**FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing**

Jason G Underwood, Andrew V Uzilov, Sol Katzman, Courtney S Onodera, Jacob E Mainzer, David H Mathews, Todd M Lowe, Sofie R Salama & David Haussler

<table>
<thead>
<tr>
<th>Supplementary Figure 1</th>
<th>Digestion of mouse U1a and 5S rRNA by P1 nuclease with or without nuclear RNA present.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Figure 2</td>
<td>Distribution of read mappings by type of annotation.</td>
</tr>
<tr>
<td>Supplementary Figure 3</td>
<td>Comparison of coverages in nuclease sample versus control sample for RNAs of several biological categories.</td>
</tr>
<tr>
<td>Supplementary Figure 4</td>
<td>Unprocessed end counts for U1a displayed in a secondary structure context.</td>
</tr>
<tr>
<td>Supplementary Figure 5</td>
<td>Correlation between cutting scores obtained from mouse embryonic stem cell versus neural precursor cell samples for select RNAs.</td>
</tr>
<tr>
<td>Supplementary Figure 6</td>
<td>Products of enzymatic structure probing of <em>in vitro</em> transcribed U15b, U22, and U97.</td>
</tr>
<tr>
<td>Supplementary Figure 7</td>
<td>FragSeq probing versus conventional nuclease probing of U22 C/D box snoRNA.</td>
</tr>
<tr>
<td>Supplementary Figure 8</td>
<td>FragSeq probing versus conventional nuclease probing of U97 C/D box snoRNA.</td>
</tr>
<tr>
<td>Supplementary Figure 9</td>
<td>Positions of RNA fragment ends relative to native ends of their parent RNA.</td>
</tr>
<tr>
<td>Supplementary Figure 10</td>
<td>Initial hypotheses for U15b secondary structure.</td>
</tr>
<tr>
<td>Supplementary Figure 11</td>
<td>Initial hypothesis for U22 secondary structure.</td>
</tr>
<tr>
<td>Supplementary Figure 12</td>
<td>Initial hypotheses for U97 secondary structure.</td>
</tr>
<tr>
<td>Supplementary Table 1</td>
<td>Summary statistics for sequencing data.</td>
</tr>
<tr>
<td>Supplementary Note 1</td>
<td>Formal description of the FragSeq algorithm.</td>
</tr>
<tr>
<td>Supplementary Note 2</td>
<td>Rationale for design of the computational pipeline and the FragSeq algorithm.</td>
</tr>
<tr>
<td>Supplementary Note 3</td>
<td>Details on data analysis and structure model building.</td>
</tr>
<tr>
<td>Supplementary Discussion</td>
<td>Interpretation of FragSeq cutting score data. Recommended sequencing coverage for a FragSeq experiment.</td>
</tr>
</tbody>
</table>

*Note: Supplementary Software and Supplementary Data 1–5 are available on the Nature Methods website.*
Supplementary Figure 1 | Digestion of mouse U1a and 5S rRNA by P1 nuclease with or without nuclear RNA present. 3’-radiolabeled, in vitro transcribed RNA was digested. Lanes showing RNA at start and end of reaction without nuclease are controls for non-specific degradation. Size markers are Ambion Decade markers; sizes of 150, 100, and 50 bases are highlighted explicitly. U1a labeling is as in Figure 2. 5S rRNA structure and helix numbers are from [31]. IL, interior loop; SL, stem-loop.
Supplementary Figure 2 | Distribution of read mappings by type of annotation. See Supplementary Note 3 for a description of how read mappings were assigned to annotations. Note that this figure shows read mappings, not reads; the number of read mappings will be greater than read counts in Supplementary Table 1 because multiple mappings are allowed. “Other ncRNA and ncRNA repeats” are ncRNAs in the UCSC Genome Browser mm9 RepeatMasker track, which include 7SK RNA, scRNA, SRP RNA, rRNA, snRNA, and tRNA, although many snRNA and tRNA mappings are removed earlier and assigned, respectively, to the “snRNA” and “GtRNAdb tRNA” categories (GtRNAdb contains a more conservative set of tRNA annotations [32] than RepeatMasker).
Supplementary Figure 3 | Comparison of coverages in nuclease sample versus control sample for RNAs of several biological categories. Each point represents a single RNA annotation. For genes that have multiple genomic copies (e.g. spliceosomal snRNA), a single copy was selected as the best representative. Coverages are for samples from UNDIFF cells. RNA categories are partitioned by size to indicate RNAs that would pass size selection for sequencing without digestion by nuclease (small), RNAs that require at least one cleavage to be put in sequencing size range (large), and RNAs spanning size selection boundary (medium). Axes are log scale; coverage (mean reads per base) plus one is shown on axes to avoid log-of-zero issues, so a value of one indicates no coverage obtained for the RNA. Three C/D box snoRNAs whose structure was probed in this study are highlighted.
Supplementary Figure 4 | Unprocessed end counts for U1a displayed in a secondary structure context. Compare with cutting scores after computational processing, shown in Figure 2G. Data is shown for UNDIFF cells. Nuclease and control sample end counts are exactly the same as in Figure 2C, except shown superimposed on the known solution secondary structure, scaled so that minimum (non-zero) and maximum scores within each sample have equal arrow sizes between the two panels. Arrow sizes corresponding to minimum (non-zero) and maximum counts are shown.
Supplementary Figure 5 | Correlation between cutting scores obtained from mouse embryonic stem cell (UNDIFF) versus neural precursor cell (D5NP) samples for select RNAs. Each dot represents a site, showing how the cutting score for that site compares between UNDIFF and D5NP samples. Only sites for which cutting scores were obtained in both UNDIFF and D5NP samples are shown; therefore, the number of sites represented here may be smaller than in Figures 2G, 3B, and 3D for U1a, U3b, and U5, respectively.
Supplementary Figure 6 | Products of enzymatic structure probing of in vitro transcribed U15b, U22, and U97. The 5’ end of RNA was radiolabeled before digestion in all cases. a, U15b. b and c, U22, resolving bases 7 through 116 and 21 through 122, respectively. d and e, U97, resolving bases 9 through 132 and 23 through 165, respectively. b and d also show results of digestion at two temperatures (25°C and 37°C); all other gels are at 37°C.
Supplementary Figure 7 | FragSeq probing versus conventional nuclease probing of U22 C/D box snoRNA.  

a, FragSeq ssRNA cutting scores (bottom, dark blue) versus band quantification readouts by Safa based on two gels resolving 5’-end-labeled probing products (shown in Supplementary Fig. 6): V1 (pink), T1 (purple), RNase A (green), P1 (light blue). X-axis shows nucleotide numbering. Gray nucleotides in sequence show areas that were outside of the reliably quantifiable area on the respective gel. Secondary structure model determined from probing (see Supplementary Note 3) is also shown, with parentheses denoting Watson/Crick base pairs and dots denoting ssRNA. For nucleases with a specific base preference, triangles denote bases where a nuclease can cut: for T1, gray triangles show its substrate G nucleotides; for RNase A, triangles show its substrate pyrimidines (black is C and red is U). Note that some outlier values were truncated (shown with red zigzag lines) to highlight lower-valued scores. Some sites at the 3’ end may have been lost in the FragSeq data due to size selection.  

b, Probing data superimposed on our structure model, with probing enzymes color-coded as in panel a. Marginal, weak, moderate, or strong enzyme activity was inferred from manual resolution of gel and Safa-quantified band intensities from panel a (see Supplementary Note 3).  

c, FragSeq cutting scores superimposed on the same structure model as in panels a and b. Box positions (green) are from [20].

(A)
Supplementary Figure 8 | FragSeq probing versus conventional nuclease probing of U97 C/D box snoRNA.  a, FragSeq ssRNA cutting scores (bottom, dark blue) versus band quantification readouts by Safa based on two gels resolving 5’-end-labeled probing products (shown in Supplementary Fig. 6): V1 (pink), T1 (purple), RNase A (green), P1 (light blue).  X-axis shows nucleotide numbering.  Gray nucleotides in sequence show areas that were outside of the reliably quantifiable area on the respective gel.  Secondary structure model determined from probing (see Supplementary Note 3) is also shown, with parentheses denoting Watson/Crick base pairs and dots denoting ssRNA.  For nucleases with a specific base preference, triangles denote bases where a nuclease can cut: for T1, gray triangles show its substrate G nucleotides; for RNase A, triangles show its substrate pyrimidines (black is C and red is U).  Note that some outlier values were truncated (shown with red zigzag lines) to highlight lower-valued scores.  Note that Safa-quantified values include some spontaneous RNA degradation that occurs in the control lane (see Supplementary Fig. 6d and 6e).  b, Probing data superimposed on our structure model, with probing enzymes color-coded as in panel a.  Marginal, weak, moderate, or strong enzyme activity was inferred from manual resolution of gel and Safa-quantified band intensities from panel a (see Supplementary Note 3).  Spontaneous degradation in the gel control lane was visually subtracted out.  c, FragSeq cutting scores superimposed on the same structure model as in panels a and b.  Regions we judge to be probable box positions are highlighted in green.
Supplementary Figure 9 | Positions of RNA fragment ends relative to native ends of their parent RNA. Shown is data for U1a, U3b, and U5 for the nuclease-treated sample from UNDIFF cells. Dot color, on a continuous scale from gray to red, indicates how many reads were observed that started and ended at that position (gray = one read, red = maximum number of reads for that RNA, the exact count shown in the top right corner of a plot). The position is expressed as the distance between the 5’ end of the read and the 5’ end of the RNA versus the distance between the 3’ end of the read and the 3’ end of the RNA. Yellow highlighting indicates fragments that can only be produced by two or more breaks (their ends are >= 5 bases away from native RNA ends), i.e. they are inconsistent with the single-hit kinetics assumption. Only trim reads (defined in Supplementary Notes 1 and 2) are plotted, because only for those reads the true 3’ ends of the sequenced RNA fragments are known (the sequencing extended up to the SOLiD 3’ adapters, enabling us to identify both ends of the fragments).
Supplementary Figure 10 | Initial hypotheses for U15b secondary structure. Highlighting is as in Figures 5b and 5c, which show the final structure model. 

a, Initial predicted MFE structure and predicted ssRNA probabilities, with structure probing data superimposed on it for comparison. 

b, Structure probing data superimposed on a prior structure model from [33].
Supplementary Figure 11 | Initial hypothesis for U22 secondary structure. Initial predicted MFE structure and predicted ssRNA probabilities, with structure probing data superimposed on it for comparison. Box highlighting is as in Supplementary Figures 7B and 7C, which show the final structure model.
Supplementary Figure 12 | Initial hypotheses for U97 secondary structure. 

**a,** Initial predicted MFE structure and predicted ssRNA probabilities, with structure probing data superimposed on it for comparison. Box highlighting is as in Supplementary Figures 8B and 8C, which show the final structure model. 

**b,** An alternative model of a conserved helix according to the Rfam database, with structure probing data and ssRNA probabilities from **a** superimposed on it for comparison.

(A)

(B)
Supplementary Table 1 | Summary statistics for sequencing data. Data is shown for the six barcoded samples used in this study (out of 10 barcoded samples in the sequencing run). Total reads, minus filtered reads and genome-mapped reads, leaves unmapped reads.

<table>
<thead>
<tr>
<th>Barcode number</th>
<th>Sample prep type</th>
<th>Cell type</th>
<th>Total number of reads</th>
<th>Number of reads filtered out (% of total)</th>
<th>Number of reads mapped to genome (% of total)</th>
<th>Number of reads uniquely mapped to genome (% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>nuclease</td>
<td>UNDIFF</td>
<td>15,306,303</td>
<td>2,139,262 (13.98%)</td>
<td>4,585,725 (29.96%)</td>
<td>2,691,118 (17.58%)</td>
</tr>
<tr>
<td>5</td>
<td>nuclease</td>
<td>D5NP</td>
<td>9,734,816</td>
<td>2,289,740 (23.52%)</td>
<td>2,376,953 (24.42%)</td>
<td>1,198,250 (12.31%)</td>
</tr>
<tr>
<td>7</td>
<td>control</td>
<td>UNDIFF</td>
<td>13,058,357</td>
<td>797,064 (6.10%)</td>
<td>5,946,923 (45.54%)</td>
<td>4,998,826 (38.28%)</td>
</tr>
<tr>
<td>8</td>
<td>control</td>
<td>D5NP</td>
<td>20,441,839</td>
<td>1,717,398 (8.40%)</td>
<td>5,570,529 (27.25%)</td>
<td>4,283,683 (20.96%)</td>
</tr>
<tr>
<td>1</td>
<td>PNK</td>
<td>UNDIFF</td>
<td>15,911,200</td>
<td>2,393,180 (15.04%)</td>
<td>3,502,921 (22.02%)</td>
<td>2,028,276 (12.75%)</td>
</tr>
<tr>
<td>4</td>
<td>PNK</td>
<td>D5NP</td>
<td>14,000,788</td>
<td>4,652,857 (33.23%)</td>
<td>2,576,880 (18.41%)</td>
<td>916,470 (6.55%)</td>
</tr>
</tbody>
</table>
Supplementary Note 1

1 Formal description of the FragSeq algorithm

This section is primarily intended to clarify all technical details about the flow of data through the FragSeq algorithm. The overview of the algorithm is given in subsection “FragSeq methodology” of Results and the justification of some aspects of algorithm design is given in Supplementary Note 2. Specific values of user-specified algorithm parameters that were used in this study are listed in sample config files distributed with version 0.0.1 of the FragSeq source code. However, some additional rationale about specific technical points is given here. Those sample config files are cross-referenced with equations in this section for clarity.

This section is also intended to help understand the algorithm implementation in Python code (version 0.0.1, which is briefly described in Methods), and is generally isomorphic to the sequence of operations and to the data structures used in the code, except that instead of tracking non-ignored sites as described here, the code tracks sites to ignore. The representation used here was chosen because it is simpler to follow, and the general meaning (that all sites are partitioned into two sets: sites to consider when computing cutting scores versus sites to ignore) is the same between this description and the code.

1.1 Definitions

The FragSeq algorithm takes as input three sets of genome annotations, $M_{\text{nucl}}, M_{\text{con}}$, and $T$, where $M_{\text{nucl}}$ and $M_{\text{con}}$ contain mappings of reads from the nuclease and control samples respectively and $T$ contains transcripts for which the user desires a structural annotation. The algorithm computes and outputs cutting scores $c(i | t) > 0$ for each non-ignored site $i$ in each transcript $t \in T$.

We define a site as a position between two adjacent bases. So, a site is an inter-base coordinate where an annotation could begin or end. Although 5’ and 3’ ends of linear chromosomes (or scaffolds, contigs, transcripts, etc.) are technically exceptions to this definition since they do not have adjacent bases, we will still regard them as sites. A break in a RNA phosphate backbone is associated with one specific site defined by the adjacent bases, regardless of whether the scission occurred 5’ or 3’ of the connecting phosphate.

Each annotation $a$ has three properties associated with it: its genomic strand $a.S \in \{+, -\}$ and its left and right end coordinates $a.L$ and $a.R$ in the genomic coordinate space and on the genomic “+” strand (so $a.L \leq a.R$, regardless of $a.S$). These coordinates are also inter-base; that is, they are indices of sites that are the start and end of an annotation. An annotation may have other information associated with it (such as sequence, quality of mapping to genome, extent of multiple mapping, etc.), but at the present we do not use any such information in our algorithm, although statistics on some of these and other information are gathered and logged during program execution (as described in Methods).

An annotation spans consecutive bases and does not contain gaps, introns, or any other
elements that break this contiguity. Therefore, the number of bases $N(a)$ in some annotation $a$ is always

$$N(a) = a.R - a.L$$ (1)

and the number of sites is always $N(a) + 1$.

Note that $M_{\text{nuc}}$, $M_{\text{con}}$, and $T$ are multisets in the sense that several annotations in a set can have the same coordinates and strand. This is valid and desirable; for example, a single highly transcribed genomic locus can give rise to many identical reads, yielding many identical mappings, and we do not want to lose information about their multiplicity.

Also note that $M_{\text{nuc}}$ and $M_{\text{con}}$ are sets of read mappings, not the reads themselves, so a single read that maps to multiple loci will be represented multiple times the respective mapping set, as it will have multiple genomic coordinates. At the present time, we do not penalize multiple mappings (as discussed in Supplementary Note 2), and each mapping contributes a single annotation to its respective set.

### 1.2 Derivation of cutting scores

We begin by computing $M_{\text{nuc}}(t)$ and $M_{\text{con}}(t)$ for each $t \in T$, which are subsets of read mappings where every mapping is completely contained within a transcript $t$. That is, the start and end positions of each read mapping are within the bounds of the transcript, but can include the transcript’s native 5’ and 3’ ends. Note that these subsets are also multisets in the sense described in Section 1.1. These subsets are proxies for RNA fragments that result from breaks in transcript $t$.

$$M_{\text{nuc}}(t) = \{ m \in M_{\text{nuc}} : (m.L \geq t.L) \land (m.R \leq t.R) \land (m.S = t.S) \}$$ (2)

$$M_{\text{con}}(t) = \{ m \in M_{\text{con}} : (m.L \geq t.L) \land (m.R \leq t.R) \land (m.S = t.S) \}$$ (3)

For each transcript $t$, we define the set $S(t)$ of all sites in it as

$$S(t) = \{ i \in \mathbb{Z} : t.L \leq i \leq t.R \}$$ (4)

where $\mathbb{Z}$ is the set of all integers.

We then compute end counts for each site $i$ in each transcript, which are simply counts of reads starting or ending at that position. All 5’ ends of read mappings are counted, but we only count a 3’ end of a read mapping if the mapping is $l$ bases or shorter (these are “trim” reads, explained in subsection “Rationale” of Supplementary Methods), with $l$ specified by the user.

$$e_{\text{nuc}}(i \mid t) = \| \{ m \in M_{\text{nuc}}(t) : (m.L = i) \lor ((m.R = i) \land (N(m) \leq l)) \} \| \quad \forall i \in S(t)$$ (5)

$$e_{\text{con}}(i \mid t) = \| \{ m \in M_{\text{con}}(t) : (m.L = i) \lor ((m.R = i) \land (N(m) \leq l)) \} \| \quad \forall i \in S(t)$$ (6)
Note that the above definition leads to a pathological edge case that occurs when two annotations \( t_1 \) and \( t_2 \) overlap on the same strand and a read mapping \( m \), from the same strand, is contained in the overlapping portion. This read mapping will then apply to both transcripts \( m \in M_{\text{sample}}(t_1) \) and \( m \in M_{\text{sample}}(t_2) \) and therefore be double-counted, contributing end counts to both \( t_1 \) and \( t_2 \). This is not an issue in the current study, as we do not analyze transcripts that are known to be alternatively spliced, alternatively processed, or have alternative transcription start or termination sites. So, we do not expect multiple isoforms. However, this may be an issue in analysis of transcripts with multiple isoforms. To circumvent this without altering the core FragSeq algorithm, an upstream algorithm can partition read mappings amongst annotations in any manner desired, then the FragSeq algorithm can be run on each annotation \( t \) separately, giving it as input only \( M_{\text{nuc}} \) and \( M_{\text{con}} \) that the user wants to apply to \( t \).

Raw end counts are used to determine sites that contain missing or low-valued data, which are masked out and ignored in further calculations. This masking out prevents us from obtaining artifactual cutting scores on sites for which there was not enough data to support any conclusion. We interpret these sites as sites for which good data is not obtainable given our conditions.

We create a set of non-ignored sites by removing from \( S(t) \) all missing and low-valued data sites, i.e. sites which have less that \( c \) ends in both the nuclease and the control samples, where \( c \) is a user-defined threshold that is applied uniformly regardless of a transcript’s coverage. In other words, we retain for further analysis only sites which have end counts of \( c \) or above in at least one of the samples.

\[
S'(t) = \{ i \in S(t) : (e_{\text{nuc}}(i \mid t) \geq c) \lor (e_{\text{con}}(i \mid t) \geq c) \} \tag{7}
\]

We then wish to remove the 5’-most and 3’-most sites of transcript \( t \), as we do not want to include native transcript ends in our calculations, being interested strictly in the relative ratios of breaks within the body of \( t \). However, since we observe a small variability in the start and end of transcripts, we use a heuristic threshold \( d \) to specify that the first and last \( d \) sites \((d \leq N(t) + 1)\) should be excluded, which captures any heterogeneity in the native ends, at the expense of losing structural information in regions close to the native ends.

\[
S''(t) = \{ i \in S'(t) : t.L + d \leq i \leq t.R - d \} \tag{8}
\]

Values of pseudocounts \( \Psi_{\text{nuc}}(t) \) and \( \Psi_{\text{con}}(t) \) are then computed separately for each transcript \( t \) in each sample (nuc or con). These pseudocounts additionally weight down low-valued data and also allow us to compute meaningful cutting scores for sites which have sufficient data in the nuclease sample \((e_{\text{nuc}}(i \mid t) \geq c)\) but have zero data in the control sample \((e_{\text{con}}(i \mid t) = 0)\).

Here, pseudocounts are an expectation of how many end counts we would observe at each non-ignored site after we know the total number of end counts at all non-ignored sites in a transcript in a sample. That is, they correspond to a belief of how many end counts would occur at a site under the simple model that every site is equally likely to produce a break, conditioned strictly on the total amount of data (end counts) obtainable under our experimental conditions for the specific transcript \( t \) in a specific sample. This is a prior belief in the sense that the amount of data is known, but the distribution of the data amongst the sites is unknown (not yet seen).
We generate the pseudocounts by simply taking the observed end counts at all non-ignored sites and evenly distributing them amongst those sites, except that we first take the square root; this is to satisfy the requirement that real counts dominate over pseudocounts as the magnitude of real counts increases:

\[
\Psi_{\text{nuc}}(t) = \sqrt{\frac{\sum_{i \in S''(t)} e_{\text{nuc}}(i|t)}{\|S''(t)\|}} \tag{9}
\]

\[
\Psi_{\text{con}}(t) = \sqrt{\frac{\sum_{i \in S(t)} e_{\text{con}}(i|t)}{\|S''(t)\|}} \tag{10}
\]

We add pseudocounts to the observed end counts and normalize to obtain probabilities of observing a break at each non-ignored site in each transcript in each sample.

\[
n_{\text{nuc}}(i|t) = \frac{e_{\text{nuc}}(i|t) + \Psi_{\text{nuc}}(t)}{\sum_{j \in S''(t)} [e_{\text{nuc}}(j|t) + \Psi_{\text{nuc}}(t)]} \quad \forall i \in S''(t) \tag{11}
\]

\[
n_{\text{con}}(i|t) = \frac{e_{\text{con}}(i|t) + \Psi_{\text{con}}(t)}{\sum_{j \in S''(t)} [e_{\text{con}}(j|t) + \Psi_{\text{con}}(t)]} \quad \forall i \in S''(t) \tag{12}
\]

Lastly, we take log-ratios of resulting probabilities to obtain cutting scores. As cutting scores below or equal to zero are meaningless in terms of providing data about enzyme susceptibility, they are replaced with a special “none” marker on output, meaning that no enzyme activity above background was observed at those sites.

\[
c(i|t) = \begin{cases} 
\ln \frac{n_{\text{nuc}}(i|t)}{n_{\text{con}}(i|t)} & \text{if } n_{\text{nuc}}(i|t) > n_{\text{con}}(i|t) \\
\text{none} & \text{if } n_{\text{nuc}}(i|t) \leq n_{\text{con}}(i|t)
\end{cases} \quad \forall i \in S''(t) \tag{13}
\]
SUPPLEMENTARY NOTE 2

Rationale for design of the computational pipeline and the FragSeq algorithm

This section describes the reasoning that led to our design choices and elaborates some aspects of the FragSeq algorithm. The exact FragSeq algorithm definition, which contains additional rationale about specific technical points, is in Supplementary Note 1.

Input of previously determined transcript coordinates

We chose to require that the user determines the genomic coordinates of transcripts for which structural information is to be inferred prior to running the FragSeq algorithm. So, the FragSeq algorithm will not automatically compute the transcript coordinates from sequencing read mappings. The decision to separate transcript inference and determination of cutting scores into two separate steps was made because we saw no advantage to coupling the two pipeline steps into a single algorithm. Additionally, the inference of transcripts from sequencing reads is an area in which much prior research has been done [34, 35], and it is not our aim here to contribute to that growing body of work; our assay aims to provide additional information (i.e. structure data) about established transcripts, not to infer their existence. Keeping transcript inference as a separate step allows modularity: before running FragSeq code, any other bioinformatics method can be used to determine transcript coordinates from sequencing data, or the study can be done on known annotations.
Handling of multiply mapped reads

Due to the presence of repetitive elements in eukaryotic genomes, it is often difficult to unambiguously determine which genomic locus gave rise to a sequenced RNA fragment. Even essential genes such as spliceosomal snRNA exist in several highly similar copies in the mouse genome. So, reads exist that map to multiple loci in the genome and raise the question of how to count such mappings.

We chose to allow all mappings of a single read to multiple loci without weighing (rewarding or penalizing) any mapping. That is, every read mapping that meets the mapping criteria (described in Methods) contributes one unit of coverage to every mapped base, regardless of the extent of multiple mappings to other genomic loci.

This results in an over-assignment of read mappings for sequences having multiple genomic copies, since only one copy can give rise to a read, yet the read is mapped to multiple copies; therefore, the coverage of each copy will be over-estimated. However, we argue that our unconventional strategy for handling multiply mapped reads is valid for our assay. This is because we are not trying to estimate or compare abundances of different transcripts, but are rather concerned about relative counts of read ends between sites of the same transcript.

Thus, we reason that over-counting bias is likely to affect most reads in a transcribed locus in a uniform fashion; it is likely that the whole locus is duplicated, so reads coming from it will be over-counted equivalently. Our normalization procedure (which normalizes each transcript in each sample separately) is therefore likely to cancel out this bias when we convert end counts to probabilities. Additionally, log-ratio scores (nuclease divided by control sample break probabilities) should generally cancel out bias that affects both samples.
We find that our strategy for handling multiple mappings, while not explicitly accounting for multiple mapping biases, still yields an accurate structure signal on spliceosomal snRNAs, which exist in many copies in the mouse genome, and is therefore an appropriate choice for this assay.

**Counting of read ends and definition of trim reads**

To obtain counts of RNA breaks for sites, we count all 5’ read ends. We also count 3’ read ends, but only for trim reads, which we define as reads shorter than or equal to 42 bases. Reads of this length range are produced when sequencing extends into the 3’ adapter; during the read mapping step, the adapter sequence is identified and trimmed off. Therefore, only trim reads contain the real 3’ end of the cloned RNA fragment. We cannot count 3’ ends of non-trim reads because we are not sure where the real 3’ end of the cloned RNA fragment is located. This approach can easily be adapted to sequencing methods that produce paired reads for each fragment, one from each end, in which case both ends of all mapped reads will be included in the set of trim reads.

Breaks at some sites will produce two fragments where both the 5’ and the 3’ ends mapping to the site will be observed. In that case, we will double-count a single break event. However, we find that counting 3’ trim ends allows us to get cutting scores for more sites and does not reduce cutting score accuracy (data not shown).

**Normalization**

We normalize the end counts (which are proxies for backbone scissions/breaks), converting them to probabilities of observing breaks, rather than using absolute end counts.
directly. We argue that comparing uncorrected end counts between nuclease and control samples yields an unreliable estimate of susceptibility to nuclease. Such a strategy results in almost every site containing a positive cutting score and produces a lot of noise (data not shown).

We believe that the main reason for this unreliability is that read coverage of two equimolar RNA fragments in two samples may show up as different in a RNA-Seq experiment if the contents of their mixtures change. We use equimolar cDNA for all barcodes. Thus, two RNAs that have the same number of copies per cell may produce different numbers of clones (and different numbers of reads) in each barcodes sample because the ratio of the abundances of their clonable/sequenceable fragments to the total cDNA abundance has changed. This can occur because the abundances of other RNAs have changed or because the fragmentation process produced different distributions of fragment lengths and end chemistries for different RNAs, resulting in two different quantities of reads being allocated to each of the two RNAs that were originally in equal copy number per cell.

Therefore, a greater number of uncorrected end counts at some site in nuclease sample versus control does not necessarily mean that this site is susceptible to cleavage by nuclease. For most RNAs we examined, most sites indeed have more uncorrected end counts in the nuclease sample than in the control regardless of structure context.

Therefore, we chose to look at cutting frequencies per site relative to other sites in the same RNA in the same sample. That is, after adding pseudocounts, we normalized end counts per site for each RNA, for each sample, separately, converting them to probabilities of observing a break at that site in that RNA in that sample. This procedure removes the effects from coverage changes.
Pseudocounts

Pseudocounts are often used in Bayesian statistical estimators of multinomial parameters to represent prior densities over these parameters. In our case, the multinomial parameters of interest are the probabilities, one for each site in a specific RNA species, that a scission occurs at that site. By adding a fixed pseudocount to the count of observed scissions at each site, using the same pseudocount for all sites in the particular RNA species, we specify a prior in which the frequency of scissions is \textit{a priori} uniform over the RNA molecule. The magnitude of this pseudocount encodes the strength of this prior. We did not follow a pure Bayesian approach in fixing the magnitude of these uniform pseudocounts, but instead let the magnitude vary for each RNA species and sample according to the total number of observed scissions for that species in that sample in such a way that RNA species with smaller numbers of observed scissions have smaller pseudocounts (see Supplementary Note 1). This approach produces cutting scores more consistent with known secondary structures of our benchmark RNAs (U1a, U3b, and U5) than assigning a uniform, fixed pseudocount value for all RNAs and samples regardless of the number of observed scissions (data not shown).

Log-ratio scores and use of a control sample

One useful feature of the control sample is that it tells us the extent of endogenous scissions generating 5'-PO₄/3'-OH chemistry, for which we must correct. However, the control sample also tells us the general susceptibility of sites for any kind of scission. Certain sites that are simply prone to being cut by any agent (e.g. by being in a flexible region of RNA) may be overrepresented in the nuclease sample. For example, the site between C71 and U72 in U1a is such an outlier site, producing many breaks in any sample (Supplementary Fig. 4). Taking log-
ratios using the control sample dampens that overrepresentation, bringing cutting scores into a more consistent range.
SUPPLEMENTARY NOTE 3
Details on data analysis and structure model building

Interpretation of probing data for U15b, U22, and U97

To convert gel band intensity to numerical data that would aid us in identifying the susceptibility of bases to various enzymes during structure probing, we used Safa software v1.1 [36, 37], which quantifies gel bands. However, to ensure accurate position assignment in order to display probing results on our structure models, we manually re-interpreted Safa output and the original gels to convert susceptibility into four discrete categories: marginal, weak, moderate, and strong.

When interpreting gel bands, we had to resolve the differences in end chemistry between different RNA fragments. P1 and V1 products have a 3’ OH, whereas T1, RNase A, and alkaline hydrolysis products have a 3’ cyclic phosphate. Therefore, the charge-to-mass ratios of P1 and V1 products are reduced compared to other products, so P1- or V1-produced RNA fragments will migrate more slowly [38]. We therefore assume that a P1 or V1 band between ladder positions of size N-1 and N should be assigned to N-1, since it would have been migrating faster if it had a charge from the 3’ cyclic phosphate. However, this simple heuristic is not reliable [39]; therefore, as in [38], we claim that our P1 and V1 cut site assignments may be off by up to two positions, and note that the migration difference due to end chemistry are smaller for fragments of longer lengths.
Building models for novel RNA secondary structures for U15b, U22, and U97

A comparison of FragSeq probing data against data from our probing experiments that used conventional techniques is most insightful in the context of a secondary structure model. However, we are not aware of any structure models for U15b, U22, or U97 that are based on experimental data. In this section, we derive such models for these RNAs using data from our probing experiments on in vitro transcribed RNA, phylogenetic information, computational predictions, and prior work. It is important to note that we do not use FragSeq data to build these models, because the purpose of these models is to show structures derived using previously existing methods, so that data obtained with our novel method can be compared against them.

As a starting point for building a secondary structure model for each RNA, we began with computational predictions by RNAstructure version 5.03 for the Windows operating system [23] (software release from 2010/03/22). We used the partition function feature to compute base pair probabilities (dot plot) for every canonical base pair in the weighed ensemble of all structures [40], then used a custom script (dpToVarna.py, submitted in Supplementary Software) to convert that data to probabilities that each base is single-stranded in the ensemble. We then plotted both the base-pairing probabilities and ssRNA probabilities on top of the predicted minimum free energy structure (MFE) and suboptimal structures, using the VARNA program [41] to make these visualizations. The ssRNA probabilities allowed us to summarize which bases are likely to be single stranded in alternative conformations. We then reviewed this pool of structures and probabilities to form secondary structure hypotheses for each RNA, then adjusted them based on in vitro probing data and phylogenetic information.

To examine each secondary structure in the context of phylogenetic information, we either used the multiz 30-way multiple genome alignment of vertebrate species downloaded from
the UCSC Genome Browser [42] or the ClustalW 2.0 web server [43] to make initial alignments for these RNAs; in both cases, alignments were manually refined afterwards and pruned to remove sequences that were too divergent. To visualize compensatory/covariant base pair substitutions and secondary structure conservation, we used the colorstock.pl Perl program [44], which color-codes and displays such information for alignments input in Stockholm 1.0 file format.

Our resulting models aim to be minimal; that is, we favor showing only the most likely (usually phylogenetically conserved in mammals) base pairs, at the expense of showing less base pairs in our model than may actually be present in mouse sequences alone. We include multiple sequence alignments containing various structure models that we considered, in Stockholm 1.0 file format, as Supplementary Data files 1 through 3.

**U15b structure model**

This RNA was first identified in HeLa cells by immunoprecipitation with an antibody against fibrillarin [33] and is predicted to guide 2’-O-methylation of 28S rRNA [26]; its HUGO Gene Nomenclature Committee “Approved Symbol” is SNORD15B.

In general, the probing data agrees with the computational prediction we used as an initial model (Supplementary Fig. 10A; probing gels in Supplementary Fig. 6A). To make the final model (Fig. 5), several base pairs had to be removed from the initial model for a combination of reasons: reactivity with ssRNA-specific enzymes, poor phylogenetic conservation, or high probability of ssRNA in the predicted ensemble of all structures. For example, the closing base of SL1 (A27/U31) was removed due to lack of conservation and high ssRNA probability, and some bases were removed in the U43 through U56 and A82 to G96 regions because of lack of
phylogenetic support. We also added to the final model base-pairing between the terminal four bases, which are common to C/D box snoRNA and have good phylogenetic support in the Rfam database (Rfam ID RF00067).

An alternative secondary structure for U15b has previously been proposed in [33]; however, the sequence identified in that study is different from the mouse genomic sequence in this study and several proposed base pairs cannot be formed. Moreover, our probing data disagrees with some parts of that model (Supplementary Fig. 10b), although both models match exactly for the stem-loop formed by bases C58 through G80, and that stem is well-supported by phylogenetic data, representing the highest-confidence part of the structure. Also, the two models agree on the closing stem.

U22 structure model

This RNA was first identified in HeLa cells after immunoprecipitation with anti-(U3)RNP antibodies and named “human RNA Y” [45]. It contains a 5’-monophosphate and a 3’-OH, has a C and D box, and associates with fibrillarin [33], but is not known to carry out 2’-O-methylation [46]; rather, it functions in 18S rRNA processing [20]. It was proposed that its termini base pair to form a 4-base-pair closing helix [33] in the same manner as many eukaryotic C/D box snoRNAs; that helix is the only secondary structure shown in the Rfam database (Rfam ID RF00099). Its HUGO Gene Nomenclature Committee “Approved Symbol” is SNORD22.

The MFE secondary structure prediction by RNAstructure is shown in Supplementary Figure 11, with probing data from gels in Supplementary Figures 6B and 6C; we used this as our initial model. We noted that bases G8 through C19 and A91 through A94 were very reactive with ssRNA-specific enzymes, so those helices were removed from our final structure model.
Additionally, those bases have a high probability of being ssRNA in the computationally predicted ensemble of all structures and do not have good phylogenetic support. U18 through U28 are also highly reactive with V1, but since V1 can cleave ssRNA [29], this evidence may indicate that this region is highly accessible and susceptible to nucleases in general, rather than being base-paired.

The remaining predicted helices were retained, although usually shortened because base pairs that were reactive were removed. Of particular note are the helices between U31 and A87 – there is good support for evolutionary pressure to conserve base-pairing in this region, evidenced by several compensating substitutions that preserve Watson/Crick base pairing, especially in the helix between C39/G78 and G48/C69. Interestingly, the highly reactive ssRNA sequence GAACCCA (bases 62 through 68) is perfectly conserved from frog to human, contrasting with base-pairing regions around it that undergo several substitutions, suggesting it could have an important function for which it is positioned by surrounding helices.

U97 structure model

This RNA was first identified, in a truncated form, as a highly abundant small non-messenger RNA in mouse brain cDNA libraries and named MBI-82 (GenBank: AF357463) [47]. It was later discovered that its full form has the canonical features of a C/D box snoRNA, although no rRNA or snRNA 2′-O-methylation target could be identified, which makes it an orphan snoRNA; the human homolog was named U97 [48] (HUGO Gene Nomenclature Committee “Approved Symbol” is SNORD97).

The MFE secondary structure prediction by RNAstructure is shown in Supplementary Figure 12A; superimposed on it is the probing data from gels in Supplementary Figures 6D and
6E. In general, the probing data is in agreement with the computational predictions. This structure served as our initial model.

However, we noted that positions 21 to 60 were particularly susceptible to cleavage by V1, which is a trend we did not observe in our U15b or U22 probing. Although V1 susceptibility is generally a proxy for dsRNA, it has been noted that V1 will also cut single-stranded RNA, as its preferred substrate is stacked bases, which may not necessarily be dsRNA [29]. This makes it difficult to interpret from our data whether the 5’ domain (positions 19 through 86) contains base pairs or not, or whether base pairing exists and is unstable, or whether several structure conformations exist. Also, the 5’ domain seems more susceptible to ssRNA nucleases as well, supporting the possibility that it lacks stable helices.

Moreover, positions 19 through 47 have a slight enrichment for purines, while positions 48 through 122 tend to be pyrimidine-rich and contain pyrimidine-rich insertions in the rodent lineage versus other mammals. This means that any base pairing in that region may simply be a side effect of sequence composition (with the sequence providing a biological function) and that the structure formed is not biologically relevant.

We then looked at whether we could detect any base pair conservation signatures in U97. We attempted to align U97 sequences of several mammalian species to the MFE structure, noting that its evolution is far more dynamic than U15b and U22. Base pairing in the 5’ domain is poorly conserved, and the pattern of substitutions in conserved canonical base pairs in any domain is not informative with regards to structure conservation. Single substitutions occur that convert between G/C to G/U base pairs, but there is a notable absence of compensating changes that are the telltale signature of evolutionary pressure to preserve base pairing. Therefore, it is
possible that conservation of base pairs obtained from the MFE structure is a side effect of sequence conservation, not selection for base pairing.

Our final model (Supplementary Fig. 8C) is taken from the MFE prediction, except that several base pairs contradicting experimental evidence were removed. Also, lack of base pairing conservation led to the removal of two helices in the 5’ domain (helices beginning at A19/U86 and G44/C57). We examined suboptimal structures (not shown) but did not find any folds that are more plausible in light of experimental evidence. It should be noted that the Rfam database entry for U97 (Rfam ID RF01291) shows an alternative, computationally predicted conserved helix beginning at base pair A31/U51 (Supplementary Fig. 12B), but like helices in our model, it also not contain compensating substitutions supporting it, so we do not find it more plausible than our structure model. The only other secondary structure shown by Rfam is the closing helix, which is in agreement with our model and in general tends to be a feature of eukaryotic C/D box snoRNA.

**Sources of known RNA structures**

When putting together a set of known mouse RNA secondary structures to assess the accuracy of FragSeq probing data (e.g. in Figures 2, 3, and 4), it was necessary to align structures from a related species to the mouse sequence. Also, we sometimes had to resolve several conflicting structure models from literature or databases; in those cases, we emphasized experimental evidence and phylogenetic support to get more reliable structures. We briefly summarize the sources of our structure data below; detailed notes on how differences were resolved are available on request. In addition to literature cited below, we also consulted the Rfam database [49] and Zwieb’s uRNA database [50] for structure models.
In all secondary structures, we represent non-canonical base pairs (e.g. G54/A85 and A55/G84 in mouse U1a) as ssRNA.

The structure for U1, human homolog of mouse U1a, has been determined by probing in solution [51] and by x-ray crystallography [52]. To this structure, we added base-pairing between the 5’ splice site recognition site and Sm protein binding site from [53].

The structure of the 5’ domain of U2 came from [54] and the remainder of the structure from [55]. However, a highly dynamic region exists that in yeast has been shown to interconvert between stem-loops IIa, IIb, and IIc at various stages in the splicing pathway [56], raising the possibility that a similar process occurs in mouse. Given this, we cannot expect that regions that we identify as ssRNA in Figure 4A are accurate, or even that there exists a unique structure of U2 in solution. This region (marked “PossFlex” in Fig. 4A) is therefore highlighted as uncertain in that figure.

The U3b structure was constructed after reconciling several structure models [15, 17, 57-59] in the context of the U3 multiple alignment from snoRNA-LBME-db [46]. The resulting structure best agrees with experimental probing data from [15], with box labeling coming from [17].

The ssRNA regions of U4b are taken from [16], but as that study probed human U4 when base-paired with U6, the ssRNA annotation is uncertain; specifically, the structures of regions marked “BpToU6” (base paired to U6) in Figure 4A are uncertain in free U4, as is the structure of the nearby hinge region.

The U5 secondary structure is derived from a comparison of [16, 60, 61].

The secondary structure model for U6 is from [16], which probes the human homolog of the RNA without protein or trans interactions with U2 or U4.
The U8 ssRNA regions are from [62]; however, that study focused on analysis of *Xenopus laevis* U8, which is highly divergent from mouse U8. The stem-loop region containing the D box (“SL3_BoxD” in Fig. 4A) does not have a discernible phylogenetically conserved structure and is therefore marked as uncertain.

The U11 secondary structure is from the Rfam database (Rfam ID RF00548), which is in turn based on computational predictions from [63] and [64]. Although we were unable to locate any experimental evidence supporting the predicted structure, the Rfam multiple sequence alignment provides good phylogenetic evidence for it, which makes us believe the model is sufficiently accurate to use as a reference for our probing data.

The U12 secondary structure is taken from [65], which a revision of an earlier structure model by [63]. Both structures are computationally predicted, but Benecke et al [65] experimentally validate the structure of the crucial SL4 stem-loop that binds the 65K protein.

**Comparison of SAFA counts (gel quantification) from follow-up probing with FragSeq cutting scores**

This sub-section describes how we compared U22 and U97 structure data from our FragSeq high-throughput experiments to data from our follow-up probing done on *in vitro* transcribed, homogeneous RNA (products resolved using gels and quantified with SAFA software).

**U22 (Supplementary Figure 7a)**

We reason that FragSeq data is accurate because moderate to high cutting scores, when present, are supported by data from nuclease in homogeneous mixture. Moderate cutting scores
at bases 54-56 and 62-64 are supported by weak to moderate cuts by T1 and P1 in both gels. High cutting scores at bases 16-26 are supported by RNase A and T1 cleavage in that region.

Interestingly, P1 nuclease cutting at bases 16-26 is present in gel 1, but very weak; it is possible that other areas of the RNA (such as bases 7-14 and 105-107) are so susceptible to the enzyme that the band intensities they produce are disproportionately high, making P1 cutting at bases 16-26 appear less pronounced by comparison.

However, in FragSeq, breaks in the 5’ and 3’ ends of the RNA are unlikely to be observed because, under the single-hit kinetics assumption, they would yield fragments that are too short and too long to be sequenced. Therefore, we cannot identify whether the 5’-most bases were highly susceptible to P1 in FragSeq. For the same reason, we cannot make claims about susceptibility to P1 in the FragSeq assay at the 3’ end of the RNA. But when moderate to high cutting scores are present, they correlate with ssRNA evidence from other enzymes, leading us to conclude that the cutting scores contain accurate structure information.

U97 (Supplementary Figure 8a)

High cutting scores for regions at bases 79-80, 95-100, 127-131, and 159-163 are all validated by cleavage by at least one other ssRNA-specific nuclease in each of those regions, leading us to believe the cutting scores are reliable even for an RNA at the bottom of the coverage range of our follow-up candidates. However, this RNA has little FragSeq data for bases 1-78. We noted in our analysis (see Supplementary Note 3) that this RNA has a lot of single-stranded character. Therefore, it may be overly susceptible to P1 in the FragSeq assay, leading to over-digestion producing fragments that are too short to pass size selection or that do not contain usable structure information. Some of this susceptibility is observed in follow-up
probing with P1, yielding a large number of low to moderate cuts. However, we judge that when high cutting scores are present at several consecutive sites, they are supported by nuclease cuts in follow-up experiments.

**Assignment of read mappings to annotations in Supplementary Figure 2**

We began with a pool of all read mappings to the mm9 genome assembly, which is larger than the count of reads in Supplementary Table 1 because each read is allowed to map to multiple genomic locations (up to 15 mappings per chromosome). We then assigned mappings to annotation types, checking one type at a time. Mappings whose genomic coordinates overlapped any annotation of a type were removed and assigned to that type, i.e. not proceeding to checking the next annotation type. We began by checking mappings against snRNA and went clockwise as shown in the pie charts. Checking for overlap with the opposite strand of non-RNA repeat annotations was therefore done last, and remaining reads were put in the “unknown” category. Overlap required that the strand of the annotation and the mapping must match, but we also checked for overlap to the opposite strands of mRNA and repetitive element annotations, as those are known to have prevalent anti-sense transcripts in eukaryotic genomes.
Interpretation of FragSeq cutting score data

An issue of great importance is how to evaluate cutting scores output by the FragSeq algorithm. Specifically, how do we figure out how large a cutting score is considered good evidence of ssRNA? We do observe that our assay occasionally produces low-valued (marginal) cutting scores that do not provide good evidence of ssRNA, which raises the question of where to draw the line between reliable and unreliable cutting score evidence.

In theory, a cutting score of 1 at a site means that we are ~2.718 times more likely to observe a break at that site in the nuclease sample than in the control sample. We generally do not observe cutting scores higher than 4, which corresponds to ~55 times higher likelihood of observing a break in the nuclease sample versus control.

We find that we cannot recommend a single threshold that separates reliable scores from unreliable scores for every RNA and every transcriptome, since distributions of cutting scores vary from RNA to RNA and between transcriptomes. Even for a specific site in a specific RNA, the magnitude of cutting scores can be different between transcriptome samples – for example, compare UNDIFF and D5NP scores for U1a in Fig. 2E, and note that scores are generally higher in D5NP, although the overall pattern of reactivity is the same relative to other sites within the same transcriptome sample. Supplementary Figure 5 shows more clearly that although there is a variance in score values, the ordering of sites from “most reactive” to “least reactive” tends to be the same between transcriptome samples. We have not carried out experiments to test whether differences in cutting scores between different RNAs, or between the same RNA from different transcriptomes, reflect a real difference in P1 cleavage rate or whether they are a consequence of
our protocol. Therefore, we recommend comparing cutting scores in a RNA to the maximum score in *that RNA, in that transcriptome sample only*, not to the maximum score across all RNAs.

How, then, do we separate low, unreliable cutting scores from high, reliable cutting scores? The best approach is to consider the context, i.e. cutting scores at adjacent sites, as suggested in [66]. A run of three or more adjacent sites with cutting scores in the moderate-to-high range of values for that RNA can be considered good evidence of ssRNA. High consecutive runs can occasionally be broken or flanked by lower values (e.g. U5 SL1 in Fig. 3D or U3b SL1 and SL2 in Fig. 3B); in the context of higher values, smaller scores are more likely to represent signal rather than noise. However, consecutive runs of only small values are potentially noise, even if they occasionally overlap with real ssRNA (e.g. bases 154 to 159, upstream of U3b box C in Fig. 3B).

Absence of ssRNA evidence (low or missing cutting scores) should not be considered evidence of RNA base pairing. We generally do not obtain high cutting scores close to the 3’ end of structured RNAs even when there is ssRNA there (Fig. 4 and data not shown). This is probably because such cuts will leave a 5’ fragment too long and 3’ fragment too short to pass size selection, and although a second cut upstream could result in a fragment of clonable size, the lack of upstream ssRNA that can be cleaved may prevent this from happening. Supplementary Note 3 describes some specific cases where missing cutting scores are likely due to cuts producing fragments that are outside of the 20-100 nt sequencing range required for ABI SOLiD3.

Lastly, to aid in rapidly producing structure models from FragSeq data, we have implemented a new feature in the RNAstructure suite of tools [23] that allows arbitrary restraints
to be applied when producing structures. Two options (\(-SSO\) and \(-DSO\)) have been added to the Fold program (for predicting low free energy structures) that allow the user to provide a list of pseudo-$\Delta G^\circ$ (pseudo free energy change) offsets for select bases that are used as “rewards” or “penalties” for folding a base as ssRNA or dsRNA, respectively. If an SSO offset was specified for a base, then every structure containing that base as ssRNA will have the offset added to its free energy change ($\Delta G^\circ$) of folding, and likewise for the DSO offset and dsRNA. Therefore, a negative SSO offset makes a base more likely to appear as ssRNA in a predicted structure, as negative free energies are stabilizing. FragSeq cutting scores can therefore be transformed and supplied to the program via the SSO option. Large (positive) cutting scores should be transformed into small (negative) offsets to increase the likelihood that nuclease-susceptible bases are ssRNA in the predicted structure.

**Recommended sequencing coverage for a FragSeq experiment**

Read coverage (mean reads per base) of a region affects cutting score reliability. We observed that lower-coverage RNAs have noisier scores and have less consecutive runs of high-scoring sites. We estimate that U97, whose UNDIFF nuclease sample coverage is 100, represents the lowest acceptable coverage upon which structural interpretations should be based. Ideally, coverage of 200 or more is desired in the nuclease sample, with higher coverage resulting in higher-quality data for more sites. For example, at a coverage of 256 in the nuclease treatment (and 115 in the control treatment), U15b yielded FragSeq data that strongly agrees with follow-up probing and is therefore reliable.
Out of 1,588 ncRNAs examined in Supplementary Figure 3, 231 ncRNAs meet the criteria of coverage >200 in the nuclease sample (309 have coverage >100). Therefore, our experiment generated reliable structure data for 231 ncRNAs by our rule of thumb.

However, the above coverage estimates are based on RNAs that take up less than 20% of successfully mapped reads in their respective barcode (see Supplementary Fig. 2), as the fraction of reads mapping to repetitive elements and mRNA is very high in our samples. So, our analysis of ncRNA is based on a small subset of the total reads. Successfully enriching for RNA of interest could therefore increase coverage per RNA by as much as five-fold, and allow the study of more RNAs at coverage of >200.

Additionally, we probed two transcriptomes, using six out of ten barcodes on a single ABI SOLiD3 sequencing run (the other four barcodes were reserved for experiments not described in this work), three barcodes per transcriptome sample. Allocating an entire sequencing run to a single transcriptome would also boost coverage by approximately three-fold.

We do not have a recommendation for minimum coverage for the control treatment, as low coverage (less than 100) in the control treatment still produces accurate cutting scores as long as nuclease sample coverage is sufficient (e.g. see U5 and other ncRNA in Fig. 4a).
SUPPLEMENTARY REFERENCES


