencing data from published SHAPE-JuMP experiments [Christy TW, Giannetti CA, Houlihan G, Smola MJ, Rice GM, Wang J, et al. Direct mapping of higher-order RNA tertiary interactions by SHAPE-JuMP. Biochemistry. 2021;60: 1971–1982. doi: 10.1021/

acs.biochem.1c00270] and for the engineered RNA (Fig 6) can be obtained from sequence read archives PRJNA687281 and PRJNA762303, respectively.

Funding: This work was supported by the US National Institutes of Health (R35 GM122532 and R01 AI068462 to K.M.W., R35 GM140844 and R01 HL111527 to A.L.). T.W.C. and C.A.G. were supported in part by NIH training grants, in Bioinformatics (T32 GM067553) and Molecular Biophysics (T32 GM008570), respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: We have read the journal's policy and the authors of this manuscript have the following competing interests: K.M.W. is an advisor to and holds equity in Ribometrix. All other authors declare that no competing interests exist.

Introduction

RNA molecules form multiple levels of intra- and inter-molecular higher order structure, and these structures often have important functions. Secondary structures form via base pairing, and secondary structures may further fold into compact tertiary structures mediated by interactions involving canonically and non-canonically interacting nucleotides [1,2]. Developing robust models of RNA secondary and tertiary structure is an important first step in understanding the underlying function of an RNA, and defining well-determined structures can lead to identification of novel functional elements [3,4]. Notable progress has been made using chemical probing experiments to broadly and accurately map biologically relevant secondary structures [5–7]. In contrast, efficient experimental mapping tertiary interactions remains a challenging, unresolved problem [8,9], although notable progress is being made [10–12].

In principle, RNA crosslinking should be able to identify short through-space interactions. Chemical probes such as psoralen analogs [13–16], formaldehyde [17,18] and bis-succinimidyl esters [18], and short wavelength ultraviolet (UV) irradiation [18–20] have been used to crosslink interacting nucleotides. In practice, identifying the precise locations of RNA crosslinks is difficult [8,9,21]. Recent, potentially high-throughput, methods to read out RNA-RNA crosslinks have used variants of proximity ligation to identify crosslinked nucleotides [13–20]. Typically, RNAs are crosslinked and then some combination of RNA fragmentation, crosslink capture, and enrichment is used to obtain linked RNAs whose ends are close to the site of the crosslink. After ligation of adapter sequences to these ends, the sequences are determined by massively parallel sequencing. These adapter-ligation methods yield a rough approximation of crosslink location with best-case resolution of plus-or-minus ten nucleotides [9,21], with the calculations of overall abundance biased by the complex multi-step ligation and library preparation steps required prior to sequencing [8,22]. In addition, commonly used crosslinking reagents and UV irradiation both have strong sequence and structural selectivity, such that observed crosslinks detect only a small fraction of intermolecular RNA interactions.

We recently introduced a strategy we call SHAPE-JuMP (for selective 2'-hydroxyl acylation analyzed by primer extension and juxtaposed merged pairs) [23] in which nucleotides in close three-dimensional proximity are crosslinked with a bi-reactive reagent (Fig 1A, *left*). Initial experiments used the crosslinker *trans* bis-isatoic anhydride (TBIA) (Fig 1B, *left*). TBIA is a SHAPE reagent and, as such, reacts with the 2'-hydroxyl group of unconstrained nucleotides, largely independent of nucleotide identity [24]. In SHAPE-JuMP, sites of crosslinking are recorded in a *single* direct step using an engineered reverse transcriptase (RT) [25] that "jumps" across the crosslink during reverse transcription, creating a deletion in the resulting cDNA [23]. Deletion sites, and thus the positions of crosslinked nucleotides, are identified by massively parallel sequencing and alignment of the deletion-containing sequences. To control for non-crosslink-mediated deletions, an experiment is performed in parallel with a reagent that yields mono-adduct containing RNAs (Fig 1A, *right*). For example, isatoic anhydride (IA) has a structure similar to TBIA, but only one reactive moiety (Fig 1B, *right*). The JuMP strategy provides, in principle, a very simple, direct and experimentally concise readout of sites of crosslinking in RNA. Nonetheless, as currently implemented, there are important limitations: The crosslink-jumping RT enzyme generates cDNAs with high levels of internal mutations, complicating accurate alignment; the "landing" site may be several nucleotides away from the site of the crosslink; and crosslinks are not always jumped consistently. We therefore developed a bioinformatic pipeline, ShapeJumper, to process SHAPE-JuMP sequencing data with the goal of mitigating these limitations.

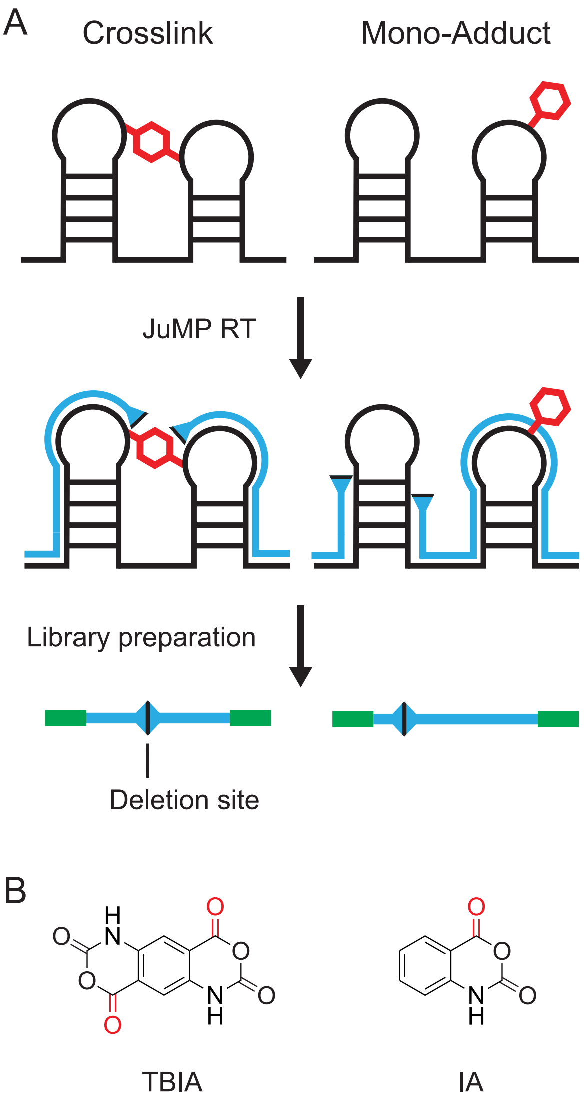

Fig 1. SHAPE-JuMP experimental overview. (**A**) RNA is crosslinked with a bi-functional SHAPE or other reagent, and the site of crosslinking is recorded as a deletion in the cDNA generated by reverse transcription under specialized RT conditions (*left*). In parallel, a control reaction that induces a mono-adduct (or no adduct) in the RNA is used to provide a control for non-crosslink-induced deletions (*right*). The cDNA is sequenced to identify deletion sites. (**B**) Examples of SHAPE reagents that form RNA crosslinks (TBIA, *left*) and monoadducts (IA, *right*). TBIA-dependent crosslinks, more frequent than the IA background, report through-space interactions in RNA.

https://doi.org/10.1371/journal.pcbi.1009632.g001

The ShapeJumper pipeline identifies crosslinked nucleotides from sequencing data (Fig 2). Sequencing reads are first processed to remove low per-nucleotide quality scores and to merge overlapping reads. Reads are aligned [26] with optimized parameters, as developed in this work. The resulting alignment file is analyzed with a custom algorithm to identify deletion sites; during this process ambiguous deletions are removed and exact alignments are enforced at deletion sites to improve accuracy. Deletion rates are then normalized by read depth, and background rates for a non-crosslinked control are subtracted to correct for crosslink-independent deletions. ShapeJumper works well for most classes of crosslinking strategies, including SHAPE-based methods, psoralen reagents, and UV irradiation.

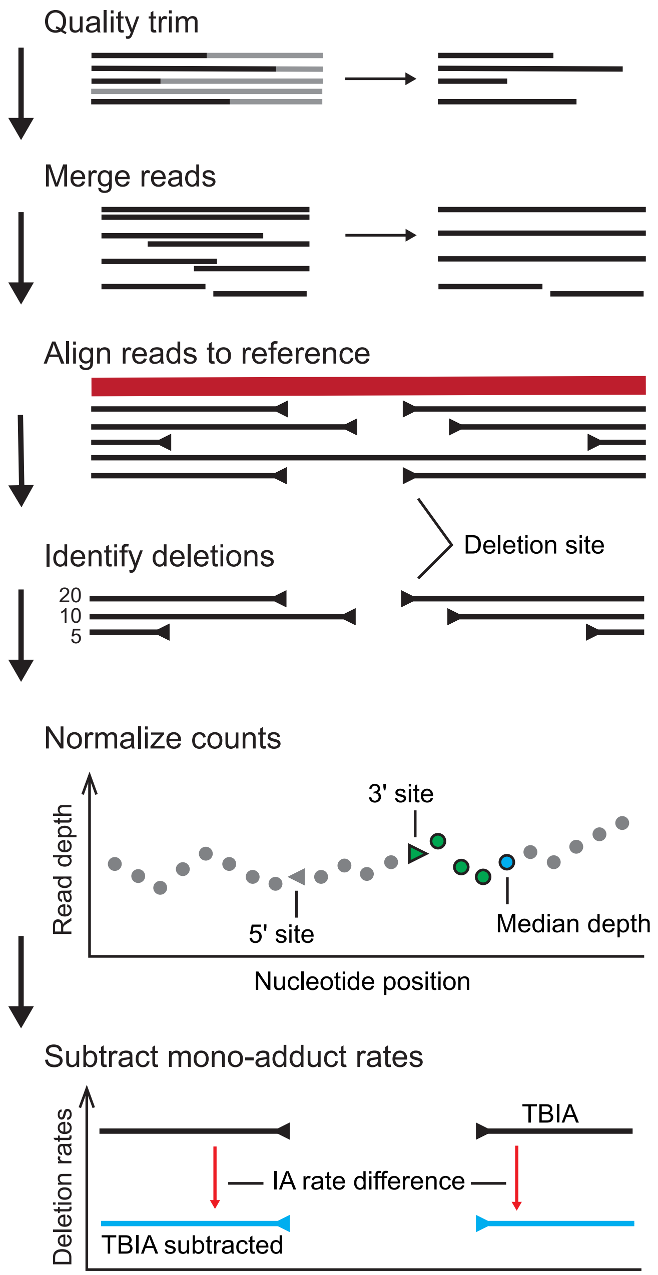

Fig 2. ShapeJumper overview. SHAPE-JuMP sequencing reads are processed for read quality, and paired reads (if used) are merged. Reads are aligned to a reference sequence, creating an initial set of candidate deletion sites. Candidate deletion sites are either identified from an alignment directly or inferred from two alignments separated by unaligned reference sequence. Deletion rates are normalized by the median read depth over the 5 nucleotides downstream of the 3' deletion site. Normalized rates are obtained by subtracting mono- or no-adduct rates.

https://doi.org/10.1371/journal.pcbi.1009632.g002

Results

An aligner for JuMP deletion analysis

Aligning SHAPE-JuMP derived reads accurately is a unique problem. Individual reads may or may not have a deletion, the deletions may vary in length, and the rates of occurrence of deletions vary. The RT enzyme currently used in the SHAPE-JuMP strategy has the special ability to read across crosslinked sites but also has a high non-crosslink-related per-nucleotide mutation rate of 3–4% [23], which makes alignment challenging (see Methods for description of background mutations). No aligner has been specifically designed or optimized to operate with this type of complex data (containing deletions of random length and random frequency, aligned with single nucleotide accuracy). Instead, established aligners are generally optimized for short-read mapping and not designed to handle reads that have longer internal insertions and deletions [27].

We evaluated five aligners as starting points for use in a SHAPE-JuMP pipeline. BLAST, a sequence-comparison-focused algorithm, was selected as an example of a basic hash-table-based aligner [28,29]. YAHA, also hash-based, was selected because it was optimized to detect genomic structural variants, including deletions [30]. Hash-table aligners are slow but perform exhaustive searches of sequence space [31,32]. We also evaluated three aligner programs based on suffix/prefix tries (based on the Burrows Wheel Transform algorithm). These aligners are faster and thus better equipped to process large numbers of inputted reads [31]. Bowtie 2 was evaluated for its ability to process gapped alignments and accept mismatches [33]. BWA-MEM [26] also allows for gapped alignments, is designed to handle sequencing errors robustly, and is optimized for reads of 100 to 1000 nucleotides. STAR was examined because it is an effective splice-site detection aligner [32,34], which share some similarity with SHAPE-JuMP deletions. Aligner programs were assessed using their default parameters, except for small changes to Bowtie 2 and STAR (see Methods).

We evaluated the ability of these aligners to detect SHAPE-JuMP deletions using datasets of synthetic sequencing reads designed to mimic SHAPE-JuMP sequencing reads that contained known deletion locations. These datasets specifically contained sequences that mirrored those observed in experimental SHAPE-JuMP reads, performed with the RT-C8 enzyme [23]. Two synthetic read datasets were created, a deletion set and a deletion-insertion set. Both datasets consist of reads with randomly placed deletions. The deletion-insertion set further contained deletions with additional random insertions of 1 to 9 nucleotides within the deletion. The frequency of each insertion length was sampled from experimental reads. Mutations included mismatches, single-nucleotide insertions, and single-nucleotide deletions, each at levels proportional to their occurrence in experimental reads. These synthetic reads were analyzed using each of the five aligners, and alignment accuracies were assessed by binning the observed deletions into one of three categories (Fig 3A): Exact matches that predict the site of the deletion correctly; close matches for which predicted 5' and 3' borders of the deletion are within three nucleotides of the actual site; and incorrect alignments that exceeded these limits. BLAST had the highest level of exact and close matches, but also had the highest level of incorrectly predicted deletion sites (Fig 3B). STAR also had a high level of exact and close matches for the deletion read set but few deletions were accurately predicted in the deletion-insertion set. Overall, BWA-MEM was the best performer in this analysis for accurately identifying sites of deletions without introducing a bias against detecting deletions in sequences containing deletion-insertions. BWA-MEM was thus used as the aligner in the SHAPE-JuMP pipeline.

Alignment and deletion detection optimization

BWA-MEM was incorporated into a proto-ShapeJumper pipeline and was optimized to address the low positive-predictive value (ppv) for a substantial subset of deletions in the synthetic deletion dataset (Fig 4A). Here ppv is defined as the fraction of predicted deletions that occur in the synthetic data set, at a given set of coordinates. Default BWA-MEM scoring parameters [26] were altered as follows: (*i*) the score penalty for mismatches (–B) was lowered from 4 to 2 to account for the high mutation rate of reads; (*ii*) the deletion score penalty (–O) was decreased from 6 to 2 to accommodate high mutation rates and to promote alignment of

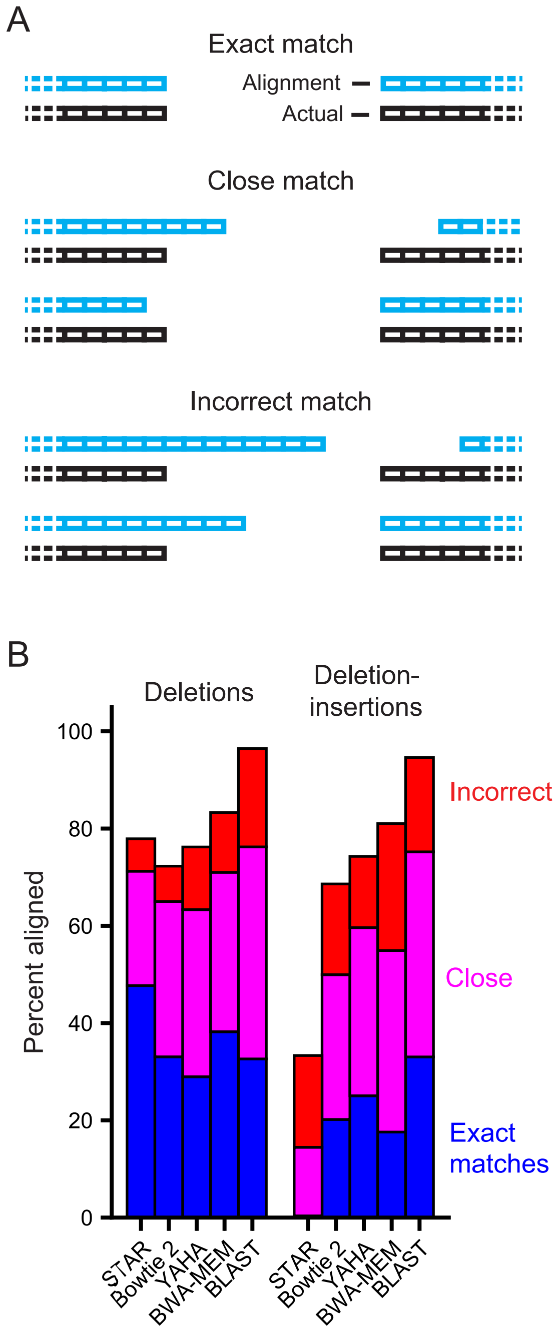

Fig 3. Accuracy analysis for candidate aligners. (**A**) Categories of aligned deletions. (**B**) Analysis of performance of a representative set of aligners on synthetic read datasets, designed to represent JuMP data. Alignments were performed using two sets of synthetic data: containing deletions and deletion-insertions. The deletions set consists of reads with a randomly placed deletion whereas the deletion-insertions dataset also contained a 1–9 nucleotide sequence insertion at the deletion site. Both synthetic datasets contain point mutations reflective of those observed in experimental reads. Each synthetic dataset contained one million reads generated from an RNase P reference sequence. Match categories reported as percentage of total reads in each category.

https://doi.org/10.1371/journal.pcbi.1009632.g003

longer deletions; and (*iii*) the scoring threshold (–T) was lowered from 30 to 15 and the initial seed (–k) shortened from 19 to 10 to allow reads with short sequences flanking a deletion to be aligned. These changes substantially increased deletion site calling accuracy, increased the number of deletions in short sequencing reads that could be aligned, and reduced the fraction of deletions that were incorrectly aligned.

Following these scoring alterations, there remained a systemic bias in the alignment of ambiguous deletions, defined as deletions where one site cannot be uniquely identified because the same nucleotide is present at both sides of the deletion (S1A Fig). The scoring function used during alignment extension from the initial seed leads to the ambiguous nucleotide always being aligned before the gap opening, resulting in a directional bias in deletion-site detection. ShapeJumper therefore removes ambiguous deletions, which results in more accurate alignment of the neighboring, unambiguous deletions (S1B Fig), and results in a roughly 20% increase in exact match detection.

Deletion-insertions also exacerbate inaccurate deletion site assignments, if the insertion includes nucleotides matching the reference sequence within the region of a deletion (S1C Fig). To mitigate insertion-induced misalignments, edge matching was enforced for all deletion sites such that three nucleotides on both sides of the deletion site are required to exactly match the reference. If this is not the case, the deletion site is moved one nucleotide to the exterior, and the removed nucleotide is identified as an insertion in the alignment. This process is repeated until all three nucleotides at the deletion site match the reference. Enforcing exact edge matching notably increased the accuracy of short deletion detection without compromising overall deletion detection (S1D Fig). The combined effect of these optimizations, custom BWA-MEM parameters, ambiguous deletion removal and exact edge matching, substantially increases deletion site detection accuracy (Fig 4B and 4C).

Optimization of the pipeline using experimental data

After optimizing the pipeline with synthetic data, the proto-ShapeJumper pipeline was used to process experimentally generated SHAPE-JuMP reads obtained from analyses of a set of small

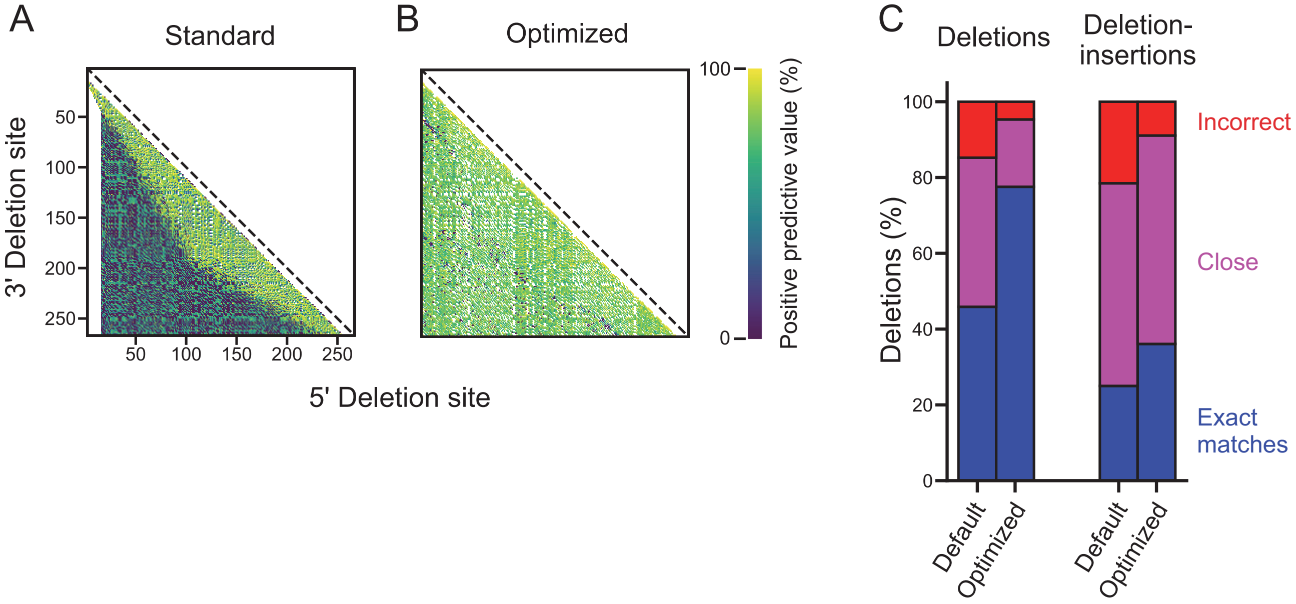

Fig 4. Alignment optimization. Interaction maps for deletion sites identified from the deletion dataset of synthetic reads for BWA-MEM alignment with (**A**) default parameters and (**B**) optimized algorithm. The optimized analysis incorporates custom BWA-MEM parameters, ambiguous site removal, and exact edge matching. Points correspond to specific 5' and 3' deletion sites and are colored by the percent of total deletion sites correctly mapped to a specific nucleotide pair (see scale). (**C**) Summary of accuracies pre- and post-optimization for synthetic deletion (*left*) and deletion-insertion (*right*) datasets. See Methods for full summary of improvement.

https://doi.org/10.1371/journal.pcbi.1009632.g004

to large RNAs (158–412 nts): the P546 group II intron domain, M-Box riboswitch, Varkud satellite ribozyme, RNase P catalytic domain, and group II intron [23]. Quality filtered and merged reads were aligned, the resulting alignments parsed to identify deletion sites, and deletion rates were normalized by the median read depth of the 5 nucleotides downstream of the 3' deletion site (Fig 2). Normalization also enables comparison between samples, including the non-crosslinked (IA) control. The normalized deletion rates observed in the non-crosslinked experiment are subtracted from those observed in the crosslinking experiment to control for non-crosslink-induced deletions. Normalization thereby also removes outliers with high deletion rates (S2 Fig). After this background subtraction step, the most frequent deletions more accurately reflect a holistic view of proximal RNA-RNA interactions (S2C Fig). Background normalization also yields increased area under curve (AUC) in receiver operating characteristic (ROC) curves for through-space interactions within 15 Å of each other for an RNA with complex higher-order structures (S2D Fig).

Long insertions in insertion-deletions are prevalent in experimental data and can contribute to alignment error. For example, for the RNase P RNA, approximately 50% of deletions contain an insertion of at least one nucleotide (S3 Fig, *blue*). Insertions were a substantial source of error in the synthetic read alignments, as evidenced by the difference in accuracy for predicting deletions compared to deletion-insertions (Fig 4C). Insertion length and deletion site assignment error are correlated. (S3 Fig, *red*). ShapeJumper therefore removes reads containing a deletion with an insertion size greater than 10, decrementing the count of deletions found at that site. Insertions longer than 10 nucleotides are infrequent so their removal had a small effect on the total number of deletions detected (S3 Fig, *blue*), and moderately improved deletion site detection.

Finally, experimental SHAPE-JuMP data were analyzed to identify additional features that might improve the precision of detecting proximal interactions. The RT enzyme jumps the crosslink in the 3' to 5' direction (Fig 5A), and it is possible that the nucleotides that physically form crosslinks are downstream of the 5' site or upstream of the 3' site. We examined this possibility by shifting the assigned 5' and 3' sites 0 to 5 nucleotides downstream and upstream, respectively, and examined the effect of these shifts on known through-space inter-nucleotide distances. Shifting the 5' crosslink site 2-nucleotides upstream both increased the detection rate for tertiary interactions and decreased the through-space distance of reported interactions (Figs 5 and S4).

Assessment of ShapeJumper using an engineered, known crosslink

The final optimized ShapeJumper algorithm and pipeline were evaluated using an RNA containing a single defined crosslink. We based this ground-truth experiment on the structure of the CR4/5 domain of a telomerase RNA, whose structure has been determined by nmr [35]. Nucleotides 17U and 38U in our construct are close in space and the RNA was synthesized with 2'-amino substitutions at these sites. The RNA was selectively crosslinked [36] at these nucleotides using an amine-selective crosslinker (N,N'-disuccinimidyl carbonate, DSC) to form a 17-to-38 crosslink (Fig 6A). A single crosslinked species was visualized by denaturing gel electrophoresis (Fig 6B). The crosslinked RNA was extracted from the gel and subjected to JuMP reverse transcription using RT-C8. cDNA products were sequenced [23], and reads were evaluated using the ShapeJumper pipeline. Three of the four most frequent 1% of deletions closely match the known crosslink site; eight of 12 of the most frequent 3% of deletions are also close matches (Fig 6C).

Several patterns are clear from this analysis. First, the 3' side of the crosslink is detected with high precision. Most of the 3' deletion sites originate at or one nucleotide prior to the site of the engineered crosslink. A smaller subset originates from a second site whose 3' position is

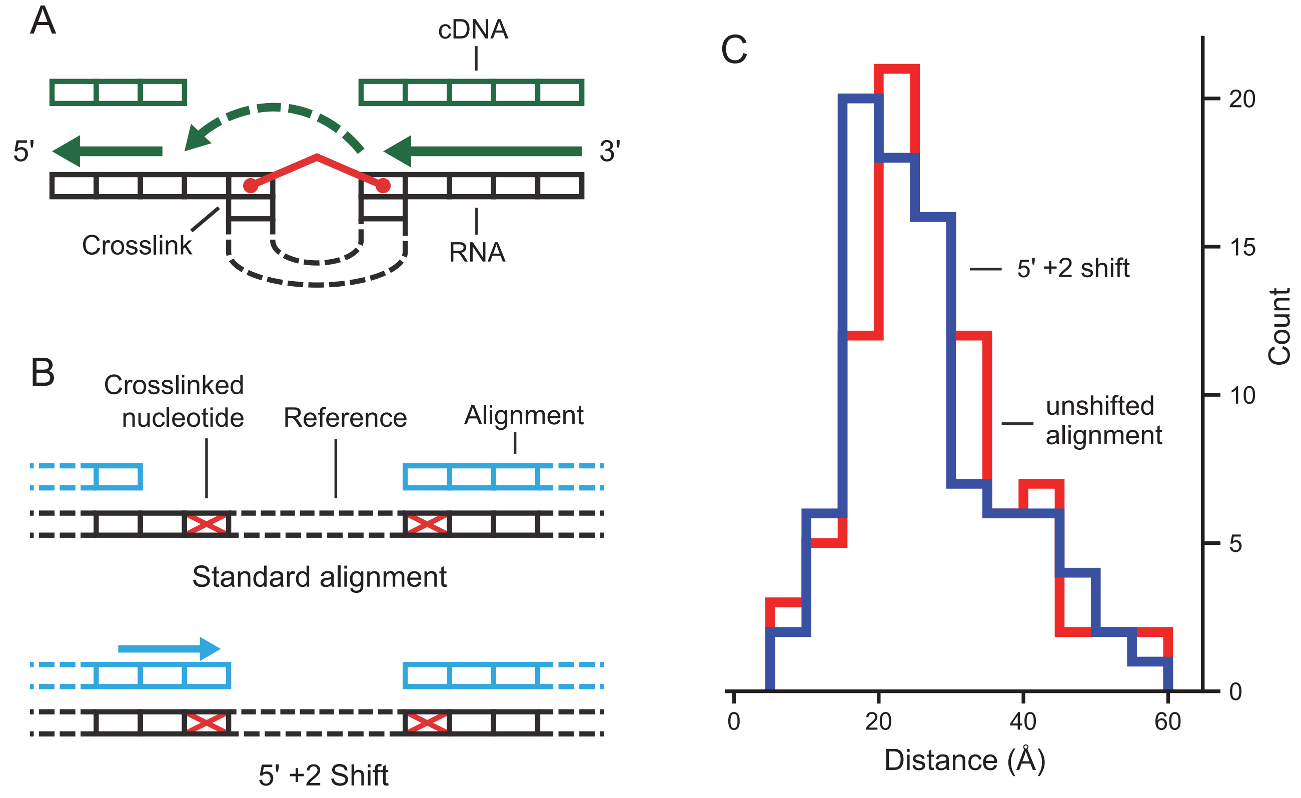

Fig 5. Effect of shifting deletion site assignment on through-space distance. (**A**) Relationship between crosslinked RNA and RT jumping. Directionality of reverse transcription and crosslink-induced steric hindrance can yield an offset at the 5' deletion site relative to crosslink position. (**B**) Deletion site adjustment to compensate for the mechanism of RT jumping. (**C**) Distance distribution of RNase P RNA SHAPE-JuMP data for unshifted (red) and shifted (blue) deletion assignments. Through-space distances are shown for the deletions corresponding to the most frequent 3% of deletions. Datasets were processed using the fully optimized ShapeJumper pipeline with the exception of the shift or not for the RT landing mechanism.

https://doi.org/10.1371/journal.pcbi.1009632.g005

nucleotide 17. One possibility is that RT-C8 traverses through the 3' position 38 and then jumps from the 17 site. Further, the 5' side of the engineered crosslink is defined less precisely than the 3' side, such that the 5' sites of detected deletions span roughly 8 nucleotide positions (Fig 6C). This imprecision for detection of the 5' side of the crosslink compared to the 3' side suggests the RT-C8 enzyme does not always "land" correctly after encountering a crosslink, and is also consistent with the median 5' +2 shift defined above (Fig 5). Overall, the ShapeJumper pipeline correctly identifies deletions that map to the area of the known crosslink site. Further improvement in crosslink site identification will likely require optimization of the experimental RT readout.

Applications of ShapeJumper

ShapeJumper includes useful tools for troubleshooting and visualizing the results of RNA crosslinking experiments. ShapeJumper tools report the distribution of deletion rates and quantifies deletions by sequence length. ShapeJumper calculates the contact distances of deletions, defined as the distance between nucleotides after omitting nested helices, which provides a measure of proximity in secondary structure versus primary sequence space [37]. ShapeJumper also provides visualization tools that facilitate efficient assessment of the quality of a crosslink strategy or experiment. Deletions can be plotted, at any level of frequency, on a secondary structure diagram (Fig

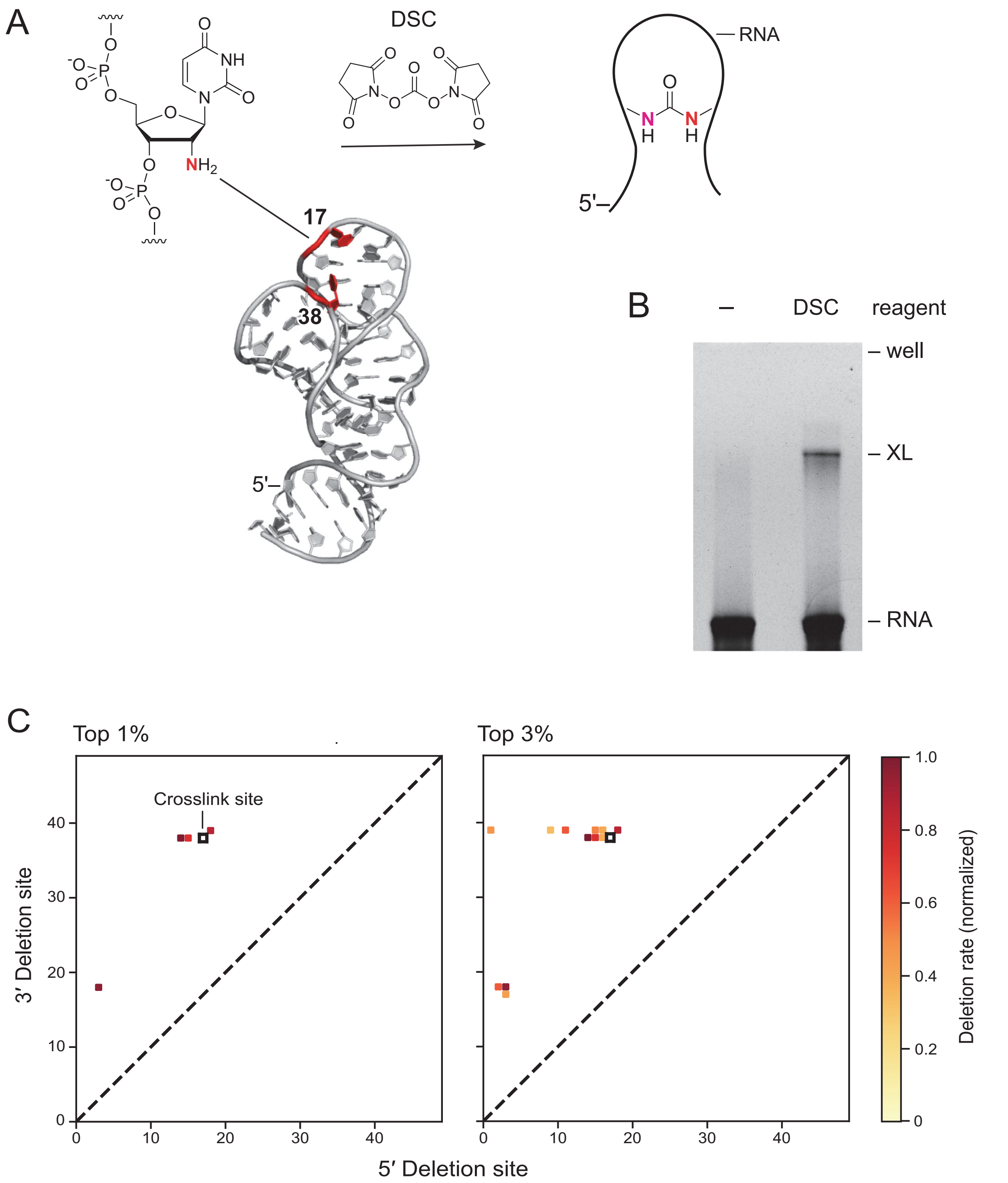

Fig 6. ShapeJumper detects deletions from known crosslink site. (**A**) Structure of the CR4/5 domain of the medaka telomerase RNA (PDB: 2mhi). 2'-amine functionalized nucleotides are red. A defined crosslinked was introduced by treatment with the amine-selective crosslinker DSC. (**B**) Visualization of crosslinking by denaturing gel electrophoresis. Experiments were performed without (–) and with DSC. Gel lanes have been straightened for clarity, but are otherwise unmanipulated. (**C**) Contact maps for most frequent 1% and 3% of deletions as detected by SHAPE-JuMP and ShapeJumper analysis of crosslink-containing RNA. Black box represents the known crosslink site; squares show deletions found by ShapeJumper. Detection rates (see scale) are normalized by the highest observed deletion rate.

https://doi.org/10.1371/journal.pcbi.1009632.g006

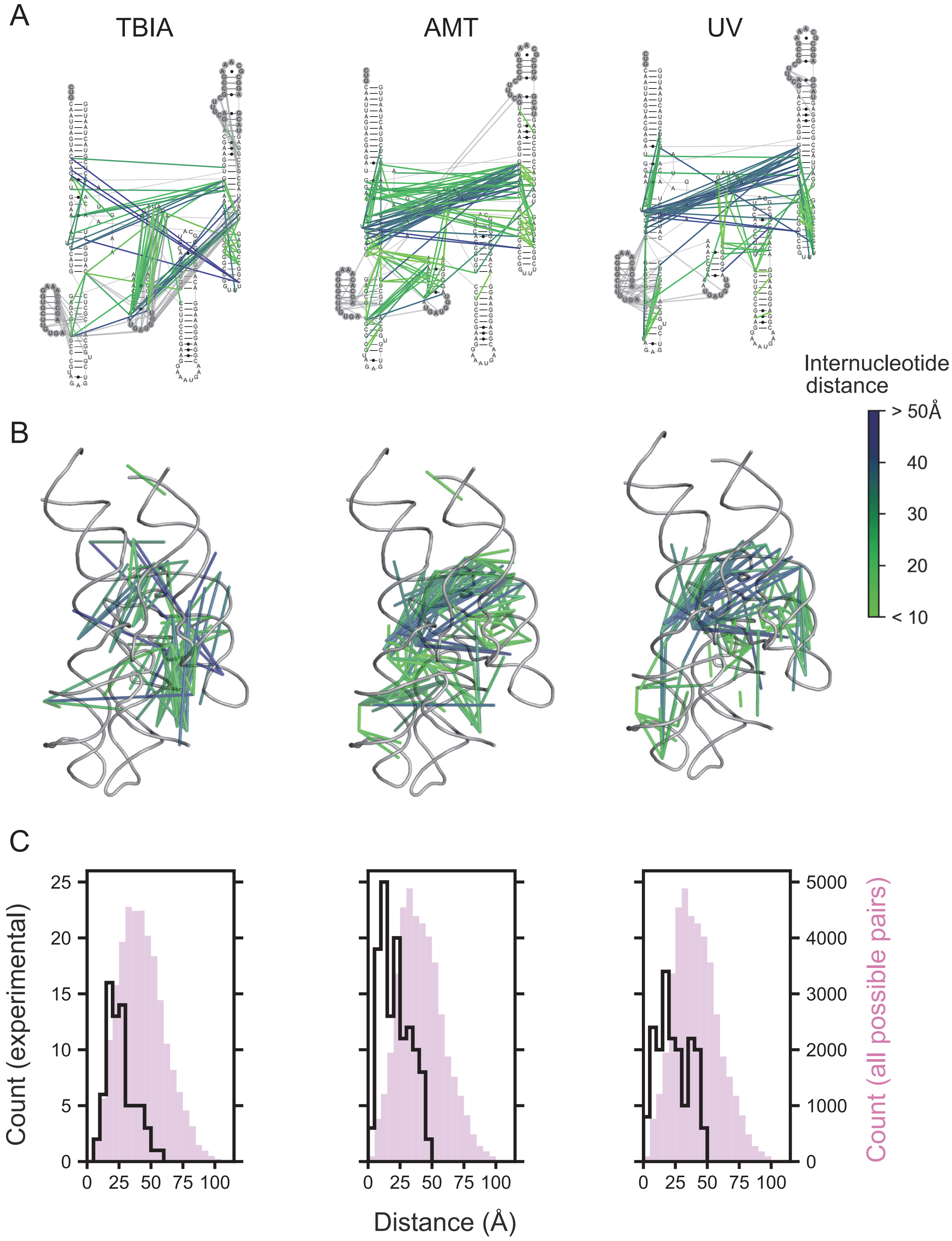
A
TBIA
AMT
UV
Internucleotide distance
> 50Å
40
30
20
< 10
B
C
Count (experimental)
25
20
15
10
5
0
0 25 50 75 100
Distance (Å)
5000
4000
3000
2000
1000
0
Count (all possible pairs)

Fig 7. ShapeJumper measures deletions obtained from diverse crosslinkers. Columns illustrate ShapeJumper analysis of experiments performed with the SHAPE reagent TBIA [23] (*left*), the psoralen derivative AMT (*middle*), and short wavelength UV (*right*). Crosslinks were obtained with the RNase P catalytic domain RNA [46]; the 3% most frequent deletions are shown. (**A**) Deletions superimposed onto the secondary structure. Deletions observed in regions not visualized in the reference three-dimensional structure are gray. (**B**) Deletions superimposed onto a tertiary structure model. In panels (A) and (B), deletions are shown as lines and are colored by through-space distance between nucleotides. (**C**) Distance distribution of deletions. Distances as measured by crosslink-induced deletions are shown as black lines; all possible distances are shown with magenta histograms. Distances were measured between ribose 2'-hydroxyl groups (*left*) or between central point of the nucleobase (*middle and right*).

https://doi.org/10.1371/journal.pcbi.1009632.g007

7A). Given a known or modeled three-dimensional structure, deletions can be visualized and colored by through-space distance (Fig 7B). Three-dimensional distances can be plotted for a given deletion rate and compared to the distance distribution expected by chance (Fig 7C).

The SHAPE-JuMP strategy and ShapeJumper software work for a wide variety of crosslinking reagents. We have used ShapeJumper to evaluate RNA crosslinking experiments performed with TBIA, the psoralen derivative 4′-aminomethyltrioxsalen hydrochloride (AMT), and short wavelength UV (Fig 7A). The patterns of observed deletions vary, and specifically report the distinct underlying chemistry of each reagent. For example, visual inspection of the crosslinking patterns induced by TBIA, AMT and UV light show enrichments (*i*) in single-stranded bases at the 3' deletion site, (*ii*) in double-stranded regions, and (*iii*) at uridine, respectively. Overall, JuMP deletions clearly map proximal sites in the large RNase P RNA (Fig 7B and 7C). We anticipate that most sequencing-based proximal-interaction identification methods [13–20] can be processed and analyzed via ShapeJumper, yielding excellent performance in the accuracy of deletion assignment sites and rates.

Discussion

Optimization of deletion site detection

In principle, crosslinking represents a simple and direct way to map through-space interactions in RNA. In practice, the power of RNA crosslinking has been difficult to realize because of numerous challenges in detecting crosslink sites accurately, and at nucleotide resolution. Identification of an RT enzyme that has the distinctive activity of extending cDNA synthesis through the sites of crosslinks in RNA, revealing these sites as deletions in the cDNA, is an important experimental advance. The cDNA signals are currently complex, however, as the RT enzyme yields cDNAs with internal mutations, the landing sites may be several nucleotides away from the site of the crosslink, and the crosslink may cause termination of polymerization. The ShapeJumper pipeline was designed to be aware of these challenges, to identify and quantify crosslink-mediated deletions, and to distinguish crosslink-induced deletions from other polymerase-mediated mutations.

Deletion rates in a SHAPE-JuMP experiment can vary substantially between RNA targets and it is therefore important to identify infrequent deletions. ShapeJumper attempts to maximally predict deletion sites by allowing low alignment-score thresholds (Fig 4B). Deletion rate variation and polymerase-mediated sequence deletions complicate reproducibility. ShapeJumper addresses these complicating features by normalizing deletion rates by read depth (Fig 2). Finally, the deletion rates of a mono-adduct experiment are subtracted from crosslinked deletion rates to control for crosslink-independent RT-mediated deletions (Figs 2 and S2C).

ShapeJumper was optimized to maximize detection accuracy for the 5' and 3' deletion sites. The crosslinker used for SHAPE-JuMP in our exploratory studies spans ~7 Å between reactive sites (Fig 1B, *left column*). Crosslinked nucleotides should be similarly close in three-dimensional space. Misidentifying the deletion site by just one nucleotide increases the inferred distance by 10–15 Å [38]. We do observe a fraction, 4%, of distances of 45 Å or greater, which likely reflect a combination of false positive measurements, conformational dynamics in these large RNAs, or other features not reflective of internucleotide distances. Accuracies of five aligners were examined, using synthetic datasets with features representative of experimental SHAPE-JuMP data, such that reads contained single-nucleotide mutations, deletions, and insertions, and insertions in the context of deletions (Fig 3). Removing ambiguous deletions and forcing exact edge matching increased assignment accuracies (Figs 4 and S1). The net effect of our aligner choice and these optimization steps is a pipeline that accurately identifies sites of crosslinking, and thus RNA through-space interactions, as shown by analysis of data from a representative set of small and medium sized RNAs (S5 Fig); using an engineered RNA with a single, defined crosslink Fig 6); and for multiple classes of crosslinking experiments (Fig 7). We note that parameter choices and optimization steps were tailored to the specific mutation and deletion characteristics of the RT-C8 [25] enzyme, characterized for SHAPE-JuMP [23]. We think the algorithm developed here will be effective for alternative jumping polymerases and reagents identified in the future, with only minor modification or optimization of the ShapeJumper pipeline.

Perspective

Melding either per-nucleotide RNA chemical probing or through-space crosslinking experiments with a readout by massively parallel sequencing enables analysis of RNA structure with unprecedented throughput and impressive detail. Among many useful applications, SHAPE-JuMP can be used to map through-space interactions in large, complex RNAs (Fig 7), and identify restraints useful for three-dimensional RNA structure modeling [23]. However, it is a challenge to convert the direct results of chemical probing or crosslinking into a form readable by massively parallel sequencing. The ongoing transition from experimentally complex -seq class experiments to much more direct mutational profiling (MaP) has simplified the experiment and increased the accuracy of per-nucleotide chemical probing [3,7,39]. Analogously, a transition from complex adapter-ligation protocols to direct JuMP experimental readouts appears poised to transform experiments that measure through-space RNA-RNA interactions via crosslinking. A key to both MaP and JuMP readouts is software that carefully accounts for the experimental idiosyncrasies of these experiments. ShapeJumper detects deletions resulting from crosslink jumping–from which RNA-RNA interactions can be inferred–with near-nucleotide resolution. The pipeline is easy to implement, requires little to no user input after execution, and works for diverse crosslinking reagents. SHAPE-JuMP and ShapeJumper inaugurate new platforms for efficient detection and analysis of through-space interactions for diverse RNA targets.

Methods

Jumping RT enzyme

Data analyzed in this work were generated by the RT-C8 enzyme, developed by directed evolution using a compartmentalized bead labelling strategy [25].

ShapeJumper pipeline

ShapeJumper is a Bash script that executes multiple python programs and is executable on most UNIX platforms. Inputs are Illumina sequencing reads of crosslinked and non-crosslinked samples in FASTQ format and a reference sequence file in FASTA format. By default, a text file with deletion locations and normalized, background-subtracted rates is output. Ambiguous deletions and deletions with insertions of 10 nucleotides or greater are removed, exact edge matching of deletion sites is enforced, and the final reported deletions have undergone a 5' +2 shift. Alignment and processing parameters can be varied, as described in the included documentation. Additional python tools are provided for analysis of measured deletions in terms of their distribution at the levels of sequence and secondary and tertiary structure. Python 2.7 and necessary third-party packages are available from the Conda package manager (https://conda.io/docs/). The following algorithms are used in the pipeline: ShapeMapper v2 [40,41] is used to trim reads by base-call quality; FLASH [42] is used for merging overlapping reads; BWA-MEM [26] is used to align reads to the reference sequence; and PYMOL (https://pymol.org/) and Biopython [43,44] are used for tertiary structure analysis.

Raw sequencing reads are trimmed by base-call quality using the read trimmer program, ShapeMapper read trimmer, part of ShapeMapper v2 [41,45]. Quality scores for each nucleotide in a read are scanned from 5' to 3'. When the first set of 5 nucleotides with an average quality score below 20 is identified, it and all downstream nucleotides are removed from the read. Reads shorter than 25 nucleotides, post trimming, are removed. The resulting trimmed reads are then merged with their pair mate using FLASH [42], which increases quality scores in the overlapping region. Reads that do not overlap are not removed. The quality trimmed and pair mate merged reads are aligned using BWA-MEM [26] with the parameters optimized in this work: Gap open penalty (–O) of 2, mismatch penalty (–B) of 2, minimum seed length (–k) of 10, score threshold for output alignments (–T) (lowered to) 15. All reads were parsed from cigar strings. Merged and unmergeable reads are parsed and aligned separately, their outputs combined, and duplicate deletions are removed.

Short deletions are directly identified by the aligner. Longer deletions generally result in two alignments per read, one each for the sequence upstream and downstream of the deletion; deletions are identified as the intervening reference sequence between the two alignments that did not align to the read. Multiple deletions can be detected in a single read. Deletions shorter than 10 nucleotides or deletions with an insertion of greater than 10 nucleotides are ignored. The 3 nucleotides upstream and the 3 nucleotides downstream of the deletion site are required to exactly match the reference. If there is a mismatch, the deletion site is shifted until there is an exact match. If these shifts involve more than 10 nucleotides total, the deletion is not reported.

Deletion counts are normalized by the median read depth of the 5 nucleotides immediately downstream of the 3' deletion site. The normalized rates of the mono-adduct control sample are subtracted from the normalized deletion rates of the crosslinked sample. Deletions detected only in the crosslinked sample are retained. Finally, 5' deletion sites are shifted two nucleotides in the 3' direction (Fig 5B). The final deletion data set is reported as each deletion 5' and 3' site, with the normalized and subtracted rate of occurrence.

Aligner evaluation

We emphasize from the outset that current aligners were not designed for our application; nonetheless, most tested aligners could be used to interpret JuMP data in a useful, exploratory way. In general, we used each aligner with default or near-default parameters, and improvement with the non-selected aligners is likely possible with additional parameter changes. In general, we downweighed aligners with higher false positive rates. Aligners were tested using two datasets, each comprised of 1,000,000 computationally generated synthetic reads: a deletion set and a deletion-insertion set. Both synthetic read sets were generated by placing deletions randomly in the sequence for the RNase P catalytic domain [46]; the sequence included flanking structure casettes [24] but deletions were not placed in the structure cassette sequences. The deletion-insertion set contains deletions generated in this manner, but the deletions also contained additional insertions. Insertion lengths were randomly sampled from the distribution of insertions observed from a SHAPE-JuMP experiment using the RNase P

RNA [23] (S3 Fig, *blue line*). Reads in both sets were randomly mutated at single nucleotides at an overall rate of 3.75%; of these mutations, 3% were insertions, 26% were deletions, and 71% were single-nucleotide changes, as found across entire reads. These rates and ratios mimic the experimentally-observed activity of the jumping RT used in this work.

Reads were aligned using the default parameters for each tested aligner with two exceptions. For Bowtie 2, the alignments reported parameter (-k) was set to 2 to enable detection of longer deletions. For STAR, the minimum intron size was set to 10 and the non-canonical junction penalty was lowered to -4 to increase the rate at which deletions were identified at exon junctions; this change was explored to take advantage of splice-site reporting in STAR and to possibly forgo the need to parse deletion sites from SAM files.

The resulting alignments were parsed for deletion-site locations. Locations were then compared to the known deletion sites encoded in the synthetic reads. Each alignment was binned into one of three deletion identification categories: exact matches, where the aligned deletion sites exactly match the encoded deletion sites; close matches, where both of the aligned deletion sites are within 3 nucleotides of the encoded site; and incorrect matches, where one or both of the aligned deletion sites are more than 3 nucleotides from the encoded site. The same synthetic reads and matching criteria were used to evaluate and develop custom BWA-MEM parameters.

As part of the BWA-MEM optimization strategy, a third increasing-insertion-length synthetic read dataset was created to evaluate the effect of insertion length on deletion-site detection accuracy. This dataset consisted of deletions that contain insertions of lengths ranging from 0 to 30. 100,000 reads were synthesized for each insertion length. Reads were created from an RNase P catalytic domain reference sequence [24,46] and mutated as described above. The increasing-insertion-length read dataset was aligned using BWA-MEM with ShapeJumper parameters. The resulting alignments were analyzed for deletion site accuracy at each insertion length (S4 Fig, *red*).

Summary of alignment optimization

Net improvements in deletion identification (shown in Fig 4) are as follows. Deletions: default parameters = 46% exact matches, BWA-MEM parameters optimized = 51%, + ambiguous deletions removed = 78%, + exact edge matching (final optimization) = 78%. Deletion-insertions: default parameters = 25% exact matches, BWA-MEM parameters optimized = 33%, + ambiguous deletions removed = 31%, + exact edge matching (final optimization) = 36%. Filtering of ambiguous deletions and exact edge matching yielded the following. Initially, we identified deletions in 83% of the synthetic reads, which all contained deletions. After optimization of alignment parameters, ambiguous deletion removal, and exact edge matching, the optimized (current) version of ShapeJumper identified 91, 52 and 50% of deletions, respectively.

Structure datasets

TBIA and IA SHAPE-JuMP datasets were collected previously [23]. The two reactive sites on TBIA react with half-lives of 30 and 180 sec; experiments are carried out for 15 min, equal to 5 half-lives of the slower reaction. Short-wavelength UV and 4′-aminomethyltrioxsalen hydrochloride (AMT) data sets were generated using a modified version of the SHAPE-JuMP protocol. Briefly, 15 pmol *in vitro* transcribed RNase P RNA was heat denatured for 1 minute and placed on ice. The RNA was incubated in folding buffer [100 mM HEPES (pH 8.0), 100 mM NaCl, 10 mM $MgCl_2$] at 37˚C for 30 minutes, divided into three 18 μL aliquots, and transferred to amber tubes. One aliquot was treated with 1/9 volume 2 mg/mL AMT (Sigma-Aldrich A4330), dissolved in water, to yield a final concentration of 200 ng/mL AMT. The other two aliquots were treated with the same volume of water, one to serve as a control and the other to be crosslinked with short wavelength UV. The samples were incubated at 37˚C for an additional 15 minutes then placed on ice for crosslinking. The control and AMT samples were exposed to 365 nm (UVP CL1000; 10 cm from light source) for 30 minutes. The short-wavelength UV sample was exposed to 295 nm (UVP Handheld UV lamp, 6 W; 15 cm from light source) for 15 minutes. The RNA was purified by size-exclusion chromatography (G50 column, GE Healthcare) and kept on ice until reverse transcription. Reverse transcription was then performed using target-specific primers under SHAPE-JuMP conditions [23] to produce a cDNA library. PCR was used to amplify cDNAs and to incorporate unique sequence barcodes [23]. Barcoded samples were sequenced (Illumina MiSeq instrument; 500 cycle v2 reagent kit). All datasets were analyzed with default ShapeJumper parameter sets (as developed in this work); for psoralen and UV crosslinking, analysis scripts were updated to define the center of the nucleobase as the site of crosslinking.

Tertiary contact ROC curve analysis

All receiver operating characteristic (ROC) curve analyses used the same classifier, the set of nucleotide pairs with a three-dimensional distance less than 15 Å, and a contact distance greater than 10, where contact distance is defined as the sequence length between two nucleotides according to the secondary structure model when nested helices are skipped. This classifier was chosen as a way to approximate pairwise interactions that correspond to tertiary interactions. The true positive rate is defined as the fraction of ShapeJumper reported contacts with deletion rates above a given threshold that match this definition of tertiary contacts. The false positive rate is defined as the fraction of ShapeJumper contacts with deletion rates above a given threshold that do not correspond to a tertiary contact.

Deletion-site shifts were assessed using data from previously described SHAPE-JuMP experiments performed on five small RNAs [23] with known three-dimensional structures: the *T*. *thermophila* group I intron P546 domain [47], *B*. *subtilis* M-box riboswitch [48], the *N*. *intermedia* Varkud satellite ribozyme [49], the catalytic domain of *B*. *stearothermophilus* RNase P [46], and the *O*. *iheyensis* group II intron [50]. To assess the effect of shifting the assignment for the 5' and 3' sites of crosslinking, SHAPE-JuMP reads were analyzed using default ShapeJumper parameters, and the resulting deletion junction sites were shifted by 0 to 5 nucleotides downstream of the 5' deletion site and/or 0 to 5 nucleotides upstream of the 3' deletion site. Shifted contacts were assessed using ROC curves and mean area under curve (AUC). ROC curve analysis was also carried out to assess the effect of each ShapeJumper analysis step (S5 Fig).

Design and purification of an RNA with engineered crosslink site

The 2mhi RNA [35] was produced by chemical synthesis (Horizon Discovery). The final sequence was 5'-*CCCCT TATTA GCGTT TGCCA* GG—GCGGC GCGGU CAGC$\underline{\text{U}}$ CGGCU GCUGC GAAGA GUUCG $\underline{\text{U}}$CUCU GUUGC—CC *GGGAA GAGGA AGAAT TAGGG* (2'-amine-substituted positions are underlined; dashes indicate 2-nt deletions relative to the sequence determined by nmr; added primer binding sequences are italicized). RNA (100 μM) was heat denatured in water for 1 minute and placed on ice. The RNA was mixed with 3.3× folding buffer [333 mM HEPES (pH 8), 333 mM NaCl, 33 mM $MgCl_2$] and incubated at 37˚C for 30 minutes. RNA was treated with 1/10 volume 500 mM N,N'-disuccinimidyl carbonate (Sigma-Aldrich) in DMSO (final concentration, 50 mM DSC, 20 μM RNA). The RNA was incubated at 4˚C for 1 hour, extracted once with phenol:chloroform:isoamyl alcohol, and recovered by precipitation with isopropanol. RNA was partitioned using a denaturing polyacrylamide gel (15% TBE-urea, ThermoFisher; in running buffer [89 mM Tris-borate (pH 8.3), and 2 mM EDTA]) at 180 V for 3.5 hours. The low mobility band was extracted and eluted into water overnight at 4˚C. The extracted, crosslinked RNA was concentrated (Centrifugal Filter 0.5 mL, Amicon) and stored at -20˚C.

Analysis of RNA containing engineered crosslink

Extracted, crosslinked RNA and non-crosslinked RNA were subject to JuMP reverse transcription [23]. The cDNA was amplified by PCR and barcoded samples were sequenced (Illumina Miseq instrument; 150 cycles, v3 reagent kit). Sequences were trimmed to remove primer binding sites and analyzed with default ShapeJumper parameters.

Three-dimensional RNA structure modeling

One application of SHAPE-JuMP and the ShapeJumper pipeline is to identify restraints useful for three-dimensional structure modeling of complex RNAs. A detailed algorithm for SHAPE-JuMP based structure modeling is provided in a companion manuscript [23].

Supporting information

S1 Fig. Effects of removing ambiguous deletions and enforcing exact edge matching. (**A**) Definition and example of an ambiguous deletion. An ambiguous deletion cannot be mapped to a unique site (*left*); an unambiguous deletion can (*right*). (**B**) Representative contact map of deletion sites from synthetic deletion read alignments containing (*left*) and without (*right*) ambiguous deletions. Ambiguous deletions enclosed in gray outline. Sites with no mapped deletions are white. Note extensive purple regions (0% ppv) are eliminated by removing ambiguous deletions. (**C**) Effect of enforcing exact edge matching (of 3 nucleotides) at a deletion site that also contain an insertion. (**D**) Representative contact map of deletion sites from synthetic deletion-insertion read alignments with ambiguous deletions removed without (*left*) and with (*right*) edge matching. Ambiguous deletions sites are removed in both cases. All contact maps (**B**, **D**) are colored on the same scale by the percent of deletions correctly mapped to a specific nucleotide pair.
(EPS)

S2 Fig. Improvement in TBIA-specific deletion rate measurement upon background subtraction. (**A**) Comparison of distributions of normalized deletion rates for crosslinked (TBIA) and mono-adduct (IA) RNase P RNA experiments. RNase P data used here show trends found in all RNAs examined to date. (**B**) Distribution of crosslink-induced deletion rates after mono-adduct subtraction. (**C**) Deletion sites corresponding to the 3% most frequent deletion rates, pre and post mono-adduct subtraction. Deletion sites are mapped onto the reference tertiary structure [46] and colored by the three-dimensional distance separating the crosslinked nucleotides. (**D**) Ability of ShapeJumper to identify short distance interactions. ROC curve comparison based on TBIA-mediated crosslinking of the RNase P RNA [23]. Tertiary contact identification is shown for raw TBIA deletion rates (blue) and for TBIA deletion rates after subtraction by the IA control (green). Classifier: Inter-nucleotide distance less than 15 Å with a contact distance > 10 (see Methods).
(EPS)

S3 Fig. Insertion length effects on deletion site assignment accuracy and experimental deletion frequency. Misalignment distance is the sequence distance between assigned and known deletion end points. Mean misalignment distance (red line) as a function of insertion length in red. Standard deviation of misalignment is shown by red shading. Observed

frequency of each insertion length in experimental SHAPE-JuMP RNase P data [23] is shown with blue line.
(EPS)

S4 Fig. Effect of 5' and 3' shifts in site assignment on through-space distances. Identification of short distance interactions, as examined by receiver operating characteristic (ROC) curve analysis. Classifier: Inter-nucleotide distance less than 15 Å with a contact distance > 10 (see Methods), based on normalized deletion rate. Mean area under curve (AUC) values for a set of SHAPE-JuMP experiments, performed using five model RNAs (see Methods), as a function of 5' or 3' shift, are shown. Red, white, and blue coloring indicate AUC below, at, or above mean AUC value.
(EPS)

S5 Fig. Progress of ShapeJumper optimization steps for through-space interaction identification. AUC values summarize the results of replicate experiments in terms of ability to measure close-in-space interactions, defined as through-space distances less than 15 Å and contact distances greater than 10. Each column represents a step in the ShapeJumper pipeline. The mono-adduct control shows the AUC for (non-crosslinked) IA samples after processing by the optimized pipeline.
(EPS)

S1 Data. Text files containing complete lists of raw and processed deletions, obtained for each of the RNAs reported in this work.
(ZIP)

Acknowledgments

We thank C.A. Lavender for the initial design of the deletion parsing script and M.J. Smola for initial design of the tertiary structure analysis scripts.

Author Contributions

Conceptualization: Thomas W. Christy, Kevin M. Weeks.

Data curation: Thomas W. Christy, Catherine A. Giannetti.

Formal analysis: Thomas W. Christy, Catherine A. Giannetti, Alain Laederach, Kevin M. Weeks.

Funding acquisition: Kevin M. Weeks.

Investigation: Thomas W. Christy, Catherine A. Giannetti, Kevin M. Weeks.

Methodology: Thomas W. Christy, Catherine A. Giannetti, Kevin M. Weeks.

Project administration: Kevin M. Weeks.

Resources: Thomas W. Christy.

Software: Thomas W. Christy.

Supervision: Alain Laederach, Kevin M. Weeks.

Validation: Thomas W. Christy, Catherine A. Giannetti.

Visualization: Thomas W. Christy, Catherine A. Giannetti.

Writing – original draft: Thomas W. Christy, Kevin M. Weeks.

Writing – review & editing: Thomas W. Christy, Catherine A. Giannetti, Alain Laederach, Kevin M. Weeks.

References

1. Woodson SA. Compact intermediates in RNA folding. Annu Rev Biophys. 2010; 39: 61–77. https://doi.org/10.1146/annurev.biophys.093008.131334 PMID: 20192764
2. Serganov A, Nudler E. A decade of riboswitches. Cell. 2013; 152: 17–24. https://doi.org/10.1016/j.cell.2012.12.024 PMID: 23332744
3. Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods. 2014; 11: 959–965. https://doi.org/10.1038/nmeth.3029 PMID: 25028896
4. Boerneke MA, Ehrhardt JE, Weeks KM. Physical and Functional Analysis of Viral RNA Genomes by SHAPE. Annu Rev Virol. 2019; 6: 93–117. https://doi.org/10.1146/annurev-virology-092917-043315 PMID: 31337286
5. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A. 2009; 106: 97–102. https://doi.org/10.1073/pnas.0806929106 PMID: 19109441
6. Mustoe AM, Busan S, Rice GM, Hajdin CE, Peterson BK, Ruda VM, et al. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. Cell. 2018; 173: 181–195.e18. https://doi.org/10.1016/j.cell.2018.02.034 PMID: 29551268
7. Mustoe AM, Lama NN, Irving PS, Olson SW, Weeks KM. RNA base-pairing complexity in living cells visualized by correlated chemical probing. Proc Natl Acad Sci U S A. 2019; 116: 24574–24582. https://doi.org/10.1073/pnas.1905491116 PMID: 31744869
8. Schonberger B, Schaal C, Schafer R, Voss B. RNA interactomics: recent advances and remaining challenges. F1000Research. 2018; 7. https://doi.org/10.12688/f1000research.16146.1 PMID: 30519453
9. Lu Z, Chang HY. The RNA Base-Pairing Problem and Base-Pairing Solutions. Cold Spring Harb Perspect Biol. 2018;10. https://doi.org/10.1101/cshperspect.a034926 PMID: 30510063
10. Ding F, Lavender CA, Weeks KM, Dokholyan N V. Three-dimensional RNA structure refinement by hydroxyl radical probing. Nat Methods. 2012; 9: 603–608. https://doi.org/10.1038/nmeth.1976 PMID: 22504587
11. Homan PJ, Favorov O V, Lavender CA, Kursun O, Ge X, Busan S, et al. Single-molecule correlated chemical probing of RNA. Proc Natl Acad Sci U S A. 2014; 111: 13858–13863. https://doi.org/10.1073/pnas.1407306111 PMID: 25205807
12. Tian S, Das R. RNA structure through multidimensional chemical mapping. Q Rev Biophys. 2016; 49: e7. https://doi.org/10.1017/S0033583516000020 PMID: 27266715
13. Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, et al. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell. 2016; 165: 1267–1279. https://doi.org/10.1016/j.cell.2016.04.028 PMID: 27180905
14. Aw JGA, Shen Y, Wilm A, Sun M, Lim XN, Boon K-L, et al. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. Mol Cell. 2016; 62: 603–617. https://doi.org/10.1016/j.molcel.2016.04.028 PMID: 27184079
15. Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ. Global Mapping of Human RNA-RNA Interactions. Mol Cell. 2016; 62: 618–626. https://doi.org/10.1016/j.molcel.2016.04.030 PMID: 27184080
16. Ziv O, Gabryelska MM, Lun ATL, Gebert LFR, Sheu-Gruttadauria J, Meredith LW, et al. COMRADES determines in vivo RNA structures and interactions. Nat Methods. 2018; 15: 785–788. https://doi.org/10.1038/s41592-018-0121-0 PMID: 30202058
17. Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell. 2014; 159: 188–199. https://doi.org/10.1016/j.cell.2014.08.018 PMID: 25259926
18. Nguyen TC, Cao X, Yu P, Xiao S, Lu J, Biase FH, et al. Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. Nat Commun. 2016; 7: 12023. https://doi.org/10.1038/ncomms12023 PMID: 27338251
19. Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, et al. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature. 2015; 519: 491–494. https://doi.org/10.1038/nature14280 PMID: 25799984

20. 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 U S A. 2011; 108: 10010–5. https://doi.org/10.1073/pnas.1017386108 PMID: 21610164

21. Weidmann CA, Mustoe AM, Weeks KM. Direct Duplex Detection: An Emerging Tool in the RNA Structure Analysis Toolbox. Trends Biochem Sci. 2016; 41: 734–736. https://doi.org/10.1016/j.tibs.2016.07.001 PMID: 27427309

22. Weeks KM. Review toward all RNA structures, concisely. Biopolymers. 2015; 103: 438–448. https://doi.org/10.1002/bip.22601 PMID: 25546503

23. Christy TW, Giannetti CA, Houlihan G, Smola MJ, Rice GM, Wang J, et al. Direct mapping of higher-order RNA tertiary interactions by SHAPE-JuMP. Biochemistry. 2021; 60: 1971–1982. https://doi.org/10.1021/acs.biochem.1c00270 PMID: 34121404

24. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J Am Chem Soc. 2005; 127: 4223–4231. https://doi.org/10.1021/ja043822v PMID: 15783204

25. Houlihan G, Arangundy-Franklin S, Porebski BT, Subramanian N, Taylor AI, Holliger P. Discovery and evolution of RNA and XNA reverse transcriptase function and fidelity. Nat Chem. 2020; 12: 683–690. https://doi.org/10.1038/s41557-020-0502-8 PMID: 32690899

26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; 1303.3997. Available: http://arxiv.org/abs/1303.3997

27. Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, et al. Technology dictates algorithms: recent developments in read alignment. Genome Biol. 2021; 22: 249. https://doi.org/10.1186/s13059-021-02443-7 PMID: 34446078

28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–3402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

30. Faust GG, Hall IM. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics. 2012; 28: 2417–2424. https://doi.org/10.1093/bioinformatics/bts456 PMID: 22829624

31. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010; 11: 473–483. https://doi.org/10.1093/bib/bbq015 PMID: 20460430

32. Borozan I, Watt SN, Ferretti V. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. PLoS One. 2013; 8: e76935. https://doi.org/10.1371/journal.pone.0076935 PMID: 24204709

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

35. Kim N-K, Zhang Q, Feigon J. Structure and sequence elements of the CR4/5 domain of medaka telomerase RNA important for telomerase function. Nucleic Acids Res. 2014; 42: 3395–3408. https://doi.org/10.1093/nar/gkt1276 PMID: 24335084

36. Chamberlin SI, Merino EJ, Weeks KM. Catalysis of amide synthesis by RNA phosphodiester and hydroxyl groups. Proc Natl Acad Sci U S A. 2002; 99: 14688–14693. https://doi.org/10.1073/pnas.212527799 PMID: 12403820

37. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc Natl Acad Sci U S A. 2013; 110: 5498–5503. https://doi.org/10.1073/pnas.1219988110 PMID: 23503844

38. Gendron P, Lemieux S, Major F. Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol. 2001; 308: 919–936. https://doi.org/10.1006/jmbi.2001.4626 PMID: 11352582

39. Weeks KM. SHAPE Directed Discovery of New Functions in Large RNAs. Acc Chem Res. 2021; 54: 2502–2517. https://doi.org/10.1021/acs.accounts.1c00118 PMID: 33960770

40. Smola M, Calabrese JM, Weeks KM. Detection of RNA-protein interactions in living cells with SHAPE. Biochemistry. 2015; 54: 6867–6875. https://doi.org/10.1021/acs.biochem.5b00977 PMID: 26544910

41. Busan S, Weeks KM. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA. 2018; 24: 143–148. https://doi.org/10.1261/rna.061945.117 PMID: 29114018

42. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27: 2957–2963. https://doi.org/10.1093/bioinformatics/btr507 PMID: 21903629

43. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

44. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. Bioinformatics. 2003; 19: 2308–2310. https://doi.org/10.1093/bioinformatics/btg299 PMID: 14630660

45. Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc. 2015; 10: 1643–1669. https://doi.org/10.1038/nprot.2015.103 PMID: 26426499

46. Kazantsev A V, Krivenko AA, Pace NR. Mapping metal-binding sites in the catalytic domain of bacterial RNase P RNA. RNA. 2009; 15: 266–276. https://doi.org/10.1261/rna.1331809 PMID: 19095619

47. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, et al. Crystal Structure of a Group I Ribozyme Domain: Principles of RNA Packing. Science. 1996; 273: 1678–1685. https://doi.org/10.1126/science.273.5282.1678 PMID: 8781224

48. Dann CE 3rd, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. Structure and mechanism of a metal-sensing regulatory RNA. Cell. 2007; 130: 878–892. https://doi.org/10.1016/j.cell.2007.06.051 PMID: 17803910

49. Suslov NB, DasGupta S, Huang H, Fuller JR, Lilley DMJ, Rice PA, et al. Crystal structure of the Varkud satellite ribozyme. Nat Chem Biol. 2015; 11: 840–846. https://doi.org/10.1038/nchembio.1929 PMID: 26414446

50. Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM. Tertiary architecture of the Oceanobacillus iheyensis group II intron. RNA. 2010; 16: 57–69. https://doi.org/10.1261/rna.1844010 PMID: 19952115