Published in final edited form as:

Biochemistry. 2017 September 05; 56(35): 4713–4721. doi:10.1021/acs.biochem.7b00323.

Interpreting reverse transcriptase termination and mutation events for greater insight into the chemical probing of RNA

Alec N. Sexton, Peter Y. Wang, Michael Rutenberg-Schoenberg, and Matthew D. Simon*

Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT 06511, Chemical Biology Institute, Yale University, West Haven, CT 06516, USA

Abstract

Chemical probing has the power to provide insight into RNA conformation *in vivo* and *in vitro*, but interpreting the results depends on methods to detect the chemically modified nucleotides. Traditionally, the presence of modified bases was inferred from their ability to halt reverse transcriptase during primer extension and the locations of termination sites observed by electrophoresis or sequencing. More recently, modification-induced mutations have been used as a readout for chemical probing data. Given variable propensity for mismatch incorporation and read-through with different reverse transcriptases, we examined how termination and mutation events compare to each other in the same chemical probing experiments. We found that mutations and terminations induced by dimethyl sulfate probing are both specific for methylated bases, but these two measures have surprisingly little correlation and represent largely non-overlapping indicators of chemical modification data. We also show that specific biases for modified bases depend partly on local sequence context, and that different reverse transcriptases show different biases toward reading a modification as a stop or a mutation. These results support approaches that incorporate analysis of both termination and mutation events into RNA probing experiments.

Keywords

Chemical probing; RNA structure; noncoding RNA

Introduction

There has been a renaissance in the use of chemical probing experiments to infer functional and structural information about RNAs.^{1–4} Advances in sequencing technology have driven high-throughput analyses of experiments detecting primer extension products that were previously performed using primer extension products separated by electrophoresis.^{5,6} These new methods allow *in vivo* and *ex vivo* transcriptome-wide or targeted probing with chemical or enzymatic methods, yielding insight into RNA structure in multiple different organisms, including RNAs that were often too long, difficult to purify, or heterogeneous.^{4,6–14} For example, we have used targeted reverse transcription primers for deep sequencing to construct the first conformational models across the entire 18 kb of mouse Xist lncRNA using Targeted Structure-Seq.¹²

* matthew.simon@yale.edu.

These probing approaches depend upon the chemistry used for probing and also the ability of a reverse transcriptase (RT) to interrogate the chemical adducts. Two of the most common reagents used in RNA chemical probing are dimethyl sulfate (DMS) and selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) reagents.^{15,16} These chemicals target different parts of the RNA structure, but both are used to infer which regions of an RNA are constrained by base pairs. Here we focus on DMS, which leads to methylation at the N1 and N3 positions of unpaired adenosine and cytosine, respectively.¹⁷

By adapting the readout of chemical probing experiments to deep sequencing, this new format provides the opportunity to probe adduct-induced mutations installed by the RT (RT mutations) in addition to the prematurely terminated reverse transcription products (RT stops). This facilitates normalization of data and creates the possibility of analyzing covariation of mutational frequencies at different sites.¹⁸ These advantages have recently led many researchers away from the use of RT stops in favor of mutational mapping strategies.^{3,19} Thus far stops and mutations have been taken as interchangeable markers of modification, but some evidence exists that, in at least naturally occurring m¹A observed in RNA-seq data, modified bases show different propensities for causing stops or mutations, and this at least partially depends on sequence context.²⁰

Having previously established an approach to examine RT stops,¹² we decided to compare the information gleaned from RT stops and RT mutations induced by chemical probing. While others have examined these chemical-induced events separately, in order to facilitate a direct comparison we focused on chemical probing data that contain both termination and mutation information from the same sequencing reads, corresponding to RNA structure in the same experiments. Unexpectedly, we found that RT mutations and stops are poorly correlated, are subject to different biases, and represent largely discrete sets of information from chemical probing of RNA structure. Thus, analyzing stops and mutations together provides a more complete picture of RNA structure. Although different RTs used for the library result in similar RNA structural information, we found that a recently characterized group II intron maturase²¹ led to identification of the maximal number of mutations. In all cases, the analysis was improved by including data from both termination and mutation events.

Materials and Methods

Cell culture and RNA preparation

MEFs were cultured in DMEM + 10% FBS (Gibco) with penicillin and streptomycin to 90% confluency in duplicate 15 cm plates. Treated cells were washed with PBS and then incubated with 0.5% DMS in PBS (5 min, 25 °C) before removing DMS solution and quenching with three washes of wash buffer (50 mM Tris pH 7.5, 100 mM NaCl, 3 mM MgCl₂, 40 mM β-mercaptoethanol). DMS-treated and untreated cells were collected in PBS by centrifugation (3 min, 1000 × *g*, 4 °C). Then cells were lysed with a Dounce homogenizer 15 times in 15 mM Tris pH 7.4, 60 mM KCl, 2 mM EDTA. The cells were collected by centrifugation (3 min, 1000 × *g*), and the supernatant was discarded. The remaining pellet was resuspended in TRIzol LS (Life Technologies) and RNA was extracted according to the manufacturer's instructions. Extracted RNA was treated with RQ1 DNase

(Promega) for 1 h at 37 °C, extracted with phenol/chloroform, and precipitated with ethanol. Denatured DMS-treated RNA was prepared by treating already-extracted cellular RNA with 0.5% DMS for 3 min at 55 °C, then quenching with equal volume of 1M DTT, with subsequent ethanol precipitation and 70% ethanol wash.

Reverse Transcription

For duplicate treated and untreated samples, RNA (1 µg) was brought to 12 µl with 10 pmol total of primers (primer sets were identical to those in Fang et al. 2015) The primer/RNA mix was brought to 70 °C for 5 min, then 4 °C for 2 min to anneal primers. Then 3 µl 5× reverse transcription buffer (for SSIII [Life Technologies]: 250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM MgCl₂, 100 mM DTT; for AMV [New England Biolabs]: 250 mM Tris-OAc, 375 mM KOAc, 40mM Mg(OAc)₂, 50 mM DTT, pH 8.3; for TGIRT-II [InGex]: 2.25M NaCl, 25 mM MgCl₂, 100 mM Tris-HCl pH 7.4, 50 mM DTT; for the Zhao et al. maturase: 250 mM Tris-HCl, pH 8.5, 500 mM KCl, 10 mM MgCl₂, 25 mM DTT) was added and mixed before incubating samples at 55 °C for 10 min. Then 5 µl RT mix was added (0.5 µl SSIII or AMV, or 1 µl TGIRT-II; 1 µl 5× RT buffer, 1 µl 100 mM DTT, 1 µl 10 mM dNTPs, to 5 µl with DEPC-treated dH₂O) and reactions were incubated for 45 min at 55 °C for SSIII, 42 °C for AMV or the maturase, or 60 °C for TGIRT-II. After reverse transcription, reactions were treated with 1 µl RNase A and 0.5 µl RNase H for 1 h at 37 °C.

Library Preparation

Reverse transcription reaction cDNA products were purified with AMPure XP (Beckman Coulter) beads by adding 1 volume beads solution (at room temperature), 3 volumes PEG solution (2.5 M NaCl, 20% PEG 8000), and 1 volume iPrOH, mixed and allowed to sit at room temperature for 15 min. Beads were immobilized for 10 min on a magnetic rack, then washed twice with fresh 80% EtOH, allowed to dry for 10 minutes, then eluted in 16 µl dH₂O. Ligation of 3' adaptor was performed by mixing 6.8 µl cDNA with 0.2 µl 100 µM 3' adaptor (/5Phos/NNNAGATCGGAAGAGCGTCGTGTAG/3Bio/), 1 µl CircLigase (Epicentre), 2 µl 10× buffer (0.5 M MOPS pH 7.5, 0.1 M KCl, 50 mM MgCl₂, 10 mM DTT), 1 µl 50 mM MnCl₂, 1 µl 1 mM ATP and 8 µl 50% PEG-8000. Ligation mixture was incubated at 65°C for 2 h, then 85°C for 15 min. For ligation of 3' adaptor using T4 RNA ligase 1, 2 µl 10× T4 RNA ligase buffer, 10 µl 50% PEG 8000, 1 µl 1 mM ATP, 1 µl 20 µM 3' adaptor, 1 µl T4 RNA ligase 1 (NEB), 5 µl cDNA. Reaction mixture was incubated overnight at 25 °C. Products were purified by adding 1 volume room temperature AMPure XP beads, mixing, and purifying as before, with elution in 16 µl dH₂O. Products were amplified for 4 cycles with Phusion (New England Biolabs) polymerase using primers matching the 5' and 3' adaptor sequences (5'-CAGACGTGTGCTCTTCCGATC-3'; 5'-CTACACGACGCTCTTCCGATCT-3'), with cycles of 98 °C, 20 s; 64 °C, 20 s; 72 °C, 90 s. Products were purified as before with 1 volume AMPure XP beads. Illumina TruSeq forward primer and indexed reverse primers were added in a final round of PCR for 4-8 cycles, and products were purified with 1 volume of AMPure XP beads as before. Multiplexed sequencing libraries were submitted for 2×75 nt paired end sequencing on an Illumina HiSeq 2500 by the Yale Center for Genome Analysis.

Analysis of stops and mutations

For both treated and untreated samples, FASTQ files resulting from sequencing were trimmed with Cutadapt to remove Illumina adaptor sequences and aligned to either mouse Xist or mouse 18S rRNA with Bowtie2 set to local alignment. For base quality filtering analyses, FASTQ files were filtered using FASTX-Toolkit. Software developed for counting stops and mutations from sequencing data (RTEventsCounter) was based on the ShapeMapper software by Smola et al.²² and the script is provided with the Supplementary Information. Paired-end reads were combined, where disagreements between paired reads were each resolved using the base call with the higher quality score. Multiple consecutive mismatches or deletions were treated as a single event on the 5'-most nucleotide of the cDNA. The 5'-most position of each combined read (start of reverse transcription) on the target sequence was matched with the expected starting position of the selected primers, with a ± 5 -nt tolerance window. Reads without primer allocations were treated as misprimed off-target reads and discarded. Base-calls beyond 800 nt from the primer start position were discarded. Events for nucleotides in the primer sites were not counted to avoid analysis of the annealing site.

To calculate the probabilities of these events, the number of read-throughs was counted for each nucleotide as the number of times a nucleotide was included in the combined reads, including the gap between paired-ends. Depth was calculated as the number of times a nucleotide was included in the combined reads, excluding the gap between paired-ends. The probability of a stop event for a given nucleotide was then measured in sequencing data on the nucleotide immediately 3' of the aligned section segment of the cDNA, where a modification would have been present and prompted a release of substrate by RT, as follows:

$$P_{\text{raw}}(\text{stop}) = \frac{n_{\text{stop}}}{n_{\text{stop}} + n_{\text{readthrough}}}$$

The number of stops divided by the number of read-through events plus the number of stops. The probability of mutation was calculated similarly:

$$P_{\text{raw}}(\text{mut}) = \frac{n_{\text{mut}}}{\text{depth}}$$

Nucleotides were subsequently filtered for total read-through above 100. Event probabilities were normalized by subtracting untreated values from treated values to generate P(mut) and P(stop) values^{23,24}:

$$\begin{aligned} P(\text{stop}) &= P_{\text{DMS}}(\text{stop}) - P_{\text{no treatment}}(\text{stop}) \\ P(\text{mut}) &= P_{\text{DMS}}(\text{mut}) - P_{\text{no treatment}}(\text{mut}) \end{aligned}$$

For analysis of local context effects, P(stop) and P(mut) values were binned by the identity of the nucleotide 3' of the modified nucleotide on the RNA. P values were determined by a

Wilcoxon test, with all A/U to G/C pairwise comparisons among A or C residues in treated samples giving $p < 0.01$.

For true positive vs. false positive comparison (receiver operator curves), $P(\text{mut})$ and $P(\text{stop})$ values were normalized to each other by calculating a Z -score for each set's probability of modification values then combining and ordering the values in descending order. For all bases for which there was modification data, true positives were considered bases with $>2\text{\AA}$ solvent accessible surface area in the crystal structure of the rabbit ribosome from 5FLX, with values transferred to homologous nucleotides for mouse.²⁵

For calling modified bases as stops, mutations, or both, we used an empirically defined threshold based on a Wald statistic ($p < 0.01$).

Results

We previously developed a targeted probing approach that is based on targeted reverse transcription analysis of termination events (Figure 1; Fang et al., 2015). Based on recent reports using mutational readouts,³ we revisited our previous chemical probing experiment. Although this had been optimized for measuring RT stops, we considered the possibility that mutations were also present in these sequencing data and could be used to complement stop information, with the ultimate goal of making structure predictions more robust.

Building on software that was created for analyzing stops or mutations separately in reads,^{12,22} we developed new software to make an integrated analysis pipeline in order to simultaneously identify and study both event types in the same high-throughput sequencing data. This allowed both stop and mutation readouts to be analyzed in parallel after filtering for reads with high quality alignments, reverse transcription length, and on-target priming. We also included filtering to consider only nucleotides with a threshold number of observations or read-throughs to ensure robust and reliable quantification. We used this approach to re-analyze data from mouse embryonic fibroblasts (MEFs) targeting mouse Xist, generating both mutation and stop probabilities. These data are normalized by subtracting untreated control probabilities from DMS-treated sample probabilities (Figure 1). DMS treatment of RNA offers a particular advantage for assessing differences between mutation and stop readouts because there are explicit expectations for what true positives are allowed in either case: only A and C bases that are single-stranded or next to single-stranded or non-Watson-Crick pairs will be modified.¹⁷

Examining the cumulative distribution of normalized stop and mutation probabilities from in-cell mouse Xist data revealed an enrichment of events at both A and C (Figure 2A). This A and C enrichment is a hallmark of DMS modification, supporting our hypothesis that information from adducts were encoded in our data as both mutation and termination events. Further, it was immediately apparent that stops and mutations carry different biases; mutations were biased toward C and stops were biased toward A (Figure 2A), consistent with what has previously been observed when measuring only DMS-induced stops.⁷

As both mutation and stop events are induced from the same DMS-modifications on the target RNA, we initially hypothesized that these two measures would be largely redundant,

Author Manuscript

differing only by a slight A-versus-C bias. To test this hypothesis, we compared the stop and mutation events by examining the modification probabilities at nucleotide level. Specifically, we examined the relationship between the probability of mutation events ($P(\text{mut})$) and the probability of termination events ($P(\text{stop})$), expecting that these would be highly correlated. In contrast to our hypothesis, we found surprisingly little correlation ($r = 0.05$, Figure 2B), demonstrating that these events are not mutually interchangeable as readouts of chemical modification. If this result were general, we reasoned that this observation would have two consequences: (1) the details of the enzymology of the reverse transcriptase and its substrate recognition may have a greater influence on RNA-probing readouts than expected, and (2) as many DMS-induced adducts were not detected by one or the other readout, combined analysis of mutation and termination events may provide insight into RNA probing experiments superior to either analysis alone.

Author Manuscript

To test whether the results from our Xist probing data were also valid for other RNAs, we chose to examine data from mouse 18S. This RNA allowed us to compare the mutation and stop probabilities to its known base-pairing structure, enabling us to test whether the DMS-induced stops and mutations both reflected *bona fide* accessible regions on the RNA. We prepared biological duplicate samples for analysis using SuperScript III (SSIII) RT, as done previously with mouse Xist. Indeed, we found the same bias for C residues at mutations and A at stops, as well as a low positive correlation between these events, for 18S ($r = 0.468$, Figure 3A,B). As expected, both the mutation and termination events were found at unpaired bases. All values were highly reproducible ($P(\text{mut})$: $r = 0.988$, $P(\text{stop})$: $r = 0.999$; Figure S1). We also looked at the untreated data to see if the bias in DMS-treated samples was an artifact of subtraction of control signal. However, untreated controls show little bias between bases, suggesting the disparity in A and C readout in DMS-treated samples is downstream of DMS-modification (Figure S2). Additionally, we considered whether this bias might be a result of either poor read qualities or noise due to nucleotides with few observations. Neither highly stringent filtering of raw sequencing reads based on base quality nor filtering of nucleotides based with greater than 10,000 readthrough events altered the A and C bias (Figure S3).

Author Manuscript

The generality of the low correlation between stops and mutations led us to consider the origin of the differences between these two read-outs. During reverse transcription, when the RT encounters an alkylated base, there is a kinetic competition between two outcomes: elongation through the adduct (likely producing a mutation), or termination of the growing cDNA (Figure 1). As suggested for naturally occurring m^1A modifications on RNA²⁰ our chemical probing results demonstrate mutation and stop events may be context dependent. We show that each nucleotide leads to reproducible preferences for the RT to either produce a mutation or terminate. We next examined what factors might influence this competition. Similar to our results from Xist, in 18S there was also a clear preference of the RT to produce mutations across from alkylated C residues, and termination events at alkylated A residues. This is likely due to a number of factors, including steric constraints within the RT active site and hydrogen bonding between the substrate and incoming bases.^{26,27} Contrary to this trend, a subset of C residues show a clear and reproducible preference for termination, with a subset of A residues showing a preference for mutation (e.g., open and closed

arrowhead in Figure 2B). This led us to wonder what other context is likely to influence the competition between stops and mutations.

We examined whether local sequence context influences how likely a polymerase is to terminate or incorporate mutations. To test this, starting in mouse Xist data, we measured the normalized probability of stops or mutations for all bases compared to the base immediately 3' in the RNA, which forms the last complete base pair in the RNA:DNA duplex in the active site of the polymerase (Figure 3C). While we did not find any trend for the influence of nearby bases on the probability of mutations, we found that stop probabilities were significantly higher in cases where there was either A or U immediately 3' in the RNA. The context bias in 18S generally agrees with that of Xist (Figure 3D) although the smaller number of nucleotides and perhaps a larger portion of paired bases in 18S make analysis less robust. Based on analogy with DNA polymerases,²⁸ these results support a model where the strength of the RNA:DNA duplex and the number of hydrogen bonds influence the propensity of the RT to dissociate, resulting in a stop. Observation of this trend in untreated controls at all bases suggests that this bias is a general property of the RT itself (Figure S3A). Context biases that are dependent on duplex formation should be attenuated when mutations or stops are analyzed by the subsequent (or 5') nucleotide in the RNA. Interestingly, while the stop bias at A was reduced in this analysis, an increase in stop events at some C nucleotides is still evident. Several points suggest this bias is generated at the level of the RT rather than by RNA structure during DMS probing. If that were the case, the bias would be expected to appear in mutations as well. Additionally, while the bias is observable at the level of raw signal in DMS-treated samples, denatured DMS-treated RNA, where structure is disrupted, still retains the same bias in the raw signal, suggesting that the phenomenon is not due to RNA structure.

We considered whether this trend also holds across different reaction conditions, as it was possible that our conditions accidentally caused RT to be precariously balanced between mutations and termination events. To see if we could tip the balance toward mutations, we examined the use of conditions containing Mn^{2+} , which has previously been used with DMS and other reagents to increase the likelihood of RT mutations.^{3,22} Under these conditions performed side-by-side with Mg^{2+} conditions, we found a shift toward mutations as expected (Figure S5). However, there were still many modifiable bases that caused RT stops but remained undetectable by mutation. Another possibility is that the library preparation inherent to a sequencing-based readout might contribute to the distribution of RT stops or mutations. Specifically, bias in specificity of the ligase used for 3' adaptor ligation can affect the profile of product when making libraries for sequencing. Although both background subtraction and the conditions used in our preparations reduce this bias in the processed data, we tested whether the A and C bias in stops and mutations would be apparent when using a different single stranded ligase for library preparation.^{23,24,29,30} When T4 RNA ligase 1 was used for 3' adaptor ligation, the prevalence of C detected as mutations and A detected as stops was consistent (Figure S6). When nucleotide-level modification probabilities were compared between T4 RNA ligase 1 and CircLigase, $P(\text{stop})$ was correlated, though with some differences in degree of modification observed, whereas $P(\text{mut})$ was more strongly correlated, consistent with mutation readout as independent of the cDNA end and therefore independent of biases from the ligase specificity.

Since the propensity for stops or mutations is inherent to the enzymology of the RT, we considered whether this phenomenon is specific to SSIII, and whether other RTs can bias the reaction toward either type of event. Indeed, recently other RTs have been used in RNA probing to increase mutational frequencies without the need for substituting with Mn^{2+} .³ Therefore, we compared mutation and stop biases across several RTs. We treated MEFs with DMS in biological duplicates and made libraries using Avian Myeloblastosis Virus (AMV) RT, thermostable group II intron (TGIRT-II) RT, or a recently studied group II intron maturase.^{21,31} Examining the cumulative distributions of event probabilities showed that, across all RTs, there was a preference for modification at A and C (Figure 4A). The additional slight preference for C in mutations and A in stops was present, but variable, between the different RTs. A comparison of $P(\text{mut})$ and $P(\text{stop})$ at individual nucleotide level showed a lack of positive correlation for all RTs, as with the Xist data set (Figure 4B). This suggests that it is a general principle that mutation and stop information has limited overlap for all RTs. Among individual RTs, AMV and SSIII mostly manifest DMS modifications as RT stops, while TGIRT-II and the Zhao et al. maturase are biased toward mutations.

Another way to assess global readout of DMS-induced modification is to measure false positive rate against true positive rate using an ROC curve. Mutation and stop probabilities were compared either separately, or by combining this information by normalizing them to each other (Figure S7). For all polymerases, the performance in the ROC curve reflected the specific RT bias for either stops or mutations; e.g. AMV performed better with stops and TGIRT-II performed better measured by mutations. In all cases, combining mutations and stops resulted in a ROC curve as good or better than either stops or mutations alone.

The diverse behaviors of different RTs led us to examine which specific nucleotides in 18S RNA were identified by stop or mutation events for each RT. In order to examine how nucleotide-level modification data mapped onto the 18S structure, we focused on the stem-loop structure between nucleotides 803-865 (Figure 5A,B). We defined nucleotides as positive for stops or mutations based on an empirical threshold (see Materials and Methods). As anticipated, all nucleotides identified as modified due to either high mutation or high stop frequencies occurred in regions of structure that are considered modifiable based on known accessibility to DMS (single-stranded, adjacent to a single-stranded nucleotide, or adjacent to a non-Watson-Crick pair). As in the overall comparison of mutations and stops presented earlier, AMV and SSIII produced mostly stops, whereas TGIRT-II and the Zhao et al. maturase showed a greater shift toward mutations. We categorized all modified nucleotides as being either a mutation, a stop, or both (Figure 5C). About 20% of nucleotides measured as positive with AMV and SSIII were seen as both stops and mutations, about 22% for TGIRT-II, and 36% for Zhao. The total number of nucleotides labeled as modified was the greatest for SSIII, and the least for TGIRT-II. The Zhao et al. maturase performed particularly well in identifying nucleotides through both mutations and stops when compared with the other enzymes, and revealed more modifiable nucleotides overall when compared with TGIRT-II, which is of the same class of enzyme. Even so, for all enzymes, considering both stops and mutations resulted in an increase in the total number of nucleotides identified as single-stranded, compared with using either mutations or stops alone.

Discussion

Chemical probing experiments are founded on the assumption that the probability of a termination or mutation event is proportional to the degree of modification. While it is clear that higher degrees of modification lead to an increase in stops or mutations, our data clearly demonstrate that there are modified bases largely invisible to either readout alone. It is not yet known what dictates the propensity of a given RT to stop or mutate when it encounters a modified base, but we show that there is a strong bias toward stops among those that occur next to an AT base pair.

As previously reported, we found that group II intron-derived enzymes are more processive than the other enzymes, and concordantly tend to misincorporate rather than dissociate at modified nucleotides. Under our conditions, AMV produced more DMS-independent stop events, and mostly resulted in stops upon encountering methylated nucleotides. SSIII similarly produced mostly stops and fewer mutations. All enzymes produced useable data that reflect the structure of a known RNA, 18S rRNA, providing confidence that a variety of enzymes are suitable for these assays. Strikingly, the recent group II intron maturase was able to identify more nucleotides as both stops and mutations, but even in that case, some nucleotides were missed when considering mutations alone.

Although both stop and mutation events faithfully report modified nucleotides, our results unambiguously demonstrate that a chemical modification does not always lead to either an RT stop event or mutation event. The probability of a modification-induced stop or mutation event is both context-dependent and dependent on the RT used in the experiment. It is necessary to consider the specific enzymology of reagents used in the assay, as these can bias the modifications that are observed. Integrating both mutations and stops in chemical probing data is one way to mitigate this bias and to provide greater insight into RNA structure from probing experiments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Antonio Giraldez, and members of the Simon lab for helpful discussions. We thank Chen Zhao and Anna Pyle for the generous gift of the maturase enzyme. This work was supported by the Anderson Fellowship (A.N.S.); Yale Dean's Research Fellowship for the Sciences (P.Y.W.); NIH New Innovator Award DP2 HD083992-01 (M.D.S.), and a Searle scholarship (M.D.S.).

References

1. Liu F, Somarowthu S, Pyle AM. Visualizing the secondary and tertiary architectural domains of lncRNA RepA. *Nat Chem Biol.* 2017; 13:282–289. [PubMed: 28068310]
2. Strobel EJ, Watters KE, Loughrey D, Lucks JB. RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs. *Curr Opin Biotechnol.* 2016; 39:182–191. [PubMed: 27132125]
3. Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS, Rouskin S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods.* 2016; 14:75–82. [PubMed: 27819661]

4. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM, Weeks KM. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. 2009; 460:711–6. [PubMed: 19661910]
5. Silverman IM, Berkowitz ND, Gosai SJ, Gregory BD. Genome-Wide Approaches for RNA Structure Probing. *Adv Exp Med Biol*. 2016; 907:29–59. [PubMed: 27256381]
6. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*. 2013; 505:696–700. [PubMed: 24270811]
7. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*. 2013; 505:701–705. [PubMed: 24336214]
8. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. Genome-wide measurement of RNA secondary structure in yeast. *Nature*. 2010; 467:103–107. [PubMed: 20811459]
9. Bevilacqua PC, Ritchey LE, Su Z, Assmann SM. Genome-Wide Analysis of RNA Secondary Structure. *Annu Rev Genet*. 2016; 50:235–266. [PubMed: 27648642]
10. Talkish J, May G, Lin Y, Woolford JL, McManus CJ. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*. 2014; 20:713–20. [PubMed: 24664469]
11. Kielpinski LJ, Boyd M, Sandelin A, Vinther J. Detection of reverse transcriptase termination sites using cDNA ligation and massive parallel sequencing. *Methods Mol Biol*. 2013; 1038:213–31. [PubMed: 23872978]
12. Fang, R., Moss, WN., Rutenberg-Schoenberg, M., Simon, MD. Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. In: Sanbonmatsu, KY., editor. *PLOS Genet*. Vol. 11. 2015. p. e1005668
13. Smola MJ, Christy TW, Inoue K, Nicholson CO, Friedersdorf M, Keene JD, Lee DM, Calabrese JM, Weeks KM. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci U S A*. 2016; 113:10322–7. [PubMed: 27578869]
14. Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, Valladares O, Yang J, Bambina S, Sabin LR, Murray JI, Lamitina T, Raj A, Cherry S, Wang LS, Gregory BD. Global Analysis of RNA Secondary Structure in Two Metazoans. *Cell Rep*. 2012; 1:69–82. [PubMed: 22832108]
15. Weeks KM. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol*. 2010; 20:295–304. [PubMed: 20447823]
16. Ziehler WA, Engelke DR. Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem*. 2001; Chapter 6(Unit 6.1)
17. Peattie DA, Gilbert W. Chemical probes for higher-order structure in RNA. *Proc Natl Acad Sci U S A*. 1980; 77:4679–82. [PubMed: 6159633]
18. Homan PJ, Favorov OV, Lavender CA, Kursun O, Ge X, Busan S, Dokholyan NV, Weeks KM. Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci U S A*. 2014; 111:13858–63. [PubMed: 25205807]
19. Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods*. 2014; 11:959–965. [PubMed: 25028896]
20. Hauenschild R, Tserovski L, Schmid K, Thüning K, Winz ML, Sharma S, Entian KD, Wacheul L, Lafontaine DLJ, Anderson J, Alfonzo J, Hildebrandt A, Jäschke A, Motorin Y, Helm M. The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res*. 2015; 43:9950–64. [PubMed: 26365242]
21. Zhao C, Pyle AM. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol*. 2016; 23:558–65. [PubMed: 27136328]
22. 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; 10:1643–69. [PubMed: 26426499]
23. Aviran S, Lucks JB, Pachter L. RNA structure characterization from chemical mapping experiments. 2011

24. Aviran S, Trapnell C, Lucks JB, Mortimer SA, Luo S, Schroth GP, Doudna JA, Arkin AP, Pachter L. Modeling and automation of sequencing-based characterization of RNA structure. *Proc Natl Acad Sci U S A*. 2011; 108:11069–74. [PubMed: 21642536]
25. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*. 2016; 5:189. [PubMed: 26973785]
26. Kunkel TA, Bebenek K. DNA Replication Fidelity. *Annu Rev Biochem*. 2000; 69:497–529. [PubMed: 10966467]
27. Yu H, Goodman MF. Comparison of HIV-1 and avian myeloblastosis virus reverse transcriptase fidelity on RNA and DNA templates. *J Biol Chem*. 1992; 267:10888–96. [PubMed: 1375233]
28. Goodman MF, Fygenon KD. DNA polymerase fidelity: from genetics toward a biochemical understanding. *Genetics*. 1998; 148:1475–82. [PubMed: 9560367]
29. Kwok CK, Ding Y, Sherlock ME, Assmann SM, Bevilacqua PC. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Anal Biochem*. 2013; 435:181–186. [PubMed: 23399535]
30. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3'-adaptor ligation. *Nucleic Acids Res*. 2012; 40:e54. [PubMed: 22241775]
31. Nottingham RM, Wu DC, Qin Y, Yao J, Hunicke-Smith S, Lambowitz AM. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA*. 2016; 22:597–613. [PubMed: 26826130]

1. Chemical probing of RNA 2. Synthesis of cDNA from modified RNA

3. Counting and normalization

Figure 1.

Detection of chemically modified bases by Targeted Structure-Seq. Cells are treated with DMS to modify RNA *in vivo*. Then cDNA is synthesized with an RT, with cDNA ends corresponding to an RT dissociation upon encountering an adduct, and mutations corresponding to RT misincorporation across from the adduct during strand extension. After library preparation and sequencing, both mutations and stops are counted from the same sample and the frequencies of these events in untreated controls are subtracted from those in the treated samples.

Figure 2.

Stops and mutations are present in Targeted Structure-Seq data. (A) Cumulative distribution functions of P(stop) or P(mut) values for all probed nucleotides in mouse Xist showing higher signal at both A and C. (B) A scatterplot of P(mut) vs. P(stop) for all probed nucleotides in mouse Xist. Pearson's r reported (95% CI). Open arrowhead marks an example of a C nucleotide that is read out only as a stop, closed arrowhead marks an example of an A nucleotide that is read out only as a mutation.

Figure 3.

Probing of mouse 18S with DMS, with SSIII RT. (A) Cumulative distribution plots that demonstrate enrichment of A and C in $P(\text{stop})$ and $P(\text{mut})$. (B) A scatterplot of all probed nucleotides in mouse 18S rRNA. (C) Distribution of $P(\text{stop})$ and $P(\text{mut})$ values binned by the nucleotide 3' of the modified nucleotide on the mouse Xist RNA. Data were analyzed by a Wilcoxon test, with all A/U to G/C pairwise comparisons among residues in treated Xist samples giving $p < 0.01$, with the exception of reference C, penultimate nucleotide A. (D) Distribution of $P(\text{stop})$ and $P(\text{mut})$ values binned by the nucleotide 3' of the modified

nucleotide on mouse 18S. Note that the low value of normalized G and U with A and U penultimate residues is due to high modification at those nucleotides in untreated samples, and underrepresentation of those residues as stops in DMS treated samples.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 4. Comparison of stop and mutation values for AMV, TGIRT-II, and the Zhao et al. maturase. (A) Cumulative distribution functions for P(stop) and P(mut). (B) Scatterplots showing relationship between P(mut) and P(stop) for all probed nucleotides in mouse 18S.

Figure 5.

P(stop) and P(mut) comparison across RTs reveals different readout biases. (A) Nucleotide-level P(mut) and P(stop) values for the mouse 18S nucleotides 803-865. Labels are provided for comparison and do not necessarily represent nucleotides that pass a given threshold. (B) Nucleotides categorized as mutation, stop, or both, are displayed on the secondary structure of the hairpin comprising nucleotides 803-865. (C) Categorized nucleotides are compared in a stacked box plot for all RTs, with absolute numbers of called nucleotides shown.