Chapter 8

High-Throughput Nuclease Probing of RNA Structures Using FragSeq

Andrew V. Uzilov and Jason G. Underwood

Abstract

High-throughput sequencing of cDNA (RNA-Seq) can be used to generate nuclease accessibility data for many distinct transcripts in the same mixture simultaneously. Such assays accelerate RNA structure analysis and provide researchers with new technologies to tackle biological questions on a transcriptome-wide scale. FragSeq is an experimental assay for transcriptome-wide RNA structure probing using RNA-Seq, coupled with data analysis tools that allow quantitative determination of nuclease accessibility at single-base resolution. We provide a practical guide to designing and carrying out FragSeq experiments and data analysis.

Key words RNA structure prediction, FragSeq, RNA-Seq, Transcriptome, Nuclease probing, Nuclease accessibility, RNA structure probing, Bioinformatics

1 Introduction

Enzymatic or chemical probing of RNA in solution provides informative data from which a structure model can be constructed. Probing agents are used to cleave the phosphate backbone or modify nucleotides in a way that provides structure information due to solvent accessibility of reactive functional groups and their structural context. Traditionally, in order to recover this information, direct end-labeling of the probed RNA or primer extension with labeled primers is used, after which the length of labeled products is inferred by means of high-resolution denaturing gel electrophoresis. A significant amount of work has gone into the development and refinement of such probing approaches over the past four decades [1, 2]. While these techniques are extremely useful and informative, the rate at which structure data can be acquired with them is slowed because they require purification of the RNA of interest or custom primer design, and at least one electrophoresis step must be done.

Over the past few years, several groups [3, 4] have adopted various RNA-Seq protocols to replace electrophoresis with high-throughput sequencing by modifying and extending existing

Douglas H. Turner and David H. Mathews (eds.), *RNA Structure Determination: Methods and Protocols*, Methods in Molecular Biology, vol. 1490, DOI 10.1007/978-1-4939-6433-8_8,

enzymatic [5–8] or chemical [9–20] structure probing approaches, thus allowing interrogation of complex mixtures of hundreds to thousands of different RNAs in a single reaction. Once calibrated, these methods allow probing of an entire transcriptome at single-nucleotide resolution in one experiment without requiring custom primer design or labeling of one specific RNA of interest; prior knowledge of sequences of the RNAs being probed is not necessary. The scale of information gained allows researchers to tackle scientific questions that were simply not possible to address with classic techniques.

FragSeq is a high-throughput enzymatic probing method that measures accessibility of RNA sites (*see* **Note 1**) to an endonuclease [5]. A complex RNA mixture (i.e., containing many different transcripts at various abundances) is subjected to partial nuclease digestion; a control sample from the same RNA mixture is prepared in parallel in the same manner except without nuclease digestion. RNA fragments in the two samples are reverse transcribed to make cDNA libraries, which are then sequenced to produce reads spanning some or all of the length of each cDNA. Reads are mapped to the reference genome or RNA sequences of interest, and the resulting mapping coordinates are input to our command-line tool to produce cutting scores, which describe nuclease accessibility at each RNA site; other useful statistics are also output. This data can be visualized in VARNA software [21] to examine it in a secondary structure context or in a genome browser to examine it in a genomic context (Fig. 1); additionally, it can be used to guide computational predictions of RNA structure (Fig. 2).

The library preparation strategy employed in ref. 5 is shown in Fig. 3. In that protocol, the P1 nuclease was used, which is specific for single-stranded RNA (ssRNA) and produces fragments with 5′ PO_4 and 3′ OH end chemistry after cleavage. Adapter ligation to ends of RNA fragments containing specifically those end chemistries allowed us to clone them and thus enrich for products of nuclease cleavage, selecting against nonspecific degradation that leaves 5′ OH and 2′,3′-cyclic phosphate. However, other nucleases can be used (Subheading 1.1). Also, although we used Applied Biosystems SOLiD sequencing in the original study, FragSeq does not require this specific sequencing technology because the key informative step is the ligation of adapters to fragment ends during sequencing library preparation; compatible library protocols for Illumina and Ion Torrent sequencing are given in Subheading 3.3. Our command-line tool can be configured to process data from alternative preparation schemes.

The FragSeq computational pipeline is outlined in Fig. 4, with real data at each step shown in Fig. 1 for mouse spliceosomal (sn)

Fig. 1 (continued) similar: undifferentiated embryonic stem cells (UESC) and day 5 neural precursor cells. Other tracks show UESC data only. *SL* stem-loop, *IL* interior loop, *MBL* multibranch loop. Figure is modified from ref. 5

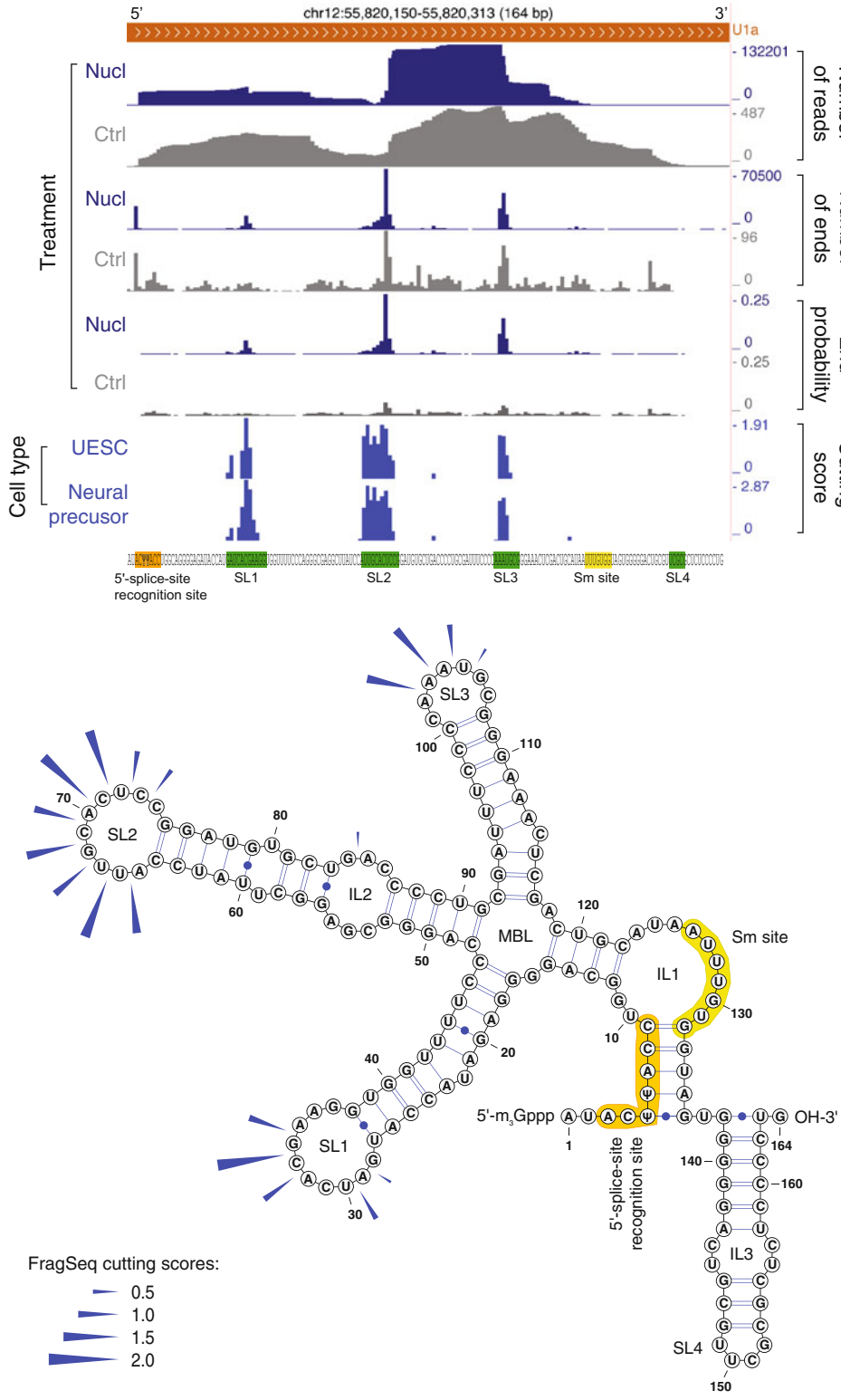


Fig. 1 Flow of data through the FragSeq pipeline (from *top* to *bottom*), displayed in the UCSC Genome Browser [37] (*top* panels) and in VARNA secondary structure viewer [21] (*bottom* panel), for mouse spliceosomal (sn)RNA U1a. Pipeline steps correspond to those shown in Fig. 4. "Cutting score" genome browser tracks compare results obtained from parallel FragSeq experiments on two cell lines where structure of this RNA ought to be

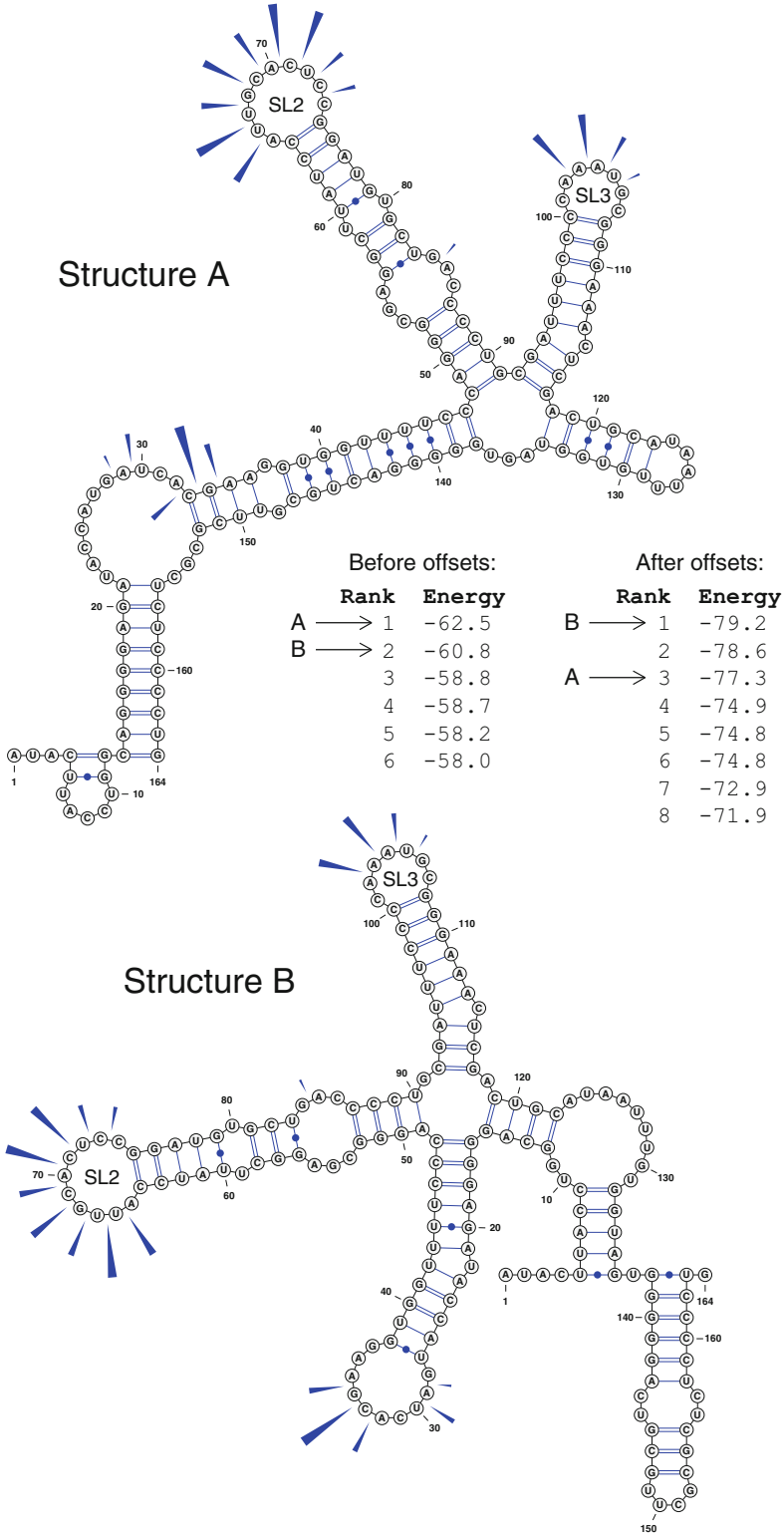


Fig. 2 Top-ranked secondary structures of mouse snRNA U1a predicted by the program Fold in the RNAstructure package [25] with (Structure B) and without (Structure A) FragSeq-derived offsets. Tables show energies (standard Gibbs free energy of folding or $\Delta G°$, units of kcal/mol) for the total set of predicted structures, ranked from most to least favorable (lower energies are more favorable); positions of Structures A and B within the ranked list is indicated. Energies are proportional to the natural log of the equilibrium constant between folded and unfolded states.

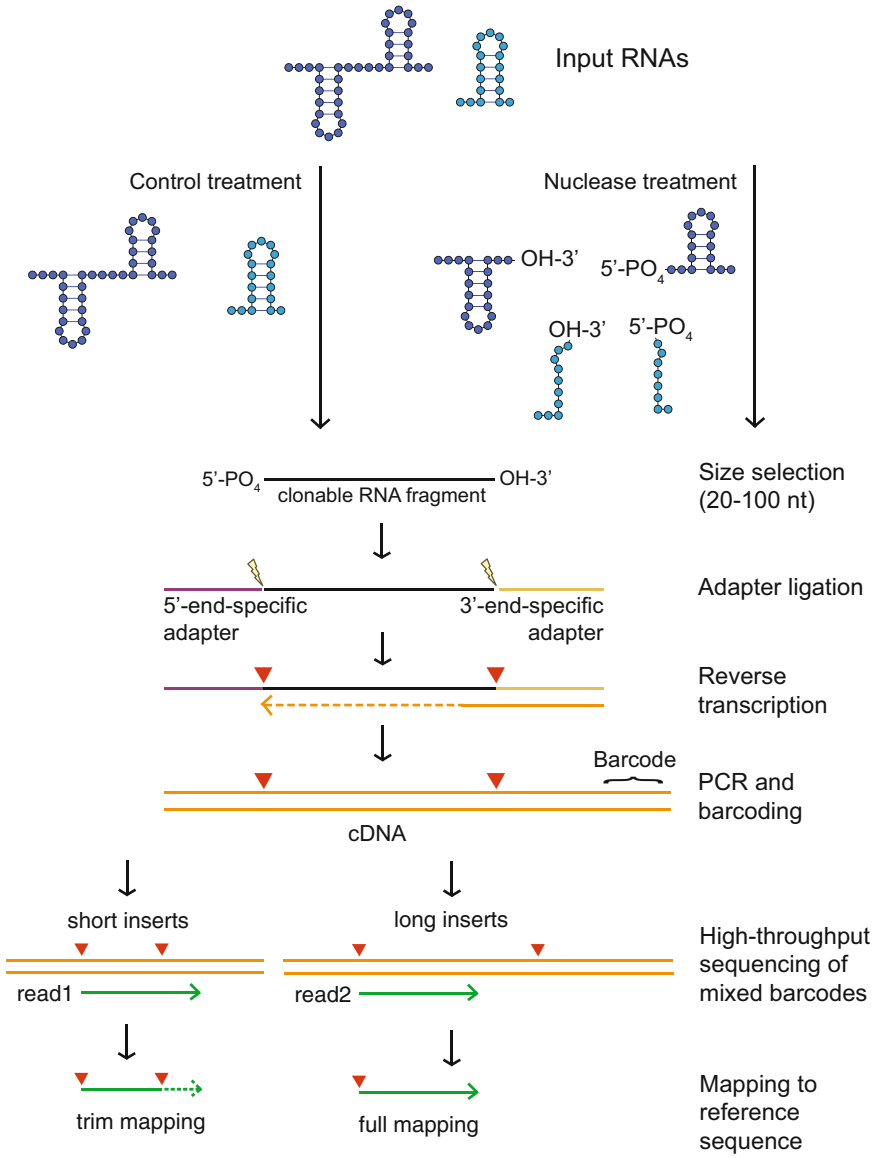


Fig. 3 Library preparation, sequencing, and read mapping strategy employed in ref. 5, which used the Applied Biosystems SOLiD Small RNA Expression Kit (SREK) for the Applied Biosystems SOLiD 3 platform. *Lightning bolts* denote ligation junctions. *Triangles* show sites that are *relevant* (*see* **Note 14**) because they correspond to ligation-competent 5′ and 3′ ends of the original RNA fragment which, in this protocol, yield structure data. An "insert" is the part of a cDNA whose sequence corresponds to the original RNA fragment. If an insert is short, a read can sequence into the opposite adapter; during mapping, adapter sequence was removed ("trimmed," dotted part of read1), producing two relevant mapping ends. If an insert is long, we could only use the 5′ end of a mapping. In this protocol, sequencing of double-stranded cDNA was initiated only from the 5′ adapter end, but sequencing can also be initiated from the 3′ end, or from both ends in two passes (paired-end sequencing), depending on the method. Figure modified from ref. 5

Fig. 2 (continued) The same cutting scores are plotted on top of both structures (*blue arrows*); these are the same as in Fig. 1, which also shows the known U1a structure. These cutting scores are linearly transformed into offsets using a slope of −1 and intercept of 0 (however, multiple slopes work in this specific case), then given to Fold using the -SSO option (other options were kept at default). Offsets lower the energy of all predicted structures because they contain correct folding of SL2 and SL3, with which cutting scores agree; the change in ranking comes from the small number of bases (C33 and G34) with which cutting scores disagree in Structure A

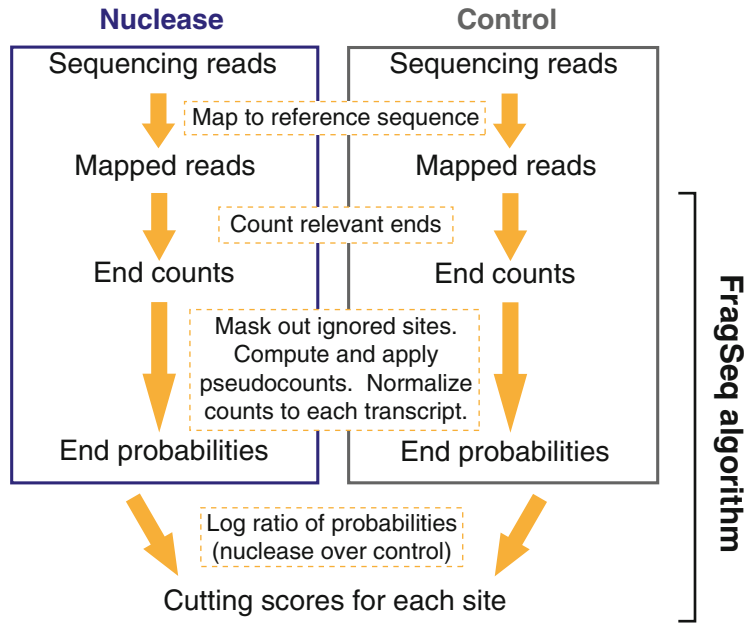


Fig. 4 The FragSeq computational pipeline. Steps carried out by the FragSeq algorithm are indicated. Key pipeline steps are in *dashed boxes*. Figure modified from ref. 5

RNA U1a. The final product is cutting scores, which indicate sites that are more susceptible to the nuclease used (*see* **Note 2**). Importantly, the normalization strategy requires that the researcher identify the RNA loci of interest and their coordinates ahead of time, as that information is input to our command-line tool (*see* **Note 3**).

Some key features distinguish FragSeq from other high-throughput enzymatic probing RNA-Seq methods. First, like in Parallel Analysis of RNA Structure (PARS) [6], FragSeq gets structure information by sequencing at or across the site where adapter ligated to the end(s) of RNA fragments, meaning the read *ends* are important (Fig. 3); in contrast, dsRNA-seq and ssRNA-seq [7] look at coverage by bases in the *body* of reads after enriching for dsRNA or ssRNA fragments. Second, FragSeq requires an explicit no-nuclease control sample for every nuclease used; in contrast, PARS compares samples digested by two different nucleases directly, without a control (*see* **Note 4**). Lastly, the FragSeq pipeline is different from PARS with respect to read count normalization (*see* **Note 3**); also, special attention is paid to dealing with missing and unreliable data so that accuracy on RNAs with lower coverage is not compromised.

In addition to this chapter, readers are encouraged to study the detailed computational and bioinformatics methods from other high-throughput RNA structure probing protocols [22, 23] for insight on experiment and pipeline design.

1.1 Considerations in Designing a FragSeq Experiment

FragSeq aims to probe a complex RNA mixture. This mixture can be total RNA purified from cells (*see* **Note 5**) or may be enriched for a specific RNA population of interest, such as by subcellular fractionation, size-selection, or using immobilized antisense oligonucleotides directed against the RNA of interest. Mixtures of in vitro-transcribed RNAs or synthetic oligonucleotides, or even a single transcript, can also be probed, as long as care is taken to make sure that all sequencing reads can be unambiguously mapped to a specific RNA in the pool (Subheadings 3.5.1 and 3.5.2). In principle, FragSeq nuclease probing can also be carried out on partially purified RNA, such as intact RNPs (more mildly extracted from in vivo or assembled in vitro), or on RNA incubated with a specific protein, ligand, or other RNA(s). These approaches would require modifications to the given sample preparation protocol, but our command-line tool can still be used to infer cutting scores.

We recommend using endonucleases that leave the 5′ PO_4 and 3′ OH end chemistry after cleavage (nucleases P1 or S1, and RNase V1), because those end chemistries are more favorable for adapter ligation and therefore allow enrichment for RNA fragments produced by the nuclease. Our command-line tool can be configured to use data from either the 5′ or the 3′ end of reads or some combination of both. Nucleases leaving other end chemistries, followed by repair to make ligation-competent ends (see below), are possible to use, though that may produce more noise. The key is to always sequence a no-nuclease control sample to control for non-nuclease-specific cleavage or degradation.

It is important to understand how enzyme treatments affect end chemistries in both nuclease and control samples, as comparison of ends between those samples is the most important aspect of the FragSeq method. The ends in the input RNA may or may not be available during subsequent adapter ligation strategies. For example, a capped, polyadenylated Pol II transcript from a eukaryotic cell would not have an available 5′ PO_4 end for ligation due to the 7-methyl guanosine cap, but it would have a free 3′ OH group available for 3′ end ligation. Similarly, an RNA produced by bacteriophage T7 transcription would have 5′ triphosphate and 3′ OH ends. Also of note should be the ends generated by cleavage of RNA by a number of common RNases (e.g., A, T1) and random base-catalyzed hydrolysis. These termini (5′ OH and 2′,3′-cyclic-phosphate), will not be captured by most ligation strategies. If information about these termini is desired, an enzymatic treatment with T4 polynucleotide kinase and ATP can generate the necessary 5′ PO_4 and 3′ OH [5, 24].

We find that cutting scores are more reliable if based on higher quantities of mapped reads. Therefore, it is important to design an experiment that maximizes the number of reads derived from RNAs of interest, which can be done by an enrichment or depletion step or by using fewer barcodes (Subheading 3.3). The length of reads is not important, as long as reads are long enough so that

they are accurately mappable to the reference sequence. Rather, it is the *count* of mapped read ends (not the mean per-base coverage) that should be maximized. We have found that we can obtain believable cutting scores for RNAs with mean mapped read ends per site as low as ~2.5 in the nuclease sample and almost no reads in the control sample. However, this number is an average over all sites as the mapped ends are not uniformly distributed (they tend to cluster near ssRNA). So, this is a very coarse estimate of the lower bound on amount of data, and we recommend aiming for at least an order of magnitude greater counts.

To assess whether the assay is producing cutting scores that are reasonable, cutting scores of RNAs with known structures should be examined. Also, it is desirable to add to the complex mixture in vitro-transcribed control RNAs whose structure is already known and preferably for which probing data is available with the specific nuclease used in the high-throughput assay. If resources are available, the different control RNAs should span a range of abundances to identify the number of mapped read ends below which data for an RNA locus becomes unreliable.

The 20–100 nucleotide (nt) size selection step employed in our study [5] is not a requirement of a FragSeq assay—that size selection was performed because that was the optimal cDNA library size for the SOLiD 3 sequencing platform. Larger sizes can be used, and current paired-end sequencing technologies can accommodate longer fragments. Alternatively, a PARS-like approach can be used to randomly shatter long nuclease-digested RNA fragments (producing 5′ OH and 2′,3′-cyclic phosphate end chemistry) so that they fall into sequencing range, followed by end-repair of only one RNA end, so that the other end is used as the tag indicating nuclease cleavage.

2 Materials

2.1 Purification of Complex RNA Mixture for Probing

Refer to the TRIzol manual for specific guidelines for the sample of interest.

1. TRIzol reagent (Sigma-Aldrich or Life Technologies).
2. Chloroform (multiple sources; molecular biology/nucleic acid extraction grade).
3. RNase-free water (multiple sources).
4. 100% isopropanol, molecular biology grade.
5. 75% ethanol, molecular biology grade.
6. Acid phenol–chloroform–isoamyl alcohol (125:24:1; Ambion; pH 4.5).
7. RNase-free DNase I and 10× digestion buffer (Ambion).
8. For resuspending the RNA: 10 mM Tris–HCl pH 8.0, 0.1 mM EDTA pH 8.0.

2.2 Nuclease Calibration and Digestion of RNA with Nuclease P1

1. P1 nuclease (Sigma-Aldrich 200U vial; dissolve vial of lyophilized powder into 250 μL of 50 mM Tris base pH 7.0, 1 mM $Zn(OAc)_2$, 50% glycerol; flash-freeze small aliquots and store at −80 °C).
2. 10× P1 nuclease digestion buffer (quasi-physiological conditions): 500 mM Tris–HCl pH 7.5, 1.5 M $NaCl_2$, 50 mM $MgCl_2$, 0.10 mM $Zn(OAc)_2$.
3. Acid phenol–chloroform–isoamyl alcohol (125:24:1; Ambion; pH 4.5).
4. Denaturing loading buffer: 95% formamide, 10 mM Tris–HCl pH 8.0, 10 mM EDTA pH 8.0, 0.1% bromophenol blue.
5. 5 M ammonium acetate (Ambion).
6. 0.5 M EDTA pH 8.0 (Ambion).
7. P1 nuclease stop solution: 10 mM Tris–HCl pH 8.0, 1 M ammonium acetate, 10 mM EDTA pH 8.0.
8. Glycogen (5 mg/mL; Ambion, molecular biology grade).
9. 100% ethanol, molecular biology grade.
10. FlashPAGE gel supplies (Ambion) or other conventional apparatus for urea–polyacrylamide gel electrophoresis.

2.3 Bioinformatics Analysis

1. Software for mapping sequencing reads to reference sequence.
2. FragSeq code version 0.2.0 (https://bitbucket.org/andrewuzilov/fragseq).
3. Python version 2.7.x, or a later 2.x version (http://python.org).
4. Cython version 0.15.x or later (http://cython.org).
5. Compiler for C and C++.
6. Optional: Java virtual machine version 1.5 or later (http://www.java.com).
7. Optional: RNAstructure [25] version 5.3 or later (http://rna.urmc.rochester.edu/RNAstructure.html).

3 Methods

3.1 Calibration of P1 Nuclease for Probing a Complex RNA Mixture

Prior to partially digesting an RNA sample for RNA-Seq library preparation (Subheading 3.2), one must carefully calibrate the properties of the digestion reaction by the nuclease. In this protocol, nuclease P1 is used, but other nucleases are applicable as well (Subheadings 1.1 and 3.3). To do this, a radioactively end-labeled homogenous RNA sample is used, which we will refer to as a "spike-in" RNA. This can be produced by in vitro transcription or by synthesis; the spike-in RNA should be one that has known structural features under probing conditions used (temperature, buffer, etc.). For each spike-in, probing must be carried out in two

parallel samples: on the spike-in radiolabeled RNA by itself and a similar reaction on the spike-in radiolabeled RNA in the unlabeled complex mixture of RNA to be used for experimental FragSeq probing. The goal is to assure that digestion of spike-in RNA(s) is the same in the complex mixture as it is by themselves, which shows that the complexity of the RNA mixture does not interfere with obtaining good probing data and that *trans* interactions are not occurring. An example of these experiments is shown in Fig. 5.

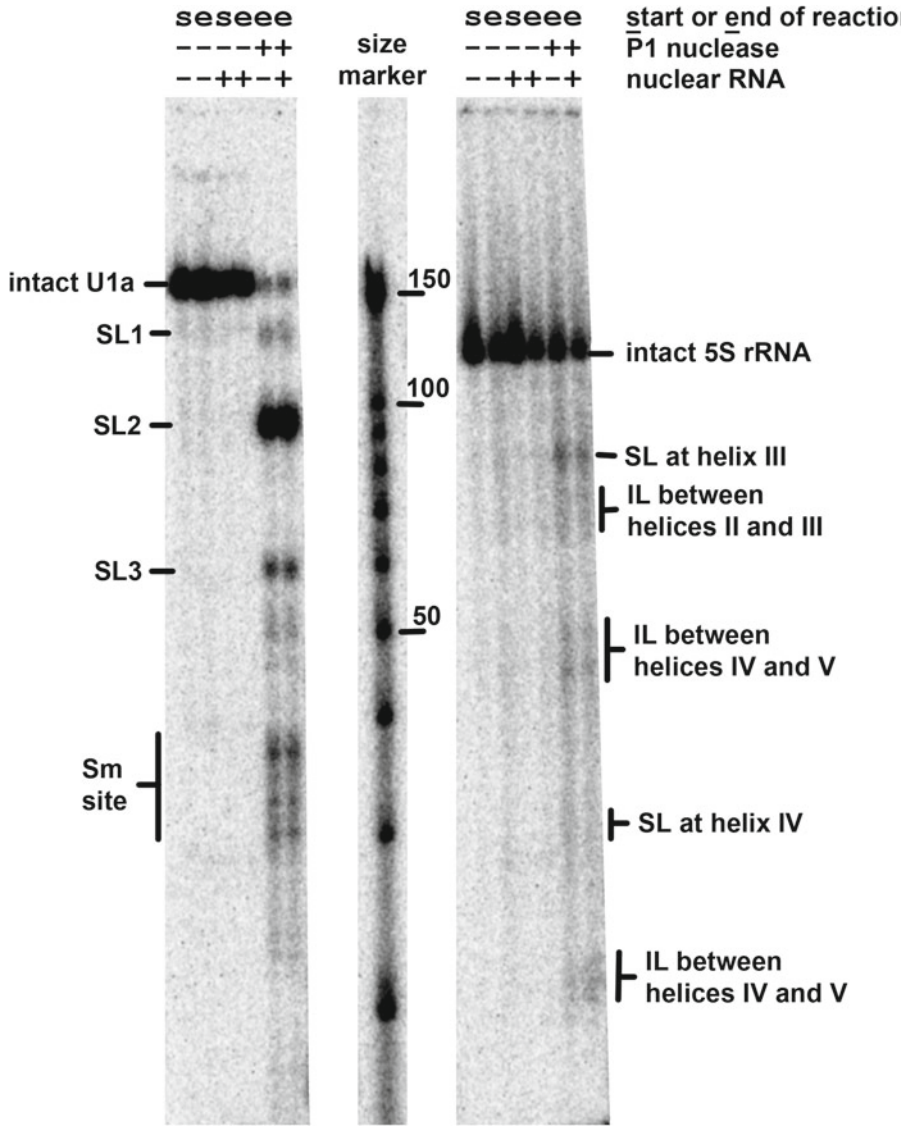


Fig. 5 Digestion of mouse snRNA U1a and 5S ribosomal (r)RNA by P1 nuclease with or without mouse nuclear RNA present. 3′-radiolabeled, in vitro-transcribed RNA was used. Lanes showing RNA at start and end of reaction without nuclease are controls for nonspecific degradation. Size markers are Ambion Decade markers; sizes of 150, 100, and 50 nt are indicated. U1a labeling is as in Fig. 1. 5S rRNA structure and helix numbers are from [46]. Figure and caption are from supplementary material in ref. 5

RNA structure probing is generally carried out under conditions that provide "single-hit kinetics" or "statistical probing"—that is, probing is carried out such that each RNA molecule is exposed to the probing agent at most once, as the first reaction of an RNA molecule may alter its structure and make subsequent reactions less informative [26]. To achieve this, probing conditions must be calibrated such that most RNAs in a sample are not cleaved by the nuclease. Therefore, the small fraction of RNAs that are cleaved will likely be cleaved only once per molecule.

Due to the PCR cycles used when amplifying a sequencing library, even a small number of non-single-hit cleavages can be observed. For certain RNAs where neither end is endogenously ligation-competent (e.g., U6, which contains a 5′ monomethyl phosphate and a 2′,3′-cyclic phosphate [27, 28]), two cleavages are required in order to produce an RNA fragment that can be cloned and amplified. While this is a violation of single-hit kinetics, we find that it tends to be a common case in our sequencing prep and that it yields useful structure information.

Spike-in RNAs should be selected so that their size is in the size range of the RNAs (or RNA domains) that one is interested in probing in the high-throughput assay, because single-hit kinetics will be calibrated for that size range. The size of the full-length RNAs being probed in the complex mixture may be larger, but structured domains of interest may fall within the calibration size range. For example, structured RNA regions in bacterial mRNAs (such as riboswitches and other regulatory elements) tend to be smaller than 200 nt. So although one would probe total, unfragmented mRNA in a FragSeq experiment, single-hit kinetics would be optimized towards domains of size 100–200 nt.

Once a suitable structured spike-in RNA has been identified, an unlabeled in vitro transcript can be produced in bulk from a PCR or plasmid template. It is recommended that this RNA be gel-purified to make sure that all of the material is full length transcript. A small amount of this purified RNA can then be radiolabeled for the calibration experiments.

If 5′ end-labeling of a transcript is desired, the RNA must be dephosphorylated to remove the triphosphate terminus by treatment with alkaline phosphatase, then kinase-labeled with γ-^{32}P-ATP and T4 polynucleotide kinase via standard methods (Sections 10.59–10.67, 11.31–11.33 in ref. 29).

For 3′ end-labeling, the RNA can be used directly after transcription and purification. There are two common methods for 3′ end-labeling with commercially available enzymes and ^{32}P nucleotides:

1. Addition of a single radioactive adenosine base to the 3′ end of the RNA by polyA polymerase and α-^{32}P-cordycepin triphosphate (3′-deoxy-ATP) [30].
2. Addition of a single radioactive cytosine base to the 3′ end of the RNA by T4 RNA ligase and 5′-^{32}P-pCp [31].

If the nuclease properties permit it, digestion conditions should be calibrated at conditions as close to physiological for the species of interest. In the protocol below, conditions are given for physiological conditions for mammalian cells. Nuclease P1 is not especially active at these conditions, but this is a desirable property in the quest for single-hit kinetics.

3.1.1 Calibration Experimental Workflow

1. For each spike-in RNA, prepare two parallel samples: 100 ng of unlabeled spike-in RNA ("homogeneous reaction") and 100 ng of unlabeled complex RNA mixture ("heterogeneous reaction"). Dilute each amount of RNA into 89 μL of water and add 10 μL of 10× P1 nuclease digestion buffer.
2. Dilute trace amount (~0.1 ng or 100,000 cpm of ^{32}P) of 5′ or 3′ end-labeled spike-in RNA into each of the two samples.
3. Heat the RNA at 55 °C for 5 min, then 37 °C for 10 min. This denatures and refolds the RNA to its lowest energy state, so that RNA structures are more consistent.
4. At this point, remove 20 μL of the reaction to serve as an "input" no-nuclease control.
5. Add 1 μL of P1 nuclease to each tube (*see* **Note 6**).
6. Incubate the tube at the desired probing temperature (for mammalian RNAs, we utilized 37 °C).
7. Remove 20 μL aliquots at desired times for optimization. We recommend 5, 15, 30, and 60 min time points for this buffer and temperature combination.
8. As each aliquot is removed, stop the reaction by bringing it to a final volume of 400 μL with P1 nuclease stop solution.
9. Add an equal volume of acid phenol–chloroform to each tube.
10. Once all of the time points are ready, process each extraction carefully and in a fume hood due to both the presence of isotope and phenol–chloroform. Transfer the aqueous portion to a fresh 1.5 mL microfuge tube. Dispose of the radioactive phenol–chloroform waste appropriately per institutional environmental health and safety regulations.
11. Add 4 μL (20 μg) glycogen to each tube.
12. Precipitate by adding 1 mL of 100% ethanol and centrifugation at 14,000 × *g* at 4 °C.
13. Resuspend and heat each sample to 95 °C in denaturing loading buffer for 5 min, then resolve in parallel lanes on a medium sized (e.g., 15 × 17 cm) denaturing PAGE gel (*see* **Note 7**). 8 M urea, 8% 19:1 acrylamide–*bis* is applicable for 100–500 nt RNAs.
14. The gel should be dried and imaged with a PhosphorImager plate for analysis.

3.1.2 Choosing a Nuclease Condition

Using the calibration assays with the spike-in RNA, one can determine the digestion parameters to be used for complex mixture probing. For true single-hit kinetics, a condition should be chosen where most of the full length molecule is still intact. One should also note if there are differences between the spike-in RNA probed on its own versus in a complex mixture, since any *trans* interactions will complicate later analysis and make using FragSeq data for guiding RNA structure prediction difficult.

The easiest parameters to alter during this calibration series include: the identity of the spike-in RNA, enzyme concentration, incubation temperature, pH, salt, and time of incubation. Nuclease P1 is a relatively thermostable enzyme, so higher temperatures (up to 70 °C) are possible, but may cause unfolding of the RNA. Increased salt will increase the stability of RNA secondary structures, but may decrease the efficiency of the nuclease. Finally, nuclease P1 is stable at pH 5–8 and shows higher activity at lower pH. As with raising temperature, this higher activity could cause over-digestion, so enzyme dilution for digestion at this pH is recommended.

3.2 Digestion of a Complex RNA Mixture with Nuclease P1

This protocol produces RNA fragments for downstream library preparation for high-throughput sequencing.

1. Suspend complex RNA mixture at 1 ng/μL concentration in P1 nuclease digestion buffer. A reaction in the range of 100–500 μL is usually applicable.
2. Separate the above master mix into two equal volumes: "nuclease" sample and "control" sample (*see* **Note 8**).
3. Heat both samples at 55 °C for 5 min, then 37 °C for 10 min.
4. Add the predetermined concentration of P1 per unit volume to the nuclease sample and incubate for the predetermined time (*see* **Note 9**). Keep the control sample at the same temperature and for the same time as the nuclease sample.
5. Stop the nuclease and control reactions at the same time by adding 1/10th of the reaction volume of 0.5 M EDTA pH 8.0 and 1/5th volume of 5 M ammonium acetate.
6. Purify the RNA in the nuclease and control reactions by acid phenol–chloroform extraction and ethanol precipitation as detailed in **steps 9** through **12** in Subheading 3.1.1 (scale up the volume of ethanol used in **step 12** per your reaction volume).
7. For both samples in parallel, select the RNA size fraction required for the specific cDNA library prep and sequencing technology. For this, we recommend the Ambion FlashPAGE system or another small PAGE system. For FlashPage, heat the samples in the included sample loading buffer and carry out initial electrophoresis per manufacturer instructions, collecting the smaller-than-desired RNA fraction in the anode cup. Then, carry out electrophoresis again to collect RNA in the desired

size fraction. We found that 30 min of this second electrophoresis step to be applicable to the 20–100 nt RNA size range (*see* **Notes 10** and **11**).

8. Recover the collected RNAs from the FlashPAGE cup by ethanol precipitation with glycogen. For a PAGE gel, the area of interest can be localized by using radiolabeled markers in parallel lanes and subsequently eluted from the gel overnight with P1 nuclease stop solution, then ethanol precipitated with glycogen. Suspend the RNA pellet at the volume and in the buffer of choice for the downstream library preparation protocol, such as 10 mM Tris–HCl pH 8.0.

3.3 Ligation and Library Prep from Nuclease and Control RNA Samples

The FragSeq methodology and command-line tool can be adapted to a variety of sequencing platforms. Since sequencing technology evolves at an astonishing rate, any specific kit recommendations made in this chapter may rapidly become obsolete. For example, the Applied Biosystems SOLiD Small RNA Expression Kit (SREK) used in ref. 5 is no longer available, though the following currently available kits may be substituted because they use the same approach of ligating to both RNA fragment ends simultaneously using adapters containing overhangs:

- SOLiD Total RNA-Seq Kit (catalog number 4445374, Applied Biosystems).
- Ion Total RNA-Seq Kit v2 (catalog number 4475936 and 4479789, Life Technologies).

Other kits that ligate to both ends of the RNA fragment using different approaches are available:

- TruSeq Small RNA Library Preparation Kit (catalog number RS-200-0012, Illumina).
- NEBNext Small RNA Library Prep Set for Illumina (catalog number E7330S or E7330L, New England Biolabs).

As kit availability and designs change, use the following guidelines when selecting a library preparation method:

1. It is critical to ligate a defined adapter sequence onto the cleaved sites within the RNA such that a sequencing read begins at or crosses the junction between that adapter and the fragment from the probed RNA. Determining the precise identity of this junction is essential for the single-base-resolution.
2. Ligation of defined sequences to *both* ends of an RNA fragment is necessary for PCR amplification. In ref. 5, end-specific adapters were ligated to both ends of RNA simultaneously using SREK. This kit, developed for miRNA and siRNA characterization, can only ligate adapters onto RNA molecules that possess a 5′ PO_4 and 3′ OH, so this was ideal as nuclease P1

cleavage produces these end chemistries. The adapter ligation strategy can be tailored to fit the nucleases of interest and also the possible products of random hydrolysis.

3. Alternately, pre-adenylated adapters can be added to the 3′ end of the RNAs, selecting for 3′-OH ends, followed by reverse transcription primed by an oligonucleotide complementary to the adapter [32]. Then, cDNA 3′ ends can be tagged with another known adapter sequence [33, 34]. The ligation efficiency of pre-adenylated adapter as given above has been criticized as having sequence bias [35]; it may be possible to work around the biases by using NEXTflex Illumina Small RNA Sequencing Kit v2 (catalog number 5132-03 or 5132-04, Bioo Scientific). However, the FragSeq algorithm normalization procedure should ameliorate the ligation bias (*see* **Note 2**) even if the method from ref. 32 is used because cutting scores are based on comparing the *same* site between two conditions.
4. Reverse transcription and PCR are used to convert the adapter-ligated RNA pool into double-stranded DNA molecules applicable to high-throughput sequencing. Barcodes can also be added during this step if desired. Barcoding is a common way to divide up a sequencing run and this is highly recommended for FragSeq methodologies since libraries from control and nuclease conditions can be multiplexed and sequenced simultaneously (e.g., on the same lane of an Illumina instrument), reducing batch effects.

3.4 Summary of Steps in the Bioinformatics Analysis Pipeline

The steps in running a computational FragSeq analysis are:

1. Prepare input files:
   (a) Identify RNA loci for which obtaining structure data is desired.
   (b) Map sequencing reads to reference sequence (Subheading 3.5).
   (c) Put coordinates of RNA loci from **step 1**(**a**) in a BED file (Subheading 3.6.2).
2. Write configuration file(s) that tell FragSeq command-line tool (`readsToStruct.py`) what to do (Subheading 3.6.3).
3. Run the FragSeq command-line tool (`readsToStruct.py`, *see* **Note 12**).
4. Examine the output:
   (a) Examine read mapping end counts, probabilities, and cutting scores for each RNA locus of interest (Subheading 3.7).
   (b) Upload wiggle tracks containing the above data to a genome browser to examine them in genomic context (Subheading 3.7.2).

(c) For RNAs for which secondary structure model(s) are available, use VARNA to plot FragSeq probing data on each structure. If no structure is available, use RNAstructure with constraints derived from cutting scores to predict secondary structures (Subheading 3.7.3).

3.5 Mapping Sequencing Reads to Reference Sequence

Sequencing reads from the cDNA library prepared in Subheading 3.3 must be mapped to a reference sequence so that we can determine which read ends correspond to which sites in our RNAs of interest. The reference sequence can be the genome assembly (e.g., hg38 for human or mm10 for mouse) or a set of RNA sequences (*see* **Note 13**), in which case `readsToStruct.py` must be run in "local" mode. We use this chapter to explain the important properties of an alignment pipeline so that the user can make their own decision in selecting the right tool, as there has been a dramatic proliferation in various alignment tools over the past few years [36] (*see* also: http://www.ebi.ac.uk/~nf/hts_mappers/).

A sequencing read is generally a tag (subsequence) of a cDNA amplicon of an RNA fragment (Fig. 3). In FragSeq, unlike in many other RNA-Seq bioinformatics analyses, we care about the *ends* of RNA fragments because those correspond to sites where the parent RNA was specifically cleaved by a nuclease or nonspecifically broke. Care must be taken that we only use reads whose *relevant* ends (Fig. 3, *see* **Note 14**) align well to reference sequence (Fig. 6). If a mapping tool fails to align the relevant ends of a read, we must not use that read for FragSeq analysis.

Reads may contain adapter sequence (*see* read1 in Fig. 3). It is very important that any such adapter sequences are stripped from the reads *before* aligning reads to reference sequence; otherwise, bases in adapter may be erroneously aligned as if they were part of the insert (*see* **Note 16**). Because single-base resolution at read ends is crucial for accuracy in FragSeq, the erroneous addition of even one or two adapter bases could distort the signal in a way to which the FragSeq algorithm is not robust.

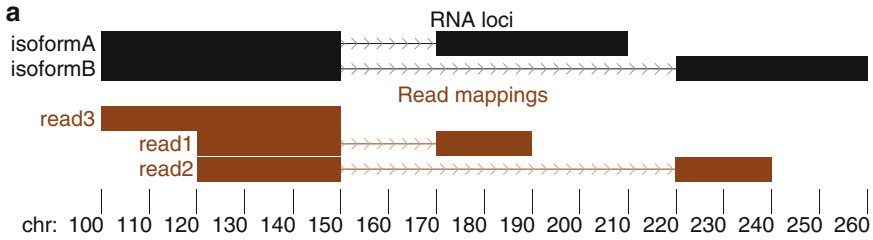
3.5.1 Special Considerations for Spliced or Overlapping RNAs

If a read originates from an exon that occurs in multiple isoforms of a spliced RNA, it is not clear to which isoform the read should be assigned. For example, `read3` in Fig. 7 cannot be unambiguously assigned to either `isoformA` or `isoformB` based on genomic annotations alone. The same issue occurs when assigning reads to *any* RNA loci whose coordinates in the reference sequence overlap. A read can map to a unique position in a reference sequence, but that position has more than one RNA locus annotation. The user must determine which reads belong to which locus; `readsToStruct.py`, although it can load reads and loci containing introns, cannot make this determination.

A simple way to partition ambiguous reads (e.g., `read3` in Fig. 7) amongst multiple loci is to randomly assign them to loci according

Fig. 6 Example alignments of read sequences (*bottom*) to reference sequence (*top*) illustrating cases when read mapping ends are (*checkmark icon*) and are not (*cross icon*) appropriate to consider by the FragSeq algorithm. To simplify the example, reads are 30 bases long and we assume that 5′ and 3′ ends are relevant (*see* **Note 14**), i.e., the reverse transcriptase copied the RNA fragment to cDNA in its entirety and adapter sequence is trimmed, like read1 in Fig. 3; these assumptions are not valid for all cDNA library preparation strategies. CIGAR alignment strings are shown (*see* **Note 21**). *Vertical lines* denote a sequence match. (**a**) A good alignment that includes both read ends. The single sequence mismatch occurs far away from the ends, so the ends are still useful for FragSeq. (**b**) The alignment algorithm was unable to align read ends, rendering them useless for FragSeq; however, the *middle* of the read is well-aligned and potentially useful for other bioinformatics analyses. This may occur if adapter sequence has been stripped incompletely, if the read is chimeric, or if the read is mapped to the wrong locus (can occur if part of the locus is a repetitive element). (**c**) Only the 5′ end of the read is reliably aligned. Although the lack of alignment at the 3′ end makes the entire mapping suspect, the aligned 5′ end may still be useful for FragSeq (*see* **Note 22**). (**d**) Neither read end is reliably aligned. Such alignments should be discarded prior to input to `readsToStruct.py` (*see* **Note 12**)

to read density observed for *unambiguous* reads (e.g., `read1` and `read2` in Fig. 7). For example, if `isoformA` has twice as many unambiguous reads mapped to it as `isoformB`, it will randomly get twice as many ambiguous reads. For splicing, a more powerful approach is to use one of several read mapping tools that have been developed specifically for dealing with multiple splicing isoforms



b

BED12:

```
chr  120  190  read1  1000  +  120  190  0  2  30,20  0,50
chr  120  240  read2  1000  +  120  240  0  2  30,20  0,100
chr  100  150  read3  1000  +  100  150  0  1  50,    0,
```

SAM:

```
read1  0  chr  121  255  30M20N20M  *  0  0  *  *
read2  0  chr  121  255  30M70N20M  *  0  0  *  *
read3  0  chr  101  255  50M        *  0  0  *  *
```

c

BED12:

```
isoformA  20  70  read1  1000  +  20  70  0  1  50,  0,
isoformB  20  70  read2  1000  +  20  70  0  1  50,  0,
isoformB  0   50  read3  1000  +  0   50  0  1  50,  0,
```

SAM:

```
read1  0  isoformA  21  255  50M  *  0  0  *  *
read2  0  isoformB  21  255  50M  *  0  0  *  *
read3  0  isoformB  1   255  50M  *  0  0  *  *
```

d

BED12:

```
chr  100  210  isoformA  1000  +  100  210  0  2  50,40  0,70
chr  100  260  isoformB  1000  +  100  260  0  2  50,40  0,120
```

SAM:

```
isoformA  0  chr  101  255  50M20N40M  *  0  0  *  *
isoformB  0  chr  101  255  50M70N40M  *  0  0  *  *
```

Fig. 7 (**a**) UCSC Genome Browser view of two splicing isoforms and genome-mapped reads that illustrate issues in assignment of reads to isoforms. chr is the name of the genomic reference sequence. (**b**) Alignments of reads using chr as the reference sequence (global mode), in BED12 and SAM formats (*see* **Note 21**). (**c**) Alignments of reads using spliced isoform sequence as the reference sequence (local mode), in BED12 and SAM formats. In this case, we arbitrarily assign `read3 to isoformB`. (**d**) Annotations of isoforms using chr as the reference sequence, in BED12 and SAM formats

(*see* the "RNA mappers" list on http://www.ebi.ac.uk/~nf/hts_mappers/). For eukaryotic analysis, this would also have the advantage that the alignment algorithms are designed for aligning short reads across large introns that occur in eukaryotic genomes.

If partitioning reads, `readsToStruct.py` must be told explicitly which read mappings are assigned to which locus. To do this, mapping coordinates must be saved in "local" mode and the

configuration file (Subheading 3.6.3) must have the `input.reads.local` setting enabled. We define "local" coordinates to be within the coordinate system of the locus to which a read is mapped (which has introns removed, if there are any; *see* also **Note 13**), as opposed to "global" coordinates within a genomic reference sequence such as chromosomes (which still contain introns). The distinction between local and global coordinates is shown in Fig. 7 (panels c versus d) using commonly used file formats for representing coordinates. The most important point is that the reference sequence (column 1 of BED12 format and column 3 for SAM format) for local coordinates specifies the name of a locus, so `readsToStruct.py` can unambiguously know which read belongs to which locus.

3.5.2 Special Considerations for Multi-copy or Repetitive RNAs

Certain RNAs (such as snRNA in eukaryotes) exist in multiple copies in a genome or have several paralogs whose sequences are similar. This can create issues when mapping reads directly to genomic reference sequence because a read may map to multiple loci equally well. Some mapping tools may discard reads that have too many multiple mappings or may assign such mappings a very low score, which causes them to fall below a cutoff threshold and become discarded. This would result in abundant RNAs having seemingly very little data. For reads that are not discarded, it is ambiguous how they should be partitioned among the multiple matching loci.

For RNAs whose sequence is multi-copy in its entirety (e.g., snRNA), we recommend creating a set of reference sequences where each multi-copy RNA occurs once and initially mapping reads to that, then mapping remaining reads to the genome. This may erroneously over-map some reads to the multi-copy RNAs, but the FragSeq algorithm is somewhat tolerant of that (*see* **Note 2**). Mappings will have to be saved in local coordinates (Subheading 3.5.1).

For RNAs where only a subsequence is multi-copy, we recommend assigning multiply mapping reads according to the number of uniquely mapping reads in the same RNA, similarly to Subheading 3.5.1.

3.6 Running the FragSeq Command-Line Tool (`readsToStruct.py`)

3.6.1 Overview

The command-line tool `readsToStruct.py` transforms read mappings from the nuclease and control RNA-Seq samples into cutting scores and other informative data, much of which can be uploaded to the UCSC Genome Browser [37] or other tools (*see* **Note 17**) or visualized in a secondary structure context using VARNA software [21] (Subheading 3.7).

These computational methods sections are written for FragSeq version 0.2.0, which has several improvements from version 0.0.1 used for ref. 5, primarily:

- Read mappings can now be input and output in SAM/BAM format.

- Reduced RAM usage.
- Improved configuration file syntax.
- Spliced input can now be handled.

FragSeq version 0.2.0 has been tested on Linux and Mac OS X operating systems. It is written in a portable way using only portable libraries and therefore should, in theory, also work on Windows operating systems; however, this has not been explicitly tested at the time of this writing.

3.6.2 Input Files

As required input, `readsToStruct.py` takes three files: coordinates of read mappings from the nuclease and control samples (Subheading 3.5) and coordinates of RNA loci for which the analysis is desired (*see* **Note 3**). The RNA loci file must be in BED format (*see* **Note 18**; http://genome.ucsc.edu/FAQ/FAQformat). Several variations of the BED file format exist; at a minimum, we require the six-column format (BED6) because the strand information in the sixth column is essential. RNA loci can be spliced, in which case the BED12 (12-column) format must be used, which gives the positions of introns or exons. A set of loci annotations from existing tracks in the UCSC Genome Browser can be downloaded using its Table Browser feature. Likewise, loci can be uploaded as a custom track to the UCSC Genome Browser for viewing (*see* **Note 17**). Read mappings are now encouraged to be in either SAM or BAM format (https://github.com/samtools/hts-specs), which are currently the de facto standard for storing sequencing read data; BED format is discouraged (*see* **Note 15**).

The RNA loci file is used by `readsToStruct.py` to figure out which reads came from which RNA, an important step because the assignment of reads to correct RNAs is crucial for normalization (*see* **Note 3**). If no RNA loci overlap, assigning reads is trivial—if a read overlaps a locus in their common coordinate space, the read is assigned to that locus (*see* **Note 19**). If RNA loci overlap, this simple procedure cannot be used, so the user must provide the reads in local mode (Subheading 3.5.1). In local mode, the assignment of reads to loci is simple and unambiguous—the algorithm just looks at the name of a read's reference sequence and looks it up in the list of RNA loci already given.

3.6.3 Configuration Files

`readsToStruct.py` is controlled by a configuration file that specifies the input files, output files, and the behavior of the algorithm. A very minimal example configuration file is given in Fig. 8; a more complex configuration file example, included with FragSeq code (Subheading 2.3), contains the exact configuration to reproduce our analysis in ref. 5.

The user must write the configuration file and feed it to `readsToStruct.py` (*see* **Note 12**); optionally, the configuration can be spread across several files, which is useful if the same piece

```
%YAML 1.1
---
##############################################################
# A simple FragSeq configuration file.
##############################################################

define:
    IN_DIR: path/to/some/input/directory
    OUT_DIR: path/to/some/output/directory

input:
    reads:
        nucl: IN_DIR/reads.nucl.bam
        ctrl: IN_DIR/reads.ctrl.bam
        type: bam
        local: False
    loci: IN_DIR/knownRnas.bed

output:
    config: OUT_DIR/analysis_log.conf.yaml
    cutscores:
        listfile: OUT_DIR/%.cutscores.list

algorithm:
    numEndSitesToIgnore: 3
    noiseCutoff: 10
```

Fig. 8 A minimal configuration file for `readsToStruct.py`. An analysis run using this file will produce only cutting scores text files and a configuration/analysis log

of configuration is reused in several analyses. You can also define variables in configuration files, which is useful if the same piece of text (e.g., a directory path) is reused several times.

All configuration settings can be saved to a log file before the analysis begins so that there is an automatic record of all parameters of an analysis (the setting `output.config` in Fig. 8). This is done after all input configuration files are merged, all defaults are applied, and all variable substitution is done. So, the saved configuration may contain settings that the user did not specify explicitly (the defaults). The saved configuration file is a valid configuration file itself that can be input to run an analysis.

3.7 Working with the Output of `readsToStruct.py`

Output files fall into two categories: per-locus and per-analysis. For per-locus files, the output file name/path in the configuration file must have a wildcard (%) character, which will be replaced with the name of the RNA locus from the input BED file of loci (Subheading 3.6.2; *see* **Note 18**; Fig. 8). Per-analysis files may not contain wildcards. If a setting is not provided in the configuration file for output of a certain type, then it will simply be skipped.

3.7.1 Interpretation of Cutting Scores

A cutting score for a site indicates how likely we are to observe read mapping end counts in the nuclease sample versus the control sample relative to other sites in that RNA, in those samples. Cutting scores are log ratios (natural log) of read mapping end probabilities

in the nuclease sample to the control sample. A positive value at a site indicates it is more likely to have relevant ends at it in the nuclease sample than control, relative to other ends in that RNA in those samples. Magnitudes of scores zero or below are not informative and are therefore filtered out (although this can be turned off for debugging) and replaced with none in text output.

When reviewing cutting scores, it is important to distinguish between "ignored" and "non-ignored" sites. Ignored sites either were masked out by the user (*see* **Note 20**) or had too little data in nuclease and control samples to be included in cutting score calculation (the end count threshold controlling this is tunable by the user via the `algorithm.noiseCutoff` configuration setting). Ignored sites will never have cutting scores, by definition. RNAs with less read mappings tend to have more ignored sites. Ignored sites are identified prior to the normalization step; in Fig. 1, the "number of ends" track shows values for all sites, but the "end probability" and "cutting scores" tracks only show values for non-ignored sites. Sites that are ignored and non-ignored, as well as how many are in each category, are logged to the per-locus stats output files (`output.stats` configuration setting).

A higher cutting score means the site is more susceptible to the nuclease, but the reverse is not true. If the site has no cutting score and is marked non-ignored (or the cutting score is small), it should *not* be interpreted as lacking susceptibility to the nuclease (put another way, absence of evidence is not evidence of absence). This could occur due to artifacts in the algorithm or the experimental method. However, when present and large, cutting scores tend to be accurate.

Moderate to high cutting score magnitudes seem to correlate with susceptibility to probing agents from other studies [5], but we find it is only possible to compare magnitudes of cutting scores between sites in the same RNA, in the same sample. For example, Fig. 1 shows that although the relative magnitudes of cutting scores in UESC and Neural Precursor samples follow the same pattern, their absolute magnitudes are different (maxima of 1.91 and 2.87, respectively).

3.7.2 Genome Browser Output

Data at every step in the FragSeq pipeline (Fig. 4) can be output in plain text wiggle format, which can be uploaded to the UCSC Genome Browser as a custom track (Fig. 1), thus allowing the user to view structure probing data in a genomic context alongside annotations that are already present in the browser or annotations that the user can upload (*see* **Note 17**).

Uncompressed, whole-transcriptome wiggle files can be tens to hundreds of megabytes in size for eukaryotic genomes and therefore may be size-prohibitive for upload to UCSC servers as custom wiggle tracks. Also, custom tracks are not guaranteed to be retained by UCSC for a long time. Lastly, uploaded wiggle data is compressed in a lossy way on UCSC servers, meaning the data

values displayed in the browser will not be exactly equal to the data uploaded. All of these issues can be circumvented if the user converts the output wiggle data to bigWig format, which is a terse, indexed, binary format storing the same information. A detailed tutorial on how to do this conversion exists (http://genome.ucsc.edu/goldenPath/help/bigWig.html; download the program `wigToBigWig`).

bigWig-format files are significantly smaller in size than wiggle files storing equivalent data, but they are required to be stored on the *user's* server—the user uploads to the UCSC Genome Browser only a track header containing an URL that points to the file on the user's server, and the browser fetches only the necessary pieces of the file as they are viewed by a user. This is more efficient than keeping all the wiggle data on UCSC servers and it is also not necessary to upload the complete data set to UCSC servers, thus circumventing the file size problem. Using the -unc flag to the `wigToBigWig` program ensures that the data is not compressed, so exact values can be seen; even without compression, there is a significant size reduction when converting wiggle to bigWig.

When input read mappings are in local coordinates (Subheading 3.5.1), output wiggle files will be in the wrong (global) coordinate system and cannot be directly uploaded to the UCSC Genome Browser. The user must write code to convert the coordinates to the genomic coordinate system before upload.

3.7.3 Output for RNA Structure Analysis

To use FragSeq data for RNA secondary structure analysis, two things can be done:

1. A secondary structure can be displayed in Varna [21], with cutting scores plotted on top of it, also highlighting non-ignored sites (Fig. 9). This is useful for evaluating secondary structure models (including comparing several models for a single RNA), producing a publication figure, seeing how well experimental results agree with a known model, etc. Raw end counts and probabilities can be displayed as well, which may be useful for troubleshooting.
2. A secondary structure can be predicted using RNAstructure software [25], using cutting scores as offsets that guide structure prediction.

The command-line version of `RNAstructure` software can predict minimum free energy (MFE) RNA secondary structures using a set of offsets provided by the user (Fig. 2). Offsets are pseudo-free-energies in units of kcal/mol; they can be ssRNA or double-stranded (ds)RNA offsets. If a base has a negative (favorable) offset, and the folding algorithm considers a structure containing that base in the corresponding configuration (i.e., base is ssRNA for an ssRNA offset or dsRNA for a dsRNA offset), the offset value is applied as a reward—the offset value is added to the

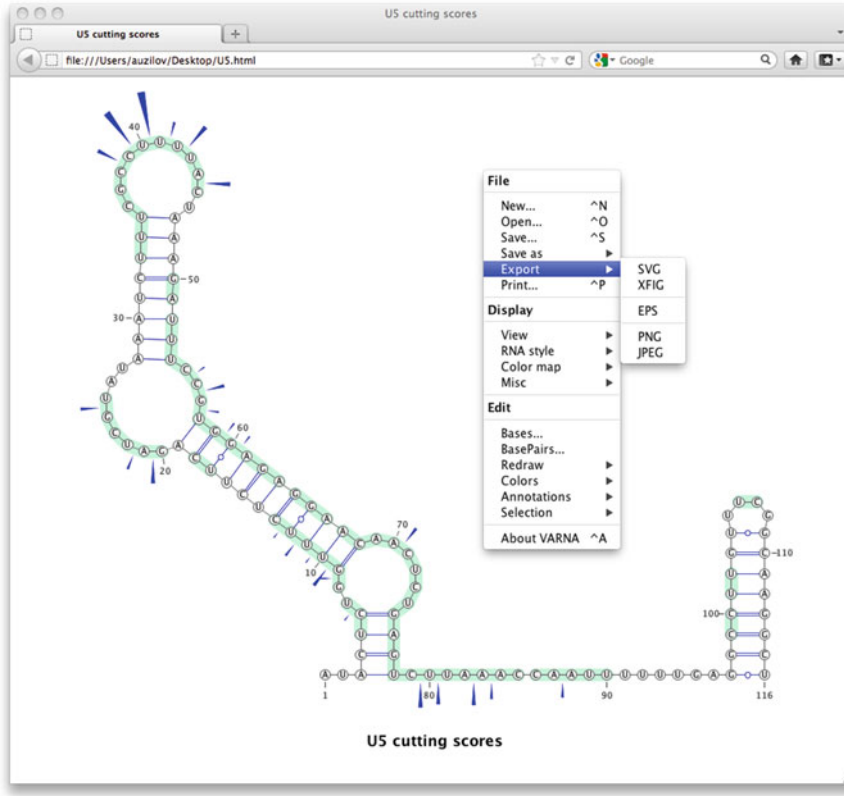


Fig. 9 Cutting scores on top of a known secondary structure, visualized using VARNA. Right-clicking opens a menu (shown) containing many useful options, such as exporting the visualization in various formats, as well as annotating the structure. *Shaded* sites are non-ignored sites

folding free energy change (ΔG°) of that structure, which makes it more likely that the structure containing that base configuration will be the MFE structure.

Cutting scores can be converted to offsets using a simple linear transformation, as has been done with SHAPE probing data [38]. Positive cutting scores must be converted to negative offsets because negative free energies are favorable. Optionally, cutting scores can be filtered to remove all scores below a certain threshold, as low cutting scores may be noise (Subheading 3.7.1).

The choice of slope, intercept, and threshold has a major impact on the predicted secondary structure. In practice, we find that there is no single slope, intercept, and threshold that can be applied uniformly to improve structure prediction for all RNAs, but that is likely due to lack of robust benchmark data, as there have not been enough FragSeq experiments done at this time. In practice, we advise that the user try several slopes, intercepts, and thresholds and compare the resulting structures in VARNA. Superimposing probing data on both "before offsets" and "after offsets" structures allows one to examine how much the structures

are changing as a function of these perturbations (e.g., *see* Fig. 2). It is not guaranteed that applying offsets will improve structure prediction in every case; the user must review plots of various secondary structures with probing data superimposed on them and making a judgment based on that and other forms of evidence to derive the best structure model.

4 Notes

1. We define a *site* as a position between two adjacent bases in any sequence (cDNA, RNA, sequencing read, reference sequence, etc.). In RNA, a site is the region between two bases where cleavage of the phosphate backbone could occur. End positions of read mappings are more conveniently described using site coordinates than base coordinates. For ease of algorithm implementation, native (i.e., pre-cleavage) 5′ and 3′ ends of transcripts are also considered sites.
2. Because cutting scores are log-ratios of per-site data between two samples, they are in theory somewhat tolerant of some experimental and computational biases and artifacts. For example, artifacts due to ligation bias at a site or multiple mappings of a read are a function of the reference sequence, so for any given locus they may affect the nuclease and control samples similarly and thus may cancel out. Adapter ligation bias may cancel out for the same reason. However, this robustness has not been rigorously assessed.
3. In FragSeq, read mapping end counts are normalized at each RNA locus independently, using only data for that locus. This normalization procedure is the reason why the user is required to identify coordinates of RNA loci to input to our tool. For each locus, normalization produces a discrete probability distribution of observing an end at a site, specifically in that transcript, in that sample. We chose this normalization strategy because it makes inference of cutting scores for a transcript independent from abundance of other transcripts—the relative probabilities of ends within one gene are not affected by the abundance of other genes. This is especially suitable when read coverage of genes relative to each other varies between the nuclease and control samples. For example, in ref. 5, the control sample was dominated by C/D box snoRNA reads because those RNAs have endogenous 5′ PO_4 and 3′ OH end chemistry which was ligation-competent and many fell within the size selection range; allocation of reads to these RNAs made other RNAs seem less abundant by comparison (an artifact), but the normalization procedure removed that effect.
4. We, as well as others [39], have observed that V1 also tends to cleave at ssRNA positions (including positions cleaved by

ssRNA nucleases); the mechanism of V1 recognition is believed to be stacked bases [26, 40], which can occur in ssRNA. Stacked ssRNA could be cleaved by both ssRNA nucleases and V1; additionally, there may be a substrate length requirement for V1 cleavage [40, 41] that may be different from the ssRNA nuclease requirement. It would therefore be difficult to interpret a scoring scheme based on a ratio of ssRNA-specific nuclease activity to V1 activity. This is why we prefer to base scores on activity of a nuclease with respect to its own control sample. The control sample allows us to get an estimate of ligation-competent RNA fragments that were not specifically produced by the desired nuclease.

5. A common method for purifying RNA from cells without selection for any specific type of RNA is guanidinium isothiocyanate–phenol–chloroform extraction, commonly referred to by the trade name TRIzol extraction [42]. This method is useful for RNA isolation from nearly all specimens (bacteria, archaea, yeast, plant, animal, etc.), although the volumes may need to be scaled to obtain the desired quantity of RNA. Refer to the manual that accompanies the TRIzol reagent for guidance on proper volumes for a particular project of interest. After purification with TRIzol, we recommend a subsequent treatment by DNase I to remove any retained DNA, followed by an acid phenol–chloroform extraction to obtain pure RNA.
6. This is a starting recommendation for enzyme that was diluted and flash-frozen in aliquots as indicated in Subheading 2.2, but this is a parameter than can and should be optimized per batch of P1. A fresh aliquot should be thawed and used each time that a probing experiment is performed to maintain consistency.
7. A size ladder can be produced by making a "G-ladder" by partial RNase T1 digestion of the spike-in radiolabeled RNA. Alternatively, many commercial RNA or DNA ladders can be easily radiolabeled with T4 polynucleotide kinase.
8. Note that a parallel reaction pair with added radiolabeled spike-in RNAs can also be performed if the researcher wants to monitor the digestion by gel in parallel. Keeping this monitoring reaction separate makes sure that radioactivity is not carried forward in sequencing library preparation.
9. These parameters were optimized in Subheading 3.1 for a 100 μL reaction, so scale up in a linear fashion if a larger reaction is desired.
10. Alternatively, a standard urea-PAGE can be used to size-select the desired RNAs.
11. Cleavage by nuclease P1 will leave a 5′ PO_4 and 3′ OH. If one desires smaller overall fragment sizes after nuclease P1 hydro-

lysis, a random hydrolysis ("RNA shattering") by magnesium, heat and alkaline pH can be utilized. This will generate 2′,3′-cyclic phosphates which cannot accept an adapter by T4 RNA ligase. These ends can be either ligated with a tRNA ligase [43] or by converting these cyclic phosphates to 3′ OH with T4 polynucleotide kinase in the absence of ATP [24].

12. It is space-prohibitive to discuss the `readsToStruct.py` configuration file in this chapter. Therefore, this chapter focuses on explaining the concepts, whereas the README file in the FragSeq source code repository (Subheading 2.3) explains the command-line usage and the configuration file syntax.
13. When aligning to RNA sequences directly, you must align to the sense strand, i.e., the reference sequence index for your alignment tool must be built from FASTA sequences of the sense strand, *not* the genomic plus strand.
14. "Relevant" sites (whether in cDNA, reads, mappings, etc.; *see* **Note 1**) are those sites that correspond to "relevant" RNA fragments ends, which are the fragment ends that yield nuclease accessibility data in the nuclease sample (or the corresponding data in the control sample). Not every RNA fragment end is relevant—this depends on the experiment design and library preparation protocol. For example, in ref. 5, both RNA *fragment* ends are relevant, but for *mappings*, only the 5′ ends are relevant in *every* mapping (the 3′ end is relevant only for trim mappings, Fig. 3). To exclude native RNA ends from the relevant end pool, *see* **Note 20**.
15. Support for BED is retained for backwards compatibility with FragSeq version 0.0.1 used in ref. 5.
16. For trimming adapter sequence from paired-end reads, we recommend SeqPrep (https://github.com/jstjohn/SeqPrep) as it takes advantage of the fact that if a read from one end sequences into the opposing adapter, its sequence will overlap the read from the other end and they can be aligned to each other to accurately identify the adapter position; however, a wide variety of other adapter trimming tools also exist.
17. Many file formats suitable for upload to the UCSC Genome Browser are also accepted by other genome browsers. If the researcher is setting up their own genome browser for data viewing, we recommend using JBrowse [44], as it is easy to set up by users themselves and does not require running a web-server, whereas the UCSC Genome Browser aims to provide a centralized data access service managed by UCSC.
18. Names of RNA loci (fourth column in the input BED file) must follow two rules. First, names must be unique—no name can be used more than once in the file. Second, names must contain only printable, non-whitespace ASCII characters. This

is because names of RNA loci will be used to create names of output files, so to avoid issues, we are restricting the set of valid name characters to a small core set. However, there are no restrictions on names of read mappings.

19. It is only useful to input reads that are already known to overlap RNA of interest, otherwise there will be a performance cost—`readsToStruct.py` will spend a lot of time parsing read data and discarding it if it does not overlap any input RNA loci. Loading read data from disk is currently the rate-limiting step. The samtools package [45] has a feature to do this filtering in a faster way (`samtools view -L` command-line invocation) than `readsToStruct.py`.
20. We found that for analysis of mouse nuclear RNA in ref. 5, masking out the first five and last five sites in each known RNA locus (using the configuration setting `algorithm.numEndSitesToIgnore`) was beneficial, especially for RNAs whose mature forms have endogenous 5′ PO_4 and 3′ OH end chemistries that are ligation-competent, such as C/D box snoRNA. This setting excludes the first and last N sites from consideration by adding them to the set of ignored sites, thus excluding counts from native 5′ and 3′ ends of an RNA; only counts from presumed phosphate backbone cleavage are considered. The value of 5 was chosen to provide padding, as transcription data did not perfectly agree with RNA locus annotation boundaries. However, this means that no cutting scores will be produced for the first and last five sites of each RNA. Users are advised to adjust this option to fit their protocol and RNA locus annotations.
21. CIGAR strings are a terse way of describing alignment of two sequences. For an explanation of CIGAR strings, consult the SAM/BAM format specification (https://github.com/samtools/hts-specs). Different alignment tools may use different CIGAR operations to describe the same alignment. Some tools may use `=` or `X` to specify sequence match or mismatch, respectively, instead of the more ambiguous `M` (alignment match). Operation `D` may be used instead of `N` to specify introns in reads. Soft clipping (`S`) operations may be used instead of hard clipping (`H`). These alternative ways are correctly interpreted by `readsToStruct.py`.
22. Whether a mapping end should be considered for FragSeq if the other end is misaligned dependents on the cDNA library preparation protocol and the sequencing technology error modes. For example, because base call quality may decrease at one end of a read, misalignment of the low-quality end is more of a concern than the high-quality end. We recommend users try several strategies for filtering ambiguous mappings and examine the effect on cutting score accuracy on spike-in controls or other RNAs of known structure.

Acknowledgements

We thank David H. Mathews for the invitation to write this chapter. We also thank John St. John, Yann Ponty, and Lukasz J. Kielpinski for helpful discussions.

References

1. Wurst RM, Vournakis JN, Maxam AM (1978) Structure mapping of 5′-32P-labeled RNA with S1 nuclease. Biochemistry 17: 4493–4499. doi: 10.1021/bi00614a021
2. Weeks KM (2010) Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol 20:295–304. doi:10.1016/j.sbi.2010.04.001
3. Mortimer SA, Kidwell MA, Doudna JA (2014) Insights into RNA structure and function from genome-wide studies. Nat Rev Genet 15:469–479. doi:10.1038/nrg3681
4. Kwok CK, Tang Y, Assmann SM, Bevilacqua PC (2015) The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. Trends Biochem Sci 40:221–232. doi:10.1016/j.tibs.2015.02.005
5. Underwood JG, Uzilov AV, Katzman S et al (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods 7:995–1001. doi:10.1038/nmeth.1529
6. Kertesz M, Wan Y, Mazor E et al (2010) Probing RNA structure genome-wide using high throughput sequencing. Protoc Exch. doi:10.1038/nprot.2010.152
7. Li F, Zheng Q, Ryvkin P et al (2012) Global analysis of RNA secondary structure in two metazoans. Cell Rep 1:69–82. doi:10.1016/j.celrep.2011.10.002
8. Sugimoto Y, Vigilante A, Darbo E et al (2015) {hiCLIP} reveals the in vivo atlas of {mRNA} secondary structures recognized by Staufen 1. Nature 519:491–494. doi:10.1038/nature14280
9. Lucks JB, Mortimer SA, Trapnell C et al (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci U S A 108:11063–11068. doi:10.1073/pnas.1106501108
10. Seetin MG, Kladwang W, Bida JP, Das R (2014) Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. In: Waldsich C (ed) Methods Mol Biol. Humana, New York, pp 95–117
11. Spitale RC, Flynn RA, Zhang QC et al (2015) Structural imprints in vivo decode RNA regulatory mechanisms. Nature 519:486–490. doi:10.1038/nature14263
12. Talkish J, May G, Lin Y et al (2014) Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA 20:713–720. doi:10.1261/rna.042218.113
13. Incarnato D, Neri F, Anselmi F, Oliviero S (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. Genome Biol 15:491. doi:10.1186/s13059-014-0491-2
14. Siegfried NA, Busan S, Rice GM et al (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods 11:959–965. doi:10.1038/nmeth.3029
15. Homan PJ, Favorov OV, Lavender CA et al (2014) Single-molecule correlated chemical probing of RNA. Proc Natl Acad Sci U S A 111:13858–13863. doi:10.1073/pnas.1407306111
16. Hector RD, Burlacu E, Aitken S et al (2014) Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. Nucleic Acids Res 42:12138–12154. doi:10.1093/nar/gku815
17. Poulsen LD, Kielpinski LJ, Salama SR et al (2015) SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. RNA 21:1042–1052. doi:10.1261/rna.047068.114
18. Rouskin S, Zubradt M, Washietl S et al (2013) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505:701–705. doi:10.1038/nature12894
19. Ding Y, Tang Y, Kwok CK et al (2013) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505:696–700. doi:10.1038/nature12756
20. Kielpinski LJ, Vinther J (2014) Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. Nucleic Acids Res 42:e70. doi:10.1093/nar/gku167
21. Darty K, Denise A, Ponty Y (2009) VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 25:1974–1975. doi:10.1093/bioinformatics/btp250

22. Kielpinski LJ, Boyd M, Sandelin A, Vinther J (2013) Detection of reverse transcriptase termination sites using cDNA ligation and massive parallel sequencing. In: Shomron N (ed) Methods Mol Biol. Humana, New York, pp 213–231
23. Kielpinski LJ, Sidiropoulos N, Vinther J (2015) Reproducible analysis of sequencing-based RNA structure-probing data with user-friendly tools. Methods Enzymol 558:153–180
24. Cameron V, Uhlenbeck OC (1977) 3′-Phosphatase activity in T4 polynucleotide kinase. Biochemistry 16:5120–5126. doi:10.1021/bi00642a027
25. Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11:129. doi:10.1186/1471-2105-11-129
26. Ehresmann C, Baudin F, Mougel M et al (1987) Probing the structure of RNAs in solution. Nucleic Acids Res 15:9109–9128
27. Gesteland R, Cech T, Atkins J (2005) The RNA World, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
28. Singh R, Reddy R (1989) Gamma-monomethyl phosphate: a cap structure in spliceosomal U6 small nuclear RNA. Proc Natl Acad Sci U S A 86:8280–8283
29. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
30. Lingner J, Keller W (1993) 3′-End labeling of RNA with recombinant yeast poly(A) polymerase. Nucleic Acids Res 21:2917–2920
31. Bruce AG, Uhlenbeck OC (1978) Reactions at the termini of tRNA with T4 RNA ligase. Nucleic Acids Res 5:3665–3677. doi:10.1093/nar/5.10.366
32. Malone C, Brennecke J, Czech B et al (2012) Preparation of small RNA libraries for high-throughput sequencing. Cold Spring Harb Protoc 2012:1067–1077. doi:10.1101/pdb.prot071431
33. Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science 315(5809):241–244. doi:10.1126/science.1132839
34. Li TW, Weeks KM (2006) Structure-independent and quantitative ligation of single-stranded DNA. Anal Biochem 349:242–246. doi:10.1016/j.ab.2005.11.002
35. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res 39:1–12. doi:10.1093/nar/gkr693
36. Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28:3169–3177. doi:10.1093/bioinformatics/bts605
37. Kent WJ, Sugnet CW, Furey TS et al (2002) The human genome browser at UCSC. Genome Res 12:996–1006. doi:10.1101/gr.229102
38. Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A 106:97–102. doi:10.1073/pnas.0806929106
39. Sobczak K, Michlewski G, de Mezer M et al (2010) Trinucleotide repeat system for sequence specificity analysis of RNA structure probing reagents. Anal Biochem 402:40–46. doi:10.1016/j.ab.2010.03.021
40. Lowman HB, Draper DE (1986) On the recognition of helical RNA by cobra venom V1 nuclease. J Biol Chem 261:5396–5403
41. Auron PE, Weber LD, Rich A (1982) Comparison of transfer ribonucleic acid structures using cobra venom and S1 endonucleases. Biochemistry 21:4700–4706. doi:10.1021/bi00262a028
42. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162:156–159. doi:10.1006/abio.1987.9999
43. Schutz K, Hesselberth JR, Fields S (2010) Capture and sequence analysis of RNAs with terminal 2′,3′-cyclic phosphates. RNA 16:621–631. doi:10.1261/rna.1934910
44. Skinner ME, Uzilov AV, Stein LD et al (2009) JBrowse: a next-generation genome browser. Genome Res 19:1630–1638. doi:10.1101/gr.094607.109
45. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352
46. Luehrsen KR, Fox GE (1981) Secondary structure of eukaryotic cytoplasmic 5S ribosomal RNA. Proc Natl Acad Sci U S A 78:2150–2154