RNA structure inference through chemical mapping after accidental or intentional mutations

Clarence Y. Cheng[a,1], Wipapat Kladwang[a,1], Joseph D. Yesselman[a], and Rhiju Das[a,b,2]

[a]Department of Biochemistry, Stanford University, Stanford, CA 94305; and [b]Department of Physics, Stanford University, Stanford, CA 94305

Edited by David P. Bartel, Massachusetts Institute of Technology, Cambridge, MA, and approved July 21, 2017 (received for review December 5, 2016)

Despite the critical roles RNA structures play in regulating gene expression, sequencing-based methods for experimentally determining RNA base pairs have remained inaccurate. Here, we describe a multidimensional chemical-mapping method called "mutate-and-map read out through next-generation sequencing" (M2-seq) that takes advantage of sparsely mutated nucleotides to induce structural perturbations at partner nucleotides and then detects these events through dimethyl sulfate (DMS) probing and mutational profiling. In special cases, fortuitous errors introduced during DNA template preparation and RNA transcription are sufficient to give M2-seq helix signatures; these signals were previously overlooked or mistaken for correlated double-DMS events. When mutations are enhanced through error-prone PCR, in vitro M2-seq experimentally resolves 33 of 68 helices in diverse structured RNAs including ribozyme domains, riboswitch aptamers, and viral RNA domains with a single false positive. These inferences do not require energy minimization algorithms and can be made by either direct visual inspection or by a neural-network–inspired algorithm called M2-net. Measurements on the P4–P6 domain of the *Tetrahymena* group I ribozyme embedded in *Xenopus* egg extract demonstrate the ability of M2-seq to detect RNA helices in a complex biological environment.

RNA structure modeling | chemical mapping | neural network | mutational profiling | *Xenopus* egg extract

Inference of RNA structures using experimental data is a crucial step in understanding RNA's biological functions throughout living organisms. Chemical-mapping methods have the potential to reveal RNA structural features in situ by probing which nucleotides are protected from attack by chemical modifiers. The resulting experimental data can be used to guide secondary-structure modeling by computational algorithms, raising the prospect of transcriptome-wide RNA structure determination (1, 2).

Despite these advances, the accuracy of RNA structure inference through chemical mapping and sequencing remains under question (3–8). For example, models of the 9-kb HIV-1 RNA genome have been repeatedly revised with updates to the selective 2′-OH acylation by primer extension (SHAPE) protocol, data processing, and computational assumptions (2, 9–11), and the majority of this RNA's helices remain uncertain. Even for small RNA domains, SHAPE and dimethyl sulfate (DMS) (methylation of N1 and N3 atoms at A and C) have produced misleading secondary structures for ribosomal domains and blind modeling challenges that have been falsified through crystallography or mutagenesis (3, 7, 12, 13). In alternative approaches based on photoactivated cross-linkers, many helix detections appear to be false positives, based on ribosome data in vitro and in vivo (14, 15).

The confidence and structural accuracy of chemical-mapping methods can be improved by applying perturbations to the RNA sequence before chemical modification. In the mutate-and-map strategy, mapping not just the target RNA sequence but also a comprehensive library of point mutants reveals which nucleotides respond to perturbations at every other nucleotide, enabling direct inference of pairs of residues that interact to form structure (16, 17). The resulting models have been consistently accurate at nucleotide resolution in RNA puzzles and other blind tests for riboswitches and ribozymes solved by crystallography, with helix recovery rates of >90% and false-positive rates under 10%, with errors typically involving minor register shifts or edge base pairs (2, 18). However, the mutate-and-map approach has required synthesis and parallel mapping of many mutant RNAs and, so far, has only been applied to RNAs under 200 nt in length probed in vitro.

Here, we introduce mutate-and-map read-out by next-generation sequencing (M2-seq), which carries out RNA preparation, mutation, and mapping in a one-pot experiment. Tests on ribozyme domains, viral domains, and riboswitch aptamers that form diverse RNA structures evaluate the ability of M2-seq to detect Watson–Crick base pairs in vitro, with signals that can be confirmed through visual inspection. We introduce a simple algorithm, M2-net, that automatically recovers these helices with a low false-positive rate (<5%) and without register shifts that have been previously problematic for chemical-mapping approaches. As a proof-of-concept for more complex biological experiments, we demonstrate direct detection of the majority of helices in the P4–P6 domain of the group I *Tetrahymena* ribozyme embedded in biologically active eukaryotic cell extract, and describe prospects for further applications in RNA structural biology.

Significance

The intricate structures of RNA molecules are crucial to their biological functions but have been difficult to accurately characterize. Multidimensional chemical-mapping methods improve accuracy but have so far involved painstaking experiments and reliance on secondary-structure prediction software. A methodology called M2-seq now lifts these limitations. Mechanistic studies clarify the origin of serendipitous M2-seq–like signals that were recently discovered but not correctly explained and also provide mutational strategies that enable robust M2-seq for new RNA transcripts. The method detects dozens of Watson–Crick helices across diverse RNA folds in vitro and within frog egg extract, with a low false-positive rate (<5%). M2-seq opens a route to unbiased discovery of RNA structures in vitro and beyond.

Author contributions: C.Y.C., W.K., and R.D. designed research; C.Y.C. and W.K. performed research; C.Y.C. contributed new reagents/analytic tools; C.Y.C. and J.D.Y. analyzed data; and C.Y.C. and R.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The M2-seq datasets reported in this study have been deposited in the RNA Mapping Database, https://rmdb.stanford.edu/ (accession nos. TRP4P6_DMS_0005–TRP4P6_DMS_0009, TRP4P6_DMS_0010–TRP4P6_DMS_0014, RNASEP_DMS_0000, and additional IDs given in *Supporting Information*).

[1]C.Y.C. and W.K. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: rhiju@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619897114/-/DCSupplemental.

Results

Workflow of M2-Seq. The M2-seq workflow tested herein is summarized in Fig. 1. First, DNA templates were prepared from PCR assembly (short constructs) or PCR from plasmids (long constructs). To ensure mutate-and-map signals, we prepared samples with a low frequency ($\sim 10^{-3}$ per nucleotide) of additional mutations as described previously (16) or using error-prone PCR (19). We also prepared samples without additional mutations to probe unexplained data correlations observed in recent high-DMS experiments (2, 20). Then, we transcribed RNAs from these DNA pools, prepared them into the desired state (e.g., Mg^{2+}-induced folding in vitro), and modified the RNA with DMS. Reverse transcription was performed under mutational profiling conditions (with SuperScript II and Mn^{2+}) to install mutations into cDNAs across from DMS modifications (21). The full-length cDNAs were amplified by PCR, and the resulting libraries were sequenced by paired-end Illumina sequencing. An initial M2-seq map was generated by recording the positions of all of the correlated mutations. The data were displayed in a 2D heatmap visualization analogous to that used for prior mutate-and-map experiments: a 1D chemical-mapping profile was estimated for each single-nucleotide variant in the RNA, each profile was normalized by the total number of reads with a mutation at that position, and the profiles were stacked according to the mutation. As described below and in *SI Results*, a more sophisticated analysis is possible that attempts to separate mutations based on their expected source (e.g., those installed during library preparation vs. those introduced later by reverse transcription across from chemical modifications). However, we will mainly describe results with a simple mutation-counting approach, which provides an initial unbiased visualization.

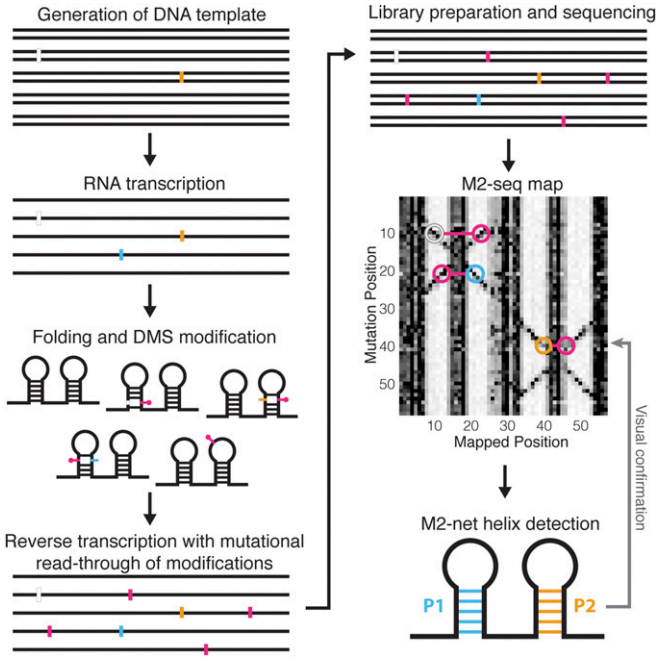

Fig. 1. Workflow for M2-seq, mutate-and-map read-out by next-generation sequencing. DNA template is generated by PCR assembly of oligonucleotides, PCR from a plasmid, or error-prone PCR, potentially introducing deletions (white) or mutations (gold). Then, RNA is transcribed (potentially introducing further mutations, cyan), folded, and DMS-modified. Reverse transcription is performed under conditions that favor mutational read-through of DMS-modified nucleotides, recording those positions as mutations (magenta), and cDNAs are PCR-amplified to generate double-stranded DNA library. Libraries are subjected to next-generation sequencing, and resulting reads are analyzed by demultiplexing, alignment to reference sequences, and correlated mutation counting to generate an M2-seq dataset (simulated here). Double-stranded RNA helices give rise to cross-diagonal features in these maps that can be automatically recognized by M2-net and confirmed visually.

Mutational Profiling Provides Precise M2-Seq Information in a Single-Pot Experiment. We first confirmed that applying the mutational profiling readout to single-mutant libraries would give secondary-structure signals similar to capillary electrophoresis (CE)-based M^2, which relies on reverse-transcription termination at modified residues rather than mutational read-through. For this comparison, we investigated the P4–P6 domain of the 158-nt *Tetrahymena* group I ribozyme (Fig. 2*A*), a widely used model system for tests of RNA chemical-mapping methods (2, 3, 22). In addition, we prepared DNA templates for the wild-type RNA and 158 point mutants of each nucleotide to its complement (16) and then pooled these molecules before the transcription step, so that all subsequent steps could be carried out in one tube. The M2-seq data for this initial pooled-mutation experiment are shown in Fig. 2*B*, after applying the pipeline described in Fig. 1 to generate 1D chemical mapping reactivity profiles for each mutation position.

Analysis of the mutational spectrum in the no-DMS samples confirmed that we had introduced the desired sequence changes at the level of ~1/158 (gray-lined outset, Fig. 2*A*). Furthermore, as expected, the M2-seq data (Fig. 2*B*) exhibit strong signals for structural elements, consistent with prior mutate-and-map data based on CE (Fig. S1). For example, M2-seq signals marking the pair C229/G245 and other pairs in the P6b helix create a visible cross-diagonal, as in prior CE data (black-lined outset, Fig. 2*B*). Base pairs for P4 and P5 (orange in Fig. 2*B*), P5a (blue), P5b (red), and P6a-b (green) are clearly visible and agree with crystallographic analysis of the RNA (Fig. 2*A*). Similarly, punctate signals reflecting the tetraloop/tetraloop receptor (TL/TLR) tertiary contact, such as between A153 and C223, also appear in both datasets. Short helices that were not observed in CE-based mutate-and-map measurements, such as the P5c helix (Fig. S1), also did not give extended cross-diagonal stripes in the M2-seq data. As expected, the no-DMS control samples did not show M2-seq signal and consisted primarily of a uniform 1D background (Fig. S2*A*).

We further tested that separate preparation of mutants was not necessary to give clear M2-seq signals of base pairs. We used error-prone PCR to generate the DNA templates for RNA transcription, giving mutations at a mean frequency of ~0.5% and mostly involving U-to-C, C-to-U, A-to-G, and G-to-A transitions (gray-lined outset, Fig. 2*C*), as expected (23). Despite having a different mutational spectrum and giving signals at different specific base pairs, we observed M2-seq signals for the same helical elements as in the pooled single mutant library experiment as well as for the TL/TLR tertiary contact (Fig. 2*C*; fine differences are better visible in black-lined magnification outsets of P6a–P6b region). The use of error-prone PCR simplified the protocol: every step of the M2-seq experiment, from DNA synthesis to final reverse transcription and sequencing, could be carried out in a single tube.

We also observed M2-seq signal in samples without mutations intentionally installed during error-prone PCR (Fig. 2*D*). We had previously noted this pattern in published sequencing data for high-DMS-modified P4–P6 RNA (2) (Fig. S3) and speculated that DMS methylation of the N1 and N3 atoms of G and U residues, respectively, could disrupt Watson–Crick base pairing, expose C and A partners, respectively, and produce a 2D signal (2, 20). Paradoxically, however, the modification reaction pH of 7.0 is too low to cause significant deprotonation at these atoms to allow DMS methylation to occur ($\sim 10^{-4}$ modification rate expected under our conditions). Furthermore, when we applied the no-mutation method to another large, highly structured RNA, the *Didymium iridis* GIR1 lariat-capping ribozyme, we

BIOCHEMISTRY

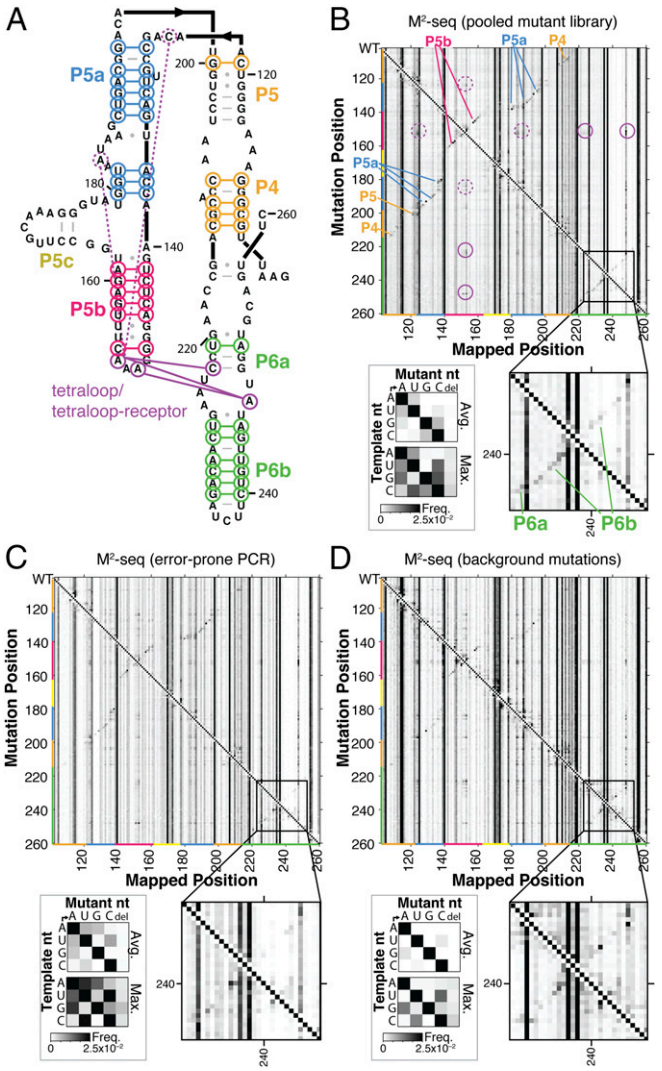

Fig. 2. M2-seq on the P4–P6 domain of *Tetrahymena* group I ribozyme. (*A*) Secondary-structure diagram of P4–P6. (*B–D*) Two-dimensional datasets from M2-seq on pooled mutate-and-map library (*B*), on RNA with mutations installed during error-prone PCR of DNA template (*C*), and with no intentionally installed mutations (*D*), all probed with DMS mapping. Each row gives frequencies of observing mutations at every position given a mutation at the row position, as indicated by the strong diagonal (*Top Left* to *Bottom Right*). In *B–D*, black-lined outsets highlight M2-seq signals in the P6a–P6b region; gray-lined outsets show average and maximum observed frequencies of each type of mutation in control RNA samples without DMS treatment. In *A* and *B*, colored lines and labels mark correspondence of structure and map signals for Watson–Crick helices, the tetraloop/tetraloop receptor contact (solid purple), and exposure of tetraloop from mutations outside its receptor (dashed purple).

observed no clear cross-diagonal stripes corresponding to long-range RNA base pairs (Fig. 3*A*).

Mechanism of "Background" RNA Base Pair Signals. To understand whether M2-seq signals could be enhanced for the GIR1 ribozyme and other RNAs, we carried out extensive experiments to understand the mechanism for the signal in the P4–P6 RNA, varying transcription templates, purification, and modification conditions. A complete description of this work is given in *SI Results* and Figs. S4–S7; a short summary follows. Briefly, we were able to discriminate between two models for how M2-seq signals might arise without intentionally preinstalled mutations. In the "double-DMS hit" model noted above, these Watson–Crick base pair signals are due to rare DMS modifications ($\sim 10^{-3}$ per nucleotide) that occur at transiently deprotonated U/G nucleotides, resulting in—or caused by—DMS modification at partner A/C nucleotides (2, 20). In an alternative "accidental mutation" model, the signals are due to background mutations (also up to $\sim 10^{-3}$ per nucleotide) introduced as errors during DNA and RNA synthesis. In the folded RNA, these mutations would expose their structural partners to DMS, as with standard mutate-and-map methods.

Favoring the accidental mutation model, differences in the M2-seq signal with different DNA preparations (PCR assembly of oligonucleotides, PCR from a plasmid stock, and synthesis in other laboratories; Fig. 2 and Fig. S3) implicated background mutations introduced in the different DNA synthesis methods, and were then confirmed by sequencing the DNA templates used for those M2-seq experiments (Figs. S4 and S5). Additional M2-seq base pair signals were traced to transitions introduced during RNA synthesis by T7 RNA polymerase and confirmed by direct sequencing of the RNA before DMS modification (gray-framed outsets, Fig. 2 and Figs. S4 and S5). Disfavoring the double-DMS model, increasing the pH, which should enhance transient deprotonation of U/G and subsequent DMS modification, did not increase the M2-seq base pairing signal except at high pH. At pH 10.0, a different, less precise signal was observed (Fig. S6). Finally, DMS dose–response measurements revealed linear dependence of the Watson–Crick base pair signals with DMS dose, as predicted by the accidental mutation model but not the double-DMS hit model, which predicts a quadratic dependence on DMS dose (Fig. S7).

Taken together, these studies traced the primary mechanism of direct base pair detection in DMS experiments to the occurrence of accidental mutations during DNA and RNA synthesis and not to double-DMS hits. Because these mutations occur in a heterogeneous and noncontrolled manner throughout the RNA molecule, they only allow detection of Watson–Crick pairs in special molecules with particular preparations. We therefore favored using error-prone PCR to seed in mutations more uniformly across transcripts. For example, in the case of the GIR1 lariat-capping ribozyme, M2-seq signals highlighting most of the RNA's helices became visible when templates were prepared with error-prone PCR (Fig. 3*B*). Even for the P4–P6 RNA, use of error-prone PCR allowed M2-seq detection of the P4–P6 helices with nearly an order of magnitude fewer sequencing reads than a protocol relying only on accidental mutations (Fig. S8).

Automated Detection of Helices Across Diverse RNA Structures. After testing the mechanism of the M2-seq signal, we evaluated the

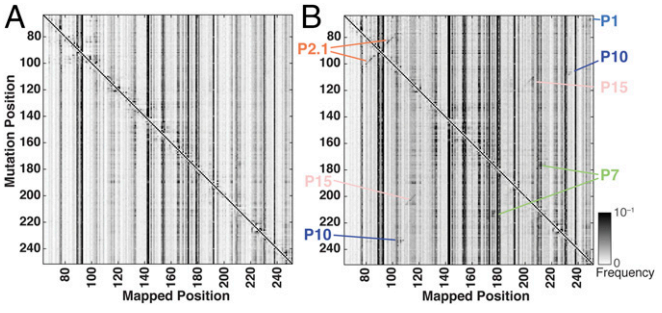

Fig. 3. M2-seq on the GIR1 lariat-capping ribozyme requires seeded mutations. M2-seq maps for DMS-treated GIR1 ribozyme prepared (*A*) without any intentionally installed mutations and (*B*) from templates seeded with mutations through error-prone PCR. Colored text labels indicate helices for which cross-diagonal helix signatures become visible and detectable by M2-net.

general applicability of the method across diverse structured RNA molecules. We chose several RNAs that have challenged prior structure modeling efforts: the P4–P6 RNA, the catalytic domain of RNase P, and the thiamine pyrophosphate (TPP) riboswitch aptamer, which were the three test cases for an earlier RING-MaP study (20, 21); and the GIR1 ribozyme, riboswitch aptamers for adenosylcobalamin (AdoCbl) and cyclic–di-AMP, and an Xrn1-exonuclease-resistant domain from the Zika virus, four targets of the RNA-puzzle community-wide trials whose secondary structures were particularly challenging for most groups and algorithms to model (Tables S1 and S2) (12, 13, 24). M2-seq gave visually apparent signals for helices in all of these cases (Fig. 4 and Fig. S9). These helices included long-range interactions connecting the most sequence-distal ends of RNA (P1 in the GIR1 and RNase P molecules), pseudoknots (P7 in GIR1; P2 in RNase P), long helices involved in tertiary contacts (9-bp P5b in P4–P6; 10-bp P8 in the AdoCBL riboswitch), and short helices (P3 in the TPP riboswitch). These signals were particularly apparent when we displayed maps of *Z*-scores, which measure how much the DMS signal at each nucleotide is enhanced over the mean at that position across all mutant variants, normalized to the standard deviation at that position. The quality of these data led us to revisit automated *Z*-score–based helix detection methods developed in early work on the mutate-and-map method (25, 26). Indeed, we discovered that our visual analysis could be automatically reproduced by a simple pipeline of *Z*-score estimation, a convolutional filter highlighting "cross-diagonal" stripes, data symmetrization, and a filter for each nucleotide having at most one partner (*SI Methods*; colored annotations in Fig. 4). We call this analysis M2-net, due to its similarity to multilayer convolutional neural nets that are now in wide use for image classification (27).

M2-net detected 34 of helices with length greater than 2 in these RNAs (Table 1). A total of 33 of these 34 helices matched the crystallographic or conventional structure available in the literature, and none of these cases involved register shifts that have been problems in prior methods (2, 7). Despite the observation of other weak signals in these data that do not correspond to helices (Fig. 4), M2-net detects only a single false positive, an altP19c helix predicted for the catalytic domain of RNase P that disagrees with the tip of the P19 domain presumed in the conventional secondary structure of this molecule (28). The region including these helices has not been directly visualized by crystallographic analysis (28). Compensatory mutagenesis experiments indicate that the region does form the predicted alt-P19 in solution (Fig. S10); we still count it as a false positive here to be conservative.

In prior work, we and others have used the RNAstructure free-energy minimization software, guided by mutate-and-map or conventional 1D chemical mapping data, to "fill in" helices

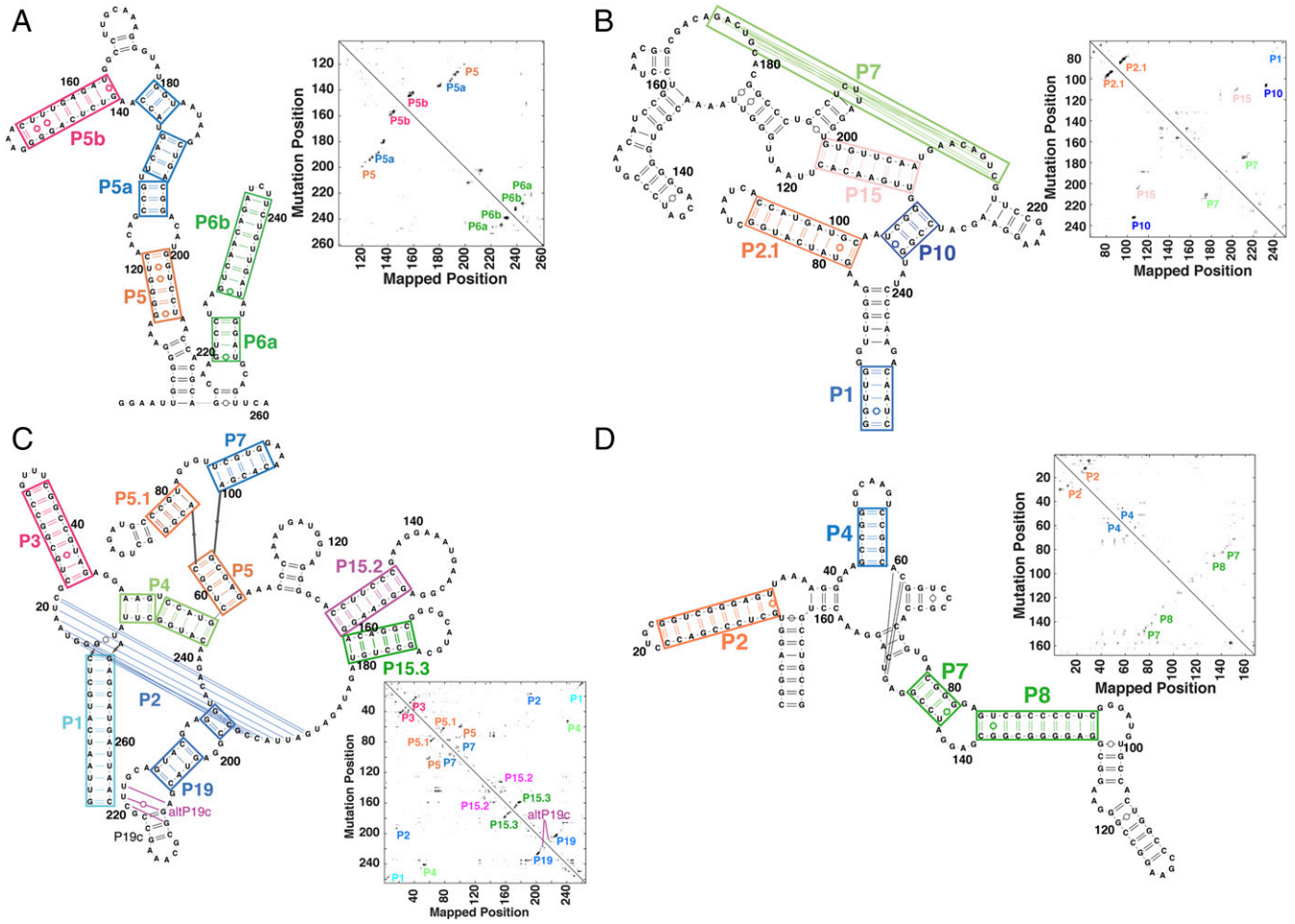

Fig. 4. M2-seq recovers helices across diverse RNA folds. Each panel shows crystallographic secondary structures and *Z*-score–transformed maps (square graphs) with colored labels (on both display items) marking helices and multihelix domains automatically identified by M2-net analysis. Differences in edge base pairs are not shown. Data sets are as follows: (*A*) P4–P6 domain of *Tetrahymena* ribozyme (background mutations), (*B*) GIR1 lariat-capping ribozyme (RNA-puzzle 5; error-prone PCR), (*C*) ribonuclease P catalytic domain (background mutations), and (*D*) adenosylcobalamin riboswitch aptamer (RNA-puzzle 6; error-prone PCR). Data for three additional RNAs of smaller length are given in Fig. S9. Table S3 compiles modeled structures.

Table 1. Recovery of helices across seven complex RNA folds from M2-seq data

RNA	No. of helices*	ShapeKnots + DMS + M2†		M2-net	
		TP	FP	TP	FP
P4–P6 domain	8	8	1	7	0
GIR1 ribozyme	11	11	0	5	0
RNase P C domain	13	11	1	10	1
AdoCbl riboswitch	10	10	0	4	0
ydaO riboswitch	7	6	1	2	0
Zika xrRNA	5	4	1	2	0
TPP riboswitch	6	6	0	3	0
Total	60	56	4	33	1
False-negative rate, %		6.7		45.0	
False-positive rate, %		6.7		2.9	
Sensitivity, %		93.3		55.0	
PPV, %		93.3		97.1	

FP, false positives; PPV, positive predictive value; TP, true positives.
*Helices with length greater than two Watson–Crick (or G•U wobble) base pairs.
†Use of one-dimensional DMS data to guide folding through energy bonuses (5) and *Z*-scores derived from 2D M2-seq experiments, applied as in ref. 16.

not directly detected by experiments (12, 20, 26). In our M2-seq benchmark, the ShapeKnots algorithm of RNAstructure guided by the M2-seq and 1D DMS data indeed increases the number of recovered crystallographic helices from 34 to 56 (out of 60 helices; 93% sensitivity). However, the higher recovery is at the expense of more false positives: 4 out of a total of 60 predicted helices are incorrect (Table 1; Table S1 also includes modeling without pseudoknots and without DMS data). The resulting false-discovery rate (7%) is similar to the rate seen in prior mutate-and-map benchmarks (16). For new RNAs where false positives would require expensive subsequent experiments to falsify, M2-net (with a false-positive rate of <5%) may be preferred over RNAstructure analysis.

RNA Base Pair Detection in *Xenopus* Egg Extract. The simplicity of the "one-pot" M2-seq protocol and the positive predictive value of the M2-net analysis motivated us to test the method in a more complex biological environment than the in vitro folding conditions typically used in benchmarking new chemical-mapping methods. We mixed the P4–P6 RNA into undiluted extract from metaphase-arrested *Xenopus* eggs, a widely used medium for reconstituting eukaryotic biological processes (29). The impact of this complex medium, compared with in vitro conditions, was apparent in a modification signature that arose in extracto but not in vitro, even in the absence of DMS treatment: A's across the transcript were mutated to G (Fig. 5*A*). These modifications likely reflect the activity of the ADAR enzyme, which targets adenosines near double-stranded RNA helices for deamination to inosine, which is, in turn, read out as guanosine by reverse transcriptase (30, 31). Even with the complexities of the *Xenopus* egg extract environment, the 2D M2-seq data gave unambiguous signals for the P4–P6 RNA secondary structure. While these signals were less visually clear than in our in vitro experiments (Fig. 5*B*), they became more apparent when the data were viewed as *Z*-score maps (Fig. 5*C*). Despite an increase in background, partially due to A-to-I mutations, M2-net detected five of the eight helices of P4–P6 and no false positives (colored labels in Fig. 5*C*).

Discussion

Rapid detection of base-pairing partners in new noncoding RNAs has been difficult, requiring structural and biochemical techniques with low throughput, limited applicability, and/or poor predictive value. To address this challenge, we have introduced and tested a method called M2-seq. Mutations introduced at a low level ($\lesssim 10^{-3}$) during DNA or RNA synthesis disrupt local structure in the folded RNA and expose interacting nucleotides to reaction with DMS. This mutation and a partner that becomes exposed to DMS methylation leave correlated imprints on single molecules, enabling readout through reverse trancription and next-generation sequencing. M2-seq permits precise detection of the major structural elements of classic model systems such as the 158-nt P4–P6 domain of the *Tetrahymena* group I ribozyme and the 265-nt *Bacillus stearothermophilus* RNase P catalytic domain. M2-seq also reveals helices that have been difficult to detect or entirely missed in recent RNA-puzzles modeling for the GIR1 lariat-capping ribozyme, the adenosylcobalamin riboswitch, the *ydaO* cyclic–di-AMP riboswitch, and the Zika virus Xrn1-resistant genomic domain. Overall, the M2-seq data recover one-half of the helices in the tested RNAs with a low false-positive rate (<5%). Finally, the method enables pairwise structure inference for the majority of helices of the P4–P6 RNA in *Xenopus* egg extract. This study reports a biochemical technique enabling direct 2D visualization of RNA base pair partners—as opposed to 1D protections of uncertain origin—in a complex biological environment.

To supplement and automate simple visual inspection of M2-seq data, we have introduced the M2-net algorithm to infer helices from cross-diagonal signatures within the data, without bias from secondary-structure modeling methods that attempt to minimize a computed free energy. The M2-net algorithm is expected to be particularly important for scenarios that are not appropriately modeled with energy minimization methods, such as cases involving nontrivial tertiary structure or multiple secondary structures, molecules with long lengths, or systems reconstituted in complex environments where protein binding partners or molecular machines prevent the RNA from reaching equilibrium. Prior studies involving visual inspection of mutate-and-map data have correctly predicted tertiary contacts as well (12), and it will be important to test whether M2-net can be expanded to inferring such 3D information.

The presented M2-seq protocol is immediately applicable to 250-nt windows of lightly mutated RNAs introduced into complex

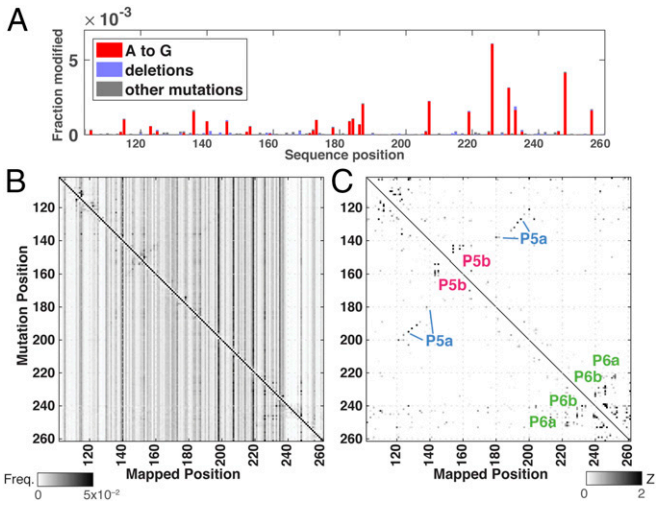

Fig. 5. M2-seq detects P4–P6 RNA base pairs in *Xenopus* egg extract. (*A*) Mutations across the P4–P6 transcript consistent with adenosine-to-inosine edits after exposure to undiluted *Xenopus* egg extract for 30 min (no DMS treatment); difference data with RNA incubated in vitro are shown. M2-seq data for P4–P6 RNA from templates prepared by error-prone PCR shown as (*B*) M2-seq map and then (*C*) transformed into *Z*-scores. Helix signatures automatically detected by M2-net are marked with colored labels.

biological environments. Synthetic long read sequencing or third-generation sequencing technologies may allow future studies to detect base pairings involving sequence separations longer than 250 nt (32–34). In terms of seeding mutations, applications to viruses and other systems that involve high-error rate RNA polymerases may obviate this step, but generally M2-seq in extracts, cells, and tissues will require transfecting DNA or RNA libraries that are prepared through error-prone PCR or other emerging techniques (33, 34). A faster and less biologically perturbing protocol would be enabled by a cell-permeable mutagen that could directly attack nucleotides initially sequestered inside RNA helices. While none of the routinely used chemical probes (e.g., DMS, SHAPE) appears appropriate, a large arsenal of mutagens remains to be tested for RNA structure mapping in vivo (35).

Methods

DMS mapping experiments on RNA were performed by modifying the RNA with DMS (170 mM final) in 10 mM $MgCl_2$ and 300 mM Na-cacodylate (pH 7.0) for 6 min at 37 °C, followed by quenching with β-mercaptoethanol and purification with ethanol precipitation. Experiments with *Xenopus* egg extract replaced ethanol precipitation with purification by TRIzol extraction and RNA Clean-and-Concentrator-5 columns (Zymo Research). Reverse transcription was performed in conditions that led to mutational read-through at methylated nucleotides (SuperScript II and Mn^{2+}), and sequencing libraries were prepared by PCR and sequenced on Illumina MiSeq instruments. ShapeMapper (36) was used to align sequencing reads to reference sequences and record mutations, and the results were converted to M2-seq data and mutation spectra using scripts available at https://github.com/ribokit/M2seq. Detailed descriptions of RNA preparation, modification experiments, map visualization, secondary-structure modeling by M2-net and RNAstructure executables, and DMS dose-dependent mutation rate analysis are provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank P. Cordero and S. Tian for initial RNAstructure analyses of previously collected MaP data; the Herschlag laboratory for the pT7L-21 plasmid; B. French and A. Straight for the generous gift of *Xenopus* egg extract; and D. Mathews and R. Watson for incorporating extensions into RNAstructure software. We acknowledge funding from the National Institutes of Health [Grant 5 T32 GM007276 (to C.Y.C.); Grant R01 GM102519 (to R.D.)] and the Burroughs Wellcome Fund [Grant CASI 1007326.01 (to R.D.)].

1. Leamy KA, Assmann SM, Mathews DH, Bevilacqua PC (2016) Bridging the gap between in vitro and in vivo RNA folding. *Q Rev Biophys* 49:e10.
2. Tian S, Das R (2016) RNA structure through multidimensional chemical mapping. *Q Rev Biophys* 49:e7.
3. Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA* 106:97–102.
4. Kladwang W, VanLang CC, Cordero P, Das R (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* 50:8049–8056.
5. Cordero P, Kladwang W, VanLang CC, Das R (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* 51: 7037–7039.
6. Hajdin CE, et al. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci USA* 110:5498–5503.
7. Tian S, Cordero P, Kladwang W, Das R (2014) High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states. *RNA* 20: 1815–1826.
8. Eddy SR (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys* 43:433–456.
9. Watts JM, et al. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460:711–716.
10. Pollom E, et al. (2013) Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. *PLoS Pathog* 9:e1003294.
11. Sükösd Z, et al. (2015) Full-length RNA structure prediction of the HIV-1 genome reveals a conserved core domain. *Nucleic Acids Res* 43:10168–10179.
12. Miao Z, et al. (2015) RNA-puzzles round II: Assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* 21:1066–1084.
13. Miao Z, et al. (2017) RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* 23:655–672.
14. Sergiev PV, Dontsova OA, Bogdanov AA (2001) [Study of ribosome structure using the biochemical methods: Judgment day]. *Mol Biol (Mosk)* 35:559–583.
15. Lu Z, et al. (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 165:1267–1279.
16. Kladwang W, VanLang CC, Cordero P, Das R (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat Chem* 3:954–962.
17. Cordero P, Das R (2015) Rich RNA structure landscapes revealed by mutate-and-map analysis. *PLoS Comput Biol* 11:e1004473.
18. Miao Z, et al. (2015) RNA-puzzles round II: Assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* 21:1066–1084.
19. Furey WS, et al. (1998) Use of fluorescence resonance energy transfer to investigate the conformation of DNA substrates bound to the Klenow fragment. *Biochemistry* 37: 2979–2990.
20. Krokhotin A, Mustoe AM, Weeks KM, Dokholyan NV (2017) Direct identification of base-paired RNA nucleotides by correlated chemical probing. *RNA* 23:6–13.
21. Homan PJ, et al. (2014) Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci USA* 111:13858–13863.
22. Cate JH, et al. (1996) Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* 273:1678–1685.
23. Wilson DS, Keefe AD (2001) Random mutagenesis by PCR. *Curr Protoc Mol Biol* Chap 8:Unit 8.3.
24. Akiyama BM, et al. (2016) Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science* 354:1148–1152.
25. Kladwang W, Das R (2010) A mutate-and-map strategy for inferring base pairs in structured nucleic acids: Proof of concept on a DNA/RNA helix. *Biochemistry* 49: 7414–7416.
26. Kladwang W, Cordero P, Das R (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA* 17:522–534.
27. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
28. Kazantsev AV, et al. (2011) Solution structure of RNase P RNA. *RNA* 17:1159–1171.
29. Desai A, Murray A, Mitchison TJ, Walczak CE (1999) The use of *Xenopus* egg extracts to study mitotic spindle assembly and function in vitro. *Methods Cell Biol* 61:385–412.
30. Bass BL, Weintraub H (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* 55:1089–1098.
31. Wagner RW, Smith JE, Cooperman BS, Nishikura K (1989) A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and *Xenopus* eggs. *Proc Natl Acad Sci USA* 86: 2647–2651.
32. Stapleton JA, et al. (2016) Haplotype-phased synthetic long reads from short-read sequencing. *PLoS One* 11:e0147229.
33. Mostovoy Y, et al. (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* 13:587–590.
34. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol* 17:239.
35. Singer B, Kuśmierek JT (1982) Chemical mutagenesis. *Annu Rev Biochem* 51:655–693.
36. Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* 11:959–965.
37. Tian S, Yesselman JD, Cordero P, Das R (2015) Primerize: Automated primer assembly for transcribing non-coding RNA domains. *Nucleic Acids Res* 43:W522–W526.
38. Kosuri S, Church GM (2014) Large-scale de novo DNA synthesis: Technologies and applications. *Nat Methods* 11:499–507.
39. Costello M, et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* 41:e67.
40. Wons E, Furmanek-Blaszk B, Sektas M (2015) RNA editing by T7 RNA polymerase bypasses InDel mutations causing unexpected phenotypic changes. *Nucleic Acids Res* 43:3950–3963.
41. Kladwang W, Hum J, Das R (2012) Ultraviolet shadowing of RNA can cause significant chemical damage in seconds. *Sci Rep* 2:517.
42. Cordero P, Kladwang W, VanLang CC, Das R (2014) The mutate-and-map protocol for inferring base pairs in structured RNA. *Methods Mol Biol* 1086:53–77.
43. Zaug AJ, Grosshans CA, Cech TR (1988) Sequence-specific endoribonuclease activity of the *Tetrahymena* ribozyme: Enhanced cleavage of certain oligonucleotide substrates that form mismatched ribozyme-substrate complexes. *Biochemistry* 27:8924–8931.
44. Guse A, Fuller CJ, Straight AF (2012) A cell-free system for functional centromere and kinetochore assembly. *Nat Protoc* 7:1847–1869.
45. Moree B, Meyer CB, Fuller CJ, Straight AF (2011) CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. *J Cell Biol* 194:855–871.
46. Seetin MG, Kladwang W, Bida JP, Das R (2014) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods Mol Biol* 1086:95–117.
47. Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* 10:1643–1669.
48. Tian S, Cordero P, Kladwang W, Das R (2014) High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states. *RNA* 20:1815–1826.