1-1-2013

Genome-Wide Analysis of RNA Secondary Structure in Eukaryotes

Fan Li
University of Pennsylvania, fanli.gcb@gmail.com

Genome-Wide Analysis of RNA Secondary Structure in Eukaryotes

Abstract
The secondary structure of an RNA molecule plays an integral role in its maturation, regulation, and function. Over the past decades, myriad studies have revealed specific examples of structural elements that direct the expression and function of both protein-coding messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs). In this work, we develop and apply a novel high-throughput, sequencing-based, structure mapping approach to study RNA secondary structure in three eukaryotic organisms.

First, we assess global patterns of secondary structure across protein-coding transcripts and identify a conserved mark of strongly reduced base pairing at transcription start and stop sites, which we hypothesize helps with ribosome recruitment and function. We also find empirical evidence for reduced base pairing within microRNA (miRNA) target sites, lending further support to the notion that even mRNAs have additional selective pressures outside of their protein coding sequence.

Next, we integrate our structure mapping approaches with transcriptome-wide sequencing of ribosomal RNA-depleted (RNA-seq), small (smRNA-seq), and ribosome-bound (ribo-seq) RNA populations to investigate the impact of RNA secondary structure on gene expression regulation in the model organism Arabidopsis thaliana. We find that secondary structure and mRNA abundance are strongly anti-correlated, which is likely due to the propensity for highly structured transcripts to be degraded and/or processed into smRNAs.

Finally, we develop a likelihood model and Bayesian Markov chain Monte Carlo (MCMC) algorithm that utilizes the sequencing data from our structure mapping approaches to generate single-nucleotide resolution predictions of RNA secondary structure. We show that this likelihood framework resolves ambiguities that arise from the sequencing protocol and leads to significantly increased prediction accuracy.

In total, our findings provide on a global scale both validation of existing hypotheses regarding RNA biology as well as new insights into the regulatory and functional consequences of RNA secondary structure. Furthermore, the development of a statistical approach to structure prediction from sequencing data offers the promise of true genome-wide determination of RNA secondary structure.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Brian D. Gregory

Second Advisor
Li-San Wang

Keywords
Computational biology, Genomics, Inference, Markov chain Monte Carlo, RNA secondary structure, RNA-seq

Subject Categories
Bioinformatics | Cell Biology | Molecular Biology

GENOME-WIDE ANALYSIS OF RNA SECONDARY STRUCTURE IN EUKARYOTES

Fan Li

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

Brian D. Gregory, Ph.D.
Assistant Professor, Biology

Co-Supervisor of Dissertation

_____

Li-San Wang, Ph.D.
Associate Professor, Pathology
and Laboratory Medicine

Graduate Group Chairperson

_____

Maja Bucan, Ph.D.
Professor, Genetics

Dissertation Committee

Harold C. Riethman, Associate Professor,
Molecular and Cellular Oncogenesis Program, The Wistar Institute

Junhyong Kim, Edmund J. and Louis W. Kahn Professor,
Biology, University of Pennsylvania

Kristen W. Lynch, Associate Professor,
Biochemistry and Biophysics, University of Pennsylvania

Arjun Raj, Assistant Professor,
Bioengineering, University of Pennsylvania

Blake C. Meyers, Edward F. and Elizabeth Goodman Rosenberg Professor,
Plant and Soil Sciences, University of Delaware

GENOME-WIDE ANALYSIS OF RNA SECONDARY STRUCTURE IN EUKARYOTES

ACKNOWLEDGEMENTS

I am extremely fortunate in having not one, but two amazingly supportive and inspiring advisors, Brian D. Gregory and Li-San Wang. Their enthusiasm and dedication to training has made the past few years a truly memorable and enriching experience. I would not have matured a tenth as much as a researcher or as a person, were it not for their exceptional guidance and support. I will truly miss the 6am coffee runs, the random not-quite-about-science meetings, and the frantic midnight emails. That last one might be a lie.

I would also like to acknowledge the fantastic camaraderie and lab environment of both the wet and dry spaces. In particular, I am eternally grateful to my fellow graduate students – Paul Ryvkin, Yih-Chii Hwang, Ian Silverman, Nate Berkowitz, and Lee Vandivier – for making Lynch (and Blockley) a second home. Many thanks to Qi Zheng, Matthew Willmann, and Isabelle Dragomir for getting me started on the right track and keeping me there, and a special thanks to Yuk Yee (Fanny) Leung for her advice and time on all matters machine learning and otherwise. To the many other members of the Gregory and Wang labs, I have enjoyed working and not working with all.

My committee – Harold Riethman, Junhyong Kim, Kristen Lynch, Blake C. Meyers, and Arjun Raj – has provided invaluable guidance and creative thought. I am extremely privileged to have learned from each and every one of you. Within the GCB program, I am grateful to the many outstanding faculty that teach and mentor us all. I also want to acknowledge Hannah Chervitz and Tiffany Barlow for their amazing work – our worlds would fall apart in your absence. I have had many wonderful collaborators including Christopher Lengner, Ning Li, Zhengquan Yu, Richard Schultz, Ralph Meyer, Paula Stein, Jun Ma, Ruth Elliott, Susan Weiss, John Rinn, Loyal Goff, Cole Trapnell, Eric Lyons, and Matt Bomhoff.

Many, many thanks to all of the friends I have made here - I hope that we will only continue to grow the bonds we have formed. Finally, I owe everything times infinity plus two to my Wifey. I cannot find the words to say anything other than how did I get so lucky?

ABSTRACT

GENOME-WIDE ANALYSIS OF RNA SECONDARY STRUCTURE IN EUKARYOTES

Fan Li

Brian D. Gregory

Li-San Wang

The secondary structure of an RNA molecule plays an integral role in its maturation, regulation, and function. Over the past decades, myriad studies have revealed specific examples of structural elements that direct the expression and function of both protein-coding messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs). In this work, we develop and apply a novel high-throughput, sequencing-based, structure mapping approach to study RNA secondary structure in three eukaryotic organisms.

First, we assess global patterns of secondary structure across protein-coding transcripts and identify a conserved mark of strongly reduced base pairing at transcription start and stop sites, which we hypothesize helps with ribosome recruitment and function. We also find empirical evidence for reduced base pairing within microRNA (miRNA) target sites, lending further support to the notion that even mRNAs have additional selective pressures outside of their protein coding sequence.

Next, we integrate our structure mapping approaches with transcriptome-wide sequencing of ribosomal RNA-depleted (RNA-seq), small (smRNA-seq), and ribosome-bound (ribo-seq) RNA populations to investigate the impact of RNA secondary structure on gene expression regulation in the model organism *Arabidopsis thaliana*. We find that secondary structure and mRNA abundance are strongly anti-correlated, which is likely due to the propensity for highly structured transcripts to be degraded and/or processed into smRNAs.

Finally, we develop a likelihood model and Bayesian Markov chain Monte Carlo (MCMC) algorithm that utilizes the sequencing data from our structure mapping approaches to generate single-nucleotide resolution predictions of RNA secondary structure. We show that this likelihood

framework resolves ambiguities that arise from the sequencing protocol and leads to significantly increased prediction accuracy.

In total, our findings provide on a global scale both validation of existing hypotheses regarding RNA biology as well as new insights into the regulatory and functional consequences of RNA secondary structure. Furthermore, the development of a statistical approach to structure prediction from sequencing data offers the promise of true genome-wide determination of RNA secondary structure.

TABLE OF CONTENTS

1. **Introduction**
   - **1.1. RNA secondary structure**
     - 1.1.1. Biochemistry of secondary structure
     - 1.1.2. Functional and regulatory roles for RNA secondary structure
   - **1.2. Determination of RNA secondary structure**
     - 1.2.1. Experimental methods
     - 1.2.2. Computational methods
   - **1.3. Outline of dissertation**
2. **A genome-wide method for structure determination**
   - **2.1. Introduction**
     - 2.1.1. Double-stranded RNA sequencing (dsRNA-seq)
     - 2.1.2. Single-stranded RNA sequencing (ssRNA-seq)
     - 2.1.3. From sequencing to structure mapping
     - 2.1.4. Datasets
   - **2.2. Validation of dsRNA- and ssRNA-seq**
     - 2.2.1. dsRNA hotspots and siRNA-mediated heterochromatin formation
     - 2.2.2. Confirmation of dsRNA and ssRNA hotspots by molecular assays
     - 2.2.3. Reproducibility between replicates
   - **2.3. Discussion**
   - **2.4. Materials and methods**
3. **Global patterns of RNA secondary structure**
   - **3.1. Introduction**
   - **3.2. Secondary structure as a marker for protein translation**
   - **3.3. Reduced base pairing at microRNA target sites**
   - **3.4. Models of mRNA secondary structure**
   - **3.5. Discussion**
   - **3.6. Materials and methods**
4. **Regulatory impact of RNA secondary structure**
   - **4.1. Introduction**
   - **4.2. Integration of multiple genomic datasets in *Arabidopsis***
     - 4.2.1. Secondary structure and mRNA abundance
     - 4.2.2. Degradation and smRNA production from structured mRNAs
     - 4.2.3. Direct processing of highly structured mRNA elements
     - 4.2.4. Secondary structure and ribosome binding
   - **4.3. Discussion**

4.4. Materials and methods

5. Sequencing-based prediction of RNA secondary structure

5.1. Introduction

5.2. A Bayesian MCMC framework for dsRNA- and ssRNA-seq

5.2.1. From experimental protocol to likelihood model

5.2.2. Metropolis-Hastings implementation

5.2.3. Generation of simulated sequencing datasets

5.3. Monte Carlo estimation of RNA secondary structure

5.3.1. Simulation results

5.3.2. Structure determination of eight *in vitro* transcribed non-coding RNAs

5.4. Discussion

5.5. Materials and methods

6. Conclusions and future directions

6.1. Summary of results

6.2. Applications to RNA biology

6.2.1. mRNA secondary structure as a regulatory feature

6.2.2. Detection of structural motifs

6.2.3. Long non-coding RNAs

6.3. Improved methods for RNA structure prediction

6.3.1. *In vivo* approaches

6.3.2. Towards genome-wide structure prediction at single base pair resolution

6.4. Concluding remarks

7. References

LIST OF TABLES

1.1 Functional non-coding RNA classes
2.1 Datasets
2.2 Functional classification of dsRNA and ssRNA hotspots
2.3 Concordance at constrained positions between three replicates
2.4 Genome-wide methods for RNA structure determination
2.5 Read processing and alignment
2.6 Histone modification datasets
5.1 Metropolis-Hastings move set
5.2 Selected non-coding RNAs
5.3 Simulation parameters
5.4 Re-estimated digestion parameters from simulated data
5.5 Comparison of RNA-seq-fold and free energy-based methods with simulated data
5.6 Read depth by locus
5.7 Estimated digestion rates from experimental data
5.8 Comparison of RNA-seq-fold and free energy-based methods with *in vitro* data
5.9 Primers used to amplify selected ncRNA loci
5.10 Definitions of sensitivity and specificity for RNA-seq-fold

LIST OF FIGURES

1.1 RNA at the molecular level
1.2 RNA base pairing
1.3 Notation for RNA secondary structure
2.1 dsRNA-seq and ssRNA-seq
2.2 Structure score
2.3 dsRNA hotspots are enriched for heterochromatic modifications
2.4 RT-PCR validation of dsRNA hotspots after RNase treatment
2.5 RT-PCR validation of novel hotspots
2.6 RNA FISH of novel dsRNA hotspots in *C. elegans*
2.7 Genomic distribution of dsRNA-seq and ssRNA-seq reads
2.8 Sliding window analysis of three replicates
2.9 Correlation between transcriptome-wide structure scores
3.1 Structure profile of eukaryotic mRNAs
3.2 Reduced base pairing at microRNA target sites
3.3 Validation of experimentally-derived mRNA structure models
3.4 Workflow of the SAVoR visualization tool
4.1 Biogenesis of small RNAs in plants
4.2 Secondary structure is negatively correlated with mRNA abundance
4.3 qPCR validation of secondary structure-mediated regulation of mRNA levels
4.4 Secondary structure is positively correlated with degradation levels
4.5 Highly structured mRNAs tend to be processed into small RNAs
4.6 Production of antisense smRNAs from highly structured mRNAs
4.7 Increased structure at regions of small RNA production
4.8 Structure is correlated with smRNA production within dsRNA hotspots
4.9 Secondary structure is positively correlated with ribosome binding
4.10 qPCR validation of ribosome binding
5.1 A cartoon representation of MCMC
5.2 Experimental motivation for the likelihood model
5.3 Eight non-coding RNAs with known secondary structures
5.4 Base pairing posteriors from simulated data
5.5 RNAfold predictions
5.6 Convergence analysis with simulated data
5.7 Power analysis with simulated data
5.8 Distribution of read endpoints in experimental versus simulated data
5.9 Base pairing posteriors from *in vitro* data

5.10 Dynamic programming and RNA-seq-fold

Chapter 1

Introduction

The central dogma of molecular biology, as originally stated by Francis Crick(17), placed RNA as an intermediary in the flow of information from genetically-encoded DNA to the functional protein form. The primary job of an RNA molecule, then, was to undergo translation into protein. Over sixty years later, we now know that a veritable alphabet soup of functional RNA species plays a multitude of roles beyond that of protein encoding. In many cases, the function of an RNA molecule is closely linked to both its primary nucleotide sequence as well as its secondary structure.

1.1 RNA secondary structure

An RNA molecule comprises a chain of nucleotides joined together much like beads on a string. Each nucleotide in the chain consists of a ribose sugar, a phosphate group attached to the 5’ carbon, and a base attached to the 1’ carbon. The string then, in our analogy, is a phosphodiester bond between the 5’ phosphate group of one nucleotide and the 3’ hydroxyl of another. As a result, the chain is directional, with the 5’ end representing the nucleotide with a free phosphate group and the 3’ end representing the nucleotide with a free hydroxyl (Figure 1.1).

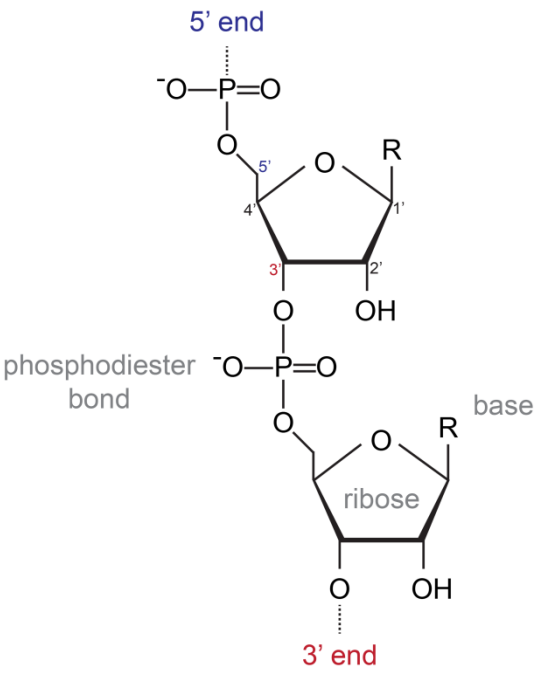


Figure 1.1: The molecular structure of ribonucleic acid (RNA).

RNA can contain four different bases at the 1' position carbon of each nucleotide – adenine (A), guanine (G), cytosine (C), and uracil (U). The ordering of bases within an RNA strand is known as the primary sequence, and specific hydrogen bond interactions between the various bases determine its secondary structure.

1.1.1 Biochemistry of secondary structure

The most common hydrogen bond interactions occur as adenine-uracil (A-U) and guanine-cytosine (G-C) interactions and are known as Watson-Crick base pairs. A third type of interaction (G-U) is also possible, but is less energetically favorable and therefore is referred to as the wobble base pair (Figure 1.2).

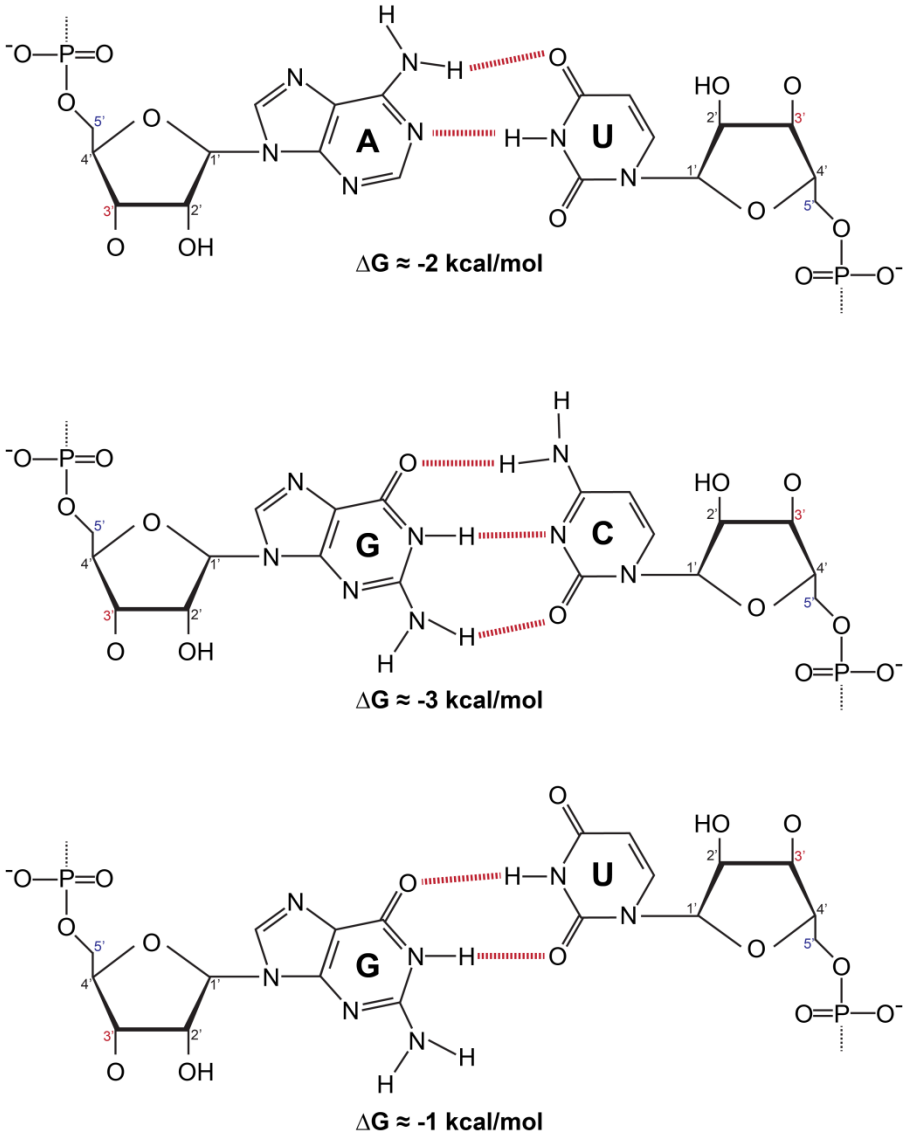


Figure 1.2: Diagram of RNA base pairing interactions. Hydrogen bonds are shown in red. ΔG values are taken from (69).

Taken together, the collection of intramolecular base pairing interactions contained within a single RNA strand is referred to as its secondary structure. Intermolecular interactions between bases on two separate strands of RNA are also common, particularly in the realm of RNA silencing (see Section 1.1.2 below).

Generally, base pairing interactions lower the free energy of an RNA molecule and are therefore preferred over the alternative of unpaired nucleotides. A natural extension of this fact is

that an RNA molecule will tend to adopt a secondary structure that maximizes the number of base paired nucleotides, which leads to complex and often stunning structures such as the tRNA cloverleaf. Other thermodynamic considerations also contribute to the overall secondary structure; for example, paired bases must be separated by at least three nucleotides in adjacent sequence space due to the rigidity of the sugar backbone. Additionally, bases must pair in a nested order such that the interactions do not overlap with one another. Of note, a number of RNAs including telomerase(16) are known to have non-nested base pairing interactions termed pseudoknots; these are essential to their proper function but are typically considered to be tertiary structural elements.

The typical notation used to represent RNA secondary structure consists of a three letter alphabet [“(“, “)”, “.”]. Matching left and right parenthesis represent base paired nucleotides, whereas dots represent unpaired nucleotides. In this way, every valid secondary structure can be uniquely represented by a dot-paren string of the same length (Figure 1.3).

GACUCCGUGGCGCAACGGUAGCGCGUCCGACUCCAGAUCGGAAGGUUGCGUGUUCAAAUCACGUCGGGGUCA

(((((((. . ((((. . . . .)))) . (((((((. . . .) .))))) . . . . . (((((. . . . .))))))))))))) .

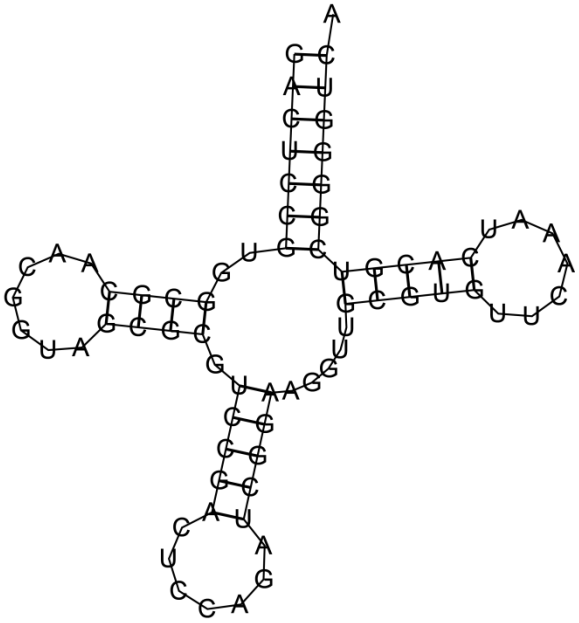


Figure 1.3: Dot-paren (middle) and 2D (bottom) representations of RNA secondary structure. The structure shown here is a tRNA from *Drosophila melanogaster*.

Further projection of the dot-paren string into a 2D structure can be done in a variety of ways, most commonly by a radial algorithm that attempts to minimize overlap between helices(98). The resultant 2D representation shows the structural backbone of the RNA molecule with base paired nucleotides connected by line segments and is the preferred method to visualize non-pseudoknotted structures.

1.1.2 Functional and regulatory roles for RNA secondary structure

The cellular RNA population can be broken down conceptually into two classes: messenger RNAs (mRNAs) that code for proteins, and many types of non-coding RNAs (ncRNAs) that do not. Among the various types of non-coding RNAs (Table 1.1), secondary structure is often crucial to proper biogenesis, maturation, and function.

Class	Functions
Transfer RNA (tRNA)	Adapter between mRNA and protein during translation
Ribosomal RNA (rRNA)	RNA component of the ribosome
Small nuclear RNA (snRNA)	Splicing, alternative polyadenylation
Small nucleolar RNA (snoRNA)	Chemical modification of rRNAs, tRNAs, and other RNAs
MicroRNA (miRNA)	Post-transcriptional gene regulation by target cleavage and/or translational inhibition
Small interfering RNA (siRNA)	Post-transcriptional and epigenetic gene regulation, transposon silencing
Piwi-interacting RNA (piRNA)	Post-transcriptional and epigenetic gene regulation
Long non-coding RNA (lncRNA)	Transcriptional, post-transcriptional, and epigenetic gene regulation

Table 1.1: Functional non-coding RNA classes

Transfer RNAs (tRNAs), as mentioned above, must fold into the canonical cloverleaf structure in order to correctly interact with the ribosome during protein translation(122). The ribosome itself is a large complex of four ribosomal RNAs (rRNAs) and approximately eighty proteins, and also requires the correct folding of the various rRNA subunits in order to assemble and function in protein translation(96). Small nuclear RNAs (snRNAs) contain an evolutionarily conserved core secondary structure that is crucial to their function in splicing as well as alternative polyadenylation(7, 11). Finally, long non-coding RNAs (lncRNAs) likely derive their regulatory functions from secondary structure, not sequence(103, 108, 112).

In the realm of protein-coding mRNAs, structural elements modulate alternative splicing(85, 112) by masking or revealing splice sites. Perhaps the best known example is that of the *Drosophila* Dscam gene, which encodes 38016 distinct transcript isoforms through mutually exclusive alternative splicing of 95 exons. In this case, conserved structural elements in the exon 6 cluster affect inclusion of the various exon variants(71). The secondary structure of mRNAs has also been shown to modulate transcript stability(33), protein translation(36), and microRNA-mediated regulation(61). A significant caveat to many of these findings is that they are derived from computational predictions of secondary structure, which suffer from reliability issues particularly for longer sequences such as mRNAs(26, 70). Thus, one major goal of this work is to provide empirical evidence for the suggested functional and regulatory roles of mRNA secondary structure (see Chapter 3).

Expounding on the topic of regulation, secondary structure is also vital to the entire repertoire of RNA-mediated silencing mechanisms. In plants, these regulatory pathways are mediated by microRNAs (miRNAs) and several classes of endogenous small interfering RNAs (siRNAs)(6, 106). miRNAs are short 21-22 nucleotide (nt) RNAs direct post-transcriptional or translational repression of specific mRNAs through direct base pairing interactions with complementary sites in the target transcript sequence. Furthermore, their biogenesis also involves formation of a hairpin stem-loop structure that is then recognized by Dicer-like (DCL) proteins for processing. Endogenous siRNAs are produced in a similar fashion by DCL-mediated cleavage of long double-stranded RNA (dsRNA). Indeed, the entire life cycle of many a small

RNA from biogenesis to function is keyed upon specific base pairing interactions either with itself (intramolecular) or with another transcript (intermolecular).

1.2 Determination of RNA secondary structure

In the previous section, we described the biochemistry of RNA secondary structure, as well as its functional and regulatory roles. Given the importance of secondary structure in the biogenesis and function of many classes of non-coding RNAs, as well as its myriad effects on mRNA splicing, stability, and translation, an immense deal of effort has been poured into the exact determination of the base pairing interactions that describe a secondary structure. Roughly speaking, the approaches can be distinguished as experimental or computational based on their primary mode of operation. In the next section, we explain the motivations and insights gained from the various studies and highlight a few key approaches in the prediction of RNA secondary structure.

1.2.1 Experimental methods

Experimental methods for studying RNA secondary structure include a host of biochemical (e.g. RNase footprinting, chemical probing) and physical (e.g. X-ray crystallography, nuclear magnetic resonance spectroscopy) approaches. Although the approaches vary widely in terms of mechanism and operation, the end results are strikingly similar: low-throughput, high-quality predictions of secondary structure. In the next few paragraphs, we highlight several of these methods as well as the insights gained from each.

X-ray crystallography

Briefly, X-ray crystallography involves generation of crystals from purified RNA followed by exposure to X-rays. Subsequent analysis of the diffraction patterns yields an electron density map that can be further decomposed into a model of the RNA in question. As experimental approaches go, X-ray crystallography is by far the most labor intensive and time consuming due

to the large number of crystallization trials needed to produce crystals that generates useful diffraction data. However, several key structures including the hammerhead ribozyme and group I self-splicing introns(27) have been determined in this manner.

Nuclear magnetic resonance (NMR) spectroscopy

NMR spectroscopy encompasses many variations of the same principle, namely that different types of nuclei give off different characteristic chemical shift frequencies when exposed to a magnetic field. Depending on the technique, these shift data can be used to study the dynamics of RNA folding in a very sensitive manner(9). It is beyond the scope of the current work to describe specific approaches in detail, and we will suffice to say that NMR spectroscopy is an extremely powerful, low-throughput technique for determination of RNA secondary structure.

Chemical probing

Many chemical reagents modify RNA in some way, and these modifications can be read out as a measure of structural properties such as hydrogen bonding, solvent accessibility, and local nucleotide accessibility(114). One popular method, selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), uses hydroxyl-selective electrophiles such as NMIA and 1M7 that preferentially form 2'-*O*-ester adducts with more flexible nucleotides(116). Sites of 2'-*O*-ester adduct formation can then be detected as stops to primer extension by reverse transcriptase. Analysis of the per-base reactivities in conjunction with free energy-based modeling techniques can then be used to infer secondary structure. The accuracy of SHAPE-based secondary structure predictions is extremely high and compares favorably with the best computational methods currently available (see next section).

SHAPE chemistry has been widely used to study the secondary structure of many RNAs including the entire HIV genome(113). More recently, SHAPE chemistry has been used in conjunction with high-throughput sequencing to simultaneously infer the secondary structure of many pooled RNAs(64). In this sense, the throughput of the SHAPE method can be tremendously

increased, albeit with the significant caveat that a correspondingly large number of custom-designed primers are still required.

RNase footprinting

Another biochemical approach to measure RNA secondary structure utilizes ribonucleases (RNases) that preferentially cleave the sugar backbone at either paired (e.g. RNase V1) or unpaired (e.g RNases ONE, T1, and A) bases. The resultant cleavage sites are then visualized by autoradiography or reverse transcription followed by gel or capillary electrophoresis. The data from enzymatic cleavage and chemical mapping experiments are very similar and a mixture of the two approaches is often used to generate complementary results(108). One key advantage of RNase footprinting is that the technique does not require tiled primers as in SHAPE chemistry, and therefore lends itself very well to the genomic-level analyses that are the subject of this work.

1.2.2 Computational methods

Complementary to the myriad experimental approaches for secondary structure determination is an equally extensive host of computational methods. Indeed, these methods have often evolved in lockstep so as to leverage additional structure mapping or modeling data. Conceptually, computational approaches for structure prediction can be broken down as either free energy-based or comparison-based.

Free energy-based modeling

The base pairing interactions that comprise RNA secondary structure decrease the free energy of an RNA molecule in a well-characterized manner. For example, a G-C base pair decreases the free energy (ΔG) by 3kcal/mol whereas an A-U base pair has a ΔG of -2kcal/mol(129). A seminal paper by Zuker and Stiegler in 1981 utilized dynamic programming to identify the combination of base pairing interactions that would result in the lowest free energy(129). This landmark work has led to a veritable explosion of improvements over the past

decades, both in terms of additions and refinements to the energy parameters used, as well as in the algorithm used to predict secondary structure(70). Current methods such as the Vienna RNAfold package(62), RNAstructure(89), and Sfold(24) include a plethora of features such as loop stability, noncanonical (G-A) base pairing, and partition function-based folding. As alluded to previously, these methods also include direct incorporation of experimental chemical mapping or nuclease cleavage data as a pseudo free-energy term(20).

Jointly, these energy-based prediction methods have become a fundamental tool in the RNA field due to their efficiency, ease of use, and acceptable reliability. However, as a trade-off to their relatively unbounded throughput, these methods suffer from mediocre accuracy particularly when long-range base pairing interactions are involved(26, 70). Additionally, free energy parameters cannot account for *in vivo* factors such as protein binding and folding dynamics that may alter the true secondary structure of an RNA molecule(66, 97).

Comparative methods

The other major class of computational prediction methods, the so-called comparative methods, has attempted to address some of the limitations of the free energy-based single sequence approaches. In principle, comparative methods leverage the tendency for homologous RNAs to form common base pairing interactions in order to produce a consensus secondary structure that likely best represents the entire family of homologous RNAs. Schematically, there are three approaches to comparative analysis. Approach 1, “align then fold”, first attempts to align the input RNA sequences and then infers a consensus structure from the multiple sequence alignment. Approach 2, “fold then align”, ignores primary sequence information and instead attempts to directly align the individually predicted secondary structures. Finally, Approach 3, “simultaneous fold and alignment”, combines classical sequence alignment and dynamic programming-based maximal base pairing.

Regardless of the approach taken, the most useful underlying implementation involves stochastic context free grammars (SCFGs), which can be used to directly represent both the primary sequence and secondary structure of RNAs. SCFG-based methods have proven

immensely useful in constructing large-scale, gold standard RNA structure databases such as Rfam(11, 81).

1.3. Outline of dissertation

On the whole, both experimental and computational approaches to secondary structure prediction have yielded important insights into the functional and regulatory outcomes of RNA secondary structure. However, the classic trade-off between performance and efficiency has limited to applicability of existing methods to true genome-wide studies. In this work, we develop a novel high-throughput, sequencing-based, structure mapping approach to study RNA secondary structure that bridges the gap between limited efficiency experimental methods and limited performance computational methods.

In Chapter 2, we describe our novel assays for RNA secondary structure termed double-stranded RNA sequencing (dsRNA-seq) and single-stranded RNA sequencing (ssRNA-seq). We also provide a meaningful and statistically robust method for transforming sequencing data into base pair resolution structure mapping scores. Finally, we validate the reliability of our method in both biological and molecular contexts.

In Chapter 3, we use the structure mapping data in three eukaryotic organisms to identify structural features that demarcate regions of protein translation and microRNA targeting. We find empirical proof of previous hypotheses of decreased secondary structure near translation start sites and within microRNA target sites. Additionally, we use our structure mapping data to produce genome-wide collections of RNA secondary structure models.

In Chapter 4, we examine the regulatory impact of RNA secondary structure in the transcriptome of the model plant *Arabidopsis thaliana*. By integration of our structure mapping data with transcriptome-wide sequencing of ribosomal RNA-depleted (RNA-seq), small (smRNA-seq), and ribosome-bound (ribo-seq) RNA populations, we find that mRNA secondary structure globally regulates the abundance of these transcripts within the cell. We also show that this

regulatory activity is likely due to the propensity for highly structured mRNAs to be degraded and/or processed into small RNAs.

In Chapter 5, we narrow our focus to the task of RNA secondary structure prediction for a single molecule. Here, we develop a likelihood model that explicitly accounts for the production of sequencing-compatible fragments during the dsRNA-/ssRNA-seq experimental protocols. We develop a Bayesian Markov chain Monte Carlo (MCMC) algorithm termed RNA-seq-fold that uses this underlying likelihood model to reconstruct a secondary structure from the observed sequencing reads. Furthermore, we show that this rigorous statistical treatment of the structure mapping data resolves ambiguities in the experimental protocol and leads to increased prediction accuracy for both simulated and real datasets.

Finally, in Chapter 6, we highlight potential applications of our genome-wide structure assays to RNA biology. We also discuss additional developments that are needed to achieve true base pair resolution secondary structure prediction at a genomic scale.

Chapter 2

A genome-wide method for structure determination

In this section, we describe a novel methodology for genome-wide studies of RNA secondary structure. We outline the statistical methods used to interpret the data generated from these experiments and validate their reliability by biological and molecular modes.

This section references work from:

- Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, et al. Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. 2010 (127)
- Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, et al. Global analysis of RNA secondary structure in two metazoans. Cell Rep. 2012 (54)
- Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. Plant Cell. 2012 (55)

2.1 Introduction

As discussed in the preceding chapter, methods for structure prediction run the gamut from single molecule approaches such as X-ray crystallography to genome-scale approaches such as free energy-based modeling. However, all of these methods are constrained by the performance versus throughput paradigm that has limited the ability to perform true genome-wide studies. To address this gap, we developed a pair of sequencing-based methodologies termed double-stranded RNA sequencing (dsRNA-seq) and single-straned RNA sequencing (ssRNA-seq).

2.1.1 Double-stranded RNA sequencing (dsRNA-seq)

dsRNA-seq marries high-throughput sequencing with classical nuclease chemistry (see Section 1.2.1 above) to generate genome-wide views of RNA secondary structure. In brief,

purified RNA is treated with RNase ONE, which specifically digests single-stranded RNA regions and leaves a population of RNA that is enriched for double-stranded molecules. The resultant fragments are then subjected to standard Illumina library preparation protocols and sequenced (Figure 2.1).

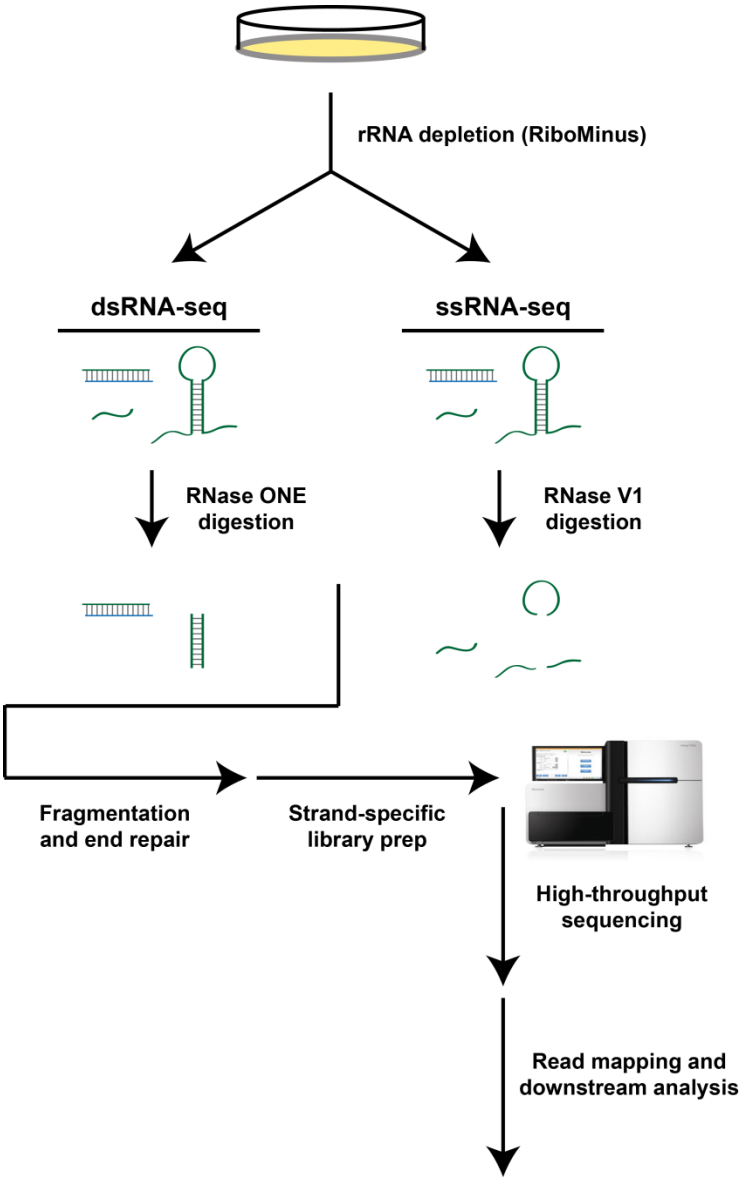


Figure 2.1: Outline of the dsRNA-seq and ssRNA-seq methods.

It is worth noting that dsRNA-seq does not distinguish between intramolecular and intermolecular base pairing interactions, as both provide the same protection against enzymatic cleavage. Additionally, the initial input consists of *in vitro* renatured RNA that may not represent the true *in vivo* species. However, this second caveat is characteristic of most RNA secondary structure assays, and the first may be addressed by lowering the concentration of input RNA or by additional computational procedures.

2.1.2 Single-stranded RNA sequencing (ssRNA-seq)

ssRNA-seq is identical to dsRNA-seq in principle, but utilizes a different enzyme (RNase V1) that specifically targets base paired RNAs. Therefore, the sequenced RNA population consists primarily of unpaired (single-stranded) RNA fragments (Figure 2.1). Taken together, the two protocols provide a complete readout of the base pairing statuses of the entire input RNA pool.

2.1.3 From sequencing to structure mapping

To interpret dsRNA-seq and ssRNA-seq data in a meaningful manner, we define a per-base structure score s_i as the generalized log-ratio (glog) of dsRNA-seq to ssRNA-seq coverage ($\boldsymbol{n_{ds}}, \boldsymbol{n_{ss}}$) after normalization by the total number of mapped reads in each library ($\boldsymbol{N_{ds}}$ and $\boldsymbol{N_{ss}}$):

$$S_i = \mathrm{glog}(ds_i) - \mathrm{glog}(ss_i) = \log_2\left(ds_i + \sqrt{1 + ds_i^2}\right) - \log_2\left(ss_i + \sqrt{1 + ss_i^2}\right)$$

where

$$ds_i = n_{ds} \times \frac{\max(N_{ds}, N_{ss})}{N_{ds}}, \quad ss_i = n_{ss} \times \frac{\max(N_{ds}, N_{ss})}{N_{ss}}$$

Roughly speaking, the structure score represents the likelihood of each base being involved in a pairing interaction. Larger (more positive) values indicate positions that are likely to be base paired, and smaller (more negative) values indicate positions likely to be unpaired (Figure 2.2).

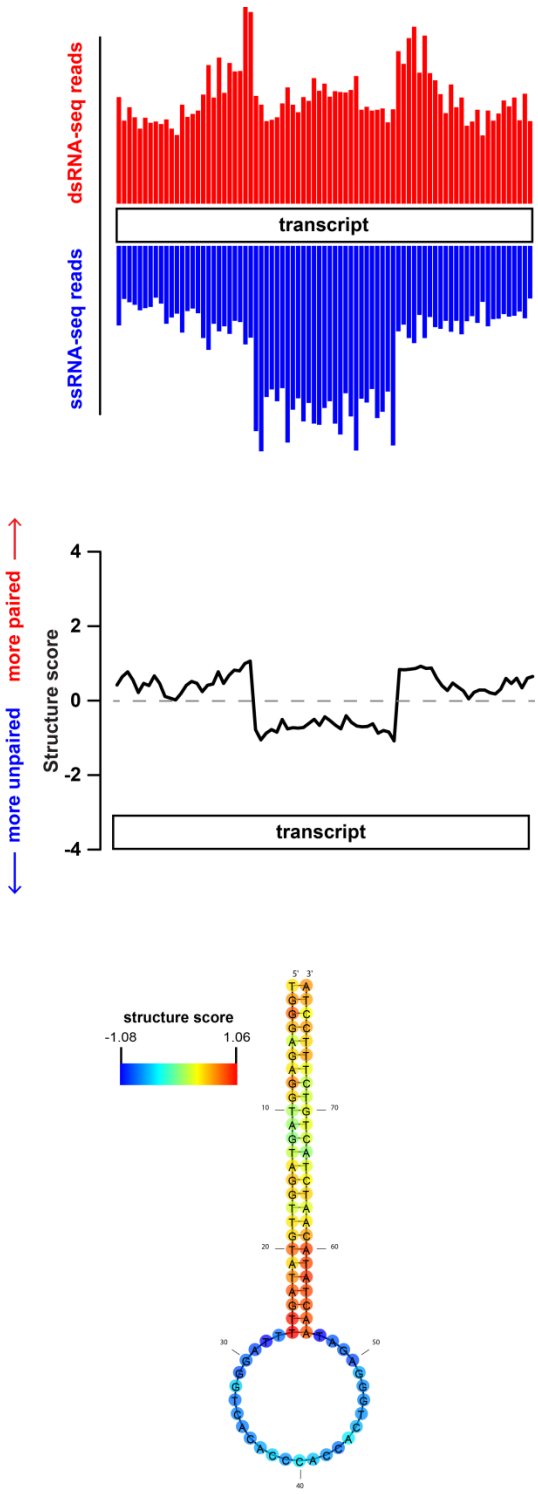


Figure 2.2: Interpretation of dsRNA-seq and ssRNA-seq data. Mapped reads (top panel) are converted into per-base structure scores, which are a normalized log-ratio of dsRNA- to ssRNA-

seq coverage. Higher scores indicate positions that are likely to be base paired, whereas lower scores indicate positions that are likely to be unpaired (bottom panel).

We can also defined a standardized z-score,

$$\boldsymbol{Z_i} = \frac{\boldsymbol{S_i} - \bar{S}}{s^2}$$

where $\bar{S}$ and $s^2$are the mean and standard deviation of scores $\boldsymbol{S_i}$ for a given transcript. These z-scores, in conjunction with permutation-based thresholding, can then be used to constrain certain bases as being either paired or unpaired (see Section 3.6 for details).

2.1.4 Datasets

Throughout the remainder of this work, we will reference datasets generated from four eukaryotic species – *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Homo sapiens* (Table 2.1).

Organism	Source	Sequenced
Arabidopsis thaliana	Columbia (Col-0) bud tissue	dsRNA, ssRNA, smRNA, total RNA, ribosome-associated RNA
Drosophila melanogaster	DL1 culture cells	dsRNA, ssRNA, smRNA
Caenorhabditis elegans	Mixed stage N2 worms	dsRNA, ssRNA, smRNA
Homo sapiens	HeLa culture cells	dsRNA, ssRNA

Table 2.1: Datasets used in this work

Most analyses described in Chapters 2-4 focus on the three model organisms, with *Homo sapiens* data being used primarily in Section 2.2.3. Read processing and alignment are described in detail in Section 2.4.

2.2 Validation of dsRNA-seq and ssRNA-seq

In the model plant *Arabidopsis thaliana*, the RNA-dependent RNA polymerase RDR6 generates double-stranded RNAs from a single-stranded RNA template(57, 118). Our initial dsRNA-seq experiment leveraged this biological process to characterize RDR6 substrates as regions of depleted dsRNA in a mutant *rdr6* plant(127). Agreement with previously known RDR6 substrates as well as RT-PCR assays provided excellent validation for the dsRNA-seq assay. In this section, we describe three additional experiments that as a whole demonstrate the reliability and accuracy of the novel dsRNA-seq and ssRNA-seq methodologies.

2.2.1 dsRNA hotspots and siRNA-mediated heterochromatin formation

In both plants and animals, RNA silencing acts to repress transposons and other parasitic genomic elements by chromatin modification of the endogenous loci(14, 67, 78, 83). Formation of heterochromatin at these target loci is directed by various small RNAs such as small interfering RNAs (siRNAs) and piRNAs that require a double-stranded intermediate for their biogenesis.

We used this aspect of smRNA biogenesis to validate our structure mapping approaches by examining the histone modifications present at highly structured genomic regions (dsRNA hotspots, see Section 2.4 for details). In all three organisms surveyed (*Arabidopsis*, *Drosophila*, and *C. elegans*), we found a significant enrichment for heterochromatic modifications (H3K9me2, H3K9me3, H3K27me1, H3K27me3, and 5mC) within dsRNA hotspots (Figure 2.3).

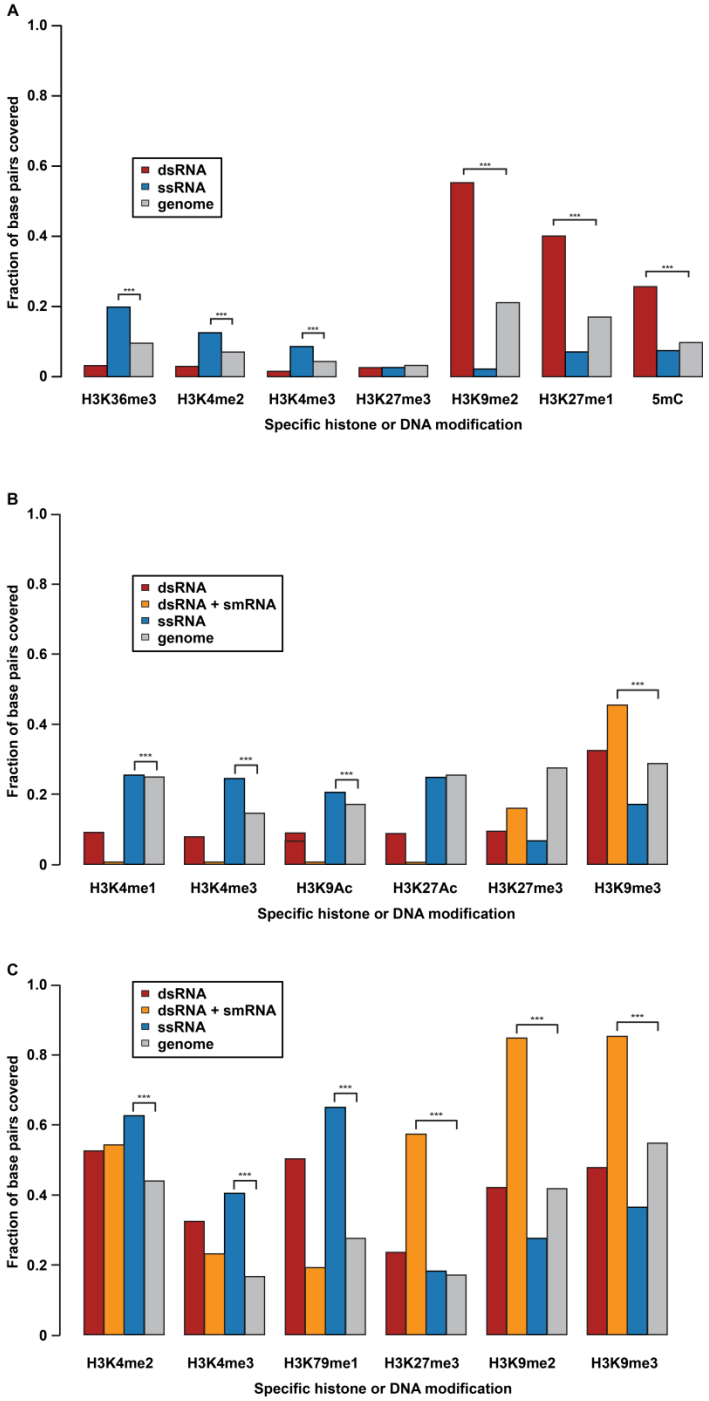


Figure 2.3: (A) Fraction of base pairs within *Arabidopsis thaliana* dsRNA (red) and ssRNA (blue) hotspots as well as the entire genome (gray) that are marked by specific histone modifications as indicated. (B) As in (A), but for *Drosophila melanogaster*. Orange indicates hotspots that also

produce a significant quantity of small RNAs. (C) As in (B), but for *C. elegans*. *** denotes p-value < 2.2e-16, X^2 test.

Further separation of dsRNA hotspots into those that produced small RNAs increased the enrichment of heterochromatic modifications, which is consistent with the known RNA silencing pathways.

2.2.2 Confirmation of dsRNA and ssRNA hotspots by molecular assays

We also wanted to validate our genome-wide protocols at a molecular level. To this end, we randomly selected highly structured and highly unstructured (dsRNA and ssRNA hotspots, respectively) regions identified by our high-throughput assays for RT-PCR follow-up. We repeated the dsRNA-seq and ssRNA-seq experimental protocols on identical input RNA samples and then amplified the regions of interest by RT-PCR. As expected, dsRNA hotspots were exceptionally susceptible to degradation by the double-stranded specific RNase (V1) but not by the single-stranded specific RNase (RNase ONE) (Figure 2.4). The converse was also true, as ssRNA hotspots were sensitive to RNase ONE but not V1.

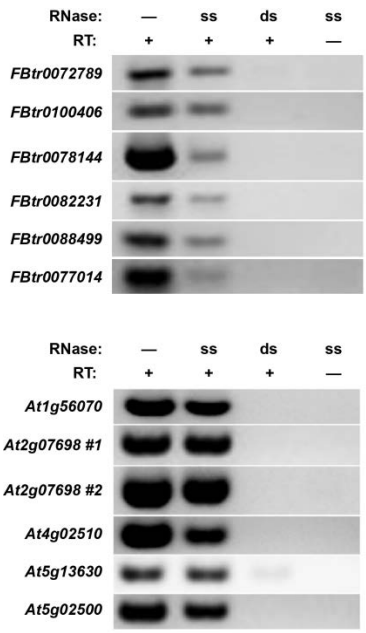


Figure 2.4: RT-PCR validation of dsRNA hotspots following RNase treatment as indicated. Note the lack of amplification following dsRNase treatment (lane 3), but not ssRNase treatment (lane 2). (Top) Six dsRNA hotspots from *Drosophila*, (bottom) six dsRNA hotspots from *Arabidopsis*.

Many of the dsRNA and ssRNA hotspots identified were localized to intergenic space, suggesting that they may represent novel transcription units (Table 2.2).

	Organism					
Class	*Arabidopsis*		*Drosophila*		*C. elegans*	
	dsRNA	ssRNA	dsRNA	ssRNA	dsRNA	ssRNA
Protein-coding	28.7%	92.1%	28.9%	41.7%	74.2%	92.0%
Non-coding RNA	1.4%	0.9%	9.1%	15.5%	2.1%	3.0%
Transposon	48.0%	1.6%	48.3%	28.9%	16.7%	1.0%
Other repeats	5.0%	1.6%	9.6%	5.7%	3.9%	2.4%
Intergenic	16.9%	3.8%	4.1%	8.2%	3.1%	1.6%

Table 2.2: Functional classification of dsRNA and ssRNA hotpots

To confirm our sequencing data and hotspot calling approach, we selected ten of these newly identified transcripts (four in *Drosophila*, six in *C. elegans*) for RT-PCR validation across a panel of tissues and developmental stages (Figure 2.5).

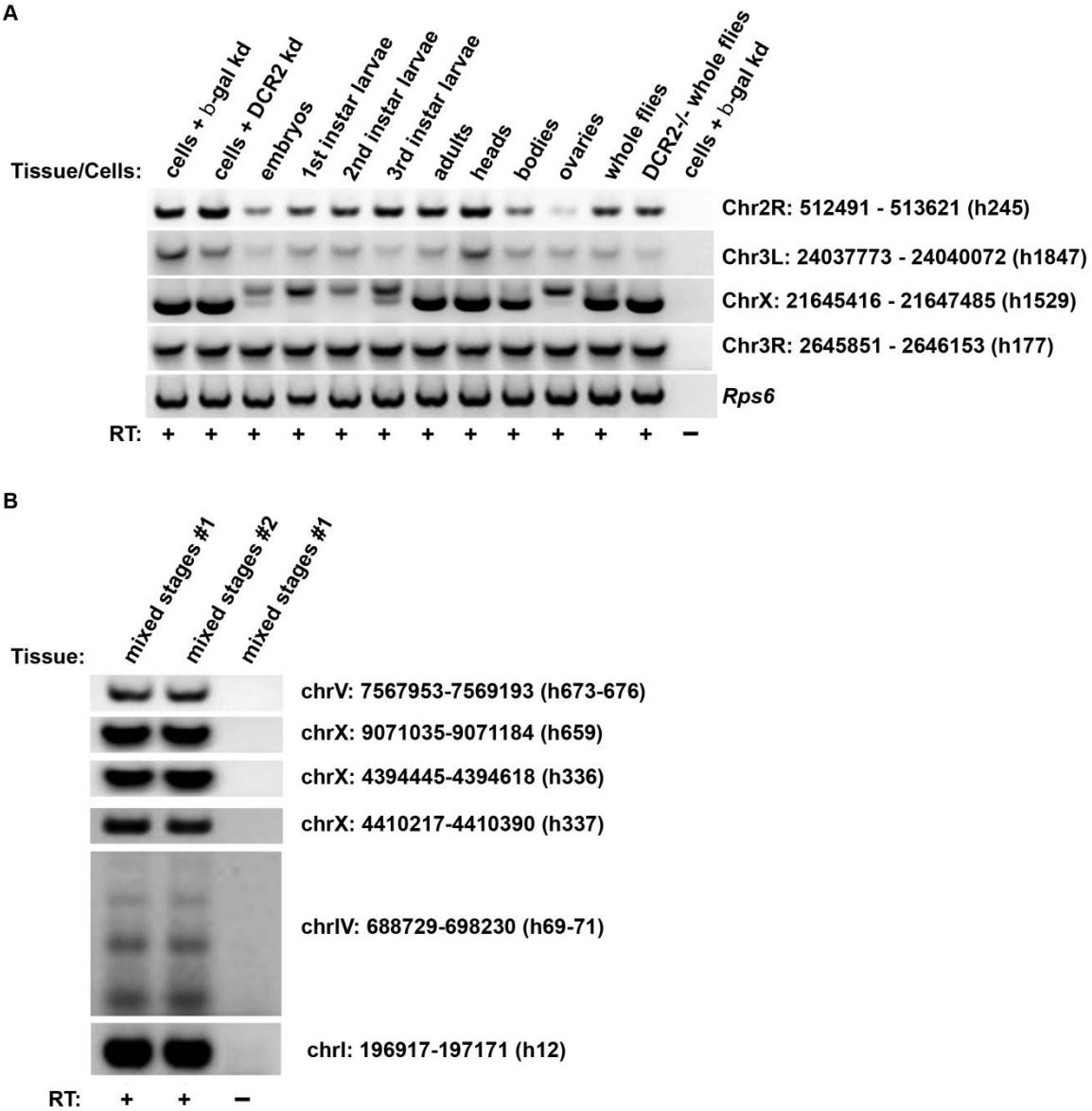
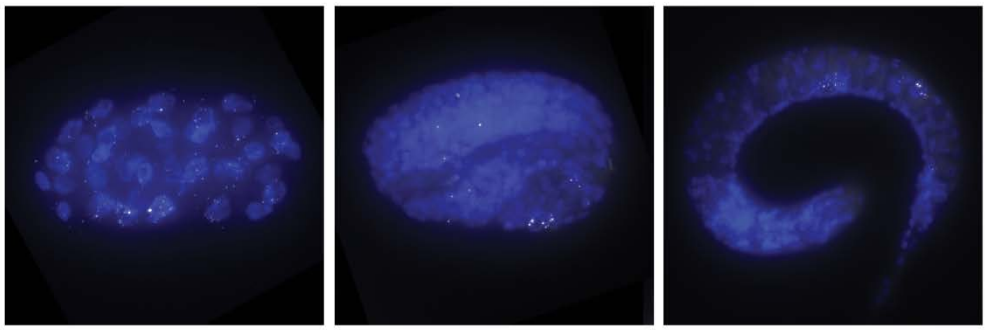


Figure 2.5: RT-PCR validation of novel hotspots in (A) *Drosophila* and (B) *C. elegans*. Tissues/cells and hotspots are as indicated. (RNA in white, nuclei stained with DAPI in blue).

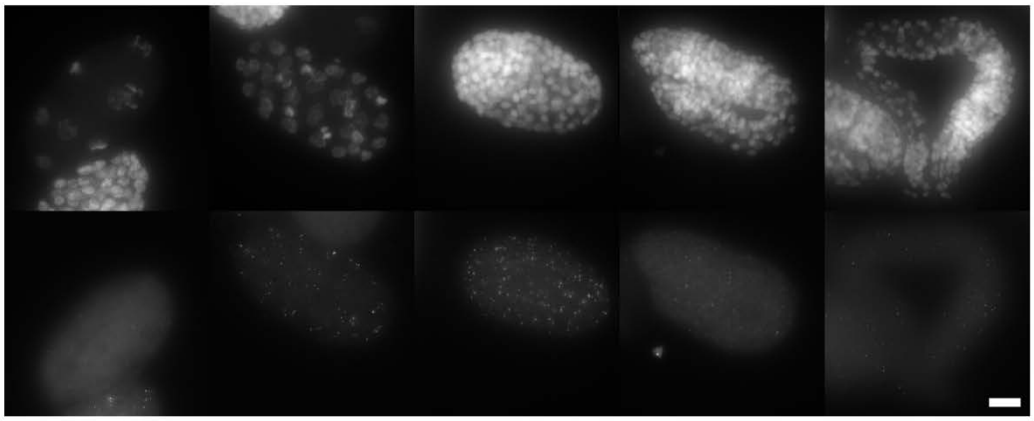
All four of the *Drosophila* hotspots were found in the original culture cell line used for our dsRNA-seq and ssRNA-seq libraries, and three of the four showed dramatic tissue- and developmental stage-specific expression patterns. We also confirmed the expression of six novel highly base paired RNAs, including three that were recently identified by high-throughput sequencing(31), in mixed stage *C. elegans*. Furthermore, we characterized the spatiotemporal expression of three

additional novel dsRNAs in *C. elegans* by single molecule RNA FISH (fluorescence *in situ* hybridization)(87). Use of this technology, which allows direct observation of single RNA molecules, revealed dynamic patterns of expression across development as well as sites of active transcription (Figure 2.6).

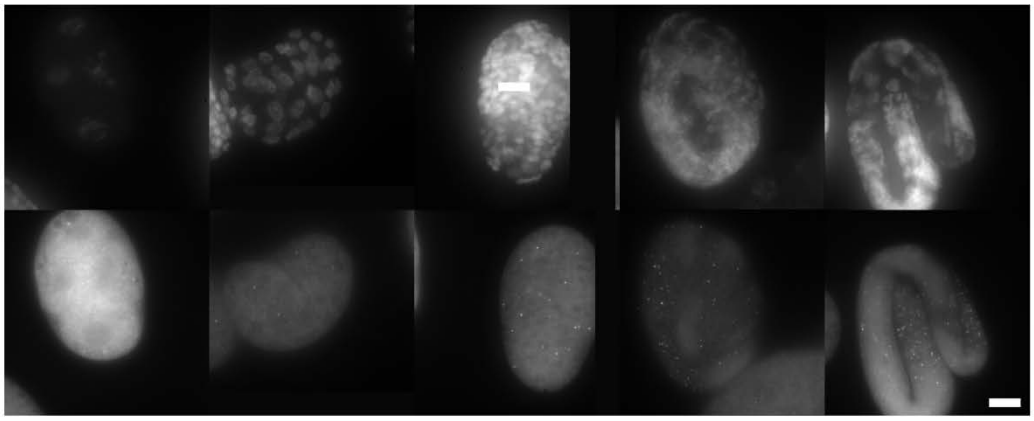
A



B



C



D

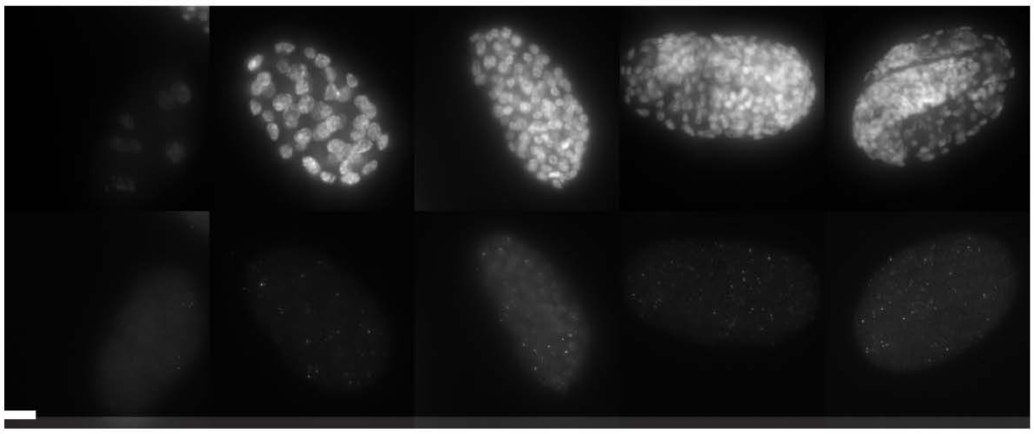


Figure 2.6: RNA FISH of the novel dsRNA hotspots in *C. elegans.* (A) dsRNA hotspot chrIV_h1804-1806 (RNA in white, nuclei stained with DAPI in blue). Images are maximum merges of a series of optical sections at a variety of developmental stages (41-cell stage, left panel; pretzel stage, middle panel; L1, right panel). Scale bars are 5 mm long. (B-D) Additional FISH images of three highly base paired RNAs of *C. elegans* (chrV_h1921 in B, chrV_h2006 in C, and chrI_h719 in D) taken at single molecule resolution at a variety of developmental stages. The top panels show the nuclei (stained with DAPI), whereas the bottom panels show maximum merges of a series of optical sections of the RNA labeled with probes coupled to the TMR fluorophore. Notice that the images contain spots of variable intensity. The dimmer spots most likely represent single dsRNA molecules (based on a comparison of spot intensity to previous acquired data(87), whereas the brighter spots mostly likely arise from the accumulation of multiple dsRNAs. We believe these agglomerations are most likely located at the site of transcription, given that we see at most 1 or two per cell and that they are located within the nucleus. All scale bars are 5 mm long.

In total, these molecular studies confirm the reliability of our dsRNA-seq and ssRNA-seq protocols both in terms of their ability to accurately measure base pairing as well as the potential to discover novel transcription units.

2.2.3 Reproducibility between replicates

Finally, in order to assess the reproducibility of dsRNA-seq and ssRNA-seq, we examined transcriptome-wide studies in three sets of replicate libraries prepared from HeLa cell-extracted RNA (see Section 2.4 for details). Initial examination of these samples revealed extremely similar distributions in terms of their genomic distribution (Figure 2.7), as well as high correlation in read coverage across the genome (Figure 2.8).

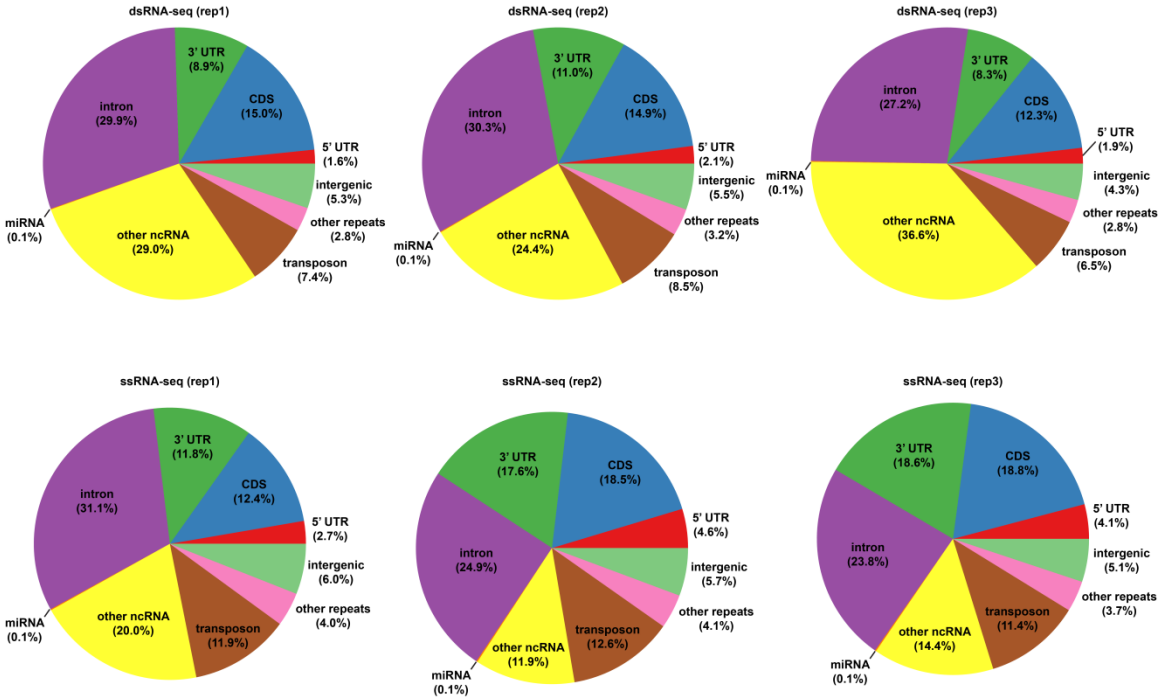


Figure 2.7: Functional classification of dsRNA-seq and ssRNA-seq reads from three HeLa cell replicates.

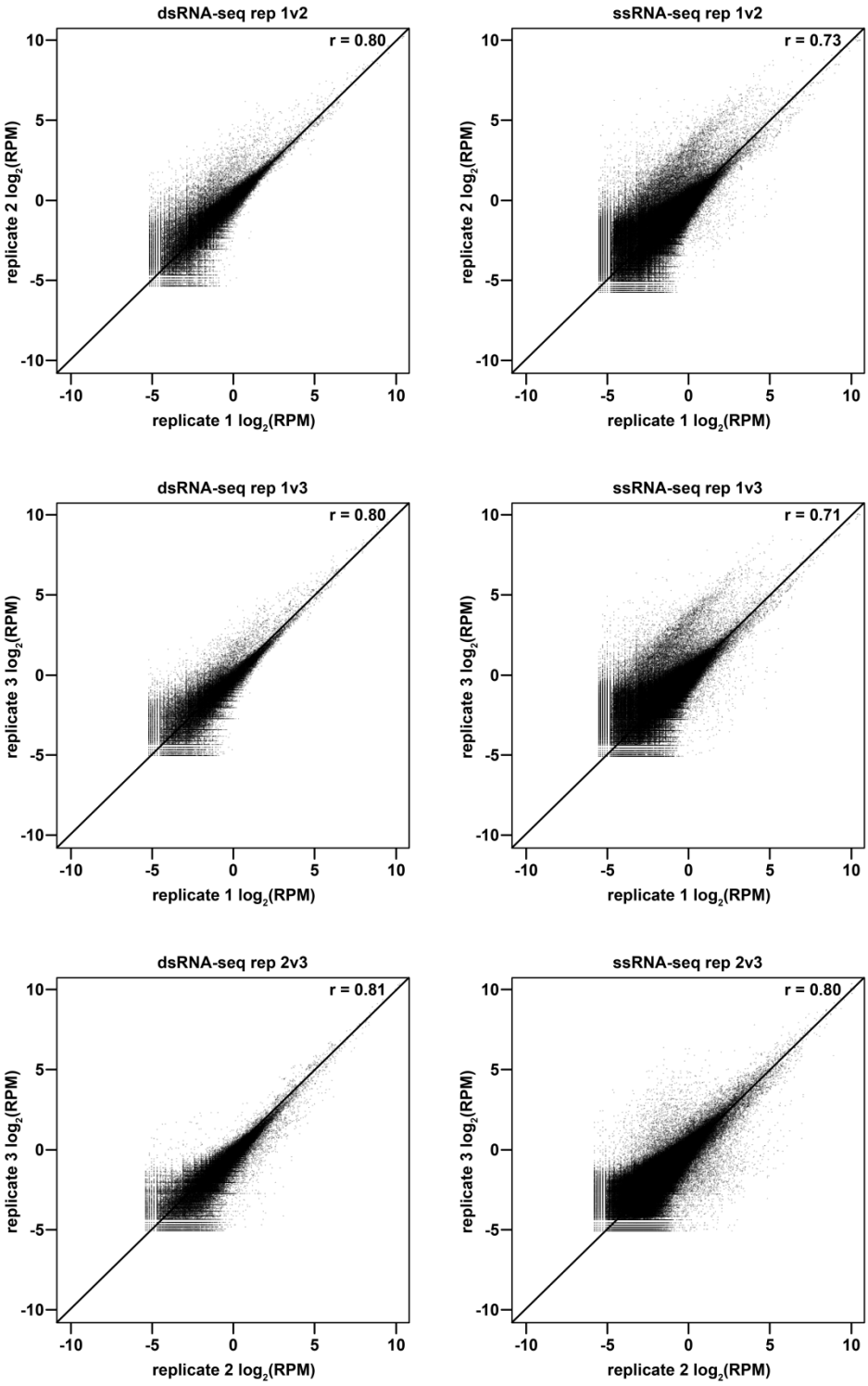


Figure 2.8: Correlation in dsRNA-seq and ssRNA-seq read counts between three HeLa cell replicates. Values are shown in $\log_2$ reads per million mapped (RPM) in 1kb bins across the genome.

To assess reproducibility at single base resolution, we first computed transcriptome-wide structure scores independently for each replicate. We then compared these scores at all informative positions (where the structure score s_i is nonzero) and found a surprisingly low correlation between all pairs of replicates (Figure 2.9, average Pearson correlation r = 0.32).

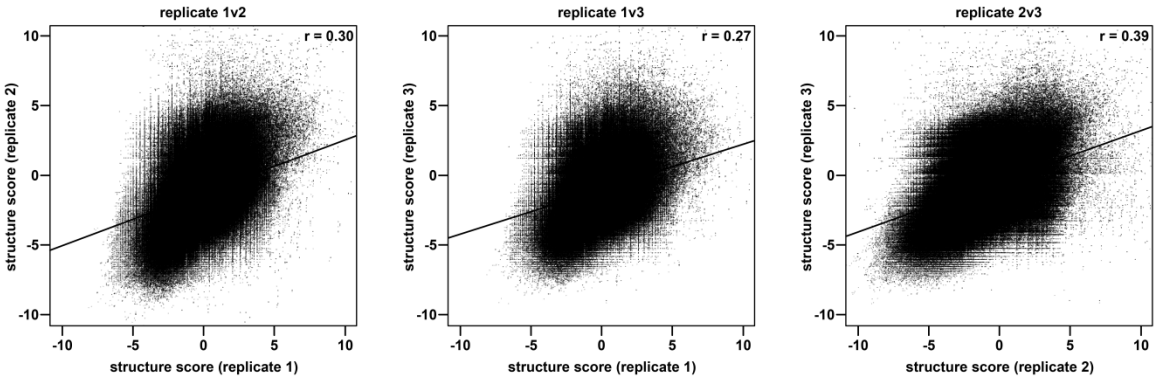


Figure 2.9: Correlation in structure score (log-ratio of dsRNA-seq to ssRNA-seq read depth) at all informative positions between three HeLa cell replicates.

However, when we restricted this analysis to high confidence positions (those with a standardized z-score z_i outside a 95% confidence interval), we found an extremely high level of agreement for constrained positions (Table 2.3**,** 92.4% average concordance).

Comparison	Number of high-confidence positions	Concordant	Discordant
Replicate 1v2	114,656	106,911 (93.2%)	7,745 (6.8%)
Replicate 1v3	80,624	73,809 (91.5%)	6,815 (8.5%)
Replicate 2v3	198,214	183,335 (92.5%)	14,879 (7.5%)

Table 2.3: Concordance at constrained positions between three replicates

These results indicate that the many of the positions with nonzero structure scores are quite noisy and highlight the importance of a stringent statistical treatment in their interpretation.

2.3 Discussion

In this section, we described novel, high-throughput, sequencing-based assays (dsRNA-seq and ssRNA-seq) to determine RNA secondary structure. We validated these protocols in a biological sense by examining their relationship to the siRNA-mediated silencing pathways. We also showed that highly structured and unstructured regions as identified by our sequencing assays are marked by significantly diverging nuclease sensitivities that correspond to their structure. Finally, we found the reproducibility of dsRNA-seq and ssRNA-seq to be extremely consistent between replicates. Taken together, these results suggest that these genome-scale methods can reliably and efficiently interrogate RNA secondary structure on a global scale.

It is worth noting that two other genome-wide methods were developed concurrently by other groups (Table 2.4).

	ds/ssRNA-seq	PARS	FragSeq	SHAPE-Seq
RNA Input	*In vitro* renatured	*In vitro* renatured	*In vitro* renatured	*In vitro* nondenatured
Probe(s)	RNases ONE and V1	RNases S1 And V1	Nuclease P1	NMIA or 1M7
Control	None	None	Untreated and PNK-treated	DMSO-treated
Base pairing readout	Both	Both	Single-stranded only	Single-stranded only
Throughput	Genome-wide	Genome-wide	Genome-wide	Limited by primer extension
Applications	Arabidopsis, Drosophila, and C. elegans whole transcriptomes	polyadenylated mRNAs from yeast	non-coding RNAs from mouse	Synthetic RNA pool

Table 2.4: Genome-wide methods for RNA structure determination

Parallel analysis of RNA structure (PARS)(45) and fragmentation sequencing (FragSeq)(104) are conceptually similar to dsRNA-/ssRNA-seq, but differ in their execution. PARS uses RNases V1 and S1, along with a single-hit kinetics model, to identify cleavage sites within paired or unpaired bases. FragSeq compares cleavage patterns between RNase P1, which cleaves single-stranded bases, and endogenous 5' OH and 5' P controls. Compared to dsRNA-/ssRNA-seq, both PARS and FragSeq are limited in their sensitivity due to the single-hit kinetics model. The two protocols are also limited in the type of RNAs that are interrogated; PARS only measures base pairing within polyadenylated mRNAs and FragSeq primarily measures structure within non-coding RNAs such as snoRNAs. In general, however, the three methods are fairly comparable and seem to agree on certain features of RNA secondary structure (see Section 3.2). In the next two chapters, we shift our focus from dsRNA-seq and ssRNA-seq as tools to the analysis and interpretation of data generated from these assays in three eukaryotic organisms.

2.4 Materials and methods

RNA materials

Arabidopsis thaliana (Columbia (Col-0) ecotype) immature flower bud clusters, *Drosophila melanogaster* DL1 culture cells, *C. elegans* mixed stage N2 worms, and HeLa culture cells were used for all experiments.

Double-stranded RNA sequencing (dsRNA-seq)

40 μg of total RNA (13.33 μg from each of three biological replicates) was subjected to two rounds (1X RiboMinus) of rRNA depletion per manufacturer's instructions (RiboMinus, Invitrogen (Carlsbad, CA)). Next, these rRNA-depleted RNA samples were treated with a single-strand specific ribonuclease per manufacturer's instructions (RNase ONE, Promega (Madison, WI)). dsRNA was then purified using a phenol:chloroform extraction. The purified dsRNA sample was subjected to a fragmentation reaction (Fragmentation Reagents, Applied Biosystems (Foster City, CA)) per manufacturer's instructions. To resolve the dsRNAs after single-stranded RNase

treatment and fragmentation, they were treated with T4 polynucleotide kinase (T4 PNK, New England Biolabs (Cambridge, MA)) as previously described(110). The fragmented RNA sample was then used as the substrate for sequencing library construction using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) per manufacturer's instructions. Sequencing was carried out on an Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA) according to manufacturer's instructions. A detailed experimental protocol follows:

dsRNA-seq protocol:

I. Start with 40 μg of RNA from desired source material, suspended in 40 μL DEPC-treated water.

II. **Ribosomal RNA (rRNA) depletion** using the RiboMinus Eukaryote Kit (manual here: http://tools.lifetechnologies.com/content/sfs/manuals/ribominus_eukaryote_man.pdf). Resuspend rRNA-depleted sample in 18 μL DEPC-treated water.

III. **RNase ONE treatment**

a. Add 2.5 μL RNase ONE Buffer, 2.5 μL 2μg/μL acetylated BSA (e.g. from Promega), and 2.0 μL RNase ONE to 18 μL sample.

b. Incubate at 37°C for one hour.

c. Bring volume up to 200 μL by adding 175 μL DEPC-treated water.

d. Phenol:chloroform extraction (e.g. http://openwetware.org/wiki/Phenol/chloroform_extraction)

e. Precipitate aqueous layer in 20 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 600 μL 100% EtOH.

f. Resuspend in 9 μL DEPC-treated water.

IV. **RNA fragmentation**

a. Add 1 μL Ambion 10X Fragmentation Reagent to 9 μL sample.

b. Incubate at 70°C for 5 minutes.

c. Add 1 μL Stop Solution to the fragmentation reaction.

d. Bring volume up to 100 μL by adding 89 μL DEPC-treated water.

e. Precipitate the fragmented RNA by adding 10 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 300 μL 100% EtOH.

f. Resuspend in 16 μL DEPC-treated water.

V. **T4 PNK treatment**

a. Add 2 μL NEB T4 DNA Ligase buffer, 1 μL T4 PNK, and 1 μL 10mM ATP to 16 μL sample.

b. Incubate at 37°C for one hour.

c. Bring volume up to 100 μL by adding 80 μL DEPC-treated water.

d. Precipitate by adding 10 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 300 μL 100% EtOH.

e. Resuspend in 10 μL DEPC-treated water.

VI. **Size selection**

a. Prepare 1000 mL 1X TBE running buffer (100 mL 10X TBE extended range + 900 mL Milli-Q water).

b. Pre-run 15% TBE-Urea polyacrylamide gel (e.g. from Invitrogen) for 25 minutes at 155 V.

c. While gel is pre-running, prepare ladder and sample:

   i. Ladder: 1.5 μL 10bp DNA ladder, 8.5 μL DEPC-treated water, and 10 μL Gel Loading Buffer (e.g. from NEB).

   ii. Add 10 μL Gel Loading Buffer to sample.

   iii. Place sample (but not ladder) at 70°C for 5 minutes, followed by 3 minutes on ice.

d. After pre-run is complete, run ladder and sample at 155 V for approximately 1.5 hours.

e. Stain gel with ethidium bromide. Add 14 μL 10 mg/mL ethidium bromide to 200 mL 1X TBE buffer in a clean RNase-free tray. Add gel and rock gently for 10 minutes.

f. Cut 20-100bp band from gel and place gel slice in a 0.5mL tube with holes (e.g. Gel Breaker Tubes #3388-100 from IST Engineering Inc.), placed inside a clean 2mL tube.

g. Spin sample at 14000RPM, 4°C for 2 minutes. Repeat until all of the gel goes through the 0.5mL tube.

h. Add 300 μL 0.3M NaCl and rotate for 4 hours.

i. Pipette entire sample into a Spin-X column and spin at 14000RPM, 4°C for 2 minutes. Transfer eluent to new 1.5mL tube.

j. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 900 μL 100% EtOH.

k. Resuspend in 5 μL DEPC-treated water.

VII. **Adapter ligation** (from TruSeq Small RNA Sample Preparation Guide)

a. Add 5 μL sample and 1 μL 5 μM RNA 3' Adapter (RA3) to a sterile, nuclease-free 200 μL PCR tube on ice.

b. Pipette mixture up and down 6-8 times to thoroughly mix and then centrifuge briefly.

c. Incubate in thermal cycle at 70°C for 2 minutes, then at 4°C for 2 minutes.

d. Add 2 μL Ligation Buffer, 1 μL RNase Inhibitor (e.g. RNaseOUT from Life Technologies), and 1 μL Epicentre T4 RNA ligase 2 deletion mutation (200 U/μL). Mix thoroughly.

e. Incubate at 28°C for 75 minutes in thermal cycler.

f. With 5 minutes left, heat 1 μL 25μM 5' Adapter (RA5) to 70°C for 2 minutes, then place on ice for 2 minutes.

g. Add 1 μL RA5, 1 μL 10mM ATP, and 1 μL T4 RNA Ligase 1 to sample tube. Mix thoroughly.

h. Incubate at 28°C for one hour in thermal cycler. Store at -20°C overnight unless proceeding directly to next step.

VIII. **Size selection to reduce adapter adapter**

a. Run sample on 15% TBE-Urea polyacrylamide gel as in Step VI. Cut 70-150bp band, taking care to avoid 50bp adapter-adapter band. Resuspend in 6 μL DEPC-treated water.

IX. **Reverse transcription**

a. Incubate 6 μL sample and 1 μL 100μM RNA RT Primer (RTP) at 70°C for 2 minutes in preheated thermal cycler. Then incubate at 4°C for 2 minutes.

b. Add 2 μL 5X First Strand Buffer, 0.5 μL 12.5mM dNTP mix (12.5mM of each nucleotide), 1 μL 100mM DTT, 1 μL RNase Inhibitor (e.g. RNaseOUT), and 1 μL SuperScript II Reverse Transcriptase. Mix thoroughly.

c. Incubate at 50°C for one hour.

X. **PCR amplification**

a. Prepare PCR master mix: 35 μL 2X Phusion Mix, 21 μL 5mM betaine, 2 μL 10μM RNA PCR Primer (RP1), and 2 μL 10μM RNA PCR Primer Index (RPI*X*).

b. Add 60 μL master mix to 12.5 μL sample, then aliquot mixture to 3 PCR tubes with approximately 25 μL in each tube.

c. PCR amplification program in thermal cycler

   i. 98°C for 30 seconds

   ii. 98°C for 10 seconds

   iii. 60°C for 30 seconds

   iv. 72°C for 15 seconds

   v. Cycle to step ii 11X

   vi. 72°C for 10 minutes

vii. Hold at 4°C

d. Precipitate by adding 10 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 300 μL 100% EtOH.

e. Resuspend in 10 μL DEPC-treated water.

XI. **Size selection**

a. Prepare 1000 mL 1X TBE running buffer (100 mL 10X TBE extended range + 900 mL Milli-Q water).

b. Prepare ladder and sample:

i. Ladder: 1.5 μL 25bp DNA ladder, 8.5 μL DEPC-treated water, and 10 μL Gel Loading Buffer (e.g. from NEB).

ii. Add 10 μL Gel Loading Buffer to sample.

c. Run ladder and sample at 155 V for approximately 30 minutes.

d. Stain gel with ethidium bromide. Add 14 μL 10 mg/mL ethidium bromide to 200 mL 1X TBE buffer in a clean RNase-free tray. Add gel and rock gently for 10 minutes.

e. Cut 138-218bp band from gel and place gel slice in a 0.5mL tube with holes (e.g. Gel Breaker Tubes #3388-100 from IST Engineering Inc.), placed inside a clean 2mL tube. Adapter-adapter is 118bp at this point.

f. Spin sample at 14000RPM, 4°C for 2 minutes. Repeat until all of the gel goes through the 0.5mL tube.

g. Add 300 μL 1X NEB Buffer 2 and rotate for 2 hours.

h. Pipette entire sample into a Spin-X column and spin at 14000RPM, 4°C for 2 minutes. Transfer eluent to new 1.5mL tube.

i. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 900 μL 100% EtOH.

j. Resuspend in 12 μL DEPC-treated water for sequencing.

Single-stranded RNA sequencing (ssRNA-seq)

40 μg of total RNA (13.33 μg from each of three biological replicates) was subjected to two rounds (1X RiboMinus) of rRNA depletion per manufacturer's instructions (RiboMinus, Invitrogen (Carlsbad, CA)). Next, these rRNA-depleted RNA samples were treated with a double-strand specific ribonuclease per manufacturer's instructions (RNase V1, Applied Biosystems (Foster City, CA)). ssRNA was then purified using a phenol:chloroform extraction. The purified ssRNA sample was subjected to a fragmentation reaction (Fragmentation Reagents, Applied Biosystems (Foster City, CA)) per manufacturer's instructions. To resolve the ssRNAs after double-stranded RNase treatment and fragmentation, they were treated with T4 polynucleotide kinase (T4 PNK, New England Biolabs (Cambridge, MA)) as previously described(110). The fragmented RNA sample was then used as the substrate for sequencing library construction using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) per manufacturer's instructions. Sequencing was carried out on an Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA) according to manufacturer's instructions. The experimental protocol for ssRNA-seq is identical to dsRNA-seq except for the following step, which replaces the RNase ONE treatment (Step III) above.

III. **RNase V1 treatment**

a. Add 3 μL 10X RNA Structure Buffer and 5 μL RNase V1 to 22 μL sample.
b. Incubate at 37°C for one hour.
c. Bring volume up to 200 μL by adding 170 μL DEPC-treated water.
d. Phenol:chloroform extraction (e.g. http://openwetware.org/wiki/Phenol/chloroform_extraction)
e. Precipitate aqueous layer in 20 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 600 μL 100% EtOH.
f. Resuspend in 9 μL DEPC-treated water.

Read processing and alignment

Details of read processing and alignment are provided in Table 2.5.

Organism	Library name	GEO accession	Platform	Adapter trimming	Mapping	Mapped reads
Ath	dsRNA-seq	GSE23439	Illumina GA2	cross_match, min 6 nt	cross_match, ≤ 8% mismatches	10,441,682
Ath	ssRNA-seq	GSE40209	Illumina HiSeq 2000	cutadapt, min 6nt	Bowtie, ≤ 4% seed-mismatches and ≤ 6% total-mismatches, up-to 100 hits per read	19,536,080
Ath	smRNA-seq	GSE28524	Illumina GA2	VectStrip, min 6 nt	cross_match, ≤ 8% mismatches	9,214,751
Ath	RNA-seq	GSE28524	ABI SOLiD	None	cross_match, ≤ 8% mismatches	21,067,985 (rep 1); 26,548,982 (rep2)
Ath	Ribo-seq	GSE40209	Illumina HiSeq 2000	cutadapt, min 8 nt	Bowtie, ≤ 6% seed-mismatches and <= 6% total-mismatches, up-to 100 hits per read	28,388,928
Ath	Degradome (GMUCT)	GSE11070	Illumina GA2	cutadapt, min 8 nt	Bowtie, ≤ 6% seed-mismatches and ≤ 6% total-mismatches, up-to 100 hits per read	Public dataset
Dme	dsRNA-seq	GSE29571	Illumina GA2 and HiSeq 2000	cross_match, min 6 nt	cross_match, ≤ 6% mismatches	86,920,519
Dme	ssRNA-seq	GSE29571	Illumina GA2 and HiSeq 2000	cross_match, min 6 nt	cross_match, ≤ 6% mismatches	20,330,923
Dme	smRNA-seq	GSE29571	Illumina GA2	cross_match, min 6 nt	cross_match, ≤ 6% mismatches	4,207,161
Cel	dsRNA-seq	GSE29571	Illumina GA2 and HiSeq 2000	cross_match, min 6 nt	cross_match, ≤ 6% mismatches	52,662,711
Cel	ssRNA-seq	GSE29571	Illumina GA2 and HiSeq	cross_match, min 6 nt	cross_match, ≤ 6%	13,177,958

			2000		mismatches	
Cel	smRNA-seq	GSE29571	Illumina GA2	cross_match, min 6 nt	cross_match, ≤ 6% mismatches	4,190,517
HeLa	dsRNA-seq	GSE49309	Illumina HiSeq 2000	cutadapt, min 6 nt	TopHat, ≤ 2 mismatches	72,498,559
HeLa	ssRNA-seq	GSE49309	Illumina HiSeq 2000	cutadapt, min 6 nt	TopHat, ≤ 2 mismatches	53,475,807

Table 2.5: Read processing and alignment

Identification of dsRNA and ssRNA hotspots

dsRNA and ssRNA hotspots were identified using a modified version of the CSAR software package(79). Specifically, structure scores were calculated for each base position in the genome and regions with significantly higher or lower than background scores at an FDR of 5% were called as dsRNA and ssRNA hotspots, respectively. Recall that higher (more positive) structure scores indicate a greater probability of being paired, whereas lower (more negative) structure scores indicate a greater probability of being unpaired. The background distribution for determining the FDR was calculated by randomly shuffling dsRNA-seq and ssRNA-seq reads and then identifying hotspots with these shuffled data.

Histone modification datasets

Various histone modification ChIP-seq and ChIP-chip data were downloaded from modENCODE (http://www.modencode.org) and other sources (Table 2.6).

Organism	Experiment type	Modification	Source
Arabidopsis thaliana	ChIP-chip	H3K9me2	(8)
Arabidopsis thaliana	ChIP-seq	H3K27me1, H3K27me3	(43)
Arabidopsis thaliana	ChIP-chip	H3K4me2, H3K4me3, H3K36me2, 5mC	(94)
Drosophila melanogaster	ChIP-seq	H3K4me1, H3K4me3, H3K9me3, H3K9ac,	(46)

		H3K27me3, H3K27ac	
Caenorhabditis elegans	ChIP-chip	H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H3K79me1	(31)

Table 2.6: Histone modification datasets

For ChIP-seq data, genomic intervals of enriched regions were directly compared to dsRNA and ssRNA hotspots. For ChIP-chip data, ChIPOTle v1.11(10) was first used to identify genomic intervals of enriched histone modifications. Genomic intervals of significantly enriched histone modifications were then overlapped with the locations of dsRNA and ssRNA hotspots.

RT-PCR analyses

RNase ONE digestion (dsRNA selection) was performed on three 20 μg total RNA samples per manufacturer's instructions. Following digestion, these three samples were pooled together and purified using a phenol:chloroform extraction. To obtain ssRNA, a dsRNase digestion (RNase V1, (Ambion, Foster City, CA)) was carried on three 20 μg total RNA samples per manufacturer's instructions. Following digestion, these three samples were pooled together and purified using a phenol:chloroform extraction. Each experiment was replicated three times.

Fluorescence in situ hybridization (FISH)

In preparation for FISH experiments, we harvested embryos and larvae from synchronized and unsynchronized cultures of N2 worms. We fixed, permeabilized, and performed single molecule FISH on *C. elegans* embryos and larvae as previously described(87). We determined the concentration of probe empirically, ending up with roughly the same concentration per fluorescently labeled oligonucleotide as used previously(87).

Reproducibility studies

Three independent sets of dsRNA-seq and ssRNA-seq libraries were prepared as described above except without rRNA depletion. Instead, duplex-specific normalization (DSN) was performed after T4 PNK treatment but prior to library construction.

Chapter 3

Global patterns of RNA secondary structure

We now proceed to explore the genomic landscape of RNA secondary structure in three eukaryotic organisms. We show empirical support for previously hypothesized roles for secondary structure and also highlight new structural features that are revealed by genome-wide analyses.

This section references work from:

- Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, et al. Global analysis of RNA secondary structure in two metazoans. Cell Rep. 2012 (54)
- Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. Plant Cell. 2012 (55)
- Li F, Ryvkin P, Childress DM, Valladares O, Gregory BD, Wang LS. SAVoR: a server for sequencing annotation and visualization of RNA structures. Nucleic Acids Res. 2012;40(Web Server issue):W59-64 (53)

3.1 Introduction

Many roles have been described for RNA secondary structure. For non-coding RNAs, their biogenesis and function often depend on their secondary structure (see Section 1.1.2). Protein-coding mRNAs also contain many structural features that modulate their stability, splicing, translation, and localization. Well-known moieties such as the AU-rich element(28), iron response element(47), and terminal 3' stem-loop of histone mRNAs(105) exist as structured hairpins. More generally, secondary structure is thought to be relaxed near the translation initiation site so as to allow easier ribosome binding. Free energy-based prediction of secondary structures showed a near-universal decrease in mRNA stability (corresponding to less secondary structure) near the translation initiation site(36). Experimental proof of these findings would address concerns about whether this is a real phenomenon or simply a byproduct of the sequence bias (e.g. Shine-Delgarno, Kozak sequences) present at these sites.

Secondary structure is also thought to affect microRNA-mediated regulatory pathways. In order for a miRNA to carry out its regulatory role, it must form base pairing interactions with a complementary sequence on the mRNA transcript. The miRNA-target interaction is thought to extend along the entire length of plant miRNAs. However, in animals, this interaction mostly involves complementary base pairing only between nucleotides 2 – 8 of a miRNA (counted from its 5' end) (seed region) and a binding site in a target transcript. In both cases, the intramolecular base pairing interactions contained within the target site must first be disrupted to allow for binding of the miRNA. The notion that perhaps miRNA target sites have evolved to be less structured so as to reduce the "cost" of miRNA-mediated regulation is quite intriguing and has been studied extensively in recent years.

The first study to incorporate target site structure in miRNA target prediction found 3-nucleotide accessible regions to be an important predictor of targeting efficiency in *Drosophila melanogaster*(91). This observation was then extended to a more general trend of decreased structural complexity and increased accessibility in regions containing miRNA target sites(126). Additional studies based on free energy-based modeling of ensemble structures highlighted the importance of target site and flanking region accessibility in miRNA targeting efficiency(44, 61). A recent genome-wide analysis of miRNA target site folding energies in four plant genomes revealed significantly higher site accessibility when compared with random sequences in genes rich in guanines and cytosines (GC-rich), but no such difference in GC-poor genes(35). However, as with the observations regarding structural complexity at translation initiation sites, these are all based on computationally predicted structures and true experimental proof is lacking.

In this section, we apply dsRNA-seq and ssRNA-seq to obtain a global view of RNA secondary structure in the three eukaryotes *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. Using these new genome-wide methodologies, we show secondary structure is greatly reduced upstream of translation initiation sites and within miRNA target sites. We also highlight distinct structural patterns that mark regions of protein translation. Finally, we provide a collection of structural models based on a combination of free energy-based modeling and our experimental data.

3.2 Secondary structure as a marker for protein translation

To identify structural features within protein coding mRNAs, we examined the average structure score (see Section 2.1.3) across the CDS and both 5' and 3' UTRs of all detected mRNA transcripts. In all three organisms, we found significant decreases (p -> 0) in structure score at both the start and stop codons of the CDS, revealing increased mRNA accessibility at the regions where protein translation begins and ends (Figure 3.1).

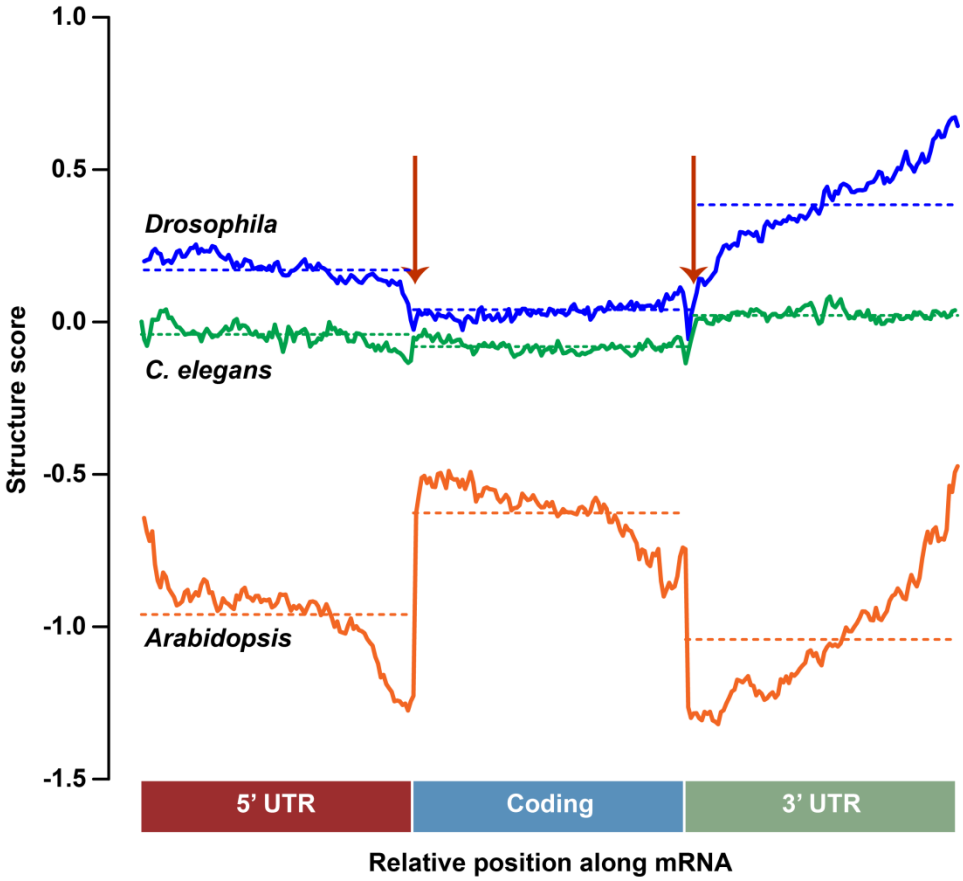


Figure 3.1: The average structure score plotted over the 5' UTR, CDS, and 3' UTR of all detectable protein-coding transcripts for *Arabidopsis* (orange), *Drosophila* (blue), and *C. elegans* (green). The overall average for each specific transcript region is shown as a dotted line. Red arrows highlight significant (p-value < 2.2e-16, t test) dips in secondary structure that occur at the junctions between the UTRs and the coding region.

A similar study of yeast mRNAs by the PARS method revealed the same trend(45); these findings, in conjunction with the computational predictions of Gu *et al*.(36), strongly suggest that structural demarcation of protein translation is a conserved feature of eukaryotic protein-coding transcripts.

Our analyses also revealed strong differences in secondary structure between the protein-coding CDS and untranslated regions (UTRs). In *Arabidopsis*, the UTRs tend to be less structured than the CDS, whereas in both animals, the UTRs tend to be more highly structured (Figure 3.1). Intriguingly, these differences may reflect the prevalence and complexity of RNA-binding protein (RBP) mediated regulation in these organisms. Animals are thought to encode a much larger repertoire of RBPs than plants(15, 63, 99), and the fact that these proteins often bind structured elements in the 3' UTR provides a possible explanation for the increased secondary structure observed within *Drosophila* and *C. elegans* UTRs.

3.3 Reduced base pairing at microRNA target sites

The global nature of the data generated by dsRNA-seq and ssRNA-seq allowed us to also interrogate the average secondary structure observed at and flanking microRNA target sites. In both *Arabidopsis* and *C. elegans*, we observed significantly ($p \rightarrow 0$ for *Arabidopsis*, $p = 2.7e\text{-}13$ for *C. elegans*) lower structure scores within predicted target sites compared to the flanking sequences 50bp up- and downstream (Figure 3.2).

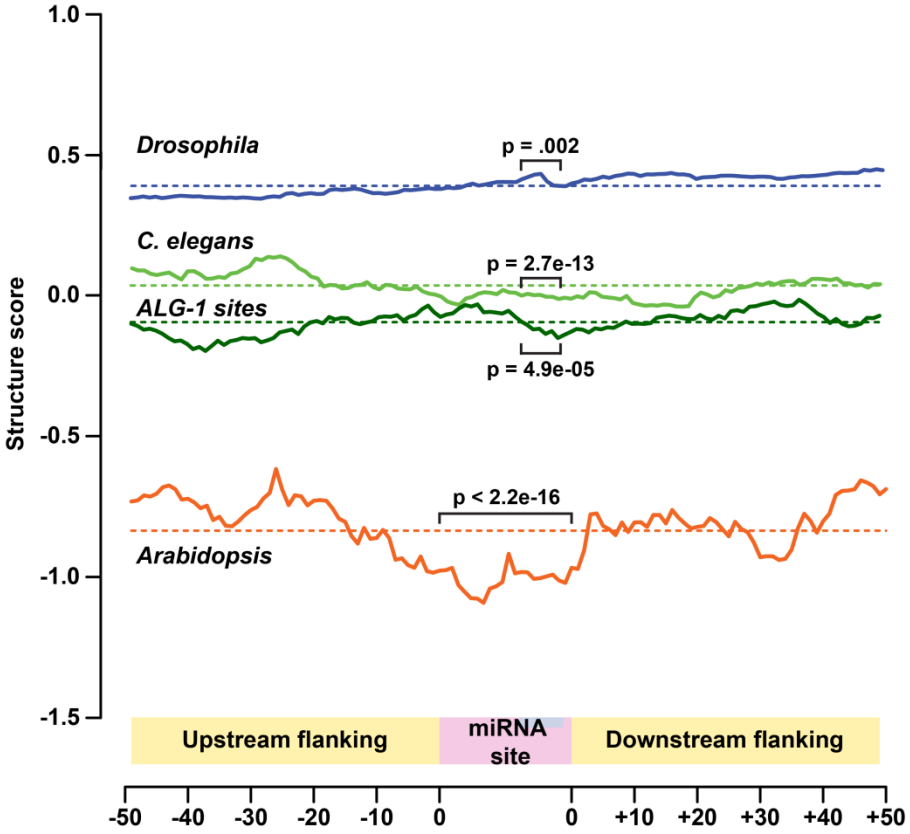


Figure 3.2: The average structure score across miRNA binding sites and for 50 bp up- and downstream flanking regions in *Arabidopsis* (orange), *Drosophila* (blue), *C. elegans* (green) target transcripts. *C. elegans* miRNA sites that are additionally bound by ALG-1 are shown in dark green. The overall structure score average for the entire ~121-bp region is shown as a dotted line. p-values were calculated by a t test.

Further analysis confined to target sites experimentally determined to be bound by ALG-1 (the ARGONAUTE (AGO) protein at the core of *C. elegans* miRISC)(128) uncovered similarly decreased base pairing within the 3' end of microRNA target sites. Notably, the structure score profile of *C. elegans* target sites appears to fit the animal model of seed pairing; that is, the decreased base pairing was largely confined to bases 2-8 of the microRNA corresponding to the 3' end of the target site. To the best of our knowledge, this is the first experimental evidence for decreased base pairing as a selective pressure within microRNA target sites, and again highlights the importance of RNA secondary structure on a genome-wide level. Interestingly, we did not

observe a decrease in secondary structure at predicted microRNA target sites in *Drosophila* (Figure 3.2), indicating that large-scale differences in microRNA targeting may be present within eukaryotes.

3.4 Models of mRNA secondary structure

Very little is known about the secondary structure of full-length mRNAs. Structured regulatory moieties such as iron response elements and AU-rich elements (see Section 3.1) have been identified by a variety of biochemical methods, but these comprise only a small fraction of the total length of mRNA sequence. Computationally predicted mRNA secondary structures are available, but they suffer from limited accuracy(26). General trends of secondary structure, such as the decreased base pairing at sites of protein translation and microRNA-mediated regulation described in the previous sections, have been identified and now validated, but these only detail the propensity of individual nucleotides to be involved in a base pairing interaction and do not capture the specific interaction itself. In other words, we can assert that a particular nucleotide is likely to be base paired, but with what other nucleotide we cannot say. Often, however, it is vitally important to know exactly the pairs of bases that comprise the secondary structure of an RNA molecule, for example in the context of mutational or comparative analyses.

To this end, we developed a method that integrates experimental data from dsRNA-seq and ssRNA-seq with free energy-based modeling to produce accurate, single-nucleotide resolution models of RNA secondary structure. In short, our method identifies nucleotides that are likely to be involved in a base pairing interaction based on a null distribution of randomly sampled sequencing reads, constrains these positions to preferentially exist in a base paired configuration, and then uses RNAfold to determine the exact pairs of interacting bases. We used this approach to generate a comprehensive collection of mRNA structure models for *Arabidopsis*, *Drosophila*, and *C. elegans*. Strikingly, experimentally-derived structure models for the *FBtr0100406* (and other) mRNAs revealed significant differences from free energy-based folding, particular with respect to the large number of ≥7nt loops present in the RNAfold model (Figure 3.3).

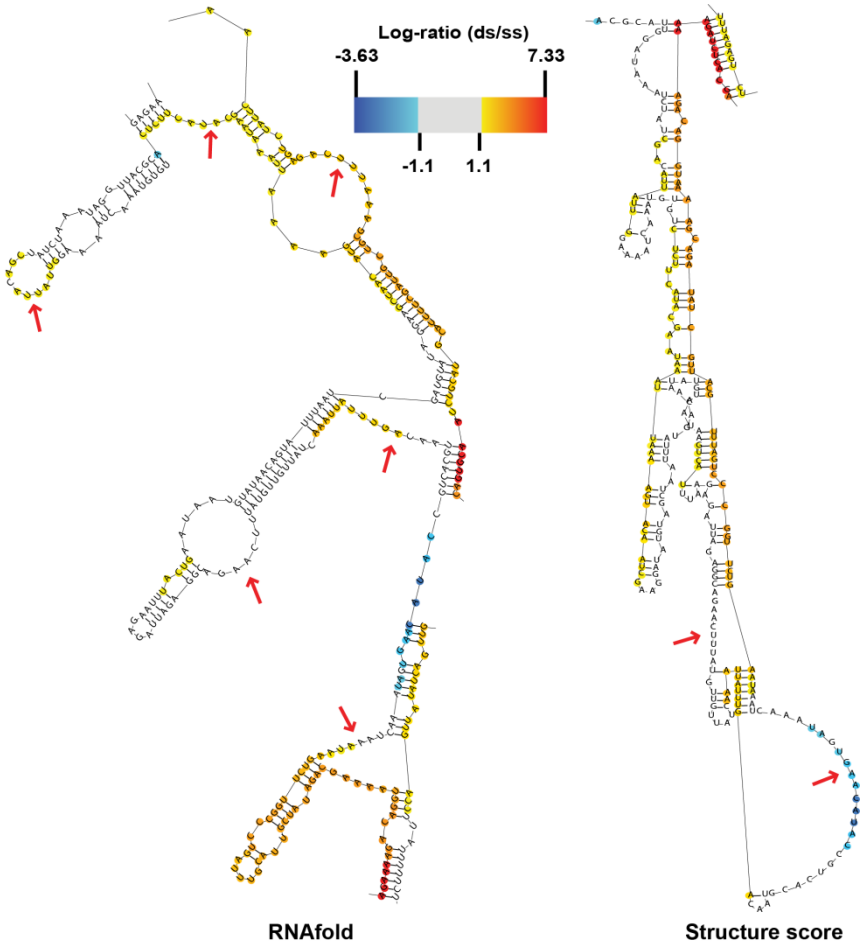


Figure 3.3: Model of secondary structure for the Drosophila FBtr0100406 transcript determined by default RNAfold (left, labeled RNAfold) or our high-throughput sequencing-based, structure mapping approach (right, labeled Structure score). The region of this RNA interrogated in Figure 2.4 is shown in this figure. The heatscale indicates the normalized log-ratio of dsRNA-seq to ssRNA-seq reads at each base position. Red arrows indicate regions of the RNA model where ~7 nt are unpaired.

RT-PCR analysis of this mRNA region showed relatively low sensitivity to ssRNase (see Section 2.2), which is not likely if the many loops predicted by free energy alone were actually present. These results and others suggest that our “constrained” models more accurately reflect the true secondary structure of transcripts in the cell.

During the process of deriving these structure models, we found that existing tools for RNA structure layout (e.g. RNAplot) lacked an effective means to visualize additional information such as our structure scores. To address this gap, we developed the Sequencing Annotation and Visualization of RNA structure (SAVoR) software tool. SAVoR combines RNA backbone layout information from RNAplot(62) with annotation values such as dsRNA-seq and ssRNA-seq derived structure scores to produce highly informative and annotated models of RNA secondary structure (Figure 3.4).

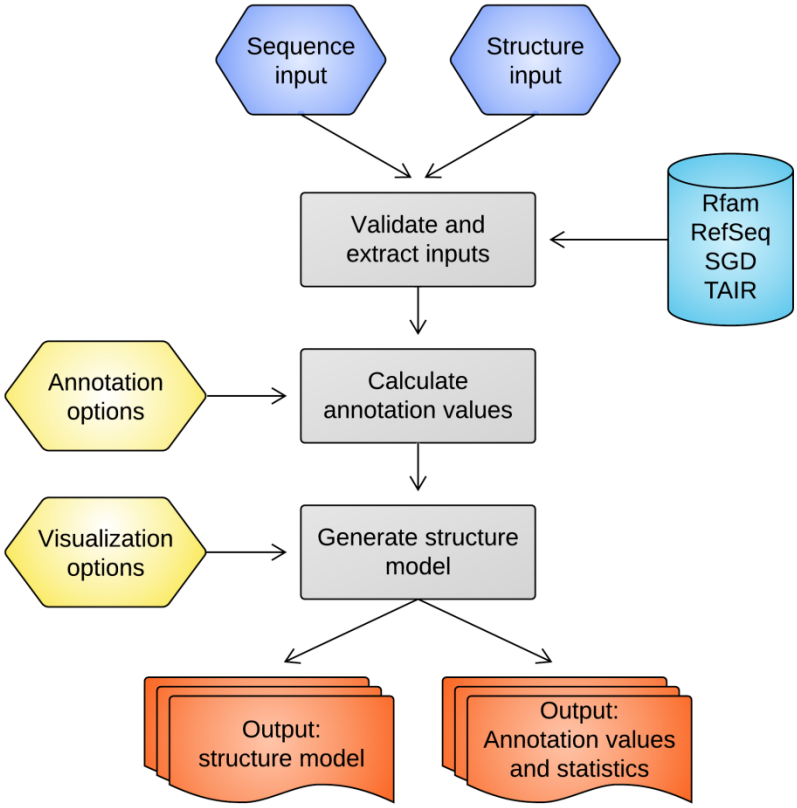


Figure 3.4: Workflow for the SAVoR structure visualization web server. Upon validation of user input, the primary sequence and genomic location of the user-submitted transcript(s) are determined, and intersecting sequence reads are converted to the desired annotation values. The secondary structure is then determined and plotted with the specified visualization options.

In addition to its usage in generating the experimentally-constrained mRNA structure models described above, we also implemented SAVoR as a publicly available web server (http://tesla.pcbi.upenn.edu/savor) with an extremely easy-to-use yet powerful user interface. SAVoR will be useful to many researchers in rapid prototyping and experimental design (e.g. oligo/primer design, structure prediction) as well as analyses of downstream data (e.g. SNPs, smRNA-seq, etc.).

3.5 Discussion

In this chapter, we described the genomic landscape of RNA secondary structure with respect to sites of protein translation and microRNA targeting. Using global measurements of secondary structure in three eukaryotic organisms, we found that a greatly decreased propensity for base pairing upstream of translation initiation sites. A link between protein translation and secondary structure has long been hypothesized(49), but we have provided the first experimental evidence of this process on a genome-wide scale. Interestingly, we also found the same decreased base pairing at stop codons; further experiments will be necessary to determine the functional role of such a mark as well as how it differs from the start codon. One hypothesis is that as the ribosome scans along an mRNA transcript(22, 86), the sudden decrease in secondary structure simply jars the ribosome loose, thereby terminating translation. Fully resolving the relationship between secondary structure and protein translation will yield important insights into this fundamental biological process and may offer particular avenues for RNA-mediated modulation of protein expression in a disease context.

From our dsRNA-seq and ssRNA-seq data, we also observed a marked decrease in base pairing within microRNA target sites relative to flanking sequences in *Arabidopsis* and *C. elegans*. These structural tendencies mirrored the known modes of action in the two organisms (full-length pairing in *Arabidopsis* and seed pairing in *C. elegans*), suggesting that base pairing is indeed a selective pressure on mRNA transcripts. Moreover, binding affinity of the ALG-1 protein in *C. elegans* was shown to be inversely proportional to the structural content of the target,

implying a direct connection between target site structure and microRNA targeting efficiency. Intriguingly, target site base pairing was not significantly reduced in *Drosophila*, suggesting that multiple modes of microRNA targeting may be active within the eukaryotic clade. Additional examination of secondary structure at true (as opposed to predicted) microRNA target sites, perhaps via immunoprecipitation of RISC-bound mRNA transcripts, is necessary to identify and characterize the action mechanisms of this important regulatory pathway. Finally, given the crucial role of microRNAs in cancer(30, 59, 120), neurodegenerative disorders(1, 19, 52, 95), and myriad other pathologies(39, 93), a thorough understanding of their targeting and regulatory principles will prove invaluable in the therapeutic setting.

To maximize the utility of our dsRNA-seq and ssRNA-seq data, we compiled a database of mRNA structure models for all three organisms studied (available at http://gregorylab.bio.upenn.edu/arabidopsisStructure/ and http://gregorylab.bio.upenn.edu/twoMetazoans/). We have also provided all of our sequencing data through the AnnoJ (http://gregorylab.bio.upenn.edu/annoj/) and JBrowse (http://gregorylab.bio.upenn.edu/jbrowse/) genome browsers as a resource for the research community. Finally, the SAVoR web server (http://tesla.pcbi.upenn.edu/savor) has also remained under active development, and recent updates include direct entry of annotation values as well as a web-enabled batch mode. It is our sincere hope that these data and tools are useful to researchers from a variety of fields and disciplines; for example, one might want to look up the secondary structure of a particular transcript of interest, identify instances of a newly discovered structural motif, or compare the structures of two orthologous transcripts.

3.6 Materials and methods

mRNA structure score profiles

mRNA annotations were downloaded from TAIR (version 9), FlyBase (r5.22), and WormBase (WS205), respectively. Structure scores were calculated as described in Section 2.1.3 but with normalization to the total number of mapped dsRNA-seq and ssRNA-seq reads per

transcript ($\boldsymbol{N_{ds}}$ and $\boldsymbol{N_{ss}}$). Each mRNA transcript was then split into 100 equally-sized bins for the 5' UTR, CDS, and 3' UTR, and the average structure score in each bin was computed. Finally, the genome-wide mRNA structure profile was calculated by averaging the profiles for all expressed mRNAs. Significance of differences in structure score within translation initiation and termination sites was determined by a Student's t-test.

Structure score at microRNA target sites

For *Arabidopsis*, microRNA target site predictions were downloaded from the psRNATarget web server(18), using the 243 published *Arabidopsis thaliana* miRNAs from miRBase(50) release 16 and all protein-coding mRNA transcripts from TAIR9. For *Drosophila* and *C. elegans*, predictions were downloaded from TargetScanFly (http://www.targetscan.org/fly_12/) and TargetScanWorm (http://www.targetscan.org/worm_12/) using 'Predicted Conserved Targets'. ALG-1 binding sites were downloaded from (128). Average structure profiles for target sites (full-length in *Arabidopsis* and seed (bases 2-8) in *Drosophila* and *C. elegans*) and 50 bps upstream and downstream were computed as described above. Significance was assessed by a Student's t-test.

Experimentally-derived models of mRNA secondary structure

A standardized version of the structure score $\boldsymbol{Z_i}$ was used to constrain RNAfold (from the Vienna package)(62) predictions of secondary structure for each transcript:

$$\boldsymbol{Z_i} = \frac{\boldsymbol{S_i} - \bar{S}}{s^2}$$

where $\bar{S}$ and s^2 are the mean and standard deviation of scores $\boldsymbol{S_i}$ for a given transcript. To determine thresholds to call paired and unpaired positions (t_{paired} and $t_{unpaired}$, respectively), a null distribution of standardized structure scores was calculated by randomly shuffling dsRNA and ssRNA reads and re-computing the standardized scores. Thus, positions with a structure score greater than t_{paired} were constrained as paired ('|' in the structural constraint input), positions with a structure score less than $t_{unpaired}$ were constrained as unpaired ('x' in the structural constraint

input), and all other positions were left unconstrained ('.' in the structural constraint input). All other RNAfold parameters were left as default.

SAVoR web server

The SAVoR webserver runs Apache 2.2.3 on a CentOS 5.7 machine with 2x Intel Xeon E5450 3.00 GHz processors and 16GB RAM. Asynchronous JavaScript and XML (AJAX) technology is used to dynamically render PHP output into formatted HTML. A local MySQL database is used to store Rfam and Refseq/SGD/TAIR entries, and a local installation of BLAST+ is used to retrieve sequence and genomic locus information. Structure prediction is optionally performed using a local installation of RNAfold [version 1.8.4] or RNAstructure [version 5.6], and backbone layout is done using RNAplot. SAMtools(56) is used to extract annotation values from BAM files, and custom Perl and Ruby scripts are used to process BED files. Inkscape [version 0.47] is used to convert from the native SVG format to publication-quality PDF and PNG output files.

Chapter 4

Regulatory impact of RNA secondary structure

In this chapter, we focus on the regulatory roles of RNA secondary structure. We integrate data from sequencing of multiple RNA subpopulations in the model plant *Arabidopsis thaliana* to identify global relationships between secondary structure and gene expression at the RNA and protein levels. Additionally, we reveal a novel mechanism by which the cellular RNA silencing pathways directly regulate mRNA abundance.

This section references work from:

- Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, et al. Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. 2010 (127)
- Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. Plant Cell. 2012 (55)

4.1 Introduction

RNA secondary structure is a critical component of many cellular regulatory processes. Proper folding is required for the biogenesis, maturation, and function of most, if not all, classes of non-coding RNAs (described in Table 1.2). In particular, the effectors of RNA silencing pathways (microRNAs and various siRNAs in plants) are generally produced from double-stranded precursors (Figure 4.1).

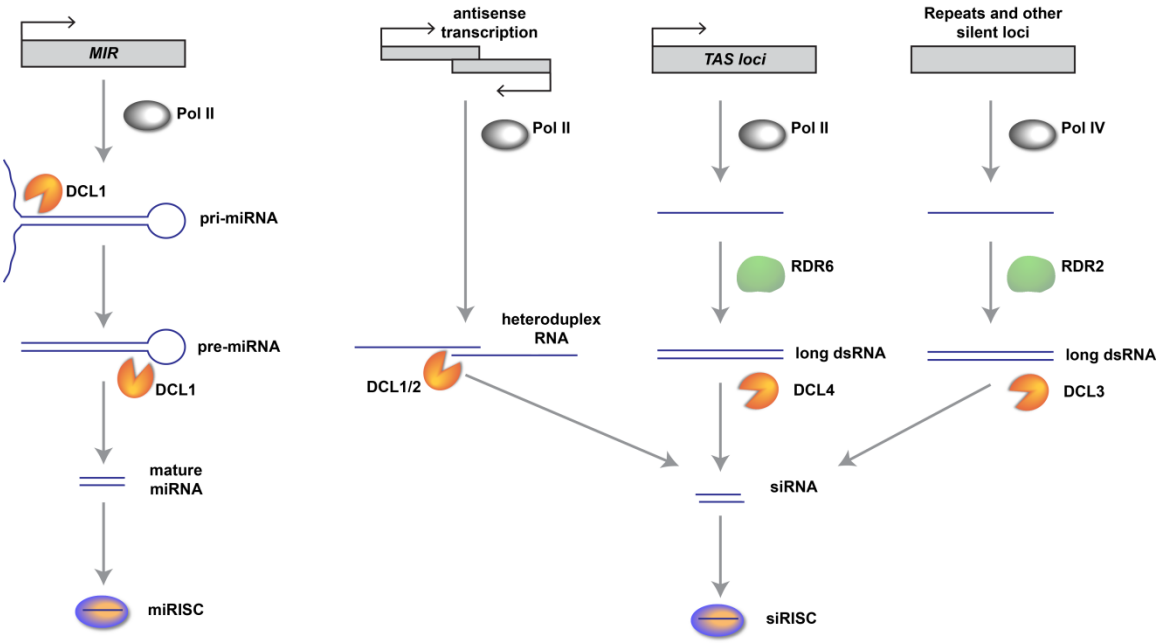


Figure 4.1: smRNA biogenesis pathways in plants.

MicroRNAs are initially transcribed by RNA Pol II, and this primary transcript (pri-miRNA) is then cleaved by DICER-LIKE 1 (DCL1) to yield a canonical stem-loop hairpin(106). An additional cleavage step then yields the mature miRNA as a ~20-21 nucleotide product that directs post-transcriptional or translational repression of specific mRNAs through direct base pairing interactions with complementary sites in the target transcript sequence. siRNAs are generated in a similar process by the three other members of the Dicer-like family (DCL2, DCL3, and DCL4) from a variety of double-stranded precursors(12). Two key differences separate microRNAs and siRNAs. First, microRNAs are defined as being exclusively endogenous, whereas siRNAs can be derived from exogenous sources such as viral, transgene, or injected dsRNA. Additionally, whereas microRNA precursors are incompletely base paired, siRNAs are thought to require perfect base pairing within their precursors. However, the line between the two small RNA classes is being increasingly blurred as more and more overlap is revealed between their modes of biogenesis and action(106).

Secondary structure is also a major player in regulation of protein-coding genes. A host of structured elements (see Section 3.1), located primarily in the untranslated regions, regulate

the stability, splicing, and localization of mRNA transcripts. Beyond these short features, however, little is known about how secondary structure is related to mRNA processing and maturation. Control of translation initiation and elongation is also tightly linked to secondary structure. Computational predictions(32, 36) as well as our results from the previous chapter have suggested a propensity for decreased structure at initiation sites to facilitate ribosome binding. Recent studies have also proposed that mRNA secondary structure impedes translation elongation(32, 115). To date, the global role of RNA secondary structure within these myriad frameworks have remained elusive primarily due to the lack of available structural data.

In this section, we attempt to ascertain the exact nature of structure-mediated control at the levels of RNA processing, abundance, and translation. We integrate the structure data from our dsRNA-seq and ssRNA-seq studies in *Arabidopsis thaliana* with transcriptome-wide maps of RNA abundance, small RNA production, and ribosome binding to reveal the many regulatory roles of secondary structure. These results uncover a particularly intriguing possibility of direct processing of highly structured mRNAs by RNA silencing machinery.

4.2 Integration of multiple genomic datasets in *Arabidopsis*

We started with the determination of secondary structure across all expressed mRNA transcripts in the model plant *Arabidopsis thaliana* as described in Chapters 2 and 3. From these dsRNA-seq and ssRNA-seq data, we computed the structure score at each position with the mature mRNA as well as the average for each transcript. We then compared the average structure score for every detected transcript with other measurements of the transcript's properties.

4.2.1 Secondary structure and mRNA abundance

We determined steady-state mRNA abundance by sequencing of the ribosomal RNA-depleted transcriptome (RNA-seq), and found that RNA folding had a significant negative effect (Pearson correlation $r = -0.45$, $p \rightarrow 0$) on total transcript levels (Figure 4.2).

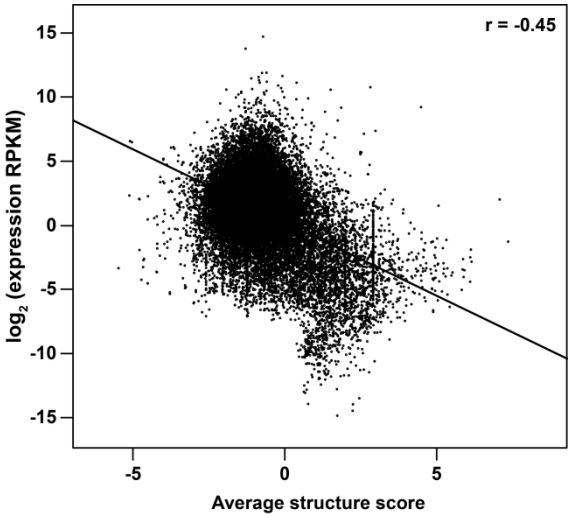


Figure 4.2: Average structure score (x axis) plotted against average expression values determined by RNA-seq (y axis) for all detectable Arabidopsis mRNAs.

We then confirmed this observation using qRT-PCR (quantitative reverse transcription PCR) on five highly and seven lowly structured mRNAs (12 total mRNAs). From this analysis, we found that the less structured mRNAs were all significantly ($p < 0.001$) more abundant than those transcripts with high levels of folding (Figure 4.3).

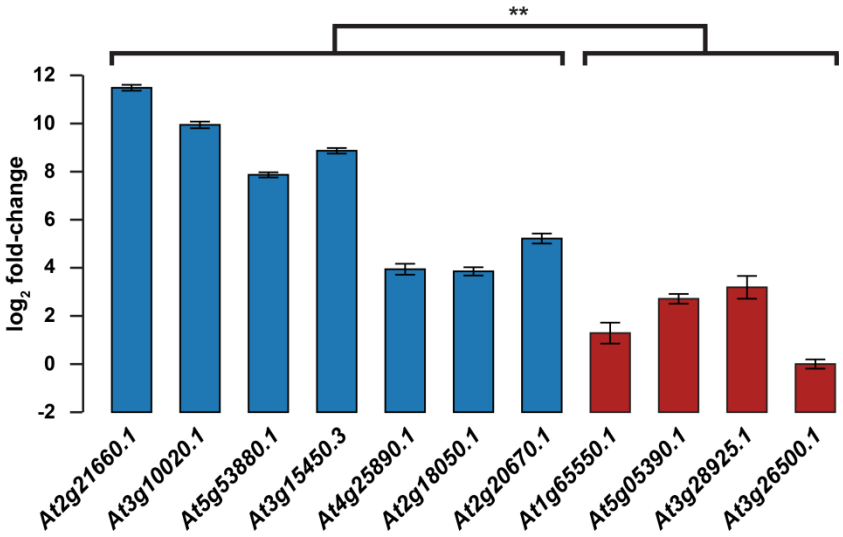


Figure 4.3: Random hexamer-primed qRT-PCR analysis of seven lowly (blue bars) and five highly (red bars) structured *Arabidopsis* mRNAs. Error bars, 6 SE. ** denotes p-value < 0.001, one-tailed t test.

In total, these findings reveal that mRNA secondary structure has a significantly negative regulatory effect on the overall abundance of mRNAs in the *Arabidopsis* transcriptome.

4.2.2 Degradation and smRNA production from structured mRNAs

Given these findings, we considered the possibility that mRNA degradation and/or smRNA processing could explain the relationship between secondary structure and overall transcript abundance. To test this, we normalized previously published genome-wide RNA degradation ('degradome') data(34) by total transcript abundance as measured by RNA-seq data to ascertain the degradation rates for every detectable mRNA. We found a significant positive correlation (Pearson correlation r = 0.21, p → 0) between the overall structure score and degradation level of *Arabidopsis* mRNAs (Figure 4.4), indicating that highly folded mRNAs tend to be degraded more frequently than less structured transcripts.

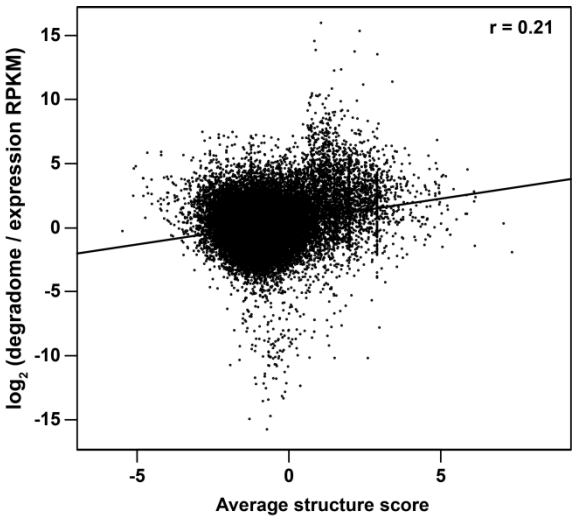


Figure 4.4: Average structure score (x axis) plotted against average degradation values determined by correcting degradome values by RNA-seq (y axis) for all detectable *Arabidopsis* mRNAs.

Interestingly, this relationship between structure and transcript degradation level is even stronger for mRNAs with predicted miRNA target sites (Pearson correlation r = 0.36, data not shown). This is likely because highly structured RNAs can be targeted for degradation both by miRNA binding events and intrinsic structural features. Taken together, these results suggested that RNA secondary structure is an intrinsically destabilizing feature of protein-coding mRNAs in *Arabidopsis*. More intriguingly, our findings also hinted at the possibility of direct smRNA processing of highly structured mRNAs as these fragments would be captured by the 'degradome' sequencing data.

To address this hypothesis, we used smRNA-seq (see Section 4.4) to assess the abundance of small RNAs that were directly processed from mRNA transcripts. Using this approach, we found a significant positive correlation (Pearson correlation r = 0.62, $p \rightarrow 0$) between increasing mRNA secondary structure and higher levels of sense smRNA production (Figure 4.5).

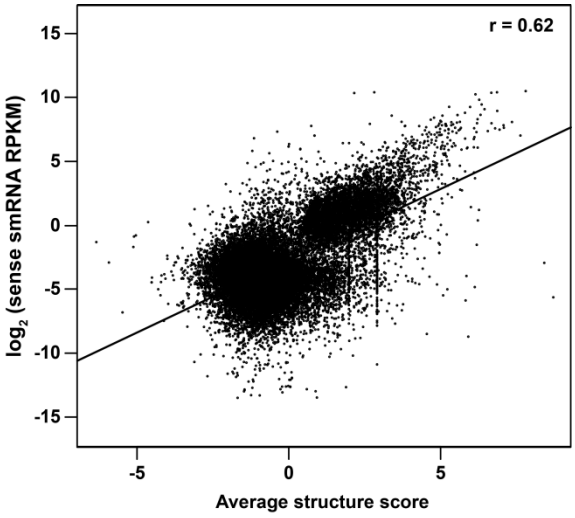


Figure 4.5: Average structure score (x axis) plotted against the total abundance of smRNAs present per transcript in the sense orientation as determined by smRNA-seq (y axis) for all detectable *Arabidopsis* mRNAs.

We also found a similar trend for production of smRNAs from the antisense strand (Pearson correlation r = 0.65, Figure 4.6), suggesting that initial processing of highly structured mRNAs leads to secondary dsRNA synthesis (likely by an RNA-dependent RNA polymerase) and subsequent production of both sense and antisense smRNAs.

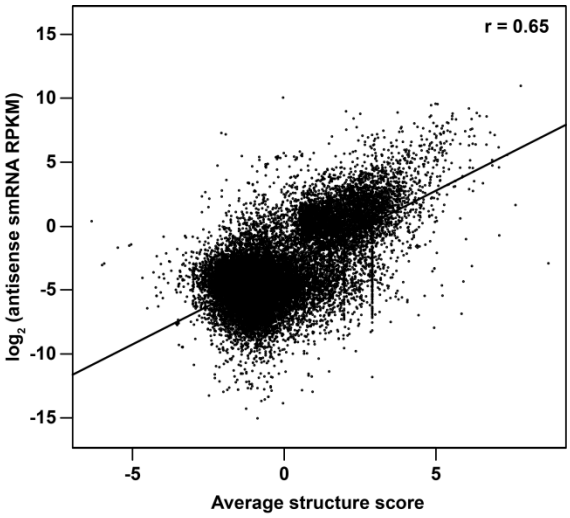


Figure 4.6: Average structure score (x axis) plotted against the total abundance of smRNAs present per transcript in the antisense orientation as determined by smRNA-seq (y axis) for all detectable *Arabidopsis* mRNAs.

4.2.3 Direct processing of highly structured mRNA elements

Our findings of increased degradation and smRNA production from structured mRNAs could be explained by bulk processes; that is, perhaps these structured transcripts tend to be lowly expressed due to turnover and rapid degradation of the entire mRNA by the exosome. We wanted to test the alternative possibility that highly structured regions of mRNAs were in fact being directly targeted by the RNA silencing machinery in a manner similar to that of the small RNA biogenesis pathways. To do so, we used our smRNA-seq data to define portions of mRNA transcripts that produced a significant amount of small RNAs (see Section 4.4). As expected under the second hypothesis, the regions of mRNAs that are processed into smRNAs were

significantly (p → 0, t-test) more structured than the regions that are not cleaved into smRNAs (Figure 4.7).

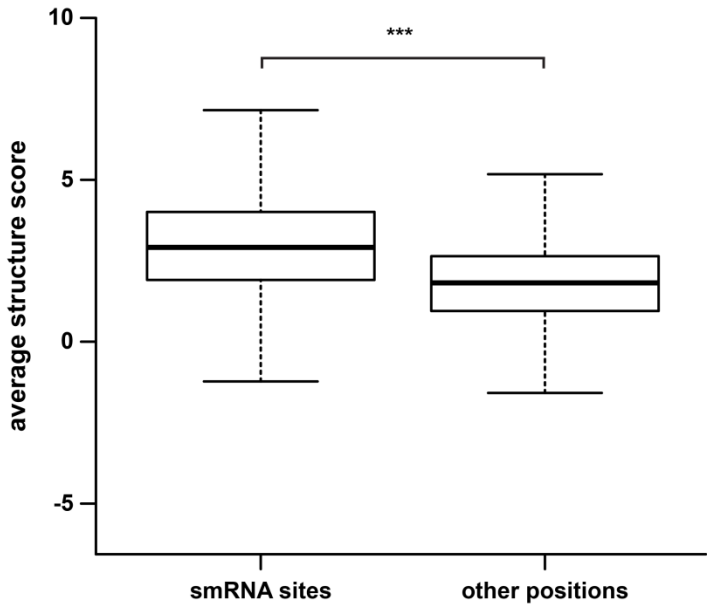


Figure 4.7: The average structure score (y axis) of mRNA regions processed into smRNAs (left box, smRNA sites) compared with those that are not (right box, other positions). *** denotes p-value < 2.2e-16, t test.

We also repeated the correlation analysis between secondary structure and smRNA production described in Section 4.2.2 but limited the calculation of smRNA levels for each transcript to those that were derived exclusively from highly structured intervals within the mRNA (dsRNA hotspots). This analysis replicated our previous findings of a strong positive correlation between mRNA structure and smRNA processing (Figure 4.8, Pearson correlation r = 0.41), suggesting a novel adaptation of the small RNA biogenesis machinery to directly process and thereby regulate mRNA levels.

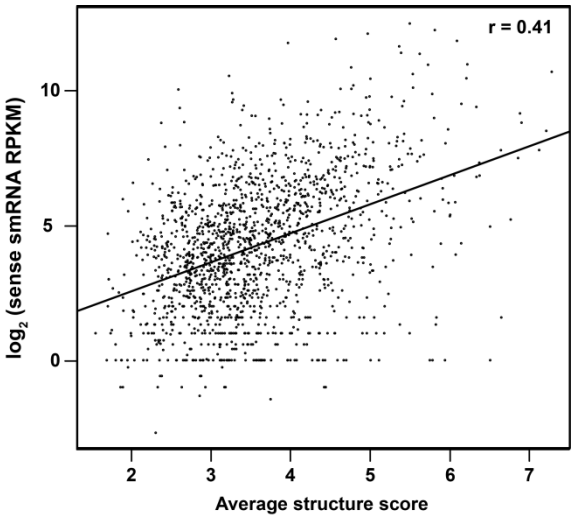


Figure 4.8: Average structure score (x axis) plotted against the total abundance of smRNAs present per dsRNA hotspot in the sense orientation as determined by smRNA-seq (y axis).

4.2.4 Secondary structure and ribosome binding

On the basis of computationally predicted base pairing as well as individual examples, mRNA secondary structure is known to be a strong impediment to translational initiation and elongation. Given our ability to measure secondary structure in a high-throughput and reliable manner, we wanted to examine the global relationship between RNA structure and translation. Therefore, we utilized the ribo-seq method(41, 80) to assess ribosome binding density across the transcriptome and found a strong positive correlation (Pearson correlation $r= 0.37$, $p \rightarrow 0$) between mRNA structure and ribosome binding (Figure 4.9). We confirmed this observation using qRT-PCR (quantitative reverse transcription PCR) on four highly and seven lowly structured mRNAs (11 total mRNAs)

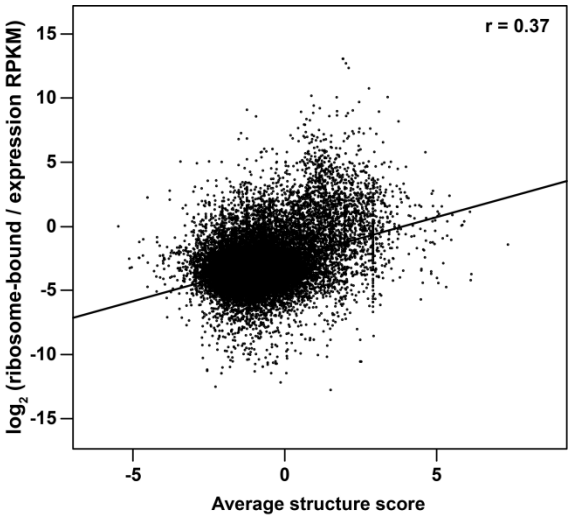


Figure 4.9: Average structure score (x axis) plotted against average ribosome association values determined by normalizing ribo-seq values by RNA-seq (y axis) for all detectable *Arabidopsis* mRNAs.

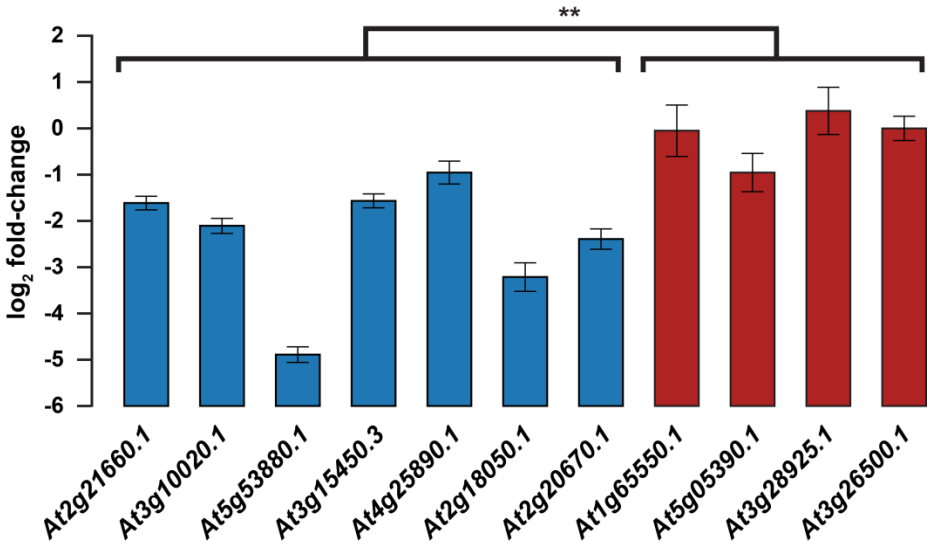


Figure 4.10: Random hexamer-primed qRT-PCR analysis of seven lowly (blue bars) and five highly (red bars) structured Arabidopsis mRNAs using ribosome-bound RNA fractions with values corrected by total RNA abundance as also measured by qRT-PCR. Error bars indicate 6SE. ** denotes p-value < 0.001, one-tailed t test.

Unfortunately, we were unable to distinguish between paused/stalled ribosomes and actively translating ribosomes using the ribo-seq approach, and therefore are left to speculate as to the underlying basis of this relationship (see Discussion below).

4.3 Discussion

In this chapter, we explored the regulatory significance of RNA secondary structure by an integrative analysis of multiple high-throughput sequencing datasets. An initial comparison of dsRNA-seq, ssRNA-seq, and total RNA-seq data revealed a strongly negative correlation between secondary structure and mRNA abundance. By adding smRNA-seq and 'degradome' sequencing, we further established that the relationship between mRNA structure and steady state levels is at least partially explained by smRNA processing and/or degradation. Finally, by restricting our analyses to highly structured elements within mRNAs, we found that these regions are indeed directly processed into small RNAs (Figure 4.8).

Many outstanding and intriguing questions yet remain. For instance, how are these structured moieties processed? Our favored hypothesis is that the canonical small RNA biogenesis pathway is co-opted; additional experiments to characterize mRNA processing in microRNA mutants (e.g. *dcl2*, *hen1*) would confirm or disprove this possibility. On a smaller scale, *in vitro* "dicing" assays with specific mRNAs that contain structure 'hotspots' but no known microRNA target sites may be able to identify the biogenesis mechanism, at least for those transcripts. If direct processing of highly structured mRNAs is indeed found to be Dicer-dependent, additional follow-up experiments would be needed to carefully tease out the differences between these structural elements and canonical pre-miRNA hairpins.

Another question relates to the potential function of these mRNA-derived small RNAs. If they are produced by the canonical smRNA biogenesis pathways, then it stands to reason that they might function as smRNAs in a regulatory sense. However, we consider this to be unlikely as we did not observe a substantial microRNA-like size pattern within smRNA reads that mapped to highly structured mRNA intervals. Given the steric constraints imposed by the PAZ domain of the

Argonaute effector proteins(100, 121), it is highly improbable that many of these mRNA-derived smRNAs could even be loaded into a RISC complex. We instead favor the hypothesis that the processing of these small RNAs *is* their function, insomuch as this processing thereby regulates mRNA levels. Interestingly, more than half of the bases in an mRNA are expected to be paired(115); this implies the existence of 'dark matter' secondary structure that is not accounted for by the existence of known structured regulatory moieties (see Section 3.1). Our findings of direct smRNA processing from newly-characterized structured elements may explain a large portion of this dark matter, pushing to further prominence the role of secondary structure in mRNA regulation and function.

We also identified a positive correlation between mRNA secondary structure as measured by dsRNA-/ssRNA-seq and ribosome binding density using the ribo-seq approach. However, because our ribo-seq data were unable to differentiate stalled and actively translating ribosomes, we are left to present possible explanations for our results. One likely possibility is that increased secondary structure leads to slowing or pausing of elongation which is then captured by ribo-seq as increased density. Increased base pairing at stop codons could also decrease the efficiency of translation termination, as suggested by the profiles of mRNA secondary structure (Figure 3.1). Additional experiments that separately measure active and inactive ribosome density (e.g. ribosome and polysome profiling) are necessary to truly assess the impact of secondary structure on translation.

Over the past three chapters, we have described the development and application of sequencing-based methodologies to assess RNA secondary structure on a global scale. Genomic analyses of these dsRNA-seq and ssRNA-seq datasets in three eukaryotic organisms – *Arabidopsis thaliana*, *Drosophila melanogaster*, and *C. elegans* – offered empirical and global evidence for many long-hypothesized features of RNA biology. For example, our structure mapping data demonstrated reduced base pairing at translation initiation and termination sites as well as microRNA target sites; these trends have been suggested based on computationally predicted base pairing models but we provide the first direct proof of their generality. Additional integration of our structure data with readouts of total RNA abundance, degradation, and small

RNA processing revealed a novel mechanism by which mRNA levels are directly regulated by processing of highly structured regions and subsequent degradation. In total, our findings have highlighted the importance and power of genome-wide studies, particularly when used in complement with classical hypothesis-driven approaches. In the next chapter, we shift our focus from genome-scale analyses to the task of single nucleotide resolution prediction of individual secondary structures.

4.4 Materials and methods

Total RNA sequencing (RNA-seq)

Two replicate RNA-seq libraries were produced using the SOLiD Total RNA-seq library preparation kit (Applied Biosystems, Carlsbad, CA). Subtraction of ribosomal RNA was carried out with the RiboMinus kit (Invitrogen, Carlsbad, CA) according to manufacturer's instructions. Both replicates were sequenced on an ABI SOLiD 3+ (ABI, Foster City, CA) according to manufacturer's instructions.

Small RNA sequencing (smRNA-seq)

smRNA-seq libraries were produced using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) per manufacturer's instructions. Sequencing was carried out on an Illumina GA2 Analyzer (Illumina Inc., San Diego, CA). A detailed protocol follows:

I. Start with 40 μg of RNA from desired source material, suspended in 40 μL DEPC-treated water.

II. **Size selection**

   a. Prepare 1000 mL 1X TBE running buffer (100 mL 10X TBE extended range + 900 mL Milli-Q water).

   b. Pre-run 15% TBE-Urea polyacrylamide gel (e.g. from Invitrogen) for 25 minutes at 155 V.

   c. While gel is pre-running, prepare ladder and sample:

i. Ladder: 1.5 μL 10bp DNA ladder, 8.5 μL DEPC-treated water, and 10 μL Gel Loading Buffer (e.g. from NEB).

ii. Add 10 μL Gel Loading Buffer to sample.

iii. Place sample (but not ladder) at 70°C for 5 minutes, followed by 3 minutes on ice.

d. After pre-run is complete, run ladder and sample at 155 V for approximately 1.5 hours.

e. Stain gel with ethidium bromide. Add 14 μL 10 mg/mL ethidium bromide to 200 mL 1X TBE buffer in a clean RNase-free tray. Add gel and rock gently for 10 minutes.

f. Cut 15-45bp band from gel and place gel slice in a 0.5mL tube with holes (e.g. Gel Breaker Tubes #3388-100 from IST Engineering Inc.), placed inside a clean 2mL tube.

g. Spin sample at 14000RPM, 4°C for 2 minutes. Repeat until all of the gel goes through the 0.5mL tube.

h. Add 300 μL 0.3M NaCl and rotate for 4 hours.

i. Pipette entire sample into a Spin-X column and spin at 14000RPM, 4°C for 2 minutes. Transfer eluent to new 1.5mL tube.

j. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 900 μL 100% EtOH.

k. Resuspend in 5 μL DEPC-treated water.

III. **Adapter ligation**

a. Add 5 μL sample and 1 μL 5 μM RNA 3' Adapter (RA3) to a sterile, nuclease-free 200 μL PCR tube on ice.

b. Pipette mixture up and down 6-8 times to thoroughly mix and then centrifuge briefly.

c. Incubate in thermal cycle at 70°C for 2 minutes, then at 4°C for 2 minutes.

d. Add 2 μL Ligation Buffer, 1 μL RNase Inhibitor (e.g. RNaseOUT from Life Technologies), and 1 μL Epicentre T4 RNA ligase 2 deletion mutation (200 U/μL). Mix thoroughly.

e. Incubate at 28°C for 75 minutes in thermal cycler.

f. With 5 minutes left, heat 1 μL 25μM 5' Adapter (RA5) to 70°C for 2 minutes, then place on ice for 2 minutes.

g. Add 1 μL RA5, 1 μL 10mM ATP, and 1 μL T4 RNA Ligase 1 to sample tube. Mix thoroughly.

h. Incubate at 28°C for one hour in thermal cycler. Store at -20°C overnight unless proceeding directly to next step.

IV. **Size selection to reduce adapter adapter**

a. Run sample on 15% TBE-Urea polyacrylamide gel as in Step VI. Cut 65-95bp band, taking care to avoid 50bp adapter-adapter band. Resuspend in 6 μL DEPC-treated water.

V. **Reverse transcription**

a. Incubate 6 μL sample and 1 μL 100μM RNA RT Primer (RTP) at 70°C for 2 minutes in preheated thermal cycler. Then incubate at 4°C for 2 minutes.

b. Add 2 μL 5X First Strand Buffer, 0.5 μL 12.5mM dNTP mix (12.5mM of each nucleotide), 1 μL 100mM DTT, 1 μL RNase Inhibitor (e.g. RNaseOUT), and 1 μL SuperScript II Reverse Transcriptase. Mix thoroughly.

c. Incubate at 50°C for one hour.

VI. **PCR amplification**

a. Prepare PCR master mix: 35 μL 2X Phusion Mix, 21 μL 5mM betaine, 2 μL 10μM RNA PCR Primer (RP1), and 2 μL 10μM RNA PCR Primer Index (RPI*X*).

b. Add 60 μL master mix to 12.5 μL sample, then aliquot mixture to 3 PCR tubes with approximately 25 μL in each tube.

c. PCR amplification program in thermal cycler

i. 98°C for 30 seconds

ii. 98°C for 10 seconds

iii. 60°C for 30 seconds

iv. 72°C for 15 seconds

v. Cycle to step ii 11X

vi. 72°C for 10 minutes

vii. Hold at 4°C

d. Precipitate by adding 10 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 300 μL 100% EtOH.

e. Resuspend in 10 μL DEPC-treated water.

VII. **Size selection**

a. Prepare 1000 mL 1X TBE running buffer (100 mL 10X TBE extended range + 900 mL Milli-Q water).

b. Prepare ladder and sample:

i. Ladder: 1.5 μL 25bp DNA ladder, 8.5 μL DEPC-treated water, and 10 μL Gel Loading Buffer (e.g. from NEB).

ii. Add 10 μL Gel Loading Buffer to sample.

c. Run ladder and sample at 155 V for approximately 30 minutes.

d. Stain gel with ethidium bromide. Add 14 μL 10 mg/mL ethidium bromide to 200 mL 1X TBE buffer in a clean RNase-free tray. Add gel and rock gently for 10 minutes.

e. Cut 133-163bp band from gel and place gel slice in a 0.5mL tube with holes (e.g. Gel Breaker Tubes #3388-100 from IST Engineering Inc.), placed inside a clean 2mL tube. Adapter-adapter is 118bp at this point.

f. Spin sample at 14000RPM, 4°C for 2 minutes. Repeat until all of the gel goes through the 0.5mL tube.

g. Add 300 μL 1X NEB Buffer 2 and rotate for 2 hours.

h. Pipette entire sample into a Spin-X column and spin at 14000RPM, 4°C for 2 minutes. Transfer eluent to new 1.5mL tube.

i. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 900 μL 100% EtOH.

j. Resuspend in 12 μL DEPC-treated water for sequencing.

Ribosome-associated sequencing (ribo-seq)

Ribo-seq libraries were made using ribosome-associated mRNAs from unopened flower buds that were isolated by differential centrifugation according to Mustroph *et al*.(80) with the following modifications. The ribosomes and associated mRNAs pelleted by centrifugation through a sucrose cushion were resuspended in 0.2 M Tris pH 8.0, 0.2 M KCl, 0.035 M $MgCl_2$, 50 μg/ml chloramphenicol, and 50 μg/ml cycloheximide. 40 μg of resuspended RNA was centrifuged over a 15-60% sucrose gradient (0.04 M Tris, pH 8.0, 0.02 M KCl, 0.02 MgCl2, 5 μg/ml chloramphenicol, and 5 μg/ml cycloheximide). Following centrifugation, 50 μl fractions of the gradient were isolated and the OD260 of each was measured. The monosomal and polysomal fractions were pooled, and the RNA was isolated using the Qiagen miRNeasy Mini Kit. Eight μg of isolated RNA were depleted of ribosomal RNA using the RiboMinus Plant Kit (Life Technologies, Carlsbad, CA), fragmented using RNA Fragmentation Reagents (Ambion, Austin, TX), treated with T4 PNK (NEB, Boston, MA) to repair 5' and 3' ends, and used for library preparation using the Illumina TruSeq smRNA-seq library preparation kit and accompanying protocols (Illumina, San Diego, CA). Sequencing was carried out on an Illumina HiSeq2000 (Illumina Inc., San Diego, CA).

Regions of small RNA production (smRNA hotspots)

Regions of significant small RNA production were determined using the following approach. First, consecutive smRNAs were identified on each chromosome and then pre-grouped into smRNA clusters (smRNA contigs). Next, a derived "per-smRNA site" abundance (PSS-abundance) was calculated for all smRNA clusters as $\frac{N_r}{L_c} \times \bar{X}_s$, where N_r and L_c are the total

number of cloned reads and length for this smRNA cluster, respectively, and $\bar{X}_s$ is the average length of all smRNA reads. Finally, the derived PSS-abundance on each chromosome was assumed to follow a Poisson distribution:

$$P(X_i = k) = \frac{\lambda_i^k}{k!} \cdot e^{-\lambda_i}$$

$$P(X_i \le k) = e^{-\lambda_i} \sum_{i=0}^{k} \frac{\lambda_i^k}{i!} = \frac{\Gamma(\lfloor k+1 \rfloor, \lambda_i)}{\lfloor k \rfloor !}$$

where X_i is the derived PSS-abundance and λ_i is the expected number of smRNA reads per smRNA-site on chromosome i. Thus, the derived PSS-abundance data can be fitted to this Poisson distribution model, the parameters λ_i estimated, and the confidence intervals for PSS-abundance of all smRNA clusters estimated for each chromosome. Finally, smRNA hotspots were identified as smRNA clusters with higher PSS-abundance than expected by chance.

Quantitative reverse transcription PCR (qRT-PCR)

RNA was isolated using the miRNeasy Mini Kit (Qiagen, Valencia, CA). Random hexamer-primed cDNA was made for at least three biological replicates per experiment. Transcripts were then quantified by qPCR using the comparative threshold cycle method ($\Delta\Delta C_t$), using *Actin 2* (*At3g18780*) as the endogenous reference and the lowest expressed transcript for renormalization.

Chapter 5

Sequencing-based prediction of RNA secondary structure

In this chapter, we present an approach for sequencing-based inference of RNA secondary structure. We develop a novel likelihood model that describes the generation of dsRNA-seq and ssRNA-seq reads from an underlying structure, as well as a corresponding simulator and Markov chain Monte Carlo (MCMC) algorithm. Application of our new method to eight known secondary structures reveals marginally increased accuracy compared with traditional free energy-based algorithms.

This section references work from:

- Li F, Ryvkin P, Silverman IS, Wang LS, Gregory BD. Sequencing-based inference of RNA secondary structure. *Unpublished results*.

5.1 Introduction

The field of RNA secondary structure is by now quite well-developed, with perhaps hundreds of methods whose throughput ranges from single molecule crystallography to genome-wide free energy-based prediction (see Section 1.2). In particular, free energy-based methods have become a mainstay in the task of structure prediction due to their simplicity and ease of use. These approaches generally use a predefined set of energy parameters (e.g. base pairing, base stacking, loop penalties, etc.) along with dynamic programming to identify the set of pairing interactions that results in the lowest free energy conformation. Although additional refinements such as non-canonical base pairs and centroid-based folding have further improved the reliability of these methods, they still cannot compete with the accuracy of more focused experimental approaches such as X-ray crystallography, chemical probing, and RNase footprinting.

Recently, several groups have attempted to increase the performance of energy-based prediction by incorporating experimental structure data. SHAPE-CR and later SHAPE-seq(64) utilized chemical probing reactivities as a pseudo-free energy term in the RNAstructure(89)

algorithm to greatly improve prediction accuracy. We and others (see Section 2.3) developed global RNase footprinting approaches that were then used to either pre-constrain RNAfold-based structure prediction (Section 3.4) or post-select from clusters of predicted secondary structures(84). Although these methods differ widely in their execution, they share the common thread of heavy reliance on free energy-based prediction. In this section, we describe a completely new paradigm of initial structure prediction based on Markov chain Monte Carlo optimization of a likelihood function that describes the generation of dsRNA-seq and ssRNA-seq data.

MCMC methods are a general class of algorithms for sampling from a posterior distribution that is difficult to directly estimate. As the name suggests, these methods work by building a Markov chain of samples which converges to the desired posterior distribution at some point along the chain. Successive moves along the chain are determined either completely at random or semi-randomly (hence the 'Monte Carlo'). A useful analogy is to imagine a hiker walking amongst a range of hills and attempting to reach the lowest point in the range (Figure 5.1, top panel). Moves along the Markov chain correspond to steps taken by the hiker in either direction (Figure 5.1, middle panel), albeit with the constraint that he/she is averse to taking large uphill steps. After an appropriate number of steps(74), the Markov chain has converged to the target distribution and our hiker has found the lowest valley (Figure 5.1, bottom panel). There are a number of conditions that may lead to extremely slow convergence; these include local optima (shallow valleys in our analogy), inefficient mixing (the hiker frequently backtracks), and bad initial estimates (the hiker starts very far from the lowest valley).

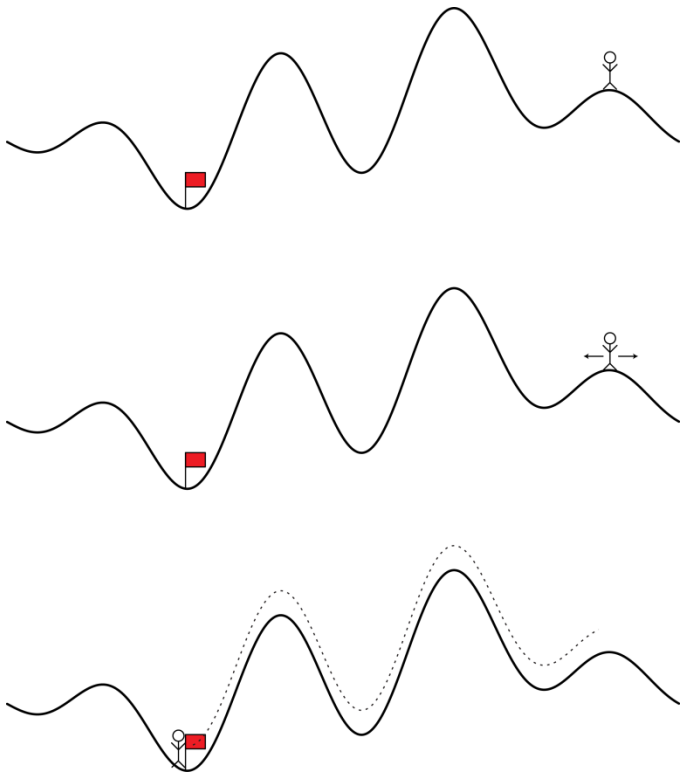


Figure 5.1: A cartoon representation of MCMC. (Top) From some initial point, our hiker must reach the lowest valley (red flag). (Middle) Move options for the hiker as indicated by arrows. Obstacles such as hills make the hiker less likely to go in that particular direction. (Bottom) After a series of moves (dotted line), our hiker reaches the target.

One popular MCMC method, the Metropolis-Hastings algorithm, attempts to mitigate these possibilities by drawing candidate steps from a proposal distribution. These candidate moves are then accepted or rejected based on the likelihood ratio between the candidate and current state. Simply put, if a candidate state x^* is more likely than the current state x_t, it is automatically accepted; however, if the candidate state is less likely, it can still be accepted with probability

$$\frac{P(x^*)Q(x_t|x^*)}{P(x_t)Q(x^*|x_t)}$$

$$\text{where } \frac{P(x^*)}{P(x_t)} \text{ is the likelihood ratio between the candidate and current states}$$

$$\text{and } \frac{Q(x_t|x^*)}{Q(x^*|x_t)} \text{ is the ratio of the proposal density}$$

In our analogy, this accept-reject paradigm allows the hiker to climb hills that he/she would other be loathe to traverse and can thereby overcome many of the obstacles described.

The ability of MCMC methods to approximate a target distribution without direct sample proves extremely useful in the context of Bayesian inference. Bayes' theorem

$$P(\theta|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\theta)P(\theta)}{P(\boldsymbol{D})}$$

expresses the posterior probability of observing parameters θ given some data $\boldsymbol{D}$ as a function of the likelihood function $P(\boldsymbol{D}|\theta)$ and a prior distribution of parameters $P(\theta)$. Additionally, the term $P(\boldsymbol{D})$ is typically equivalent for all parameters θ and therefore becomes a constant:

$$P(\theta|\boldsymbol{D}) \propto P(\boldsymbol{D}|\theta)P(\theta)$$

At this point, analytical optimization of the posterior $P(\theta|\boldsymbol{D})$ is difficult, but MCMC methods such as the Metropolis-Hastings algorithm can be used to sample from this distribution and eventually converge upon an approximately optimal solution.

As such, MCMC methods are widely used in many applications that can be framed in a Bayesian context but are too complex to solve analytically. In the realm of RNA secondary structure prediction, several algorithms have utilized so-called Bayesian MCMC to address various tasks. SimulFold(77) uses the Metropolis-Hastings algorithm to simultaneously infer RNA structures, alignments, and trees from unaligned multiple sequence data. McQFold(75) attempts to predict pseudoknotted RNA secondary structures by a similar MCMC approach. Our algorithm, termed RNA-seq-fold, also implements the Metropolis-Hastings algorithm, but with a very different likelihood function that is based on the production of dsRNA-seq and ssRNA-seq reads rather than thermodynamic or sequence considerations. In the next section, we describe our likelihood model in greater detail and provide direct experimental motivation for each of its mathematical terms. We then apply our novel approach to eight non-coding RNAs with known secondary structures.

5.2 A Bayesian framework for dsRNA- and ssRNA-seq

The Bayesian interpretation is a natural fit for dsRNA-seq and ssRNA-seq data, as we have observations (sequencing reads) as well as unknown parameters (the underlying secondary structure, enzyme digestion rates, etc.). As before, we have:

$$P(\theta|\boldsymbol{D}) \propto P(\boldsymbol{D}|\theta)P(\theta)$$

which can be rewritten as:

$$P(\boldsymbol{s},\boldsymbol{u},\boldsymbol{v},N,r_{min},r_{max}|\boldsymbol{R}) \propto P(\boldsymbol{R}|\boldsymbol{s},\boldsymbol{u},\boldsymbol{v},N,r_{min},r_{max})P(\boldsymbol{s},\boldsymbol{u},\boldsymbol{v},N,r_{min},r_{max})$$

where $\boldsymbol{s}$ is a secondary structure of length l, $\boldsymbol{u}$ and $\boldsymbol{v}$ are enzyme digestion rates, N is the number of enzymatic events per molecule, r_{min} is the minimum cloneable fragment size, r_{max} is the maximum cloneable fragment size, and $\boldsymbol{R}$ is our sequencing data. Importantly, this formulation is directly motivated by the experimental protocol that is used to generate dsRNA-seq and ssRNA-seq libraries.

5.2.1 From experimental protocol to likelihood model

Let us first consider the likelihood term $P(\boldsymbol{R}|\boldsymbol{s},\boldsymbol{u},\boldsymbol{v},N,r_{min},r_{max})$, which describes the probability of observing a set of sequencing reads $\boldsymbol{R}$ from an underlying structure $\boldsymbol{s}$ of length l. We start by assuming independence of individual reads, such that

$$P(\boldsymbol{R}) = \prod_{k=1}^{m} P(R_k)$$

for an experiment with *m* total reads. To derive $P(R_k)$, let us consider our experimental setup. We start with a dilute solution of RNA molecules and RNase (Figure 5.2, step 1). In this situation, each individual RNA molecule will be subjected to stochastic interaction with free RNase; conditioning on the number of such 'enzymatic events' (Figure 5.2, step 2) gives:

$$P(R_k) = \sum_{N} P(R_k|N)P(N)$$

Next, we define a set of cleavage patterns $\boldsymbol{C^N}$ where each $C_{\boldsymbol{i}}^N = [c_a, c_b, \dots]$ is a vector of length $\binom{l}{N}$ representing the positions that are cleaved by RNase (Figure 5.2, step 3). This gives:

$$P(R_k) = \sum_{N} \sum_{C_i} P(R_k|C_i)P(C_i|N)P(N)$$

The left-most term $P(R_k|C_i)$ is simply an indicator variable

$$P(R_k|C_i) = \begin{cases} 1, & \text{if fragment } R_k \text{ is contained in } C_i \\ 0, & \text{otherwise} \end{cases}$$

that represents the possibility of cloning and sequencing the given fragment if enzymatic cleavage were to occur at the specified positions (Figure 5.2, step 4).

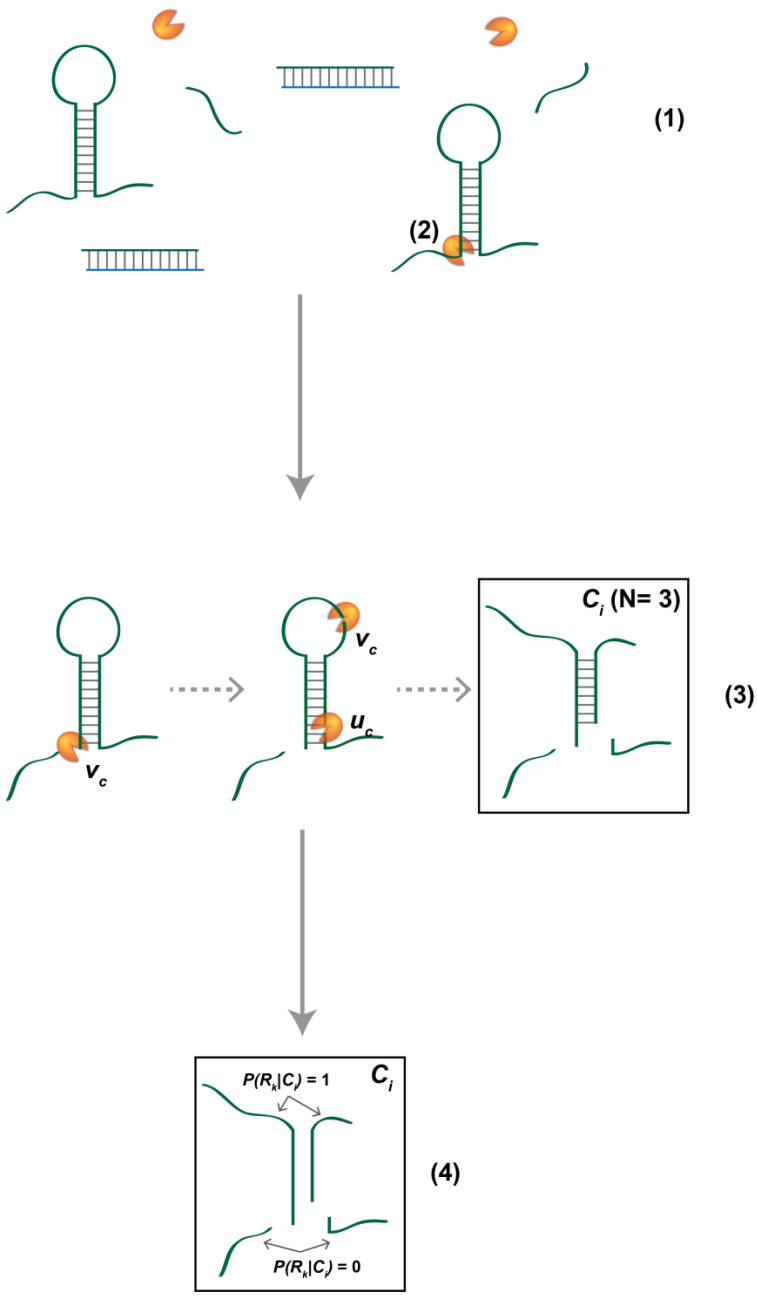


Figure 5.2: Generation of cloneable sequence fragments using the dsRNA-seq protocol. In a pool of RNA molecules and enzymes (1), cleavage events occur in a stochastic manner (2). The possible cleavage patterns $\boldsymbol{C^N}$ for a given number N of cleavage events per individual RNA molecule have probabilities that reflect the digestion rates at each of the N cleavage events (3). Finally, the probability of observing fragments is encoded as an indicator function (4) given the cleavage pattern C_i^N and the minimum and maximum fragment size (r_{min} and r_{max}).

$P(R_k|C_{\boldsymbol{i}})$ is also restricted by the parameters r_{min} and r_{max} as fragments that do not fall within the allowable size distribution are treated as non-cloneable. The middle term $P(C_{\boldsymbol{i}}|N)$ depends on the enzyme digestion rates $\boldsymbol{u}$ and $\boldsymbol{v}$ (where u_c and v_c are the probability of digestion occurring 3' of nucleotide c in the structure if the position is paired or unpaired, respectively) (Figure 5.2, step 3).

$$P(C_{\boldsymbol{i}}|N) = \prod_{i \in \{a,b,\ldots\}}^{N} D_i$$

where

$$D_i = s_i u_{c_i} + (1 - s_i) v_{c_i} \qquad \text{and} \qquad s = \begin{cases} 1, \text{if position } i \text{ is paired} \\ 0, \text{if position } i \text{ is unpaired} \end{cases}$$

The right-most term $P(N)$ describes the probability of having N enzymatic events per RNA molecule and is determined by the relative concentrations of RNA and RNase in the experimental setup. To summarize, the probability of observing a single read R_k is the sum of all cleavage pattern probabilities that yield a compatible fragment, and we obtain the final full probability of all reads $\boldsymbol{R}$ simply as the product of probabilities for all $R_k \in \boldsymbol{R}$.

The prior term $P(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max})$ is not well characterized and we therefore rely on several assumptions to arrive at a reasonable estimate. First, we treat all parameters as independent such that

$$P(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}) = P(\boldsymbol{s})P(\boldsymbol{u})P(\boldsymbol{v})P(N)P(r_{min}, r_{max})$$

Next, we assume that all valid secondary structures $\boldsymbol{s}$ are equally likely, although another possibility would be to use free energy-based methods to assign prior weights. For the digestion rates $\boldsymbol{u}$ and $\boldsymbol{v}$, we use the total number of read endpoints that fall on paired or unpaired positions within some subset of known secondary structures. Of note, good initial estimates of $\boldsymbol{u}$ and $\boldsymbol{v}$ are therefore inherently reliant on the presence of known structures in the dataset at hand. This can be addressed experimentally by inclusion of a spike-in RNA with a known secondary structure. We estimate the number of enzymatic events N by comparing the distribution of read lengths relative to the size of the initial full-length RNA molecule. Finally, r_{min} and r_{max} are set based on the actual fragment lengths that are excised during the experimental procedures.

5.2.2 Metropolis-Hastings implementation

Using the framework described above, we now turn to the task of sampling from the posterior distribution $P(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max} | \boldsymbol{R})$ by random walk Metropolis-Hastings on the parameter space $\{\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}$. We define a move set $\boldsymbol{M}$ that simply and comprehensively explores the entire parameter space (Table 5.1).

Move	Example	Parameter	Constraints
Add a pairing interaction	...((.....))... ↓ ..(((.....))))..	$s \rightarrow s^*$	Must be a valid base pair, follow steric hindrance rules, and result in a fully nested structure.
Delete a pairing interaction	...((.....))... ↓ ...(.....)...	$s \rightarrow s^*$	None

Table 5.1: Metropolis-Hastings move set

The ratio of the proposal density between two structures $\boldsymbol{s}$ and $\boldsymbol{s}^*$ is given by:

$$\frac{Q(\boldsymbol{s}|\boldsymbol{s}^*)}{Q(\boldsymbol{s}^*|\boldsymbol{s})} = \frac{I\big(\boldsymbol{M}(\boldsymbol{x} \rightarrow \boldsymbol{s})\big)}{I\big(\boldsymbol{M}(\boldsymbol{x} \rightarrow \boldsymbol{s}^*)\big)}$$

where $I\big(\boldsymbol{M}(\boldsymbol{x} \rightarrow \boldsymbol{s})\big)$ is the number of valid structures that can yield structure $\boldsymbol{s}$ in a single move. Note however that the symmetrical nature of the move set $\boldsymbol{M}$ allows us to calculate $I\big(\boldsymbol{M}(\boldsymbol{x} \rightarrow \boldsymbol{s})\big)$ simply as the number of valid moves from the structure $\boldsymbol{s}$. Taken together, we have the following pseudocode for the Metropolis-Hastings implementation:

1. Initialize $\{\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}$ based on prior distribution $P(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max})$
2. Generate candidate state $\{\boldsymbol{s}^*, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}$ using move set $\boldsymbol{M}$
3. Compute Metropolis-Hastings likelihood ratio:

$$L = \frac{P(\boldsymbol{R}|\boldsymbol{s}^*, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max})Q(\boldsymbol{s}|\boldsymbol{s}^*)}{P(\boldsymbol{R}|\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max})Q(\boldsymbol{s}^*|\boldsymbol{s})}$$

4. Assign new state $\{\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}$:

$$\{\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\} = \begin{cases} \{\boldsymbol{s}^*, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}, & \text{if } L > \text{runif}(0,1) \\ \{\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}\}, & \text{otherwise} \end{cases}$$

5. Repeat steps 2-4 until convergence or maximum number of iterations reached

The resultant samples from the posterior distribution $P(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, N, r_{min}, r_{max}|\boldsymbol{R})$ can be interpreted in terms of base pairing probabilities along our structure of interest. Formally, we define a base pairing probability vector $\boldsymbol{b}$ of length l:

$$\boldsymbol{b} = \sum_{t=\{0,1,\ldots,j\}} \frac{I\left(\boldsymbol{s}_{\boldsymbol{t_0+wt}}\right)}{j-1} \qquad I(s_t) = \begin{cases} 1, & s_t \text{ is paired} \\ 0, & s_t \text{ is unpaired} \end{cases}$$

where $I\left(\boldsymbol{s}_{\boldsymbol{t_0+wt}}\right)$ is an indicator on the structure $\boldsymbol{s}$ at time $t_0 + wt$, with t_0 and w being the burn-in period and sampling frequency, respectively. Conceptually, this is analogous to the base pairing probabilities that are derived from free energy models via a marginal distribution on the ensemble of secondary structures:

$$\boldsymbol{p} = \sum_{\boldsymbol{s} \in \boldsymbol{S}(x)} I(\boldsymbol{s}) p(\boldsymbol{s}|x)$$

where $I(\boldsymbol{s})$ is the indicator function described above and $p(\boldsymbol{s}|x)$ is a probability distribution on the set $\boldsymbol{S}(x)$ of all possible secondary structures for sequence x given by the Boltzmann distribution(23, 42, 72). However, we have replaced $p(\boldsymbol{s}|x)$ with a distribution of equally weighted MCMC samples at the time points $t_0 + wt$.

5.2.3 Generation of simulated sequencing datasets

To accurately infer secondary structure from dsRNA-seq and ssRNA-seq data using the described framework, two conditions (corresponding to the previous two sections) need to be met. First, the likelihood model must fit the actual experimental process of generating sequencing data. Additionally, the Metropolis-Hastings algorithm must converge on the desired target distribution (e.g. the true structure) within a reasonable number of iterations. We decided to initially address the second condition by implementing a read simulator with the exact likelihood

model used in the MCMC algorithm. In this ideal environment, sequencing noise and other variations in model fit are eliminated and we can focus exclusively on ability of our Metropolis-Hastings implementation to find the true secondary structure.

For a given structure $\boldsymbol{s}$ of length l, we define an $l \times l$ matrix $\boldsymbol{L}$ where L_{ij} is the probability of generating the fragment $[i, j]$. To set the values of $\boldsymbol{L}$, we iterate through all possible cleavage patterns C_i and increment the appropriate entries:

$$L_{ij} += \sum_{N} \sum_{C_{\boldsymbol{i}}} P\left(R_{ij} \middle| C_{\boldsymbol{i}}\right) P(C_{\boldsymbol{i}} | N) P(N)$$

where R_{ij} is the fragment $[i, j]$. Once the values of $\boldsymbol{L}$ are set, direct sampling can be used to generate faux sequencing reads that perfectly fit the likelihood model used in our algorithm.

5.3 Monte Carlo estimation of RNA secondary structure

With our simulation and inference framework in hand, we next turned to selection of an appropriate dataset on which to test RNA-seq-fold. Given the novelty of our approach, we wanted to limit the initial testing to RNAs with previously-determined secondary structures so that we could have a reasonable 'gold standard' with which to compare our inferred structures. We also wanted to test our algorithm on a reasonably complex mixture of structures with differences in stem and loop sizes and composition. To meet these criteria, we chose a total of eight non-coding RNA species (Table 5.2) with known secondary structures from the Rfam(11) and miRBase(50) databases.

RNA	Length (nt)	Structure source
U1 snRNA	158	Rfam (RF00003); based on chemical probing(51)
U3 snRNA	216	Rfam (RF00012); based on phylogenetic comparison(68)
U5 snRNA	114	Rfam (RF00020); based on

		phylogenetic comparison(102)
SNORD15 (U15) snoRNA	146	(104)
U22 snoRNA	126	(104)
U97 snoRNA	142	(104)
hsa-let-7a-1	80	miRBase (MI0000060)
hsa-mir-17	84	miRBase (MI0000071)

Table 5.2: Selected non-coding RNAs

The selected RNAs vary widely in terms of their overall size (80nt hsa-let-7a-1 to 216nt U3 snoRNA) and base pairing composition (Figure 5.3, compare the small loops in hsa-mir-17 to the large loops in U97 snoRNA) and therefore provide a wide spectrum of structures along which both the sensitivity and specificity of RNA-seq-fold can be tested.

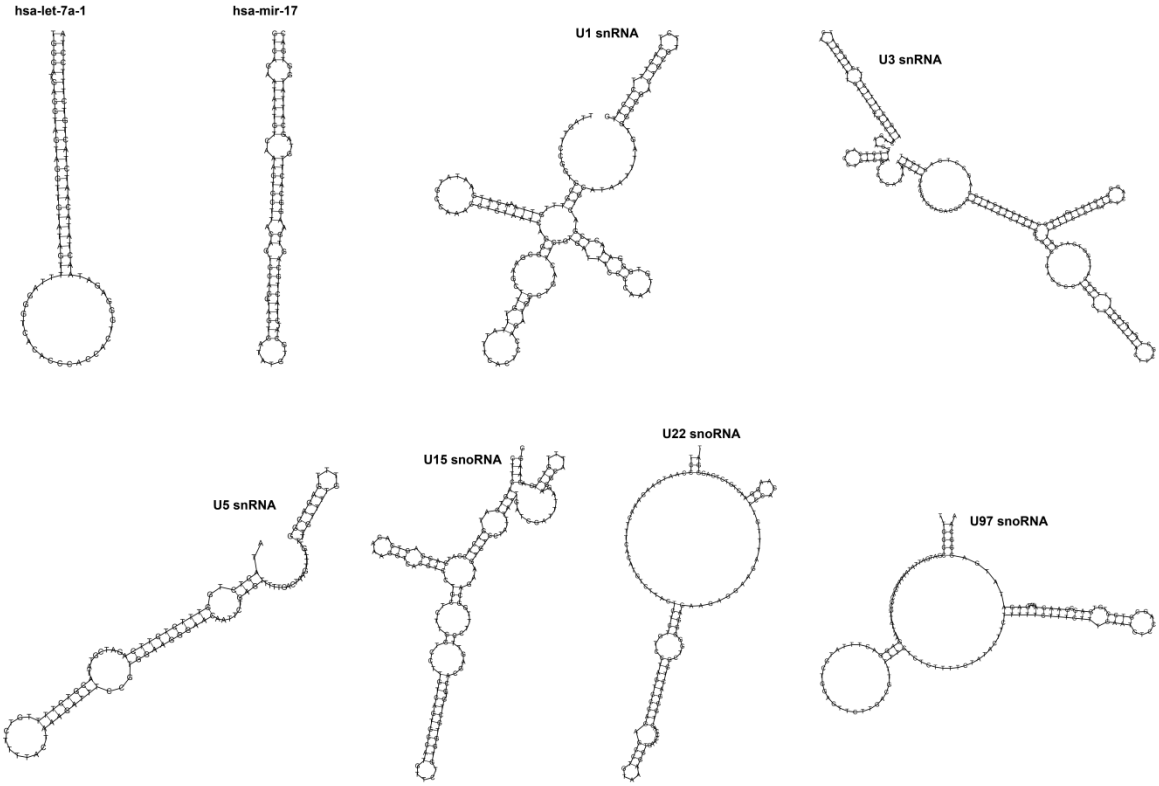


Figure 5.3: Known secondary structures for eight non-coding RNAs.

5.3.1 Simulation results

We used the simulator described in Section 5.2.3 to generate 100,000 simulated dsRNA-seq reads for each of the eight ncRNA loci under parameters that roughly approximate our observations during the experimental protocol (Table 5.3).

Parameter	Description	Value
s	Secondary structure	s_{known}
$u_{\{A,C,T,G\}}$	Digestion rates at paired positions	$\{0.04, 0.045, 0.05, 0.04\}$
$v_{\{A,C,T,G\}}$	Digestion rates at unpaired positions	$\{0.08, 0.09, 0.10, 0.08, \}$
N	Number of cleavage events per molecule	$\{1,2,3\}$
r_{min}	Minimum fragment size	10
r_{max}	Maximum fragment size	40

Table 5.3: Simulation parameters

We then ran RNA-seq-fold for 100,000 iterations using RNAfold structure predictions and re-estimated digestion parameters from all eight RNAs as the initial parameter values (Table 5.4).

Digestion rate	Original values	Re-estimated values
$u_{\{A,C,T,G\}}$	$\{0.04, 0.045, 0.05, 0.04\}$	$\{0.04, 0.045, 0.05, 0.04\}$
$v_{\{A,C,T,G\}}$	$\{0.08, 0.09, 0.10, 0.08, \}$	$\{0.082, 0.074, 0.088, 0.053\}$

Table 5.4: Re-estimated digestion rates from simulated data

Note that only one set of digestion rates is changed during the re-estimation process as we are simply comparing the ratio of digestion at paired versus unpaired positions. After discarding the first 10,000 iterations as burn-in, we computed the posterior base pairing probability $\boldsymbol{b}$ with a sampling frequency of 100 (e.g. using every 100th MCMC iteration) (Figure 5.4).

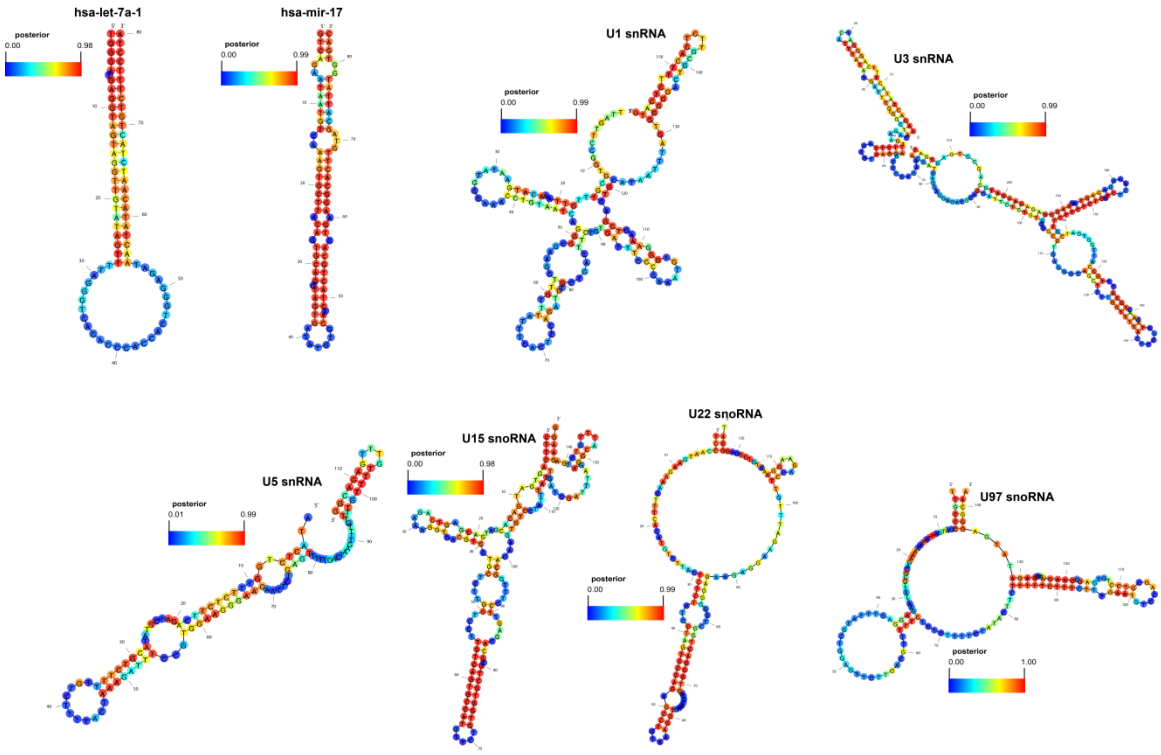


Figure 5.4: Base pairing posteriors estimated from simulated dsRNA-seq data. Shaded circles represent posterior values (as indicated by the color scale) drawn on the known secondary structure for each locus.

We also used the free energy-based methods RNAfold and RNAstructure to predict secondary structures for these loci (Figure 5.5).

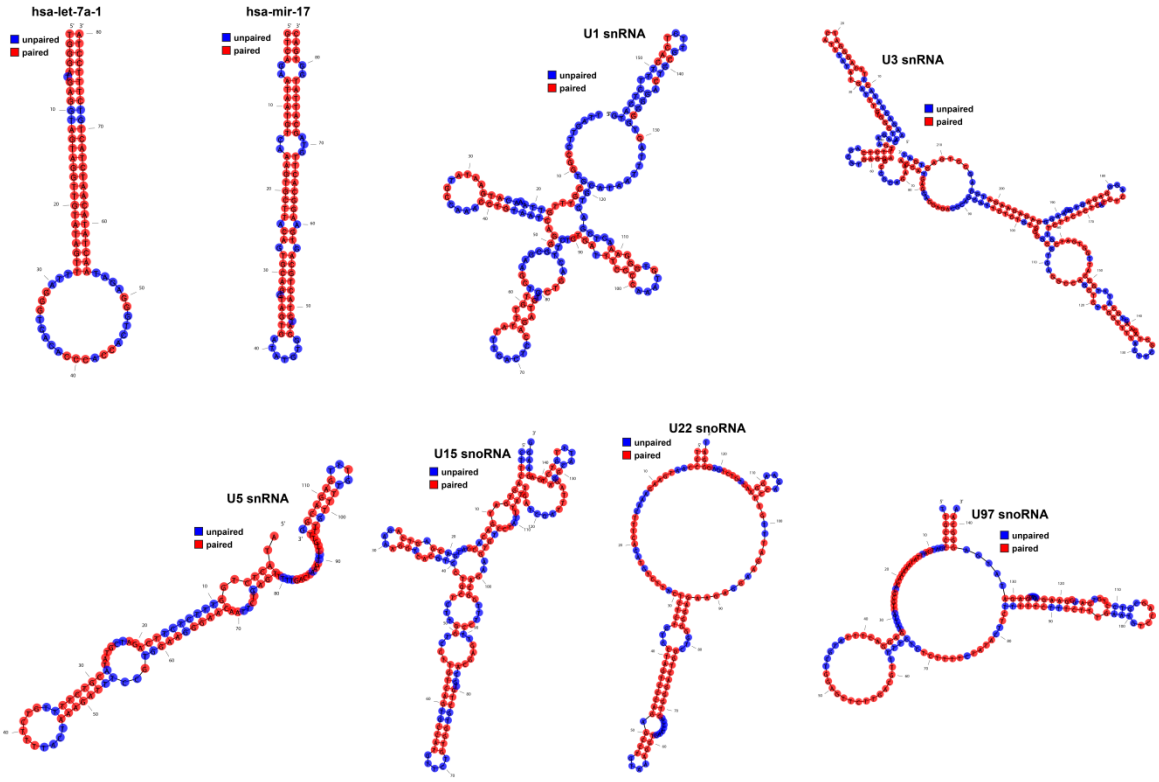


Figure 5.5: RNAfold predicted secondary structures overlayed with the known secondary structure for each locus. Red and blue circles indicate paired and unpaired positions, respectively.

To compare our method to the free energy-based structure predictions from these algorithms, we counted positions i where $b_i > thresh$ and $b_i \leq thresh$ as paired and unpaired, respectively, across a range of threshold values. For 7 of the 8 loci, RNA-seq-fold outperformed the free energy methods across almost the entire range of threshold values, with the lone exception being hsa-mir-17 whose structure was predicted perfectly by RNAfold and RNAstructure (Table 5.5).

Locus	Length (nt)	Num. correct	Num. incorrect	F-score	MCC	Method	$thresh$
hsa-let-7a-1	80	77 (96.2%)	3 (3.8%)	0.97	0.92	MCMC	0.5
		77 (96.2%)	3 (3.8%)	0.97	0.92	MCMC	0.55
		76 (95.0%)	4 (5.0%)	0.96	0.9	MCMC	0.6

		72 (90.0%)	8 (10.0%)	0.92	0.81	MCMC	0.65
		70 (87.5%)	10 (12.5%)	0.89	0.77	MCMC	0.7
		68 (85.0%)	12 (15.0%)	0.87	0.73	MCMC	0.75
		62 (77.5%)	18 (22.5%)	0.79	0.63	MCMC	0.8
		57 (71.2%)	23 (28.8%)	0.72	0.55	MCMC	0.85
		51 (63.8%)	29 (36.2%)	0.61	0.47	MCMC	0.9
		45 (56.2%)	35 (43.8%)	0.49	0.38	MCMC	0.95
		66 (82.5%)	14 (17.5%)	0.87	0.6	RNAfold	NA
		66 (82.5%)	14 (17.5%)	0.88	0.63	RNAstructure	NA
hsa-mir-17	84	73 (86.9%)	11 (13.1%)	0.91	0.66	MCMC	0.5
		73 (86.9%)	11 (13.1%)	0.91	0.66	MCMC	0.55
		75 (89.3%)	9 (10.7%)	0.93	0.73	MCMC	0.6
		74 (88.1%)	10 (11.9%)	0.92	0.7	MCMC	0.65
		74 (88.1%)	10 (11.9%)	0.92	0.7	MCMC	0.7
		75 (89.3%)	9 (10.7%)	0.93	0.74	MCMC	0.75
		73 (86.9%)	11 (13.1%)	0.9	0.72	MCMC	0.8
		72 (85.7%)	12 (14.3%)	0.89	0.72	MCMC	0.85
		67 (79.8%)	17 (20.2%)	0.84	0.64	MCMC	0.9
		55 (65.5%)	29 (34.5%)	0.69	0.48	MCMC	0.95
		84 (100.0%)	0 (0.0%)	1	1	RNAfold	NA
		84 (100.0%)	0 (0.0%)	1	1	RNAstructure	NA
U1_snRNA	158	124 (78.5%)	34 (21.5%)	0.78	0.57	MCMC	0.5
		123 (77.8%)	35 (22.2%)	0.77	0.56	MCMC	0.55
		122 (77.2%)	36 (22.8%)	0.75	0.54	MCMC	0.6
		117 (74.1%)	41 (25.9%)	0.71	0.48	MCMC	0.65
		115 (72.8%)	43 (27.2%)	0.68	0.46	MCMC	0.7
		118 (74.7%)	40 (25.3%)	0.68	0.52	MCMC	0.75
		114 (72.2%)	44 (27.8%)	0.63	0.48	MCMC	0.8
		112 (70.9%)	46 (29.1%)	0.6	0.47	MCMC	0.85
		111 (70.3%)	47 (29.7%)	0.57	0.47	MCMC	0.9
		95 (60.1%)	63 (39.9%)	0.31	0.29	MCMC	0.95
		92 (58.2%)	66 (41.8%)	0.54	0.16	RNAfold	NA

		90 (57.0%)	68 (43.0%)	0.6	0.15	RNAstructure	NA
U3_snRNA	216	189 (87.5%)	27 (12.5%)	0.89	0.75	MCMC	0.5
		189 (87.5%)	27 (12.5%)	0.89	0.75	MCMC	0.55
		179 (82.9%)	37 (17.1%)	0.85	0.67	MCMC	0.6
		176 (81.5%)	40 (18.5%)	0.83	0.65	MCMC	0.65
		174 (80.6%)	42 (19.4%)	0.81	0.65	MCMC	0.7
		169 (78.2%)	47 (21.8%)	0.79	0.62	MCMC	0.75
		162 (75.0%)	54 (25.0%)	0.74	0.58	MCMC	0.8
		153 (70.8%)	63 (29.2%)	0.68	0.55	MCMC	0.85
		147 (68.1%)	69 (31.9%)	0.64	0.51	MCMC	0.9
		120 (55.6%)	96 (44.4%)	0.41	0.35	MCMC	0.95
		124 (57.4%)	92 (42.6%)	0.65	0.11	RNAfold	NA
		128 (59.3%)	88 (40.7%)	0.66	0.15	RNAstructure	NA
U5_snRNA	114	93 (81.6%)	21 (18.4%)	0.82	0.63	MCMC	0.5
		91 (79.8%)	23 (20.2%)	0.8	0.6	MCMC	0.55
		90 (78.9%)	24 (21.1%)	0.79	0.58	MCMC	0.6
		90 (78.9%)	24 (21.1%)	0.78	0.59	MCMC	0.65
		85 (74.6%)	29 (25.4%)	0.72	0.51	MCMC	0.7
		83 (72.8%)	31 (27.2%)	0.69	0.49	MCMC	0.75
		80 (70.2%)	34 (29.8%)	0.65	0.45	MCMC	0.8
		79 (69.3%)	35 (30.7%)	0.62	0.45	MCMC	0.85
		73 (64.0%)	41 (36.0%)	0.52	0.38	MCMC	0.9
		65 (57.0%)	49 (43.0%)	0.35	0.27	MCMC	0.95
		62 (54.4%)	52 (45.6%)	0.62	0.07	RNAfold	NA
		66 (57.9%)	48 (42.1%)	0.65	0.15	RNAstructure	NA
U15_snoRNA	146	128 (87.7%)	18 (12.3%)	0.89	0.75	MCMC	0.5
		128 (87.7%)	18 (12.3%)	0.89	0.75	MCMC	0.55
		125 (85.6%)	21 (14.4%)	0.87	0.71	MCMC	0.6
		124	22 (15.1%)	0.86	0.7	MCMC	0.65

		(84.9%)					
		120 (82.2%)	26 (17.8%)	0.84	0.64	MCMC	0.7
		119 (81.5%)	27 (18.5%)	0.82	0.64	MCMC	0.75
		114 (78.1%)	32 (21.9%)	0.78	0.59	MCMC	0.8
		109 (74.7%)	37 (25.3%)	0.73	0.54	MCMC	0.85
		96 (65.8%)	50 (34.2%)	0.59	0.42	MCMC	0.9
		92 (63.0%)	54 (37.0%)	0.53	0.4	MCMC	0.95
		92 (63.0%)	54 (37.0%)	0.69	0.24	RNAfold	NA
		100 (68.5%)	46 (31.5%)	0.74	0.35	RNAstructure	NA
U22_snoRNA	126	91 (72.2%)	35 (27.8%)	0.73	0.54	MCMC	0.5
		88 (69.8%)	38 (30.2%)	0.7	0.47	MCMC	0.55
		95 (75.4%)	31 (24.6%)	0.74	0.55	MCMC	0.6
		95 (75.4%)	31 (24.6%)	0.74	0.53	MCMC	0.65
		98 (77.8%)	28 (22.2%)	0.75	0.56	MCMC	0.7
		100 (79.4%)	26 (20.6%)	0.75	0.58	MCMC	0.75
		101 (80.2%)	25 (19.8%)	0.75	0.59	MCMC	0.8
		102 (81.0%)	24 (19.0%)	0.76	0.6	MCMC	0.85
		100 (79.4%)	26 (20.6%)	0.72	0.56	MCMC	0.9
		88 (69.8%)	38 (30.2%)	0.47	0.35	MCMC	0.95
		90 (71.4%)	36 (28.6%)	0.74	0.55	RNAfold	NA
		92 (73.0%)	34 (27.0%)	0.75	0.57	RNAstructure	NA
U97_snoRNA	142	110 (77.5%)	32 (22.5%)	0.73	0.57	MCMC	0.5
		113 (79.6%)	29 (20.4%)	0.75	0.6	MCMC	0.55
		115 (81.0%)	27 (19.0%)	0.77	0.62	MCMC	0.6
		117 (82.4%)	25 (17.6%)	0.77	0.64	MCMC	0.65
		117 (82.4%)	25 (17.6%)	0.77	0.64	MCMC	0.7
		117 (82.4%)	25 (17.6%)	0.76	0.62	MCMC	0.75
		117 (82.4%)	25 (17.6%)	0.75	0.61	MCMC	0.8

		118 (83.1%)	24 (16.9%)	0.74	0.62	MCMC	0.85
		117 (82.4%)	25 (17.6%)	0.7	0.61	MCMC	0.9
		109 (76.8%)	33 (23.2%)	0.54	0.48	MCMC	0.95
		70 (49.3%)	72 (50.7%)	0.49	0.08	RNAfold	NA
		74 (52.1%)	68 (47.9%)	0.44	0.05	RNAstructure	NA

Table 5.5: Comparison of RNA-seq-fold and free energy-based methods with simulated data. MCC = Matthews correlation coefficient.

Notably, our method outperformed RNAfold and RNAstructure by a substantial margin on U97_snoRNA, likely due to extraneous base pairing in the large loops that is favored by a free energy minimization method (compare U97_snoRNA in Figures 5.4 and 5.5).

RNA-seq-fold, as with most MCMC algorithms, is computationally demanding due to the large parameter space and stochastic nature of its exploration. This expense is further multiplied by the fact that likelihood estimation in this case cannot be written in a closed form and therefore must be computed numerically. In fact, the number of contributions to the likelihood term

$$P(R_k) = \sum_{N} \sum_{C_i} P(R_k|C_i)P(C_i|N)P(N)$$

grows as $\binom{l}{N}$ where l is the length of the RNA molecule and N is the number of cleavage events per locus. Using the hsa-mir-17 locus as a test case, we analyzed the running time of RNA-seq-fold under a variety of conditions (Table 5.6).

	dsRNA-seq replicate 1	dsRNA-seq replicate 2
hsa-let-7a-1	35,538	64,665
hsa-mir-17	38,967	101,251
U1_snRNA	375,922	764,652
U3_snRNA	986,115	1,531,236

U5_snRNA	150,877	241,020
U15_snoRNA	222,781	250,871
U22_snoRNA	142,825	201,929
U97_snoRNA	1,527,835	2,073,011

Table 5.6: Number of mapped reads per locus

As expected, running time scaled linearly with the number of MCMC iterations whereas read depth had little effect on the computational expense. The choice of $N = \{1,2,3\}, \{1,2,3,4\}$, $\text{and } \{1,2,3,4,5\}$ demonstrated near-factorial growth due to the number of terms in the likelihood calculation; therefore, we chose to limit our subsequent analyses with the condition $N = \{1,2,3\}$. In the future, optimization of the likelihood calculation should enable this constraint to be dropped (see Discussion).

Another major consideration in MCMC approaches is chain convergence (i.e. if the sampling distribution approximates the target distribution within some error tolerance). To assess convergence, we computed the posterior base pairing probability $\boldsymbol{b}$ from successively shorter MCMC chains and then compared the performance of these subsampled chains to that of the full length posterior. We observed almost prediction accuracy at 10% of the original chain length (Figure 5.6), suggesting that RNA-seq-fold converges rapidly to the most likely secondary structure.

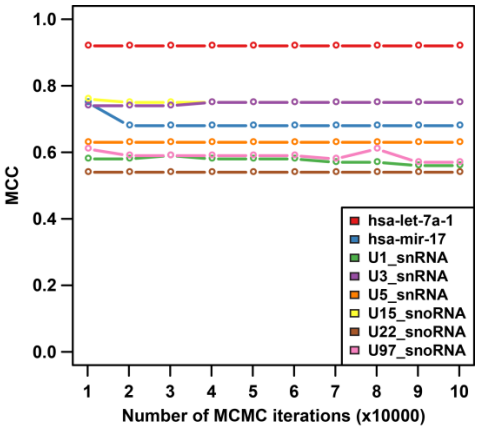


Figure 5.6: Convergence of RNA-seq-fold with simulated dsRNA-seq data. MCC (y-axis) is plotted against chain length (x-axis) for each locus as indicated in the legend.

We also examined the effect of sequencing depth on prediction accuracy by running on RNA-seq-fold with subsampled dsRNA-seq read data. As with chain length, sequencing depth appeared to have little to no effect on performance (Figure 5.7), although some locus-dependent variation was observed.

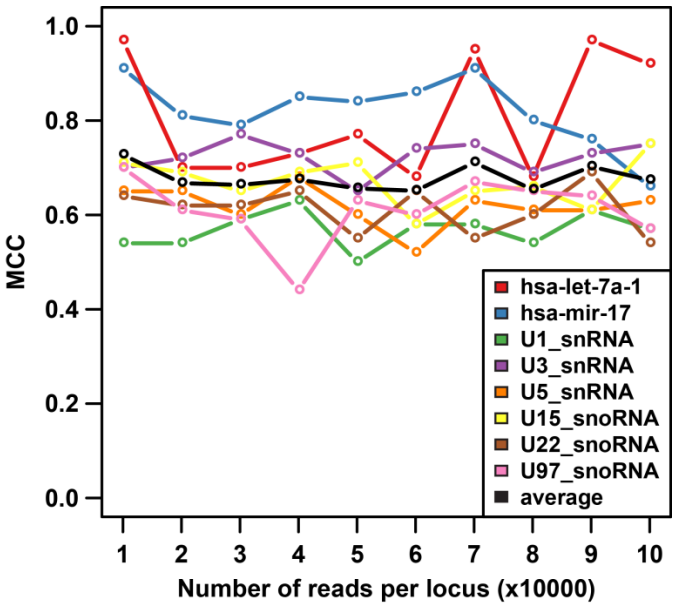


Figure 5.7: Power analysis of RNA-seq-fold. MCC (y-axis) is plotted against sequencing depth (x-axis) for each locus as indicated, as well as the average across all loci (black line).

It is possible that these differences would be minimized by additional sampling trials, but we did not test this hypothesis due to computational limitations. In general, our simulations demonstrated consistent and reliable inference of known secondary structures using the RNA-seq-fold framework across a range of parameters.

5.3.2 Structure determination of eight *in vitro* transcribed non-coding RNAs

Given the promising results achieved with our simulated data, we next set out to test RNA-seq-fold on real data generated by performing a modified dsRNA-seq protocol (see Section 5.5) on a pool of the eight selected RNAs. Two independent replicates yielded an average of ~544,000 mapped reads per locus per replicate (Table 5.6), with no locus having fewer than ~35,000 reads. An initial diagnostic analysis of enzyme digestion rates revealed a surprisingly high level of noise with little separation between paired and unpaired positions (Table 5.7, compare v values to those in Table 5.4).

Digestion rate	Estimated values
$u_{\{A,C,T,G\}}$	$\{0.04, 0.045, 0.05, 0.04\}$
$v_{\{A,C,T,G\}}$	$\{0.055, 0.052, 0.079, 0.062\}$

Table 5.7: Estimated digestion rates from experimental data. Note that u is arbitrarily fixed as the baseline digestion rate and cannot be directly estimated from sequencing data.

However, the relative digestion rates trended according to the known enzyme specificities, suggesting that we could still distinguish the pairing status of each nucleotide position based on the pattern of cleavage events. We also examined the distribution of read endpoints in our experimental data and found a significant bias due to nonlinear PCR amplification (Figure 5.8). Therefore, to offset the exponential clonal amplification that resulted from the PCR step, we used a $\log_2$ transform on our mapped read counts for all subsequent analyses.

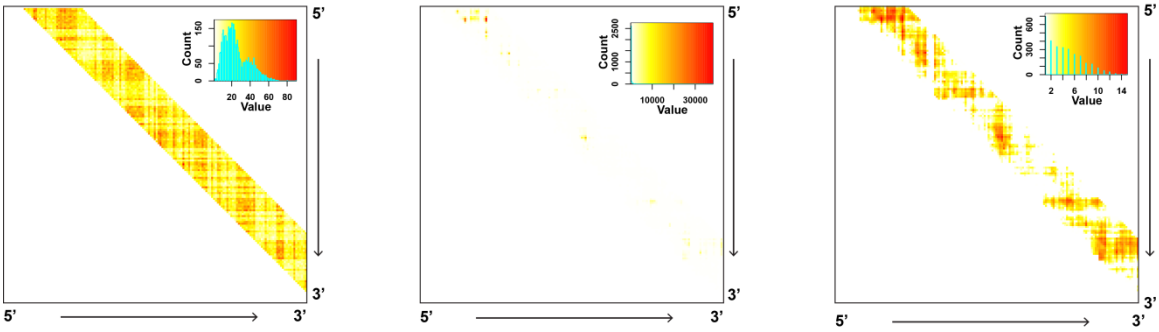


Figure 5.8: Distribution of read endpoints from simulated data (left), raw experimental data (middle), and $\log_2$ transformed experimental data (right). Each cell in the heatmap represents the number of reads whose 5' and 3' endpoints are located at the column and row values, respectively. Data are shown for the U1_snRNA locus as a representative example.

As with the simulated data, we ran RNA-seq-fold for 100,000 iterations and then calculated the base pairing posterior probabilities $\boldsymbol{b}$ following a burn-in period of 10,000 iterations and with a sampling frequency of 100 (Figure 5.9).

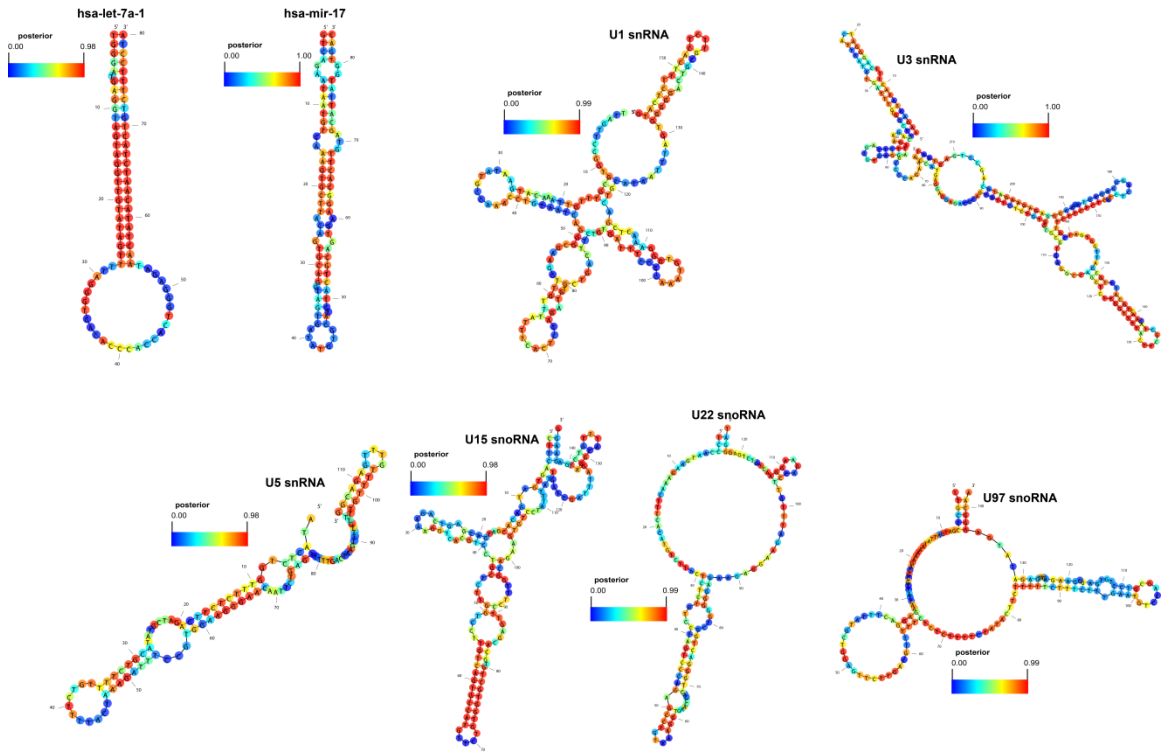


Figure 5.9: Base pairing posteriors estimated from *in vitro* dsRNA-seq data. Shaded circles represent posterior values (as indicated by the color scale) drawn on the known secondary structure for each locus.

Using the same thresholding approach described for simulated data, we found marginal to no improvement of our method versus free energy-based predictions (Table 5.8).

Locus	Length (nt)	Num. correct	Num. incorrect	F-score	MCC	Method	$thresh$
hsa-let-7a-1	80	64 (80.0%)	16 (20.0%)	0.85	0.55	MCMC	0.5
		63 (78.8%)	17 (21.2%)	0.84	0.52	MCMC	0.55
		63 (78.8%)	17 (21.2%)	0.84	0.52	MCMC	0.6
		64 (80.0%)	16 (20.0%)	0.85	0.55	MCMC	0.65
		65 (81.2%)	15 (18.8%)	0.85	0.59	MCMC	0.7

		65 (81.2%)	15 (18.8%)	0.85	0.59	MCMC	0.75
		63 (78.8%)	17 (21.2%)	0.83	0.55	MCMC	0.8
		64 (80.0%)	16 (20.0%)	0.84	0.58	MCMC	0.85
		58 (72.5%)	22 (27.5%)	0.77	0.46	MCMC	0.9
		51 (63.8%)	29 (36.2%)	0.65	0.37	MCMC	0.95
		66 (82.5%)	14 (17.5%)	0.87	0.6	RNAfold	NA
		66 (82.5%)	14 (17.5%)	0.88	0.63	RNAstructure	NA
hsa-mir-17	84	60 (71.4%)	24 (28.6%)	0.78	0.44	MCMC	0.5
		60 (71.4%)	24 (28.6%)	0.78	0.44	MCMC	0.55
		61 (72.6%)	23 (27.4%)	0.78	0.5	MCMC	0.6
		60 (71.4%)	24 (28.6%)	0.77	0.49	MCMC	0.65
		59 (70.2%)	25 (29.8%)	0.76	0.47	MCMC	0.7
		59 (70.2%)	25 (29.8%)	0.76	0.47	MCMC	0.75
		54 (64.3%)	30 (35.7%)	0.69	0.41	MCMC	0.8
		55 (65.5%)	29 (34.5%)	0.7	0.45	MCMC	0.85
		47 (56.0%)	37 (44.0%)	0.58	0.35	MCMC	0.9
		36 (42.9%)	48 (57.1%)	0.37	0.27	MCMC	0.95
		84 (100.0%)	0 (0.0%)	1	1	RNAfold	NA
		84 (100.0%)	0 (0.0%)	1	1	RNAstructure	NA
U1_snRNA	158	85 (53.8%)	73 (46.2%)	0.59	0.09	MCMC	0.5
		87 (55.1%)	71 (44.9%)	0.59	0.11	MCMC	0.55

		85 (53.8%)	73 (46.2%)	0.57	0.08	MCMC	0.6
		83 (52.5%)	75 (47.5%)	0.53	0.05	MCMC	0.65
		84 (53.2%)	74 (46.8%)	0.53	0.06	MCMC	0.7
		84 (53.2%)	74 (46.8%)	0.52	0.06	MCMC	0.75
		84 (53.2%)	74 (46.8%)	0.49	0.06	MCMC	0.8
		84 (53.2%)	74 (46.8%)	0.46	0.06	MCMC	0.85
		85 (53.8%)	73 (46.2%)	0.39	0.07	MCMC	0.9
		85 (53.8%)	73 (46.2%)	0.26	0.07	MCMC	0.95
		92 (58.2%)	66 (41.8%)	0.54	0.16	RNAfold	NA
		90 (57.0%)	68 (43.0%)	0.6	0.15	RNAstructure	NA
U3_snRNA	216	124 (57.4%)	92 (42.6%)	0.65	0.1	MCMC	0.5
		121 (56.0%)	95 (44.0%)	0.63	0.08	MCMC	0.55
		119 (55.1%)	97 (44.9%)	0.62	0.08	MCMC	0.6
		117 (54.2%)	99 (45.8%)	0.6	0.06	MCMC	0.65
		120 (55.6%)	96 (44.4%)	0.6	0.11	MCMC	0.7
		126 (58.3%)	90 (41.7%)	0.6	0.19	MCMC	0.75
		126	90 (41.7%)	0.59	0.21	MCMC	0.8

		(58.3%)					
		121 (56.0%)	95 (44.0%)	0.54	0.19	MCMC	0.85
		109 (50.5%)	107 (49.5%)	0.44	0.11	MCMC	0.9
		104 (48.1%)	112 (51.9%)	0.35	0.11	MCMC	0.95
		124 (57.4%)	92 (42.6%)	0.65	0.11	RNAfold	NA
		128 (59.3%)	88 (40.7%)	0.66	0.15	RNAstructure	NA
U5_snRNA	114	65 (57.0%)	49 (43.0%)	0.63	0.13	MCMC	0.5
		67 (58.8%)	47 (41.2%)	0.62	0.17	MCMC	0.55
		71 (62.3%)	43 (37.7%)	0.64	0.24	MCMC	0.6
		70 (61.4%)	44 (38.6%)	0.63	0.23	MCMC	0.65
		67 (58.8%)	47 (41.2%)	0.58	0.18	MCMC	0.7
		68 (59.6%)	46 (40.4%)	0.55	0.21	MCMC	0.75
		67 (58.8%)	47 (41.2%)	0.53	0.2	MCMC	0.8
		67 (58.8%)	47 (41.2%)	0.51	0.21	MCMC	0.85
		62 (54.4%)	52 (45.6%)	0.4	0.14	MCMC	0.9
		59 (51.8%)	55 (48.2%)	0.18	0.17	MCMC	0.95
		62 (54.4%)	52 (45.6%)	0.62	0.07	RNAfold	NA
		66 (57.9%)	48 (42.1%)	0.65	0.15	RNAstructure	NA

U15_snoRNA	146	90 (61.6%)	56 (38.4%)	0.67	0.22	MCMC	0.5
		92 (63.0%)	54 (37.0%)	0.67	0.25	MCMC	0.55
		91 (62.3%)	55 (37.7%)	0.65	0.25	MCMC	0.6
		90 (61.6%)	56 (38.4%)	0.63	0.24	MCMC	0.65
		90 (61.6%)	56 (38.4%)	0.62	0.25	MCMC	0.7
		93 (63.7%)	53 (36.3%)	0.63	0.3	MCMC	0.75
		90 (61.6%)	56 (38.4%)	0.6	0.27	MCMC	0.8
		86 (58.9%)	60 (41.1%)	0.55	0.22	MCMC	0.85
		86 (58.9%)	60 (41.1%)	0.54	0.24	MCMC	0.9
		75 (51.4%)	71 (48.6%)	0.36	0.13	MCMC	0.95
		92 (63.0%)	54 (37.0%)	0.69	0.24	RNAfold	NA
		100 (68.5%)	46 (31.5%)	0.74	0.35	RNAstructure	NA
U22_snoRNA	126	73 (57.9%)	53 (42.1%)	0.57	0.2	MCMC	0.5
		77 (61.1%)	49 (38.9%)	0.59	0.25	MCMC	0.55
		77 (61.1%)	49 (38.9%)	0.57	0.23	MCMC	0.6
		78 (61.9%)	48 (38.1%)	0.56	0.23	MCMC	0.65
		79 (62.7%)	47 (37.3%)	0.54	0.23	MCMC	0.7
		76 (60.3%)	50 (39.7%)	0.47	0.15	MCMC	0.75
		77 (61.1%)	49 (38.9%)	0.46	0.16	MCMC	0.8
		79 (62.7%)	47 (37.3%)	0.46	0.19	MCMC	0.85
		73 (57.9%)	53 (42.1%)	0.29	0.04	MCMC	0.9
		73 (57.9%)	53 (42.1%)	0.23	0.02	MCMC	0.95

		90 (71.4%)	36 (28.6%)	0.74	0.55	RNAfold	NA
		92 (73.0%)	34 (27.0%)	0.75	0.57	RNAstructure	NA
U97_snoRNA	142	61 (43.0%)	81 (57.0%)	0.42	-0.07	MCMC	0.5
		60 (42.3%)	82 (57.7%)	0.4	-0.1	MCMC	0.55
		62 (43.7%)	80 (56.3%)	0.4	-0.08	MCMC	0.6
		62 (43.7%)	80 (56.3%)	0.39	-0.09	MCMC	0.65
		65 (45.8%)	77 (54.2%)	0.38	-0.07	MCMC	0.7
		65 (45.8%)	77 (54.2%)	0.33	-0.12	MCMC	0.75
		68 (47.9%)	74 (52.1%)	0.29	-0.12	MCMC	0.8
		75 (52.8%)	67 (47.2%)	0.29	-0.06	MCMC	0.85
		74 (52.1%)	68 (47.9%)	0.19	-0.14	MCMC	0.9
		83 (58.5%)	59 (41.5%)	0.17	-0.06	MCMC	0.95
		70 (49.3%)	72 (50.7%)	0.49	0.08	RNAfold	NA

Table 5.8: Comparison of RNA-seq-fold and free energy-based methods with *in vitro* data. MCC = Matthews correlation coefficient

This unexpected result is likely due to several factors including overdigestion and local RNA folding (see Section 5.4 for detailed discussion).

5.4 Discussion

In this chapter, we developed a Bayesian Markov chain Monte Carlo (MCMC) approach to infer secondary structure from the dsRNA-seq and ssRNA-seq protocols. We tested our likelihood model and estimator on simulated sequencing data from eight non-coding loci with

known secondary structure and found a substantial improvement in prediction accuracy versus free energy-based methods. However, analysis of dsRNA-seq data generated from *in vitro* transcribed RNA showed only marginally better performance. We propose several possibilities for our findings and suggest alternative approaches that may address these issues in future studies.

Based on our past experiences, we decided here to size select fragments from 10-40 nucleotides in length after RNase treatment. In retrospect, given the size range of the full-length RNA molecules (80-216 nt), the selected fragments are likely the result of multiple cleavage events per RNA molecule. We hypothesize that these experimental conditions have resulted in nonspecific overdigestion at positions that do not necessarily reflect the structure-sensitive nature of the RNase used. Future studies to determine the specificity of RNase ONE as a function of its concentration and digestion time are needed to test this idea. Of note, we do not expect anticipate large-scale conformational changes to occur as a result of sequential cleavage events(109) as long as these cleavages occur in single-stranded regions. The reasoning here is that such events are unlikely to cause spontaneous unfolding of base paired regions, although the converse is probably not true.

Another explanation for the lack of agreement between our predictions and the gold standard structures is simply that they were obtained under different conditions. Importantly, the three snoRNA structures are based on *in vitro* transcribed and denatured RNA with subsequent renaturation(104), whereas we did not denature our transcription products before enzyme treatment. It is possible that our data reflect a conformation that is suboptimal on the global structure landscape, but rather forms as a result of co-transcriptional folding(3, 76, 119). To address this possibility, future experiments should be performed on renatured and non-renatured RNA popluations to specifically interrogate the differences between global and local RNA folding pathways.

In these initial studies, we utilized simple parallelization of individual MCMC chains to offset the computational expense of RNA-seq-fold. However, application of our approach to longer RNAs such as mRNAs will require more extensive measures to ensure convergence within a reasonable time frame. Empirically, we observed $\binom{l}{N}$ growth in the computational cost as a

function of RNA length l and the number of enzymatic cleavage events N. Fortunately, dynamic programming can be used to reduce growth to a manageable polynomial function. Such an approach works because any given problem of size (N, l) is reducible to two subproblems of size $(N_1, l - i)$ and (N_2, i) where $N_1 + N_2 = N$ and i is the position of the Nth cleavage event (Figure 5.10).

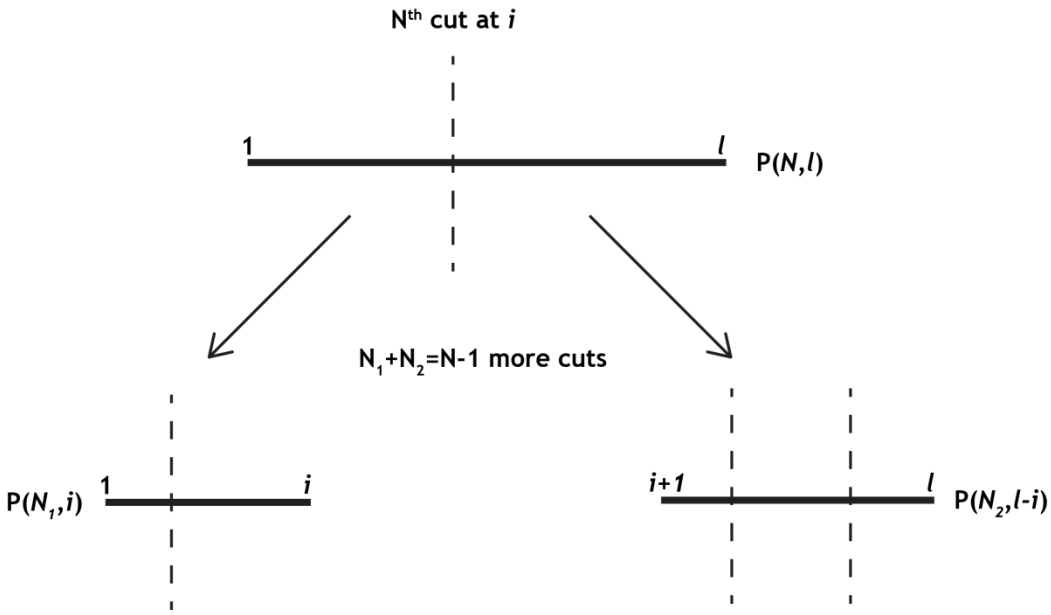


Figure 5.10: A dynamic programming approach to RNA-seq-fold. The problem of N cleavage events along an RNA of length l is reducible to subproblems for each of the two fragments generated by the Nth cleavage event. $N = 4$, $N_1 = 1$, and $N_2 = 2$ in this example, with cleavage events marked by dotted lines.

5.5 Materials and methods

In vitro transcription

Sequence-specific primers with a T7 promoter (Table 5.9) were designed for the eight selected ncRNA loci and used to selectively amplify these regions from genomic DNA (gDNA). These PCR products were then transcribed using an *in vitro* system.

Locus	Primers
U1_snRNA	Forward: TAATACGACTCACTATAGGTTAGTTCCGGTGCGTTTGTT Reverse: CATGAGAAAGTGAGAACGCAGT
U3_snRNA	Forward: TAATACGACTCACTATAGGAAGACTATACTTTCAGGGATCATTTAT Reverse: ATCACTCAGGCTGCATCTT
U5_snRNA	Forward: TAATACGACTCACTATAGGATACTCTGGTTTCTCTTCAGATCGT Reverse: CCGTCTCAAACAAAACAAAAC
U15_snoRNA	Forward: TAATACGACTCACTATAGGCTTCAGTGATGACACGATGACG Reverse: CCTTCTCAGACAAATGCCTCTAAAT
U22_snoRNA	Forward: TAATACGACTCACTATAGGTCCCAATGAAGAAACTTTCAC Reverse: ATCCCTCAGACAGTTCCTTCT
U97_snoRNA	Forward: TAATACGACTCACTATAGGTTGCCCGATGATTATAAAAAGAC Reverse: TTGCCCTCATATCTCATAATCTTC
hsa-let-7a-1	Forward: TAATACGACTCACTATAGGTGGGATGAGGTAGTAGGTTGTATAG Reverse: TAGGAAAGACAGTAGATTGTATAGTTATCTC
hsa-mir-17	Forward: TAATACGACTCACTATAGGGTCAGAATAATGTCAAAGTGCTTACA Reverse: GTCACCATAATGCTACAAGTGC

Table 5.9: Primers used to amplify selected ncRNA loci. Note that the forward primers contain the T7 promoter sequence.

A detailed protocol follows:

I. Start with 0.2 μg of genomic DNA, suspended in 12 μL nuclease-free water.

II. **PCR amplification**

a. Add genomic DNA sample, 2 μL 10X Ex Taq buffer, 1.6 μL 25mM $MgCl_2$, 1.6 μL 2.5mM dNTP mix, 0.1 μL Ex Taq, 1 μL forward primer, and 1 μL reverse primer to a sterile, nuclease-free PCR tube. Note: Ex Taq is available from http://www.millipore.com/catalogue/item/RR001A.

b. PCR amplification program in thermal cycler:

i. 98°C for 30 seconds

ii. 98°C for 10 seconds

iii. 61°C for 30 seconds

iv. 72°C for 15 seconds

v. Cycle to step ii 24X

vi. 72°C for 10 minutes

vii. Hold at 4°C

c. Recover product using a PCR purification kit (e.g. QIAquick PCR Purification Kit).

d. Resuspend PCR product in 11.5 μL DEPC-treated water and quantify.

III. ***In vitro* transcription**

a. Aliquot 1 μg of PCR template into a new sterile, nuclease-free PCR tube. Add sufficient DEPC-treated water to bring total volume up to 154 μL.

b. Add 8 μL 25mM rNTP mix, 20 μL 10X transcription buffer (e.g. 500mM Tris-HCl pH 7.5, 150mM $MgCl_2$, 50mM DTT, 20mM spermidine), 10 μL 2μg/μL acetylated BSA, 4 μL RNaseOUT, and 4 μL T7 RNA polymerase (https://www.neb.com/products/m0251-t7-rna-polymerase). Mix thoroughly.

c. Incubate at 37°C for 4 hours.

d. Add 4 μL Turbo DNase (http://www.lifetechnologies.com/order/catalog/product/AM2238) and incubate at 37°C for an additional 30 minutes.

e. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 2 μL glycogen, and 1000 μL 100% EtOH.

f. Resuspend in 10 μL DEPC-treated water.

IV. **Gel purification**

a. Prepare 1000 mL 1X TBE running buffer (100 mL 10X TBE extended range + 900 mL Milli-Q water).

b. Pre-run 15% TBE-Urea polyacrylamide gel (e.g. from Invitrogen) for 25 minutes at 155 V.

c. While gel is pre-running, prepare ladder and sample:

    i. Ladder: 1.5 μL 10bp DNA ladder, 8.5 μL DEPC-treated water, and 10 μL Gel Loading Buffer (e.g. from NEB).

    ii. Add 10 μL Gel Loading Buffer to sample.

    iii. Place sample (but not ladder) at 70°C for 5 minutes, followed by 3 minutes on ice.

d. After pre-run is complete, run ladder and sample at 155 V for approximately 1.5 hours.

e. Stain gel with ethidium bromide. Add 14 μL 10 mg/mL ethidium bromide to 200 mL 1X TBE buffer in a clean RNase-free tray. Add gel and rock gently for 10 minutes.

f. Cut 20-100bp band from gel and place gel slice in a 0.5mL tube with holes (e.g. Gel Breaker Tubes #3388-100 from IST Engineering Inc.), placed inside a clean 2mL tube.

g. Spin sample at 14000RPM, 4°C for 2 minutes. Repeat until all of the gel goes through the 0.5mL tube.

h. Add 300 μL 0.3M NaCl and rotate for 4 hours.

i. Pipette entire sample into a Spin-X column and spin at 14000RPM, 4°C for 2 minutes. Transfer eluent to new 1.5mL tube.

j. Precipitate by adding 30 μL 3M NaOAc (pH 5.5), 3 μL glycogen, and 900 μL 100% EtOH.

k. Resuspend in 21.5 μL DEPC-treated water and quantify.

RNase ONE and RNase V1 treatment

RNase digestions and subsequent library preparations were performed as described in Section 2.4 with the following modifications.

- 0.1 μg of each of the eight transcribed RNAs was combined for a total of 0.8 μg of starting RNA.
- Digestions were performed with 1 μL of 0.3 U/μL (3:10 dilution of manufacturer stock) RNase ONE and 1 μL of 0.004 U/μL (1:250 dilution of manufacturer stock) RNase V1, respectively. These concentrations were selected by extensive testing of enzyme dilutions to achieve the desired digestion fragment sizes of 10-40nt.
- No RiboMinus or fragmentation was performed.

Libraries were sequenced on a single lane of an Illumina HiSeq 2000 to a length of 100 bases.

Data processing and mapping

Adapter sequences were removed with cutadapt -a TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCATGGCATCTCGTATGCCGTCTTCTGCTTG -e 0 -O 63 -m 6, which required a perfect adapter sequence to be matched at the 3' end of each sequence. Trimmed reads were then mapped using bowtie with options '-v 0 -m 1 -y --norc --all --best –strata'. To remove PCR amplification biases, we used a $\log_2$ transform on the mapped read counts (rounding up to the nearest integer value).

Estimation of enzyme efficiency

Inference of base pairing status is based on the differential sensitivity of paired versus unpaired positions to the specific ribonuclease used, which can be estimated by simply counting the ratio of read endpoints that fall in paired and unpaired positions according to the gold standard structure. Therefore, for fixed values of $\boldsymbol{u}$, we estimated $\boldsymbol{v}$ as:

$$v_c = \left(\frac{e_c^u}{e_c^p}\right) u_c$$

where e_c^p and e_c^u are the number of read endpoints that fall in paired and unpaired positions with the given nucleotide c, respectively.

RNA-seq-fold implementation

RNA-seq-fold is written in C++ and requires both STL and Boost libraries. The read simulator is coded as an R script, and both are available from [insert site here]. Running time analyses were performed on a single CPU core of an Intel Xeon.

MCMC performance analysis

Each position of the pairing posterior $\boldsymbol{b}$ was considered as paired if $b_i > thresh$ for $thresh \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$ and unpaired otherwise. A 2x2 contingency table was then calculated to assess sensitivity and specificity measures (Table 5.10).

	Paired in known structure	Unpaired in known structure
$b_i > thresh$	True positive	False positive
$b_i \leq thresh$	False negative	True negative

Table 5.10: Definitions of sensitivity and specificity for RNA-seq-fold

The F-score was calculated as:

$$F = 2 \times \frac{\left(\frac{TP}{TP+FP}\right) \times \left(\frac{TP}{TP+FN}\right)}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} = 2 \times \frac{\text{PPV} \times \text{recall}}{\text{PPV} + \text{recall}}$$

The Matthews correlation coefficient was calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

RNAfold version 2.1.1 and Fold version 5.6 (from the RNAstructure package) were used to generate free energy-based structure predictions. All parameters were left as default.

To assess MCMC convergence, chains of length $l_{sub} \in \{10000,\ 20000, \ldots, 90000\}$ were taken from the full length chain with a burn-in period of $\frac{l_{sub}}{10}$ and a sampling frequency of $\frac{l_{sub}}{1000}$.

Performance of these subsampled chains was then calculated as described above with $thresh = 0.5$ as this appeared to generally produce the best predictive accuracy with the full length chain.

Chapter 6

Conclusions and future directions

In this work, we have described a novel approach to the task of genome-wide determination of RNA secondary structure. We applied these methods to study the global patterns as well as regulatory functionalities of secondary structure in four eukaryotic species. We also developed a Bayesian model and optimization framework to infer base pair resolution secondary structures from our structure-sensitive sequencing datasets.

6.1 Summary of results

In Chapter 2, we introduced a pair of high-throughput, structure-sensitive sequencing approaches termed dsRNA-seq and ssRNA-seq to assay RNA secondary structure on a global scale. To interpret these data, we developed a per-base structure score that captures the relative tendency of each nucleotide to be base paired. We then validated the reliability and reproducibility of our methods in three ways. First, we assessed the prevalence of various heterochromatic histone modifications within regions of high base pairing (dsRNA hotspots). Based on the requirement for base paired intermediates in the biogenesis pathways of small RNAs that direct heterochromatin formation, we expected to find significant enrichment for heterochromatic marks within our dsRNA hotspots. As expected, we found that dsRNA hotspots identified in three eukaryotic species (*Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) were all enriched for heterochromatic marks. We also validated the reliability of dsRNA-seq and ssRNA-seq by more direct molecular assays. Using nuclease digestion coupled with RT-PCR, we showed that regions of high base pairing as determined by our genome-wide approaches were extremely sensitive to double-stranded RNase (dsRNase) but not single-stranded RNase. Finally, we repeated our structure mapping approach on three replicates of HeLa cell RNA and found that positions of high predictive confidence were in almost perfect agreement across all three samples.

With our dsRNA-seq and ssRNA-seq techniques in hand, we next set out to explore the global landscapes of RNA secondary structure in three eukaryotes (Chapter 3). By mapping profiles of secondary structure across protein-coding mRNAs, we revealed a striking reduction in base pairing at sites of translational initiation and termination that was conserved across all three species. We also found large-scale differences in overall 3' UTR structure content between animals and plants, which may reflect the complexity of RBP-mediated regulation in the various organisms. Finally, we assessed the relationship between microRNA targeting and secondary structure, and found a strong inhibitory effect of target site structure on microRNA binding affinity in *C. elegans*. Although this effect has long been suggested by computational predictions of secondary structure, our data provided the first global experimental evidence as such. Surprisingly, we did not observe a similar relationship in *Drosophila*, suggesting that there may be general differences in microRNA targeting modes within animals.

In addition to identifying global patterns of secondary structure, we also addressed the regulatory functions and mechanisms of this important feature (Chapter 4). To do so, we performed an integrative analysis of several genomic datasets (RNA-seq, smRNA-seq, degradome sequencing, and ribo-seq, as well as our dsRNA-seq and ssRNA-seq data) in the model plant *Arabidopsis thaliana*. In general, we found that highly structured mRNA transcripts tended to be lower in overall abundance, were more likely to be degraded, and produced more smRNA species in both sense and antisense directions. Taken together, these results hinted at the possibility of direct processing of highly structured transcripts by the RNA silencing machinery. Additional findings of increased structure within regions of high smRNA production as well as positive correlation between smRNA production and structure score within regions of high base pairing provided further support for this hypothesis. Further studies are necessary to definitely prove our model and elucidate the exact mechanism by which structured mRNAs and "proper" silencing precursors are delineated.

In Chapter 5, we shifted our focus from genome-wide analyses of RNA secondary structure to smaller scale but higher resolution studies. We developed a Bayesian framework and Markov chain Monte Carlo (MCMC) algorithm termed RNA-seq-fold to predict the secondary

structure of individual RNA molecules based on dsRNA-seq and ssRNA-seq data. Starting with simulated dsRNA-seq reads, we showed that RNA-seq-fold outperforms free energy-based methods on most of the tested structures, particularly for those containing large loop segments. We also observed quick and reliable convergence to the correct secondary structure even with fairly shallow sequencing depth. However, when tested with *in vitro* datasets, RNA-seq-fold did not greatly outperform free energy-based methods. The primary impediment to high predictive accuracy was found to be nonspecific digestion at both paired and unpaired nucleotides. To address this shortcoming, we are currently in the process of repeating the *in vitro* structure mapping experiments with a reduced enzyme concentration and a modified protocol that preferentially selects for longer digestion fragments.

6.2 Applications to RNA biology

One of the major contributions of this work has been to provide a resource of structural data for future RNA-centric studies of cellular gene expression and functionality. In this next section, we highlight two areas to which our datasets are particularly well suited and suggest approaches to their study.

6.2.1 mRNA secondary structure as a regulatory feature

Our findings from Chapter 4 point to a novel mode of gene regulation via smRNA processing of highly structured mRNA regions. We proposed as a mechanism the co-opting of small RNA pathways to directly cleave and thereby regulate mRNA transcripts, which may not be surprising given the relaxed binding specificities of Dicer-like (DCL) proteins in plants and Drosha-DGCR8 and Dicer in animals(38, 48, 125). Of note, the main requirement for pri-miRNA recognition appears to be a ~33nt stem with single-stranded flanking sequences(38); this suggests that wayward processing of similar stem-loop structures contained within mRNA transcripts is not uncommon. Additionally, DGCR8 was recently shown to bind non-specifically to single-stranded, double-stranded, and random hairpin transcripts(92), thereby leading the authors

to conclude that Drosha-DGCR8 heterodimers impart specificity to the detection of true substrates. In light of these findings, free DGCR8 may be the most likely candidate for direct processing of mRNA transcripts assuming that such a mechanism exists in animals. As our findings were from *Arabidopsis* and plant DCL proteins carry out the functions of both Drosha-DGCR8 and Dicer in animals, we can only speculate that the DCL proteins are key players in the plant pathway.

To address this question, as well as those of regulatory functionality and secondary effects, we propose the following studies. First, *in vitro* dicing assays can be used to identify the protein(s) responsible for directing cleavage of these structured mRNA regions. To show the same result *in vivo* is a bit more difficult as miRNA-mediated regulation and secondary transcriptional effects must be taken into consideration. A reasonable start would be to select mRNA transcripts containing candidate regions of high secondary structure, but no known miRNA target sites. Abundance of these transcripts as well as the candidate smRNAs could then be measured in wild-type and DCL mutant plants. Techniques that specifically capture cleaved RNA fragments(117, 124) could also be used to identify sites of DCL-mediated cleavage within the candidate regions. If it is indeed the case that a Dicer-like protein is responsible for direct processing of stem-loops within mRNAs, then subsequent follow-up studies to assess the functionality of the smRNAs produced from these loci would be desirable. For example, one possible approach may be to look for these RNA species in RISC (e.g. by Argonaute CLIP). Additionally, target transcripts of these small RNAs could be examined for evidence of miRNA-like regulation. Finally, comparison of structured mRNA regions that are shown to be processed by DCL with known miRNA precursors may yield incredible insights into the specificity determinants of the small RNA biogenesis pathways. To close this section, we note that parallel studies may need to be performed in plants and animals as the protein players and smRNA maturation pathways are not identical between the two clades.

6.2.2 Detection of structural motifs

Another topic that may benefit substantially from our genome-wide structure datasets is the detection and characterization of structural motifs. Existing instances of these moieties (e.g. AU-rich element, iron response element, etc.) have been identified primarily by targeted study(13, 73) or computational approaches such as TEISER, MEMERIS, RNAMotif, and RNAMotifModeler(33, 40, 65, 111). The major caveat of existing computational methods is that they rely on predictions of secondary structure, such that their reliability is inherently capped by the performance of the underlying structure prediction. In fact, TEISER discards structure prediction entirely and operates on the basis of possible stem-loop structures, although this assumption is ameliorated somewhat by the requirement for functional effect of a detected motif(33). Our genome-wide structure data may prove useful for improving computational motif identification as it combines the accuracy of more laborious studies with the throughput of the methods described above. We propose an approach that builds upon the expectation-maximization (EM) framework popularized by the MEME(4) algorithm in a manner similar to that of MEMERIS(40).

Given a set of input sequences $\boldsymbol{X} = \{X_1, X_2, \dots, X_n\}$, MEME operates on the two quantities $\boldsymbol{Z}$ and $\boldsymbol{\rho}$, where Z_{ij} is the probability of a given motif starting at position j in sequence i and ρ_{ck} is the probability of having character c at position k. The probability of observing any given sequence X_i is given by:

$$Pr\left(X_i \middle| Z_{ij} = 1, \rho\right) = \prod_{k=1}^{j-1} \rho_{c_k,0} \prod_{k=j}^{j+W-1} \rho_{c_k,k-j+1} \prod_{k=j+W}^{L} \rho_{c_k,0}$$

In the E-step, $\boldsymbol{Z}$ is estimated from $\boldsymbol{\rho}$ by:

$$Z_{ij}^{(t)} = \frac{Pr\left(X_i \middle| Z_{ij} = 1, \rho^{(t)}\right)}{\sum_{k=1}^{L-W+1} Pr(X_i | Z_{ik} = 1, \rho^{(t)})}$$

Intuitively, the probability of having a motif at position j is the probability of observing the particular sequence that contains the motif at position j divided by the sum of probabilities of all motif positions. Similarly, for the M-step, $\boldsymbol{\rho}$ is estimated from $\boldsymbol{Z}$:

$$\rho_{c,k}^{(t)} = \frac{n_{c,k} + d_{c,k}}{\sum_b n_{b,k} + d_{b,k}} \text{ where } n_{c,k} = \begin{cases} \sum_i \sum_{\{j|X_{i,j+k-1=c}\}} Z_{ij}, & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j}, & k = 0 \end{cases}$$

As with the E-step, the M-step is quite intuitive – the probability of observing character c at position k in the motif is simply the fraction of all instances of the character that is contained within the motif locations $\boldsymbol{Z}$. MEME thus proceeds by alternating between the E-step and M-step until some convergence criterion is reached. A straightforward modification of the basic MEME approach could incorporate our genome-wide structure scores (Section 2.1.3) as continuous-valued vectors $\boldsymbol{Y} = \{Y_1, Y_2, \dots, Y_n\}$. The joint sequence-structure probability function is then:

$$Pr\left(X_i, Y_i \middle| Z_{ij} = 1, \rho, \tau\right) = \prod_{k=1}^{j-1} \rho_{c_k,0} \prod_{k=j}^{j+W-1} \rho_{c_k,k-j+1} s(\tau_k) \prod_{k=j+W}^{L} \rho_{c_k,0}$$

where τ_k is the average value of the vectors $\boldsymbol{Y}$ at position k and $s(\tau_k)$ is some scoring function for how closely the given sequence resembles the current motif τ. The E-step is modified only to include a scoring function for τ:

$$Z_{ij}^{(t)} = \frac{Pr\left(X_i \middle| Z_{ij} = 1, \rho^{(t)}, \tau^{(t)}\right)}{\sum_{k=1}^{L-W+1} Pr(X_i | Z_{ik} = 1, \rho^{(t)}, \tau^{(t)})}$$

For the M-step, we add the following calculation:

$$\tau_k^{(t)} = \frac{\sum_i \sum_k^{L-W+1} Z_{ik} Y_{ik}}{n}$$

which represents the weighted average profile of continuous data values $\boldsymbol{Y}$ at the current motif locations $\boldsymbol{Z}$. This approach is similar to that of the MEMERIS algorithm, except that the free energy-based modeling has been replaced by our experimental structure data. Alternatively, as $\boldsymbol{Y}$ is simply a vector of continuous-valued data, they could be replaced with the pairing posteriors derived from RNA-seq-fold (Chapter 5).

Regardless of the data source used, integration of sequence and experimentally-derived structure data is likely to increase the sensitivity and accuracy of structural motif prediction. Improved prediction of structural motifs would have far-reaching implications in a number of research areas. For example, the known role of secondary structure in alternative splicing(85,

112) suggests that splicing predictors(5) may benefit from incorporation of structural motifs. Structure-sensitive analysis would also be useful in the study of RNA-binding proteins (RBPs), many of which bind to specific structural elements within their target RNAs(99). Finally, single nucleotide polymorphisms (SNPs) detected by genome-wide association studies (GWAS) could be screened against a database of structural motifs to help prioritize and interpret these mutations. Such a tool would be extremely valuable in mechanistic, pharmacogenomic, and therapeutic studies of disease-associated polymorphisms.

6.2.3 Long non-coding RNAs

A third application of the dsRNA-seq and ssRNA-seq methodologies is the characterization of long non-coding RNAs (lncRNAs). lncRNAs are a diverse class of transcripts that biochemically resemble protein-coding mRNAs but are distinguished by their length (> 200 nt), lack of coding potential, and high level of secondary structure(29, 82, 90, 107). These RNAs are thought to function primarily as regulators of gene expression and are almost uniformly expressed at very low levels in extremely spatiotemporal specific patterns(21, 88). To characterize lncRNAs, recent studies have variously utilized chromatin structure(37), manual curation(21), and custom tiling arrays(60, 88) as a means of focusing on these elusive transcripts. Given the relatively high structural content of lncRNAs, it is likely that dsRNA-seq and ssRNA-seq could be used to selectively interrogate the lncRNA population while simultaneously generating the first comprehensive map of lncRNA secondary structure. Furthermore, as these transcripts are thought to function through their structure rather than sequence(29, 82, 107), such studies may also provide substantial insight into lncRNA function, a topic that as of yet remains mostly unexplored.

Taken together, the dsRNA-seq and ssRNA-seq protocols, in conjunction with the analysis methods presented in this work, hold considerable promise for future studies of many aspects of RNA biology. General and extensive application of our novel structure mapping approaches to a multitude of organisms, cell types, and conditions (in particular the three areas mentioned above) should prove exceptionally useful to their respective researchers.

6.3 Improved methods for RNA structure prediction

As our data suggests, the work herein is only a first step towards the ultimate goal of genome-wide secondary structure prediction at base pair resolution. Therefore, continued development of both experimental and computational aspects of our approaches concomitant with their widespread application, will be crucial to future RNA structural studies. In this next section, we consider new experimental approaches that will enable measurement of *in vivo* secondary structure. We also examine the generalizability of RNA-seq-fold as it pertains to large-scale predictions of RNA secondary structure and address several potential pitfalls.

6.3.1 *In vivo* approaches

To date, most RNA structural studies have been carried out *in vitro* on denatured and renatured RNAs. A prominent concern, therefore, is that these assays do not measure the true *in vivo* structure as it may be affected by other factors such as protein binding, cellular localization, and co-transcriptional folding(25, 97, 123). Several methods have been developed to probe *in vivo* secondary structure(2, 58, 101), but none of these can be used to feasibly perform genome-wide studies. In contrast, dsRNA-seq and ssRNA-seq can be performed on *in vivo* cross-linked RNA populations; the cross-linking in effect holds RNA molecules in their native conformation and thereby allows our mapping techniques to detect true cellular structure. In fact, we recently used this approach to study the global landscape of RNA-protein interactions based on formaldehyde cross-linking of nucleic acids and proteins, with additional follow-up studies of the secondary structure at these interaction sites currently in the works. These future investigations will provide the first genome-wide characterization of *in vivo* secondary structure and should contribute substantially to our current understanding of RNA structure and its functionality.

6.3.2 Towards genome-wide structure prediction at single base pair resolution

In this work, we have provided global structure-sensitive assays (dsRNA- and ssRNA-seq) and the tools to infer secondary structure from these data (RNA-seq-fold). Our initial proof-of-principle study of eight *in vitro* transcribed non-coding RNAs achieved moderate predictive accuracy under reasonable sequencing depth, suggesting that the approach can be scaled up to genome-wide studies. Before such a study is undertaken, several topics should be taken under careful consideration. First and foremost, the experimental conditions (e.g. concentration of input RNA, extent of RNase treatment, etc.) must be optimized to generate a range of cloneable fragments that can be used to accurately infer the secondary structure. In our pilot study, an extremely dilute enzyme concentration was used in an attempt to maintain high cleavage specificity; however, the nonspecific digestion that we observed suggests that even more dilute conditions are required. In addition, it remains unclear if such digestion conditions are suitable for genome-wide experiments in which the more varied RNA population likely results in a broader range of enzyme affinities. On the other hand, as the RNases used in our protocols are insensitive to intramolecular versus intermolecular base pairing, it is imperative to maintain the RNA pool at a dilute concentration so as to avoid heteroduplex formation. Careful investigation of the differences between *in vivo* and *in vitro* structures is a challenge that needs to be addressed. On the computational side, additional model parameters may be needed to interpret the RNA population complexity as well as the corresponding increase in stochasticity.

Even with these caveats, our approach promises substantial advances in the study of RNA secondary structure. Extensive application of our methods to different RNA populations (e.g. poly(A)+, size-selected) can be used to generate a comprehensive atlas of secondary structure. Such a resource would be of great value to all RNA-related fields ranging from detailed mechanistic studies to high-throughput drug and RNA therapeutic screening. Our methods could also be used to study multiple related species, thereby allowing insight into the evolution of RNA secondary structure.

6.4 Concluding remarks

Secondary structure is an intrinsic feature of all cellular RNAs and plays a fundamental role throughout their biogenesis, regulation, and function. In this work, we have established a novel high-throughput, sequencing-based, structure mapping approach to study RNA secondary structure on a genome-wide scale. We also developed a Bayesian Markov chain Monte Carlo algorithm to infer base pair resolution secondary structures from our global structure-sensitive sequencing data. With the ever-increasing throughput and proliferation of sequencing technologies, the methods described in this work present a unique opportunity to vastly expand the scope and breadth of RNA structural studies. Widespread application of our novel structure mapping approaches, in conjunction with additional development of computational methods to interpret these data, will undoubtedly increase our understanding of RNA secondary structure and its many functional roles.

Chapter 7

References

1. Abe M, Bonini NM. MicroRNAs and neurodegeneration: role and impact. Trends Cell Biol. 2013;23(1):30-6. Epub 2012/10/03. doi: 10.1016/j.tcb.2012.08.013. PubMed PMID: 23026030; PubMed Central PMCID: PMC3540990.

2. Adilakshmi T, Lease RA, Woodson SA. Hydroxyl radical footprinting in vivo: mapping macromolecular structures with synchrotron radiation. Nucleic Acids Res. 2006;34(8):e64. Epub 2006/05/10. doi: 10.1093/nar/gkl291. PubMed PMID: 16682443; PubMed Central PMCID: PMC1458516.

3. Al-Hashimi HM, Walter NG. RNA dynamics: it is about time. Current opinion in structural biology. 2008;18(3):321-9. Epub 2008/06/13. doi: 10.1016/j.sbi.2008.04.004. PubMed PMID: 18547802; PubMed Central PMCID: PMC2580758.

4. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1994;2:28-36. Epub 1994/01/01. PubMed PMID: 7584402.

5. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465(7294):53-9. Epub 2010/05/07. doi: 10.1038/nature09000. PubMed PMID: 20445623.

6. Baulcombe D. RNA silencing in plants. Nature. 2004;431(7006):356-63. PubMed PMID: 15372043.

7. Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, et al. U1 snRNP determines mRNA length and regulates isoform expression. Cell. 2012;150(1):53-64. Epub 2012/07/10. doi: 10.1016/j.cell.2012.05.029. PubMed PMID: 22770214; PubMed Central PMCID: PMC3412174.

8. Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, Jacobsen SE. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. PloS one. 2008;3(9):e3156. Epub 2008/09/09. doi: 10.1371/journal.pone.0003156. PubMed PMID: 18776934; PubMed Central PMCID: PMC2522283.

9. Bothe JR, Nikolova EN, Eichhorn CD, Chugh J, Hansen AL, Al-Hashimi HM. Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nat Methods. 2011;8(11):919-31. Epub 2011/11/01. doi: 10.1038/nmeth.1735. PubMed PMID: 22036746; PubMed Central PMCID: PMC3320163.

10. Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol. 2005;6(11):R97. Epub 2005/11/10. doi: 10.1186/gb-2005-6-11-r97. PubMed PMID: 16277752; PubMed Central PMCID: PMC1297653.

11. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2013;41(Database issue):D226-32. Epub 2012/11/06. doi: 10.1093/nar/gks1005. PubMed PMID: 23125362; PubMed Central PMCID: PMC3531072.

12. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009;136(4):642-55. Epub 2009/02/26. doi: 10.1016/j.cell.2009.01.035. PubMed PMID: 19239886; PubMed Central PMCID: PMC2675692.

13. Casey JL, Koeller DM, Ramin VC, Klausner RD, Harford JB. Iron regulation of transferrin receptor mRNA levels requires iron-responsive elements and a rapid turnover determinant in the 3' untranslated region of the mRNA. EMBO J. 1989;8(12):3693-9. Epub 1989/12/01. PubMed PMID: 2583116; PubMed Central PMCID: PMC402052.

14. Castel SE, Martienssen RA. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat Rev Genet. 2013;14(2):100-12. Epub 2013/01/19. doi: 10.1038/nrg3355. PubMed PMID: 23329111.

15. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012;149(6):1393-406. Epub 2012/06/05. doi: 10.1016/j.cell.2012.04.031. PubMed PMID: 22658674.

16. Chen JL, Greider CW. Functional analysis of the pseudoknot structure in human telomerase RNA. Proc Natl Acad Sci U S A. 2005;102(23):8080-5; discussion 77-9. Epub 2005/04/26. doi: 10.1073/pnas.0502259102. PubMed PMID: 15849264; PubMed Central PMCID: PMC1149427.

17. Crick FH. On protein synthesis. Symposia of the Society for Experimental Biology. 1958;12:138-63. Epub 1958/01/01. PubMed PMID: 13580867.

18. Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 2011;39(Web Server issue):W155-9. Epub 2011/05/31. doi: 10.1093/nar/gkr319. PubMed PMID: 21622958; PubMed Central PMCID: PMC3125753.

19. De Smaele E, Ferretti E, Gulino A. MicroRNAs as biomarkers for CNS cancer and other disorders. Brain research. 2010;1338:100-11. Epub 2010/04/13. doi: 10.1016/j.brainres.2010.03.103. PubMed PMID: 20380821.

20. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A. 2009;106(1):97-102. Epub 2008/12/26. doi: 10.1073/pnas.0806929106. PubMed PMID: 19109441; PubMed Central PMCID: PMC2629221.

21. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775-89. Epub 2012/09/08. doi: 10.1101/gr.132159.111. PubMed PMID: 22955988; PubMed Central PMCID: PMC3431493.

22. Dever TE, Green R. The elongation, termination, and recycling phases of translation in eukaryotes. Cold Spring Harbor perspectives in biology. 2012;4(7):a013706. Epub 2012/07/04. doi: 10.1101/cshperspect.a013706. PubMed PMID: 22751155; PubMed Central PMCID: PMC3385960.

23. Ding Y. Statistical and Bayesian approaches to RNA secondary structure prediction. RNA. 2006;12(3):323-31. Epub 2006/02/24. doi: 10.1261/rna.2274106. PubMed PMID: 16495231; PubMed Central PMCID: PMC1383571.

24. Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. 2004;32(Web Server issue):W135-41. Epub 2004/06/25. doi: 10.1093/nar/gkh449. PubMed PMID: 15215366; PubMed Central PMCID: PMC441587.

25. Doetsch M, Schroeder R, Furtig B. Transient RNA-protein interactions in RNA folding. FEBS J. 2011;278(10):1634-42. Epub 2011/03/18. doi: 10.1111/j.1742-4658.2011.08094.x. PubMed PMID: 21410645; PubMed Central PMCID: PMC3123464.

26. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC bioinformatics. 2004;5:105. Epub 2004/08/07. doi: 10.1186/1471-2105-5-105. PubMed PMID: 15296519; PubMed Central PMCID: PMC514602.

27. Edwards AL, Garst AD, Batey RT. Determining structures of RNA aptamers and riboswitches by X-ray crystallography. Methods Mol Biol. 2009;535:135-63. Epub 2009/04/21. doi: 10.1007/978-1-59745-557-2_9. PubMed PMID: 19377976; PubMed Central PMCID: PMC3156247.

28. Fialcowitz EJ, Brewer BY, Keenan BP, Wilson GM. A hairpin-like structure within an AU-rich mRNA-destabilizing element regulates trans-factor binding selectivity and mRNA decay kinetics. J Biol Chem. 2005;280(23):22406-17. Epub 2005/04/06. doi: 10.1074/jbc.M500618200. PubMed PMID: 15809297; PubMed Central PMCID: PMC1553220.

29. Flintoft L. Non-coding RNA: Structure and function for lncRNAs. Nat Rev Genet. 2013. Epub 2013/08/07. doi: 10.1038/nrg3561. PubMed PMID: 23917630.

30. Garzon R, Marcucci G. Potential of microRNAs for cancer diagnostics, prognostication and therapy. Current opinion in oncology. 2012;24(6):655-9. Epub 2012/10/20. doi: 10.1097/CCO.0b013e328358522c. PubMed PMID: 23079782.

31. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330(6012):1775-87. Epub 2010/12/24. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PubMed Central PMCID: PMC3142569.

32. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Molecular systems biology. 2011;7:481. Epub 2011/04/14. doi: 10.1038/msb.2011.14. PubMed PMID: 21487400; PubMed Central PMCID: PMC3101949.

33. Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, Salavati R, et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature. 2012;485(7397):264-8. Epub 2012/04/13. doi: 10.1038/nature11013. PubMed PMID: 22495308; PubMed Central PMCID: PMC3350620.

34. Gregory BD, O'Malley RC, Lister R, Urich MA, Tonti-Filippini J, Chen H, et al. A link between RNA metabolism and silencing affecting Arabidopsis development. Dev Cell. 2008;14(6):854-66. Epub 2008/05/20. doi: 10.1016/j.devcel.2008.04.005. PubMed PMID: 18486559.

35. Gu W, Wang X, Zhai C, Xie X, Zhou T. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Molecular biology and evolution.

2012;29(10):3037-44. Epub 2012/04/12. doi: 10.1093/molbev/mss109. PubMed PMID: 22490819.

36. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS computational biology. 2010;6(2):e1000664. Epub 2010/02/09. doi: 10.1371/journal.pcbi.1000664. PubMed PMID: 20140241; PubMed Central PMCID: PMC2816680.

37. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223-7. Epub 2009/02/03. doi: 10.1038/nature07672. PubMed PMID: 19182780; PubMed Central PMCID: PMC2754849.

38. Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, et al. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell. 2006;125(5):887-901. Epub 2006/06/06. doi: 10.1016/j.cell.2006.03.043. PubMed PMID: 16751099.

39. Hata A. Functions of microRNAs in cardiovascular biology and disease. Annual review of physiology. 2013;75:69-93. Epub 2012/11/20. doi: 10.1146/annurev-physiol-030212-183737. PubMed PMID: 23157557.

40. Hiller M, Pudimat R, Busch A, Backofen R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. Nucleic Acids Res. 2006;34(17):e117. Epub 2006/09/22. doi: 10.1093/nar/gkl544. PubMed PMID: 16987907; PubMed Central PMCID: PMC1903381.

41. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012;7(8):1534-50. Epub 2012/07/28. doi: 10.1038/nprot.2012.086. PubMed PMID: 22836135; PubMed Central PMCID: PMC3535016.

42. Iwakiri J, Kameda T, Asai K, Hamada M. Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. Bioinformatics. 2013. Epub 2013/08/13. doi: 10.1093/bioinformatics/btt453. PubMed PMID: 23933973.

43. Jacob Y, Stroud H, Leblanc C, Feng S, Zhuo L, Caro E, et al. Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature. 2010;466(7309):987-91. Epub 2010/07/16. doi: 10.1038/nature09290. PubMed PMID: 20631708; PubMed Central PMCID: PMC2964344.

44. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39(10):1278-84. Epub 2007/09/26. doi: 10.1038/ng2135. PubMed PMID: 17893677.

45. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010;467(7311):103-7. Epub 2010/09/03. doi: 10.1038/nature09322. PubMed PMID: 20811459.

46. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 2011;471(7339):480-5. Epub 2010/12/24. doi: 10.1038/nature09725. PubMed PMID: 21179089; PubMed Central PMCID: PMC3109908.

47. Kikinis Z, Eisenstein RS, Bettany AJ, Munro HN. Role of RNA secondary structure of the iron-responsive element in translational regulation of ferritin synthesis. Nucleic Acids Res. 1995;23(20):4190-5. Epub 1995/10/25. PubMed PMID: 7479083; PubMed Central PMCID: PMC307361.

48. Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 2005;6(5):376-85. Epub 2005/04/27. doi: 10.1038/nrm1644. PubMed PMID: 15852042.

49. Kozak M. Influence of mRNA secondary structure on binding and migration of 40S ribosomal subunits. Cell. 1980;19(1):79-90. Epub 1980/01/01. PubMed PMID: 7357609.

50. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(Database issue):D152-7. Epub 2010/11/03. doi: 10.1093/nar/gkq1027. PubMed PMID: 21037258; PubMed Central PMCID: PMC3013655.

51. Kretzner L, Krol A, Rosbash M. Saccharomyces cerevisiae U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. Proc Natl Acad Sci U S A. 1990;87(2):851-5. Epub 1990/01/01. PubMed PMID: 2405391; PubMed Central PMCID: PMC53364.

52. Lau P, de Strooper B. Dysregulated microRNAs in neurodegenerative disorders. Semin Cell Dev Biol. 2010;21(7):768-73. Epub 2010/01/19. doi: 10.1016/j.semcdb.2010.01.009. PubMed PMID: 20080199.

53. Li F, Ryvkin P, Childress DM, Valladares O, Gregory BD, Wang LS. SAVoR: a server for sequencing annotation and visualization of RNA structures. Nucleic Acids Res. 2012;40(Web Server issue):W59-64. Epub 2012/04/12. doi: 10.1093/nar/gks310. PubMed PMID: 22492627; PubMed Central PMCID: PMC3394343.

54. Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, Aiyer S, et al. Global analysis of RNA secondary structure in two metazoans. Cell Rep. 2012;1(1):69-82. Epub 2012/07/27. doi: 10.1016/j.celrep.2011.10.002. PubMed PMID: 22832108.

55. Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. Plant Cell. 2012. Epub 2012/11/15. doi: 10.1105/tpc.112.104232. PubMed PMID: 23150631.

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMC2723002.

57. Li H, Xu L, Wang H, Yuan Z, Cao X, Yang Z, et al. The Putative RNA-dependent RNA polymerase RDR6 acts synergistically with ASYMMETRIC LEAVES1 and 2 to repress BREVIPEDICELLUS and MicroRNA165/166 in Arabidopsis leaf development. Plant Cell. 2005;17(8):2157-71. Epub 2005/07/12. doi: 10.1105/tpc.105.033449. PubMed PMID: 16006579; PubMed Central PMCID: PMC1182480.

58. Liebeg A, Waldsich C. Probing RNA structure within living cells. Methods in enzymology. 2009;468:219-38. Epub 2009/01/01. doi: 10.1016/S0076-6879(09)68011-3. PubMed PMID: 20946772.

59. Liu C, Tang DG. MicroRNA regulation of cancer stem cells. Cancer Res. 2011;71(18):5950-4. Epub 2011/09/16. doi: 10.1158/0008-5472.CAN-11-1035. PubMed PMID: 21917736; PubMed Central PMCID: PMC3177108.

60. Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet. 2010;42(12):1113-7. Epub 2010/11/09. doi: 10.1038/ng.710. PubMed PMID: 21057500; PubMed Central PMCID: PMC3040650.

61. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol. 2007;14(4):287-94. Epub 2007/04/03. doi: 10.1038/nsmb1226. PubMed PMID: 17401373.

62. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for molecular biology : AMB. 2011;6:26. Epub 2011/11/26. doi: 10.1186/1748-7188-6-26. PubMed PMID: 22115189; PubMed Central PMCID: PMC3319429.

63. Lorkovic ZJ. Role of plant RNA-binding proteins in development, stress response and genome organization. Trends Plant Sci. 2009;14(4):229-36. Epub 2009/03/17. doi: 10.1016/j.tplants.2009.01.007. PubMed PMID: 19285908.

64. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, et al. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci U S A. 2011;108(27):11063-8. Epub 2011/06/07. doi: 10.1073/pnas.1106501108. PubMed PMID: 21642531; PubMed Central PMCID: PMC3131332.

65. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 2001;29(22):4724-35. Epub 2001/11/20. PubMed PMID: 11713323; PubMed Central PMCID: PMC92549.

66. Mahen EM, Watson PY, Cottrell JW, Fedor MJ. mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS biology. 2010;8(2):e1000307. Epub 2010/02/18. doi: 10.1371/journal.pbio.1000307. PubMed PMID: 20161716; PubMed Central PMCID: PMC2817708.

67. Malecova B, Morris KV. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. Current opinion in molecular therapeutics. 2010;12(2):214-22. Epub 2010/04/08. PubMed PMID: 20373265; PubMed Central PMCID: PMC2861437.

68. Marz M, Stadler PF. Comparative analysis of eukaryotic U3 snoRNA. RNA biology. 2009;6(5):503-7. Epub 2009/10/31. PubMed PMID: 19875933.

69. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999;288(5):911-40. Epub 1999/05/18. doi: 10.1006/jmbi.1999.2700. PubMed PMID: 10329189.

70. Mathews DH, Turner DH. Prediction of RNA secondary structure by free energy minimization. Current opinion in structural biology. 2006;16(3):270-8. Epub 2006/05/23. doi: 10.1016/j.sbi.2006.05.010. PubMed PMID: 16713706.

71. May GE, Olson S, McManus CJ, Graveley BR. Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. RNA. 2011;17(2):222-9. Epub 2010/12/17. doi: 10.1261/rna.2521311. PubMed PMID: 21159795; PubMed Central PMCID: PMC3022272.

72. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29(6-7):1105-19. Epub 1990/05/01. doi: 10.1002/bip.360290621. PubMed PMID: 1695107.

73. Meisner NC, Hackermuller J, Uhl V, Aszodi A, Jaritz M, Auer M. mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure. Chembiochem : a European journal of chemical biology. 2004;5(10):1432-47. Epub 2004/10/01. doi: 10.1002/cbic.200400219. PubMed PMID: 15457527.

74. Mengersen KL, Tweedie RL. Rates of convergence of the Hastings and Metropolis algorithms. Ann Stat. 1996;24(1):101-21. PubMed PMID: ISI:A1996UV51100006.

75. Metzler D, Nebel ME. Predicting RNA secondary structures with pseudoknots by MCMC sampling. Journal of mathematical biology. 2008;56(1-2):161-81. Epub 2007/06/26. doi: 10.1007/s00285-007-0106-6. PubMed PMID: 17589847.

76. Meyer IM, Miklos I. Co-transcriptional folding is encoded within RNA genes. BMC molecular biology. 2004;5:10. Epub 2004/08/10. doi: 10.1186/1471-2199-5-10. PubMed PMID: 15298702; PubMed Central PMCID: PMC514895.

77. Meyer IM, Miklos I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. PLoS computational biology. 2007;3(8):e149. Epub 2007/08/19. doi: 10.1371/journal.pcbi.0030149. PubMed PMID: 17696604; PubMed Central PMCID: PMC1941756.

78. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature. 2009;457(7228):413-20. Epub 2009/01/23. doi: 10.1038/nature07756. PubMed PMID: 19158787; PubMed Central PMCID: PMC3246369.

79. Muino JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. Plant Methods. 2011;7:11. Epub 2011/05/11. doi: 10.1186/1746-4811-7-11. PubMed PMID: 21554688; PubMed Central PMCID: PMC3114017.

80. Mustroph A, Juntawong P, Bailey-Serres J. Isolation of plant polysomal mRNA by differential centrifugation and ribosome immunopurification methods. Methods Mol Biol. 2009;553:109-26. Epub 2009/07/10. doi: 10.1007/978-1-60327-563-7_6. PubMed PMID: 19588103.

81. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25(10):1335-7. Epub 2009/03/25. doi: 10.1093/bioinformatics/btp157. PubMed PMID: 19307242; PubMed Central PMCID: PMC2732312.

82. Novikova IV, Hennelly SP, Sanbonmatsu KY. Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? Bioarchitecture. 2012;2(6):189-99. Epub

2012/12/26. doi: 10.4161/bioa.22592. PubMed PMID: 23267412; PubMed Central PMCID: PMC3527312.

83. Olovnikov I, Aravin AA, Fejes Toth K. Small RNA in the nucleus: the RNA-chromatin ping-pong. Current opinion in genetics & development. 2012;22(2):164-71. Epub 2012/02/22. doi: 10.1016/j.gde.2012.01.002. PubMed PMID: 22349141; PubMed Central PMCID: PMC3345048.

84. Ouyang Z, Snyder MP, Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res. 2013;23(2):377-87. Epub 2012/10/16. doi: 10.1101/gr.138545.112. PubMed PMID: 23064747; PubMed Central PMCID: PMC3561878.

85. Pervouchine DD, Khrameeva EE, Pichugina MY, Nikolaienko OV, Gelfand MS, Rubtsov PM, et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA. 2012;18(1):1-15. Epub 2011/12/01. doi: 10.1261/rna.029249.111. PubMed PMID: 22128342; PubMed Central PMCID: PMC3261731.

86. Petrov A, Kornberg G, O'Leary S, Tsai A, Uemura S, Puglisi JD. Dynamics of the translational machinery. Current opinion in structural biology. 2011;21(1):137-45. Epub 2011/01/25. doi: 10.1016/j.sbi.2010.11.007. PubMed PMID: 21256733.

87. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008;5(10):877-9. Epub 2008/09/23. doi: 10.1038/nmeth.1253. PubMed PMID: 18806792; PubMed Central PMCID: PMC3126653.

88. Ramos AD, Diaz A, Nellore A, Delgado RN, Park KY, Gonzales-Roybal G, et al. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. Cell stem cell. 2013;12(5):616-28. Epub 2013/04/16. doi: 10.1016/j.stem.2013.03.003. PubMed PMID: 23583100; PubMed Central PMCID: PMC3662805.

89. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC bioinformatics. 2010;11:129. Epub 2010/03/17. doi: 10.1186/1471-2105-11-129. PubMed PMID: 20230624; PubMed Central PMCID: PMC2984261.

90. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annual review of biochemistry. 2012;81:145-66. Epub 2012/06/06. doi: 10.1146/annurev-biochem-051410-092902. PubMed PMID: 22663078.

91. Robins H, Li Y, Padgett RW. Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A. 2005;102(11):4006-9. Epub 2005/03/02. doi: 10.1073/pnas.0500775102. PubMed PMID: 15738385; PubMed Central PMCID: PMC554828.

92. Roth BM, Ishimaru D, Hennig M. The core Microprocessor component DiGeorge syndrome critical region 8 (DGCR8) is a non-specific RNA-binding protein. J Biol Chem. 2013. Epub 2013/07/31. doi: 10.1074/jbc.M112.446880. PubMed PMID: 23893406.

93. Rottiers V, Naar AM. MicroRNAs in metabolism and metabolic disorders. Nat Rev Mol Cell Biol. 2012;13(4):239-50. Epub 2012/03/23. doi: 10.1038/nrm3313. PubMed PMID: 22436747.

94. Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S, et al. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. 2011;30(10):1928-38. Epub 2011/04/14. doi: 10.1038/emboj.2011.103. PubMed PMID: 21487388; PubMed Central PMCID: PMC3098477.

95. Salta E, De Strooper B. Non-coding RNAs with essential roles in neurodegenerative disorders. Lancet neurology. 2012;11(2):189-200. Epub 2012/01/24. doi: 10.1016/S1474-4422(11)70286-1. PubMed PMID: 22265214.

96. Schmeing TM, Ramakrishnan V. What recent ribosome structures have revealed about the mechanism of translation. Nature. 2009;461(7268):1234-42. Epub 2009/10/20. doi: 10.1038/nature08403. PubMed PMID: 19838167.

97. Schroeder R, Grossberger R, Pichler A, Waldsich C. RNA folding in vivo. Current opinion in structural biology. 2002;12(3):296-300. Epub 2002/07/20. PubMed PMID: 12127447.

98. Shapiro BA, Maizel J, Lipkin LE, Currey K, Whitney C. Generating non-overlapping displays of nucleic acid secondary structure. Nucleic Acids Res. 1984;12(1 Pt 1):75-88. Epub 1984/01/11. PubMed PMID: 6694904; PubMed Central PMCID: PMC320985.

99. Silverman IM, Li F, Gregory BD. Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. Plant Sci. 2013;205-206:55-62. Epub 2013/03/19. doi: 10.1016/j.plantsci.2013.01.009. PubMed PMID: 23498863.

100. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L. Crystal structure of Argonaute and its implications for RISC slicer activity. Science. 2004;305(5689):1434-7. Epub 2004/07/31. doi: 10.1126/science.1102514. PubMed PMID: 15284453.

101. Spitale RC, Crisalli P, Flynn RA, Torre EA, Kool ET, Chang HY. RNA SHAPE analysis in living cells. Nature chemical biology. 2013;9(1):18-20. Epub 2012/11/28. doi: 10.1038/nchembio.1131. PubMed PMID: 23178934; PubMed Central PMCID: PMC3706714.

102. Thomas J, Lea K, Zucker-Aprison E, Blumenthal T. The spliceosomal snRNAs of Caenorhabditis elegans. Nucleic Acids Res. 1990;18(9):2633-42. Epub 1990/05/11. PubMed PMID: 2339054; PubMed Central PMCID: PMC330746.

103. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011;147(7):1537-50. Epub 2011/12/27. doi: 10.1016/j.cell.2011.11.055. PubMed PMID: 22196729; PubMed Central PMCID: PMC3376356.

104. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods. 2010;7(12):995-1001. Epub 2010/11/09. doi: 10.1038/nmeth.1529. PubMed PMID: 21057495; PubMed Central PMCID: PMC3247016.

105. Vasserot AP, Schaufele FJ, Birnstiel ML. Conserved terminal hairpin sequences of histone mRNA precursors are not involved in duplex formation with the U7 RNA but act as a target site for a distinct processing factor. Proc Natl Acad Sci U S A. 1989;86(12):4345-9. Epub 1989/06/01. PubMed PMID: 2734288; PubMed Central PMCID: PMC287265.

106. Voinnet O. Origin, biogenesis, and activity of plant microRNAs. Cell. 2009;136(4):669-87. Epub 2009/02/26. doi: 10.1016/j.cell.2009.01.046. PubMed PMID: 19239888.

107. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res. 2013;41(Database issue):D246-51. Epub 2012/10/09. doi: 10.1093/nar/gks915. PubMed PMID: 23042674; PubMed Central PMCID: PMC3531107.

108. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. Nat Rev Genet. 2011;12(9):641-55. Epub 2011/08/19. doi: 10.1038/nrg3049. PubMed PMID: 21850044.

109. Wan Y, Qu K, Ouyang Z, Chang HY. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. Nat Protoc. 2013;8(5):849-69. Epub 2013/04/06. doi: 10.1038/nprot.2013.045. PubMed PMID: 23558785.

110. Wang LK, Shuman S. Mutational analysis defines the 5'-kinase and 3'-phosphatase active sites of T4 polynucleotide kinase. Nucleic Acids Res. 2002;30(4):1073-80. Epub 2002/02/14. PubMed PMID: 11842120; PubMed Central PMCID: PMC100346.

111. Wang X, Juan L, Lv J, Wang K, Sanford JR, Liu Y. Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. BMC genomics. 2011;12 Suppl 5:S8. Epub 2012/03/06. doi: 10.1186/1471-2164-12-S5-S8. PubMed PMID: 22369183; PubMed Central PMCID: PMC3287504.

112. Warf MB, Berglund JA. Role of RNA structure in regulating pre-mRNA splicing. Trends in biochemical sciences. 2010;35(3):169-78. Epub 2009/12/05. doi: 10.1016/j.tibs.2009.10.004. PubMed PMID: 19959365; PubMed Central PMCID: PMC2834840.

113. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature. 2009;460(7256):711-6. Epub 2009/08/08. doi: 10.1038/nature08237. PubMed PMID: 19661910; PubMed Central PMCID: PMC2724670.

114. Weeks KM. Advances in RNA structure analysis by chemical probing. Current opinion in structural biology. 2010;20(3):295-304. Epub 2010/05/08. doi: 10.1016/j.sbi.2010.04.001. PubMed PMID: 20447823; PubMed Central PMCID: PMC2916962.

115. Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, et al. Following translation by single ribosomes one codon at a time. Nature. 2008;452(7187):598-603. Epub 2008/03/11. doi: 10.1038/nature06716. PubMed PMID: 18327250; PubMed Central PMCID: PMC2556548.

116. Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc. 2006;1(3):1610-6. Epub 2007/04/05. doi: 10.1038/nprot.2006.249. PubMed PMID: 17406453.

117. Willmann MR, Berkowitz ND, Gregory BD. Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes-GMUCT 2.0. Methods. 2013. Epub 2013/07/23. doi: 10.1016/j.ymeth.2013.07.003. PubMed PMID: 23867340.

118. Willmann MR, Endres MW, Cook RT, Gregory BD. The Functions of RNA-Dependent RNA Polymerases in Arabidopsis. Arabidopsis Book. 2011;9:e0146. Epub 2012/02/04. doi: 10.1199/tab.0146. PubMed PMID: 22303271; PubMed Central PMCID: PMC3268507.

119. Wong TN, Sosnick TR, Pan T. Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. Proc Natl Acad Sci U S A. 2007;104(46):17995-8000. Epub 2007/11/08. doi: 10.1073/pnas.0705038104. PubMed PMID: 17986617; PubMed Central PMCID: PMC2084285.

120. You JS, Jones PA. Cancer genetics and epigenetics: two sides of the same coin? Cancer cell. 2012;22(1):9-20. Epub 2012/07/14. doi: 10.1016/j.ccr.2012.06.008. PubMed PMID: 22789535; PubMed Central PMCID: PMC3396881.

121. Yuan YR, Pei Y, Ma JB, Kuryavyi V, Zhadina M, Meister G, et al. Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. Mol Cell. 2005;19(3):405-19. Epub 2005/08/03. doi: 10.1016/j.molcel.2005.07.011. PubMed PMID: 16061186.

122. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, et al. Crystal structure of the ribosome at 5.5 A resolution. Science. 2001;292(5518):883-96. Epub 2001/04/03. doi: 10.1126/science.1060089. PubMed PMID: 11283358.

123. Zemora G, Waldsich C. RNA folding in living cells. RNA biology. 2010;7(6):634-41. Epub 2010/11/04. PubMed PMID: 21045541; PubMed Central PMCID: PMC3073324.

124. Zhai J, Arikit S, Simon SA, Kingham BF, Meyers BC. Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. Methods. 2013. Epub 2013/07/03. doi: 10.1016/j.ymeth.2013.06.025. PubMed PMID: 23810899.

125. Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. Single processing center models for human Dicer and bacterial RNase III. Cell. 2004;118(1):57-68. Epub 2004/07/10. doi: 10.1016/j.cell.2004.06.017. PubMed PMID: 15242644.

126. Zhao Y, Samal E, Srivastava D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. Nature. 2005;436(7048):214-20. Epub 2005/06/14. doi: 10.1038/nature03817. PubMed PMID: 15951802.

127. Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, et al. Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. 2010;6(9):e1001141. Epub 2010/09/30. doi: 10.1371/journal.pgen.1001141

128. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, et al. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat Struct Mol Biol. 2010;17(2):173-9. Epub 2010/01/12. doi: 10.1038/nsmb.1745. PubMed PMID: 20062054; PubMed Central PMCID: PMC2834287.

129. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9(1):133-48. Epub 1981/01/10. PubMed PMID: 6163133; PubMed Central PMCID: PMC326673.