2016

Genome-Wide Approaches To Study Rna Secondary Structure

Nathan Daniel Berkowitz
University of Pennsylvania, nberk@upenn.edu

Genome-Wide Approaches To Study Rna Secondary Structure

Abstract
The central hypothesis of molecular biology depicts RNA as an intermediary conveyor of genetic information. RNA is transcribed from DNA and translated to proteins, the molecular machines of the cell. However, many RNAs do not encode protein and instead function as molecular machines themselves. The most famous examples are ribosomal RNAs and transfer RNAs, which together form the core translational machinery of the cell. Many other non-coding RNAs have been discovered including catalytic and regulatory RNAs. In many cases RNA function is tightly linked to its secondary structure, which is the collection of hydrogen bonds between complimentary RNA sequences that drives these molecules into their three dimensional structure.

Over the last decade, technology for determining the sequence of DNA and RNA has advanced rapidly, making transcriptome-wide expression profiling fast and widely available. In this dissertation, I discuss recent efforts to leverage this powerful technology to study, not just RNA expression, but several other aspects of RNA function. In particular, I focus on three tightly linked aspects of RNA biology: RNA-secondary structure, RNA cleavage, and regulatory small RNAs. I introduce a database for integrating, comparing, and contrasting techniques for determining RNA secondary structure including a technique developed in my dissertation laboratory. Additionally, I discuss a newly improved technology capable of detecting RNA cleavage events. Finally, I integrate RNA secondary structure probing and RNA cleavage detection to interrogate a family of genes important for eukaryotic small RNA-mediated silencing. These diverse analyses are just a few examples of the vast promises offered by adapting RNA-sequencing technology to probe RNA function across many cellular processes.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Brian D. Gregory

Second Advisor
John I. Murray

Keywords
miRNA, RNA, RNA Dependent RNA Polymerase, Secondary Structure, Sequencing, siRNA

Subject Categories
Bioinformatics | Molecular Biology

GENOME-WIDE APPROACHES TO STUDY RNA SECONDARY STRUCTURE

Nathan Berkowitz

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

2016

Supervisor of Dissertation

____

Brian D Gregory, PhD

Associate Professor of Biology, University of Pennsylvania

Graduate Group Chairperson

____

Li-San Wang, PhD, Associate Professor of Pathology and Laboratory Medicine, Perelman School of Medicine

Dissertation Committee

Russ Carstens, MD, Associate Professor of Medicine, Perelman School of Medicine
Pamela Green, PhD, Crawford H. Greenewalt Chair, Professor of Plant and Soil Sciences, University of Delaware
John Murray, PhD, Associate Professor of Genetics, Perelman School of Medicine
Li-San Wang, PhD, Associate Professor of Pathology and Laboratory Medicine

This dissertation is dedicated to Donna and Gary Berkowitz

ACKNOWLEDGMENT

First I must thank my Advisor, Brian Gregory for his guidance and support on all of my projects. His drive and attention to detail have helped me become a better writer and scientist. I also have to thank my dissertation committee whose time and expertise have been invaluable improving the ideas behind my projects and keeping me on track. In addition, I thank the community of plant researchers here at Penn. Their discussions have helped me to shape and refine many of my projects.

I would also like to acknowledge all of my current and former labmates in the Gregory lab for their friendship and for all of the scientific discussion we have had. When I first joined the lab, Fan Li's peer mentorship was invaluable for orienting myself in the landscape of bioinformatics tools. Lunch discussions with Ian Silverman, who joined the lab when I did, have been a great forum for exchanging ideas, troubleshooting projects and maintaining sanity. I have enjoyed bouncing ideas off of Lee Vandivier and I have also enjoyed his cooking. Everyone else who has joined the lab has contributed to great discussions in the lab and fun moments outside. Thank you to Matthew Willmann who prepared many of the libraries I analyzed for this work.

I have had the pleasure seeing my graduate group grow under two terrific chairs, Maja Bucan, who endless cheers for all GCB students present and past, and Li-San Wang, who has constantly works to improve the program. I also thank Li-San and his group for productive discussions about the computational aspects of my projects and for hosting my database. In Li-San's group, I thank Micah Childress for his help designing the website which makes the database public, and Otto Valladares for helping to maintain it.

Finally, I have to thank all of the GCB students I've met during my time here at Penn. We are a small community, which is why it's so important that we've taken time to get to know each other, whether exchanging ideas at chalk talk, gorging on dim sum or trying to outwit each other at board game nights. The students who came before me made me feel welcome from when I first interviewed here and through my first two years. As larger classes have arrived, the community has gotten even more active and fun.

ABSTRACT

GENOME-WIDE APPROACHES TO STUDY RNA SECONDARY STRUCTURE

Nathan Berkowitz

Brian Gregory

The central hypothesis of molecular biology depicts RNA as an intermediary conveyor of genetic information. RNA is transcribed from DNA and translated to proteins, the molecular machines of the cell. However, many RNAs do not encode protein and instead function as molecular machines themselves. The most famous examples are ribosomal

RNAs and transfer RNAs, which together form the core translational machinery of the cell. Many other non-coding RNAs have been discovered including catalytic and regulatory RNAs. In many cases RNA function is tightly linked to its secondary structure, which is the collection of hydrogen bonds between complimentary RNA sequences that drives these molecules into their three dimensional structure.

Over the last decade, technology for determining the sequence of DNA and RNA has advanced rapidly, making transcriptome-wide expression profiling fast and widely available. In this dissertation, I discuss recent efforts to leverage this powerful technology to study, not just RNA expression, but several other aspects of RNA function. In particular, I focus on three tightly linked aspects of RNA biology: RNA-secondary structure, RNA cleavage, and regulatory small RNAs. I introduce a database for integrating, comparing, and contrasting techniques for determining RNA secondary structure including a technique developed in my dissertation laboratory. Additionally, I discuss a newly improved technology capable of detecting RNA cleavage events. Finally, I integrate RNA secondary structure probing and RNA cleavage detection to interrogate a family of genes important for eukaryotic small RNA-mediated silencing. These diverse analyses are just a few examples of the vast promises offered by adapting RNA-sequencing technology to probe RNA function across many cellular processes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1: Introduction

1.1 Post-transcriptional Regulation

1.1.1 The Central Dogma

Proteins are often thought of as molecular machines that do the work required for cellular life. They include motors, enzymes, signal receptors, structural elements and more. The Central Dogma, coined by Francis Crick, describes a model for protein synthesis.[1] Specifically, the information required to build a protein is stored as a linear sequence of deoxyribonucleic acid (DNA) nucleotides. This sequence is then transcribed to a chemically similar ribonucleic acid (RNA) molecule. The RNA is then translated to protein. Since it was first proposed, many nuances and exceptions have been described. However, the Central Dogma has been, and still is, a powerful model for describing gene expression.

1.1.2 Post-transcriptional Regulatory Processes

Historically, there have been many important advances in understanding gene regulation at the transcriptional level. These regulatory processes include the binding of transcription factor proteins to DNA that subsequently control gene expression from that locus. Transcription factors increase expression and transcriptional repressors inhibit expression from bound loci. The interplay of these proteins with DNA can explain many changes in gene expression. However, especially in eukaryotes, it has become increasingly clear that expression can also be controlled after RNAs are made by post-transcriptional regulatory processes (PTGR).[2–4]

A major factor in PTGR is transcript diversity. In eukaryotes, a single gene locus can give rise to multiple diverse transcripts. One straightforward consequence of this is that multiple different protein products can be made. More subtly, transcript variants can differ in the regulatory elements they contain leading to differences in stability and transcription. There are several ways in which transcript diversity can be introduced.[5]

Before a protein coding transcript can be translated, it must undergo several processing steps to become a mature mRNA. Genes are not encoded on the chromosome contiguously. The exons, regions to be included in the mature transcript, are separated by introns, sequences that are not included. The full length of the gene is transcribed including both introns and exons as a pre-mRNA transcript. Exons are joined together and introns are excised by the spliceosome, a large catalytic complex composed of proteins and RNA molecules.[5]

In this splicing process, not all exons are included in all transcripts. Optional cassette exons may be included or excluded resulting in multiple possible versions of the mature mRNA for the same gene. Additional, more complex exon inclusion patterns also affect the mature sequence. Exons may have alternate 5' or 3' splice junctions. Also, a pair of exons may be mutually exclusive. The combination of all of these diverse splicing choices leads to many possible mature mRNAs for a given pre-mRNA. This added diversity is an important mechanism of PTGR for protein coding genes as it leads to many possible protein products, which can serve different functions. Additionally, alternative splicing can be important for PTGR as it may result in inclusion or exclusion of regulatory elements that affect translation or turnover.[5,6]

In addition to being spliced, pre-mRNAs must also undergo modification to their 5' and 3' ends before they are mature. Both of these modifications increase the stability of the transcript. A modified guanosine, referred to as the 7-methyguanlate cap, is added to the 5' end. Additionally, a region of the 3' end is cleaved and a polyadenosine (polyA) tail is added to the cleavage site. The addition of the tail is directed by a specific sequence on the pre-mRNA, the polyadeylation signal (PAS): AAUAAA. A given pre-mRNA may contain more than one PAS and the polyadenylation machinery may select different PAS's on individual transcript molecules. Like alternative splicing, alternative polyadenylation results in transcript diversity as different PAS choices lead to transcripts with different 3' ends and potentially different protein products. Also, like alternative splicing, it can control the inclusion or exclusion of regulatory elements in the sequence of the mature mRNA.

Another source of transcript variation is alternative initiation of transcription. If there is more than one transcriptional initiation site, the transcriptional machinery may use them all. Just as alternative polyadenylation can lead to differences in the 3' ends of transcripts, alternative initiation can lead to differences in the 5' ends. These differences can lead to PTGR by generating variants that are translated with different efficiency.[7]

There are many possible ways in which regulatory elements can affect gene expression. One general mechanism is that they can act as recognition sites for sequence specific repressive factors. This is the case for transcripts regulated by a class of small RNAs (smRNAs) called micro-RNAs (miRNAs).

From transcription to degradation, RNAs are constantly bound to proteins. They affect all aspects of a transcript's life cycle including splicing, translation, trafficking and stability. Many RBPs are sequence specific and bind to recurring nucleotide patterns on RNAs known as motifs. Discovery of motifs has been an appealing goal for computational study.[4,8,9] Researchers have developed expectation maximization algorithms to find motifs without prior knowledge of the motif locations or motif sequences.[10] One drawback is that the algorithms do not determine the RBP associated with the motif. More recently, motifs for a large number of RBPs were determined experimentally.[11]

In addition to sequence motifs, RBPs may identify transcripts by recognizing RNA structural elements. One example of this type of structure based recognition is the mammalian iron response element, a short hairpin structure formed at the 5' end of several genes involved in iron metabolism. In the absence of iron, an RBP binds the hairpin and represses translation.[12]

1.2 RNA Secondary Structure

1.2.1 Non-coding RNAs and RNA Structure

One added complexity to the Central Dogma is the existence of RNAs that do not encode protein, but perform cellular functions without being translated (ncRNAs). Two of the most famous examples of such non-coding RNAs are transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) that are both required for translation in cells. While these ancient genes were once thought to be special exceptions to the general linear nature of the Central Dogma, many examples of important, functional ncRNAs have since been uncovered. For instance, RNAs have been found that act as catalysts[13] as well as RNAs that fulfill a wide variety of regulatory roles.

One broad category of ncRNAs which are intimately linked to secondary structure are small RNAs (smRNA). Many classes of small RNAs are processed from longer structured precursors.

Perhaps the most famous class of smRNAs are miRNAs. These begin as Pol II transcripts which are capped and polyadenylated. They fold back on themselves forming a hairpin loop with a structured stem. This stem is then cleaved two or more times to produce short (~21-22 nt) non-coding RNAs that target other transcripts using base pairing interactions.

In addition to miRNAs there are several other classes of smRNAs that function in PTGR. For example, small interfering RNAs (siRNAs), which are similar in size to miRNAs but differ from miRNAs in their biogenesis. They have a broader range of precursors than miRNA but they are also generated through cleavage of double stranded regions.

siRNAs can post-transcriptionally regulate their targets. One class of siRNAs acts similarly to plant miRNAs, triggering cleavage and degradation. Another class is processed from transcripts that are encoded at repetitive and heterochromatic regions. These siRNAs epigenetically silence the same types of regions from which they originate.

Another broad class of ncRNAs are long non-coding RNAs (lncRNAs). These transcripts are capped and polyadenylated, but they do not encode protein. There are many known lncRNAs and many more suspected, although not all have been assigned functions.

One well known example is Xist, a transcript from the X inactivations locus which is important in silencing one of the two X chromosomes in female mammals. Another is HOTAIR, which also has a role in regulating chromatin state. Both of these examples have high secondary structures and secondary structure is thought to be important to lncRNA function in general.

In many cases, the functions of ncRNAs depend on both their sequence and their structure. I will discuss some examples of regulatory ncRNAs and their structural requirements in more detail, but first it is worth defining RNA structure.

1.2.2 Levels of RNA Structure

In addition to being a conveyor of information, RNA is also a physical molecule with its own chemical properties. In fact, every RNA is composed of a linear sequence of ribonucleotide monomers linked together through covalent bonds. This chain is flexible and can fold into complex three-dimensional (3D) shapes. The 3D shape of an RNA depends on intra-molecular interactions to hold it into place.

Based on this model, RNA structure can be dissected into distinct layers of information. The *primary structure* of a RNA is simply its sequence of nucleotides from beginning to end. The *secondary structure* is the pattern of intra-molecular interactions between these nucleotides. The *tertiary structure* of a RNA is its 3D geometry that is directed by its underlying secondary structure.

Secondary structure is mostly composed of the hydrogen bonds between the nucleotide bases. Two of these bond types, the Watson-Crick pairs, mirror the bonds allowed in genomic DNA, Guanine (G) with Cytosine (C) and Adenine (A) with the Thymine (T) analog Uracil (U). Additionally, RNA can form hydrogen bonds between G and U. These are referred to as “wobble” base pairs because of their lower thermodynamic stability.

The three layers of structure information are not independent of one another. The tertiary structure depends on the secondary structure, which, in turn depends on the set of possible interactions dictated by the primary sequence. However, each layer seems to have unique information. Later in this chapter, I will discuss some current methods for predicting secondary structure from primary structure, and in subsequent chapters I will compare them to direct secondary structure measurements.

1.2.3 Biological roles of RNA secondary structure

For many RNAs, structure is linked to function. Examples of the importance of structure can be found in both splicing, the removal of introns from transcripts, and translation. The spliceosome, which is responsible for catalyzing the splicing reactions, is composed of proteins and small nuclear RNAs (snRNAs). These snRNAs must adopt specific structures for these catalytic reactions to occur properly.[14,15] In translation, the RNAs that make up the ribosome (rRNAs) must adopt complex 3D structures in order to function properly.[16–21]

Riboswitches are another example of the importance of RNA structure. These structured elements in mRNAs act as molecular detectors, typically for metabolites. In the presence of the riboswitch's preferred binding metabolite it changes its structural conformation resulting in a change in gene expression.[22–25] There are many examples of riboswitches in bacteria that respond to diverse organic and inorganic metabolites[24,26–30] and they have also been identified in eukaryotes as well.[22,31,32]

RNA secondary structure is also an important aspect of intra-molecular RNA-RNA interactions. Many classes of smRNAs use base pairing to identify target transcripts. Throughout this dissertation, I briefly discussed two such classes, small-interfering RNAs and microRNAs (siRNAs and miRNAs, respectively). In later chapters I will revisit them in more detail. Another such class is the small nucleolar RNAs, which guide RNA modification enzymes to various classes of substrate RNA molecules.[33–35] Secondary structure also plays a role in this targeting step as more highly structured transcripts are less accessible to miRNA binding.

Structure is important for regulation of translation. I discussed one example, the iron response element in which a specific secondary structure plays a role in PTGR by acting as a binding site for a repressive RBP. It is also possible for RNA secondary structures to directly repress translation by blocking the ribosome.[36,37]

In addition to these well characterized examples, there are intriguing findings that hint at even more roles yet to be discovered for RNA secondary structure. Genome wide analysis of structural trends has revealed global secondary structure patterns across mRNAs. One observation is that regions of relatively low structure tend to be found at start and stop codons. It has been hypothesized that these structural patterns are involved in ribosome recruitment and release.[38,39]

It has also been observed that the coding sequence of mRNAs has a different average structure than the 5' and 3' UTRs, however the relative differences in average structure between the regions differ between species. In two animal species, the UTRs tend to be more structured than the coding sequence while the opposite is found in the model plant *Arabidopsis thaliana.*[38] It's not obvious what this trend reflects, however more detailed structure probing will likely reveal additional details.

Another study, performed in the model plant *Arabidopsis thaliana*, investigated the genome wide relationship between secondary structure and total RNA expression. It found that mRNAs with extensive secondary structure were much less abundant that unstructured transcripts, suggesting that structure decreases abundance. The proposed mechanism for this decrease is that structured RNAs act as substrates for the RNA silencing machinery and are cleaved into smRNAs.[40]

It's unclear if this structure dependent repression is limited to plants or if it is a more general mechanism across organisms. It's also not clear whether the smRNAs themselves are functional or if they are simply byproducts of mRNA PTGR.

Ultimately, it is clear that RNA secondary structure is pervasively important in many aspects of biology and new roles for it are likely to be discovered. This is especially exciting given the recent advances in technology for studying secondary structure. Next, I will discuss some of the techniques that have been developed historically and how they have been advanced and updated.

1.3 Techniques for determining secondary structure

Efforts to determine RNA secondary structure have been diverse and inventive, including both computational and empirical approaches. One of the earliest technologies, X-ray crystallography, is still in use today and can produce very high quality three-dimensional structures. It works by interpreting the diffraction pattern of X-rays as they pass through crystallized RNA. Two groups were independently able to capture the structure of a yeast tRNA in 1974[16,41] using this technique. Since then, the structures of many important RNAs have been have been discovered using X-ray crystallography including the RNA components of the ribosome.[42,43] As powerful as this technique is, it does have some caveats. Obtaining high quality crystals of RNA is not trivial. In particular, crystallizing RNAs that are long or unstructured is difficult and in some cases may not be possible.

Nuclear magnetic resonance (NMR) is another technique that employs electromagnetic radiation to probe the three-dimensional structure of RNA. Unlike X-ray crystallography, NMR does not require crystals and can inspect RNA in solution. However NMR requires large quantities of pure, homogeneous RNA and usually relies on synthetic RNAs or RNAs produced by cloning. Such molecules are not guaranteed to be identical to their *in vivo* counterparts.

Both NMR and X-ray crystallography have been productive sources of RNA structures. As of this writing, the Protein Data Bank (PDB),[44] a central resource for three-dimensional structures of biomacromolecules, contains 472 RNA structures discovered through NMR and 652 from X-ray crystallography. It should be noted that this impressive collection represents decades of work. The oldest NMR entry is from 1991 while the first X-ray crystallography structure was deposited in 1978. This time scale illustrates the central drawback of these two powerful techniques: throughput. Both approaches produce highly informative three-dimensional structures, but each structure is a time and labor intensive enterprise.

Technical improvement are constantly making both techniques faster and more tractable, but technological growth has been much more rapid in two other fields, namely computing and nucleotide sequencing. Increases in power and decreases in cost have led several groups to develop approaches for studying RNA structure using purely *in silico* methods or transcriptome scale methods which employ high throughput RNA sequencing.

In silico methods use the nucleotide sequences of RNAs to predict thermodynamically favorable base pairings. The high-throughput sequencing methods generally adapt and expand existing smaller scale techniques to work with sequencing technology as explained below.

1.3.1 *In silico* techniques

The simple rules that govern RNA folding have made it an appealing target for computational prediction. For a given RNA sequence, there is a finite set of valid secondary structures since only three types of base pairs are possible, the two Watson-Crick pairings {A,U}, {C,G} and the RNA-specific pairing {G,U}. However, not all of these valid structures are equally likely. Each structure has a free energy which can be estimated using its unpaired bases. Higher free energies represent more unlikely structures, so many algorithms search for the valid structure with the minimum free energy (MFE).

It has been shown that, in the most general case, finding the MFE structure from a sequence belongs to a class of problems known as NP-complete.[45,46] It is widely suspected, although not proven, that such problems cannot be solved by polynomial time algorithms. In practical terms, this means that small increases in input sequence size lead to very large increases in computing time. Consequently, folding even small molecules *in silico* may be inherently intractable.

However, algorithms have been developed to find MFE solutions in polynomial time with some simplifying assumptions.[47,48] One major source of complexity is pseudoknots. Pseudoknots are structural features in which a series of continuous nucleotides participates in two different stem loops which are *not* continuous. Algorithms which do not consider structures containing pseudoknots are able to calculate MFEs efficiently. Some programs are even able to consider structures containing pseudoknots, in a limited way, without losing polynomial time efficiency.[46,49,50]

A general approach for computing structures has been to use dynamic programming to compute a table of pairwise bonding probabilities.[47,48,51] Algorithms differ in the models they use for estimating the individual free energy of each pair and in the features they offer. Support has been added for circular RNAs[52] and there are optimizations for parallel distributed computing.[53,54]

1.3.2 Chemical probing

The purpose of chemical probing techniques is to determine which of an RNA molecule's nucleotides are involved in pairing interactions and which ones are unpaired. This classification is useful when evaluating the output of *in silico* secondary structures. It can also be used to inform folding algorithms by imposing constraints about which bases are allowed to be paired. Additionally, the pairing state of a nucleotide likely affects its affinity for RNA-binding proteins.

There are multiple reactive groups on both the nucleotide bases of RNA and on its sugar backbone. A wide range of chemical adducts has been explored for labeling them. One straightforward strategy is to selectively label unpaired nucleotides with a chemical that targets the atoms involved in pairing. The corresponding atoms on paired nucleotides are inaccessible and are not labeled. There are several reagents useful for this purpose including dimethyl sulphate[55], klethoxal and carbodiimide metho-p-toluene sulfonate.[56] Each prefers to label a different set of atoms and is useful for probing different nucleotides (Figure 1.1).

Another approach involves acylating the 2' hydroxyl of the RNA's sugar backbone. The principle behind this strategy is not obvious. The reactive group being labeled is not directly protected by base pairing. Instead, the technique takes advantage of the fact that paired nucleotides are constrained in space in such a way that the negatively charged phosphodiester of the backbone is close the 2' hydroxyl protecting it from

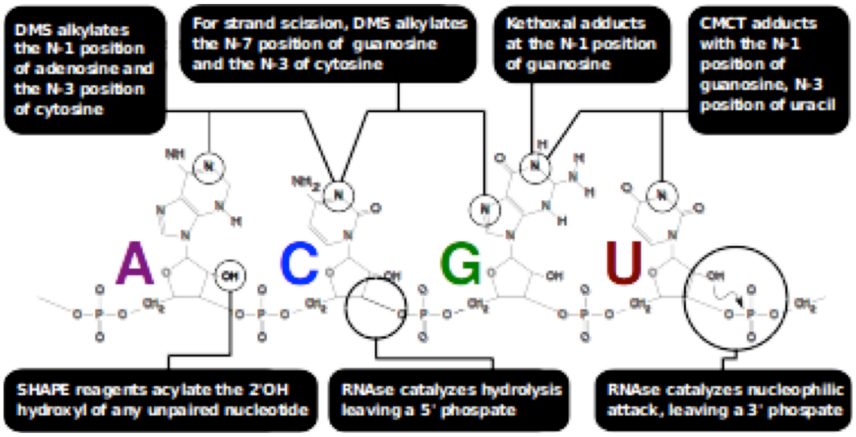

DMS alkylates the N-1 position of adenosine and the N-3 position of cytosine
For strand scission, DMS alkylates the N-7 position of guanosine and the N-3 of cytosine
Kethoxal adducts at the N-1 position of guanosine
CMCT adducts with the N-1 position of guanosine, N-3 position of uracil
A
C
G
U
N
NH
NH_2
H
O
OH
P
CH_2
SHAPE reagents acylate the 2'OH hydroxyl of any unpaired nucleotide
RNAse catalyzes hydrolysis leaving a 5' phospate
RNAse catalyzes nucleophilic attack, leaving a 3' phospate

Figure 1.1: Examples of labeling strategies. Many chemical probes can label the reactive Nitrogens on the Watson Crick edge of the RNA. Each has a target preference. SHAPE reagents acylate the 2' hydroxyl of the backbone in a nucleotide agnostic way. RNAses cleave the backbone. Some prefer to cleave between nucleotides involved in base pairs. Others prefer unpaired regions.

acylating agents (Figure 1.1).[57,58]

1.3.3 Enzymatic probing

An alternate approach to chemical probing is to take advantage of ribonucleases that preferentially cut RNA depending on it secondary structure. Methods that use this strategy use one or more nucleases to partially digest the RNA and then inspect the resulting pattern of fragments using polyacrylamide gel electrophoresis and autoradiography. The resulting pattern of bands reveals details about the actives of the nucleases. Because the enzymes have structure preferences, secondary structure information can be inferred from the pattern of cuts.[59–61]

Many structure specific nucleases prefer not just structural contexts but also specific sequences. While such enzymes have been successfully used to probe structure, they must be used in combination to effectively probe entire RNA molecules.[61,62] Nucleases that do not have a sequence preference can be used to probe in a more unbiased way. S1 Nuclease, Mung Bean (MB) Nuclease, P1 Nuclease, and RNase I are all able to preferentially cut single stranded RNAs without sequence specificity.[63,64] RNase V1, cleaves 4-6 nucleotide stretches of double-stranded RNAs without sequence bias.[65,66] Each of these enzymes, alone or with other reagents, has been used to learn the secondary structures of individual RNAs.

1.3.4 Adaptation of probing methods to high-throughput

The increasing availability of high throughput sequencing has made it possible to adapt both chemical structure probing and enzymatic structure probing to transcriptome-wide studies. I will discuss four current methods in more detail in the next chapter. The underlying idea is that a pool of RNAs from the full transcriptome is treated with one of the chemical or enzymatic probe discussed above prior to library preparation (Figure 1.2).

The chemical probes affect this step by blocking reverse transcriptase and casing it to fall off. This results in reads that end abruptly at the position occupied by the probe. Because all of the chemical probes discussed prefer unpaired nucleotides, positions where multiple reads end can be considered to be unpaired. There are some caveats to these approaches, which I will discuss, and it is important to choose an appropriate control as RT stops can be difficult to interpret.

Enzymatic probing approaches use RNA cleavage to reveal structure in high throughput using two libraries, one treated with an enzyme that prefers to cleave paired nucleotides and another treated with an enzyme that prefers unpaired bases.

One strategy is to generate nicks in the RNA with enzymes. 5' sequencing adapters can be ligated onto these nicks during library preparation. After sequencing, the ligation sites, and thus the nicks, are represented by the 5' read ends. If a given position was more commonly nicked by an RNAse that prefers paired or unpaired nucleotides, that position's structure can be inferred.

Alternately, it is possible to more fully digest away either dsRNA or ssRNA using similar reagents but different treatment conditions. The surviving regions of RNA can then be subjected to library preparation and sequencing. Under this strategy, positions that are more represented in the library depleted for ssRNAs can be considered to come from double stranded regions. Similarly, positions that are more represented in the library depleted for dsRNAs can be considered to come from single stranded regions.

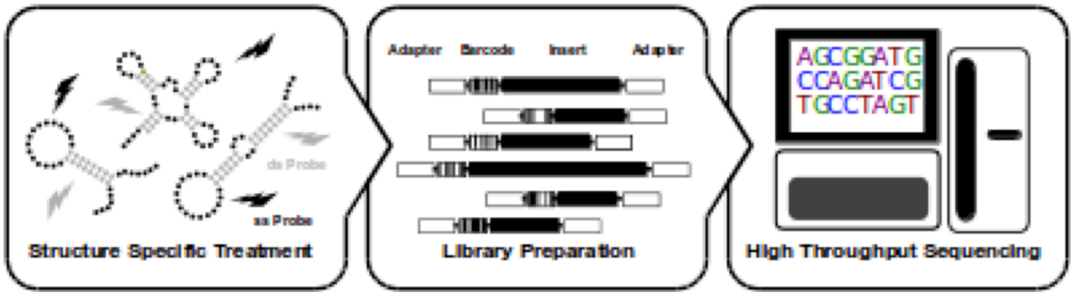


Figure 1.2: General Strategy for determining RNA secondary structure by high throughput sequencing. First RNA is treated with a reagent that is sensitive to structure and visible by sequencing. Treated RNAs are subjected to library preparation and sequencing

1.4 Small RNAs

As mentioned above, there are many known mechanisms of post-transcriptional gene regulation (PTGR), and most likely, more will be discovered. In general, PTGR can affect the stability, translation, and localization of a transcript. One major category of PTGR is RNA silencing mediated by various classes of small RNAs (smRNAs).

1.4.1 Classes of Small RNAs and their biogenesis

Two major classes of smRNAs important for RNA silencing are siRNAs and miRNAs and they differ in their biogenesis. miRNAs begin as Pol II transcripts. They are capped and polyadenylated. In some cases they have introns which are spliced out. This transcript, the primary miRNA (pri-miRNA) folds back on it self creating a stem loop structure. An endonuclease makes a double stranded cut in the stem, typically removing the 5' and 3' end of the transcript and leaving the precursor miRNA (pre-miRNA)[67,68] although some pri-miRNAs are initially cut to remove the loop.[69] Alternatively, pre-miRNAs may begin as introns which are spliced directly out of genes without an initial cleavage step. In plants, the loop is sometimes removed first resulting in a pre-miRNA with intact 5' and 3' ends. Finally one or more double stranded cuts are made in the stem resulting in a duplex of smRNAs. The two strands dissociate and one or both of them are mature miRNAs. These mature miRNAs are loaded onto ARGONAUT proteins and, with other factors, they form RNA-induced silencing complexes (RISCs). Some of the details differ between organisms. For example, animal pri-miRNAs are cleaved by the enzyme DROSHA and pre-miRNAs are cleaved by DICER. In plants both cleavage steps are done by a DICER ortholog.

There are a variety of mechanisms that produce siRNAs each involving a double stranded RNA precursor.[70] Once type of precursor is created by two promoters facing eachother in close proximity causing transcription of two complementary transcripts. Another mechanism relies on a class of enzymes called RNA dependent RNA polymerases (RDRs). These enzymes use an RNA molecule as a template to create its reverse complement resulting in a double stranded RNA. In some mechanisms, this RNA molecule is long and the siRNAs are created through cleavage, however some RDRs are able to create siRNAs of the correct length simply by arresting polymerization.[71–73]

Both miRNAs and siRNAs act as specificity factors, using base pairing interactions with target RNAs to guide effector complexes to silence specific transcripts. miRNAs are ~22-nt RNAs. Along with ARGONAUT proteins and other factors they form RNA-induced silencing complexes (RISCs). These complexes use their miRNAs to detect fully or partially complementary transcripts and silence them.[74–76]

1.4.2 Mechanisms of smRNA-mediated RNA silencing

In animals, miRNAs base pair imperfectly with their targets, often in the 3' UTR of the transcript. They regulate their targets by inhibiting translation.[76] In some cases, this translational silencing seems not to affect the abundance of the target[74,75,77,78] although there is evidence that in some cases targeting leads to transcript degradation.[79–81]

In plants, The primary mechanism of miRNAs is to direct endonucleases to cleave the target site. This results in a 5' cleavage product without a polyA tail and a 3' product without a cap, both of which are quickly degraded.[82–85] The full range of complexity of miRNA silencing mechanisms is outside the scope of this dissertation, however it is important to note that there are multiple related mechanisms and many are current topics of research.

Similar to miRNAs, siRNAs incorporate into RISC complexes and effect PTGS, however a more diverse set of mechanisms has been described for them. Multiple classes of siRNAs found in *Arabidopsis* will be discussed in more detail in Chapter 5.

1.4.4 Methods for identifying targets for smRNA-mediated RNA silencing

Because smRNAs identify their targets through base pairing interactions, the sequence of the target region must be partially or fully complementary to the sequence of the smRNA. This relationship has made it possible to predict targeting interactions *in silico*. Much of this work has focused on predicting animal miRNA targets. Algorithms specialized for this application often take into account the fact that such miRNAs have a “seed” region, nucleotides 2-8 from the 5' end of the miRNA, which must be highly complementary to the target site.

A wide variety of algorithms has been used to search for matches including deterministic methods including Support Vector Machines[86], Hidden Markov Models,[87] boosted genetic programming algorithms.[88] In general, they then evaluate the predicted thermodynamic stability of the full match.[87,89–93] In addition to the match quality, some algorithms use additional information such as the thermodynamically predicted secondary structure of the target site[94] and its evolutionary conservation.[95]

In contrast to animal miRNAs, the miRNAs found in plants are thought to require near perfect sequence matches to their targets. Consequently, approaches for detecting such targets typically do not take into account any particular sub-sequence of the miRNA. Plant miRNA have been identified *in silico* by more general matching algorithms such as matching with regular expressions[96,97] and BLAST.[98] Computational techniques have been developed specifically to detect plant miRNAs and some are available as online tools.[99,100]

Although these were developed with miRNAs in mind, it is reasonable to expect them to detect the targets of plant siRNAs as well. Some siRNA producing transcripts are targets of their own siRNAs suggesting that siRNA target matching in plants is as precise as miRNA matching. In a later chapter, I will discuss using the target prediction software psRNATarget[100] to detect targets of plant siRNAs.

CHAPTER 2: An Integrated Resource for Secondary Structure Data

This section was adapted from:
Berkowitz N.D., Silverman I.M., Childress D.M., Kazan H., Wang L.S., et al. A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). BMC Bioinformatics. (2016) May 17;17(1):215.

Abstract

RNA molecules fold into complex three-dimensional shapes, guided by the pattern of hydrogen bonding between nucleotides. This pattern of base pairing, known as RNA secondary structure, is critical to their cellular function. Recently several diverse

methods have been developed to assay RNA secondary structure on a transcriptome-wide scale using high-throughput sequencing. Each approach has its own strengths and caveats, however there is no widely available tool for visualizing and comparing the results from these varied methods. To address this, I developed Structure Surfer, a database and visualization tool for inspecting RNA secondary structure in six transcriptome-wide data sets from human and mouse (http://tesla.pcbi.upenn.edu/structuresurfer/). Users have the ability to query individual loci as well as detect trends across multiple sites. Here, I describe the included data and illustrate its function with known structural elements and example use cases in which combined data is used to detect structural trends.

2.1 Background

RNA molecules serve as both conveyors of genetic information and as molecular machines with specific structural and catalytic functions in the cell. The function and regulation of every RNA molecule depends on its specific secondary structure, the intricate pattern of hydrogen bonds between complementary ribonucleotides that forms in its specific cellular environment. For instance, the ribosome, the central enzymatic complex in protein translation, is the classic example of an RNA-based machine, and thus the structure of its RNA subunits (ribosomal RNAs (rRNAs)) has been carefully dissected using detailed analyses. However, thousands of other structural RNA elements and catalytic RNAs exist in the cell, and the resources required to study them in more detail are mostly unavailable for large-scale use by the broader research community.

Advances in high-throughput sequencing technologies have allowed a significant increase in technical development of methods for studying RNA secondary structure on a transcriptome-wide scale. This has led to a diverse collection of sequencing-based approaches available for interrogating RNA secondary structure, and thus there are a number of large-scale data sets that are currently publicly available.[39,101–103] There are important methodological differences between these high-throughput structure-probing techniques, but the unifying principle is that they involve treating RNA samples with a reagent that selectively reacts with nucleotides depending on their base pairing status and then interrogating the treated RNA by high-throughput sequencing. I will discuss all of the methods that I curated for this project. In summary, it includes two chemical data from two chemical probing approaches and two enzymatic probing approaches. The enzymatic approaches both involve proteins that cleave RNA.

There are two methods that take advantage of ribonuclease (RNase)-mediated cleavage of RNA bases that are either double- or single-stranded (ds- and ssRNase, respectively). The first example is Parallel Analysis of RNA Structures (PARS), which requires two high-throughput sequencing libraries per sample. One library is treated with the ssRNase-specific RNase S1, while the other involves cleavage by the dsRNase-specific RNase V1. Both RNase treatments are titrated for single hit kinetics, meaning that each RNA molecule is cleaved only once by the nuclease used for treatment and thus it is not fully digested. The resulting singly cleaved RNA ends are immediately used as the substrate for ligation of a 5’ adapter molecule as the first step in high-throughput sequencing library preparation. Sequencing libraries prepared in this way produce reads whose 5' ends directly correspond to the site of nuclease cleavage. The structure of an RNA molecule can then be inferred from the relative number of RNase S1 (unpaired) and V1 (paired) cuts at each nucleotide position[39].

Similar reagents are used in ds/ssRNA sequencing (ds/ssRNA-seq) but to a different effect. As with PARS, each RNA sample is split into two aliquots, which are then treated with either an ssRNase (RNaseONE) or dsRNase (RNase V1). However, instead of utilizing single hit kinetics on the RNA samples, the nucleases are allowed to proceed to full digestion. The resulting RNase-resistant regions from each treatment are sequenced, and a structure score is then computed for each detectable nucleotide position by directly comparing the sequencing read coverage between the dsRNA- and ssRNA-seq libraries[40].

Two other approaches whose data I curated (see Methods) have combined chemical probing of RNA secondary structure with high-throughput sequencing technologies. For these approaches, unpaired RNA bases are labeled with a small molecule that inhibits elongation by reverse transcriptase (RT) used for cDNA synthesis during sequencing library preparation. This block in RT elongation results in termination of the cDNA molecules at the sites of these modified single-stranded nucleotides. Therefore, the resulting sequencing reads have 5' ends at the site that was labeled by addition of the chemical adducts.

DMS-seq is named for the labeling reagent that it employs, dimethyl sulfate (DMS). This small molecule labels unpaired adenosines and cytosines, but does not react efficiently with these nucleotides when they are base paired with another nucleotide[104,105]. Unlike the nuclease-based methods, DMS-seq does not include a reagent that specifically labels paired nucleotides. Instead, it directly assesses unpaired bases by measuring the DMS reactivity of nucleotides in natively folded RNA molecules compared to a control library where purified, denatured RNAs are treated with DMS and used as substrates in sequencing library preparation[101]. Double-stranded RNA regions are then inferred based on absence of DMS-seq signal at those nucleotides.

The other chemical-based structure probing method is selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq)[106–108], which uses any of several reagents that selectively label the 2' hydroxyl of unpaired nucleotides. Like DMS, this label causes RT to terminate due to the inhibition of its ability to elongate, which ultimately results in sequencing reads whose 5’ terminal nucleotide corresponds to the labeled position. The 5’ end read depth of each position in the treated library can then be compared to the corresponding read depth in an untreated DMSO control. This approach has recently been updated to allow higher resolution of RNA secondary structure, especially in mammalian transcriptomes. Specifically, the recently developed in vivo click SHAPE (icSHAPE) added an additional improvement to this general approach, in which the 2’ hydroxyl-labeling reagent also contains a biotin moiety, allowing enrichment of labeled RNA fragments in the final sequencing libraries.[102]

Although these techniques have been used to generate large-scale, broadly useful RNA structure probing data, there is no available resource that provides convenient access to these important data sets. Furthermore, there is no easy way to directly compare the results from these disparate approaches. To address this gap, I developed Structure Surfer, a database for exploring and comparing data generated by these new high-throughput structure-probing techniques (http://tesla.pcbi.upenn.edu/structuresurfer/). To do this, I have curated a comprehensive database of RNA secondary structure scores produced by the described experimental approaches. Structure Surfer allows users to query individual genomic loci of interest and visualize the local structural environment to directly compare the various methods. Additionally, I have included a tool for aggregating data across multiple genomic loci that allows users to query transcriptome-wide structural trends in a collection of regions of interest (e.g. all transcript start codons). In total, Structure Surfer provides an important and easy-to-use resource for querying and comparing the high-throughput RNA secondary structure probing data that is available for mammalian transcriptomes.

2.2 Methods

ds/ssRNA-seq

HEK293T cells were seeded in 15 cm standard Corning tissue culture treated culture dishes (Sigma, St Louis, MO), grown to 90% confluence (approximately 18 million cells) in DMEM media (Life Technologies, San Diego, CA) supplemented with L-glutamine, 4.5 g/L D-glucose, 10% fetal bovine serum (FBS (Atlanta Biologics, Atlanta, GA)), and Pen/Strep (Fisher Scientific, Waltham, MA).

RNA was isolated using the Qiagen miRNeasy RNA isolation kit following the manufacturer's protocol (Qiagen, Valencia, CA). Two aliquots of 50 mg were used to make two replicates each of dsRNA-seq and ssRNA-Seq libraries. These two types of structure-specific libraries were constructed as previously described[38,40].

Data Resources

I curated RNA secondary structure data from two published studies of the human transcriptome: DMS-Seq[101] and PARS[103], as well as previously unpublished structure scores from our ds/ssRNA-Seq data set for human HEK293T cells. Additionally, I compiled the scores from both in vitro and in vivo icSHAPE experiments in mouse[102]. The icSHAPE scores were reformatted and loaded directly into a mySQL database. For the other methods, I obtained the raw high-throughput sequencing reads and calculated the structure scores similarly to the published method specific to each one. All scoring functions are summarized below.

Genome Coverage

For DMS-seq, PARS, and ds/ssRNA-seq data sets, raw reads were trimmed using cutadapt[109]. This step removes any contaminating 3’ adapter sequences caused by inserts shorter than the sequencing read length. Trimmed and untrimmed reads were combined and mapped to the human genome using TopHat[110]. Reads that could not be trimmed or mapped were discarded. I allowed up to two mismatches per read and a maximum edit distance of two. I discarded reads that mapped to more than five locations. For DMS-Seq and PARS data, we computed the read coverage at each position in the genome with bedtools[111] using only the 5' most nucleotide of each read. When calculating coverage for ds/ssRNA-seq, the entire read was used.

DMS-seq Scores

DMS labeling of a nucleotide causes RT to stall during the cDNA synthesis step of RNA-seq library construction. Unstructured nucleotides, those that are not involved in base pairing, are more highly reactive with DMS and thus they are more likely to be the site of such a stall. Thus, the resulting RNA-seq reads from this type of high-throughput structure probing technique have 5' ends corresponding to the reactive, unpaired position. However, DMS labeling is not the only possible explanation for positions with a high tendency to cause RT stalls. For this reason, DMS-seq scores are expressed as nucleotide reactivity compared to a denatured control. The signal at each position is calculated based on the normalized number of 5' read ends mapping to that position in the native structure library compared to the control[101].

$$R_i = \frac{D_i \,/\, D_{max}}{C_i \,/\, C_{max}}$$

The reactivity R for position i is computed by first dividing the 5' read end coverage at that position, Di by the maximum 5' read end coverage in the library, Dmax. The resulting ratio is divided by Ci, the 5’ end read coverage at position i in the denatured control library normalized to the maximum 5’ end read coverage of the control library, Cmax. This reactivity score represents the degree of over-representation of RT stops in the DMS treated library compared with the control. High scores indicate positions where RT stops were frequent suggesting an unpaired nucleotide labeled by DMS.

icSHAPE Scores

As with DMS-seq, icSHAPE scoring reflects the higher reactivity of unpaired nucleotides compared to nucleotides involved in pairing. Reactivity is calculated from the count of 5' read ends covering each position. These counts are normalized to counts from a no-reagent background library and adjusted according to a background base density[102].

$$R_i = \left(D_i - C_i\right) / \left(B\right)$$

Reactivity R for position i is based on the 5' read end coverage at that position, Di, minus the coverage in the DMSO treated control library, Ci. The background base density profile for each transcript, B, is defined as the sequencing depth of each base in the DMSO library.

PARS Scores

PARS scores reflect the differential cleavage of paired and unpaired regions to ribonucleases. Unpaired regions are more cleaved by RNase S1 while paired regions are more cleaved by RNase V1. Both enzymes create RNA fragments with 5' phosphate groups by cleaving in their respective preferred regions. These ends are directly ligated onto sequencing primers. After cDNA synthesis and sequencing, each read has a 5' end corresponding to a cleavage site. Scores were calculated from the count of 5' read ends covering each position in the two nuclease treated libraries. Each score is based on the log ratio of the two coverage scores. The generalized log ratio is calculated by adding one count per position to both the numerator and the denominator before calculating the log ratio. This allows scoring of positions with positive counts in one of the two input libraries but no counts in the other library. Such positions are of interest because there is evidence that they are in a particular structural state, but the standard log ratio for them is undefined. A 5' nucleotide (nt) rolling average is applied for smoothing. Positions with no coverage in either library were omitted[103].

PARS structure score S for position i is the generalized log ratio of the normalized 5' end coverage for that position in the RNase V1 library and the corresponding coverage in the RNase S1 library. For each position, this value is calculated across the surrounding 5 nt window.

$$S_i = \log_2\left(\sum_{j=i-2}^{j=i+2} \frac{V1_j + 5}{5}\right) - \log_2\left(\sum_{j=i-2}^{j=i+2} \frac{S1_j + 5}{5}\right)$$

Ds/ssRNA-seq scores

Unlike scores from the other methods, ds/ssRNA-seq scoring takes into account all positions from each read rather than the 5' end coverage only. It employs similar reagents to PARS, but uses a longer enzyme treatment resulting in more complete digestion of each enzyme's preferred structure type. After cDNA synthesis and sequencing, reads represent regions that were protected from structure specific digestion. For each position the score is the generalized log ratio of the normalized counts in the two libraries[38].

$$S_i = log_2\left(ONE_i + 1 / V1_i + 1\right)$$

Visualization

The database’s plotting tool is implemented using the Python package PyGal. For plotting purposes, scores are scaled and re-centered to reveal local structural patterns and to make the data sets visually comparable. For the same reason, DMS and icSHAPE scores, which represent nucleotide reactivity as opposed to degree of structure, are inverted when displayed such that high scores indicate evidence of paired nucleotides in all data sets. Raw scores are available for download alongside the plots.

Availability of data and materials

All of the ssRNA- and dsRNA-seq libraries included in this study were made by my colleague, Ian Silverman, in Brian Gregory's lab. These data, from HEK293T cells, were deposited in GEO under the accession GSE72681. The remaining data were obtained from published studies. PARS, DMS-seq, and icSHAPE data were downloaded from GEO using the accession numbers GSE50676, GSE45803, and GSE60034, respectively. The complete Structure Surfer database is available as a MySQL dump file at PennBox, https://upenn.app.box.com/s/1kj2f1w994sp3jmaakqhy9cw2w11vajk. The Python search tool and database schema can be found at GitHub, https://github.com/nberkow/StructureSurfer. The structure score profiles for ~100 RBPs (as shown in Figure 2.5) calculated by Structure Surfer are available for download at http://tesla.pcbi.upenn.edu/structuresurfer. No login is required to access these resources.

2.3 Utility And Discussion

Database content

The database contains structure scores from four methods including six individual experiments across human and mouse. The score coverage varies greatly between methods. Despite having the lowest sequencing depth, the ds/ssRNA-Seq experiment produces the greatest score density. However this is not surprising given that the method uses all nucleotides covered by each read to generate scores while all of the other methods use only a single nucleotide per read when calculating scores. The most sparse scores come from the human PARS data set which covers only ~1 megabase of the transcriptome.

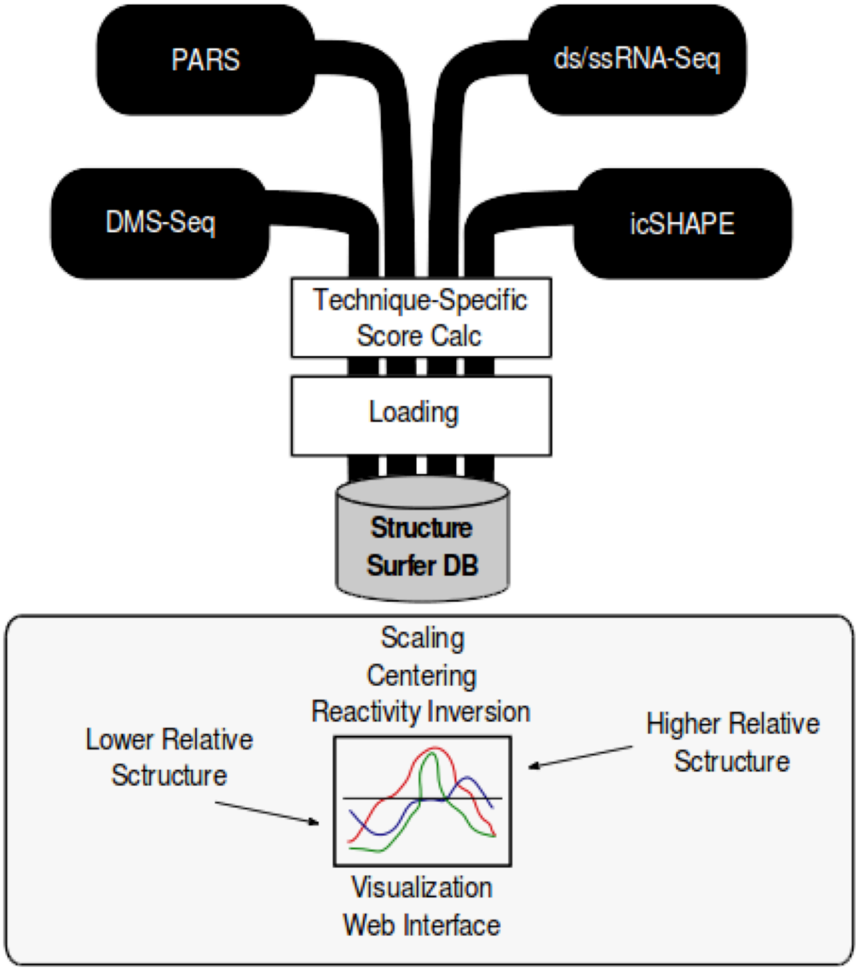

PARS
ds/ssRNA-Seq
DMS-Seq
icSHAPE
Technique-Specific Score Calc
Loading
Structure Surfer DB
Scaling
Centering
Reactivity Inversion
Lower Relative Sctructure
Higher Relative Sctructure
Visualization Web Interface

Figure 2.1: Diagram of the Structure Surfer database. Four different experiment types were acquired from GEO. Each one was analyzed with one of the technique specific scoring methods summarized above. The scores were loaded into the database and can be downloaded directly or can be visualized using a web tool. The visualization scales and centers the scores and ensures that high scores indicate more nucleotide pairing.

PARS, DMS-Seq, and icSHAPE all use a single base pair per read to calculate scores but the libraries used in the DMS experiment were sequenced to a higher depth which likely explains its greater score density (Table S2). The two icSHAPE experiments, which were sequenced to the highest depth of all the data sets included, produced an intermediate number of scored positions indicating that each scored position represents a greater number of reads on average. Each of the different methodologies produces scores that follow a distinct distribution (Figure 2.2) making it difficult to draw direct comparisons between them. These differences are likely due in part to differences in reagent kinetics. PARS and ds/ssRNA-seq, for example, employ similar reagents but PARS digests RNA very mildly resulting in single hit kinetics while ds/ssRNA-seq involves digesting regions of RNA to near completion. Other differences may arise from normalization strategy, as with the two nucleotide labeling techniques. DMS uses, as a normalization control, a denatured RNA sample, which is more highly reactive to DMS. In contrast, icSHAPE uses an RNA sample treated with solvent only, which reflects absence of icSHAPE reactivity. Structure Surfer addresses this by allowing users to focus on local structure patterns and transcriptome-wide structure trends.

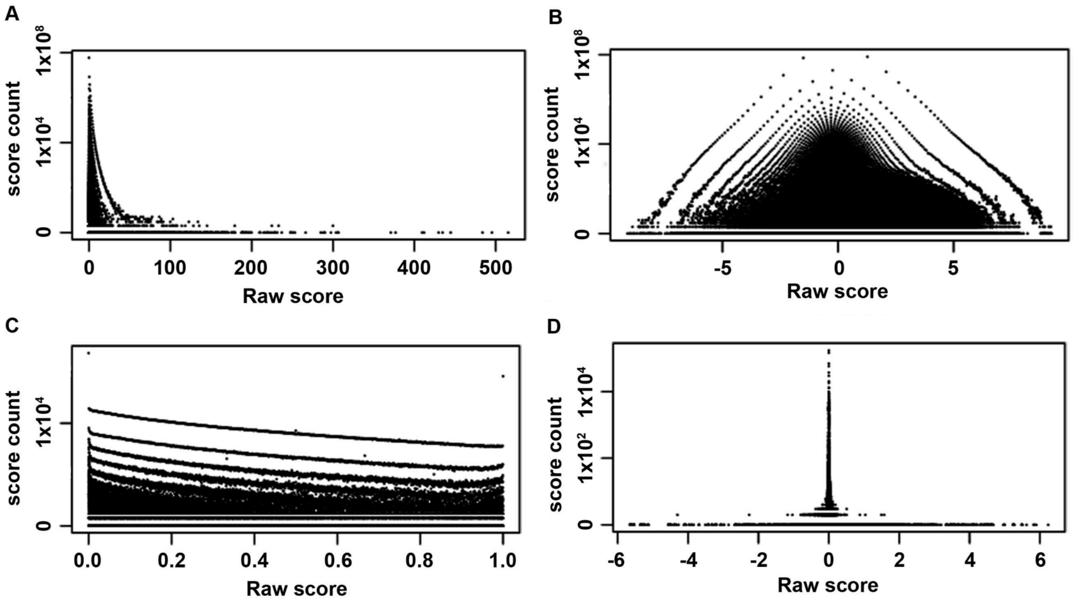


Figure 2.2: Score distributions for the four data types. Raw scores are given on the x axis, the y axis represents the number of times each score appears in the database for a) PARS b) ds/ssRNA-Seq c) icSHAPE and d) DMS.

Structure examples

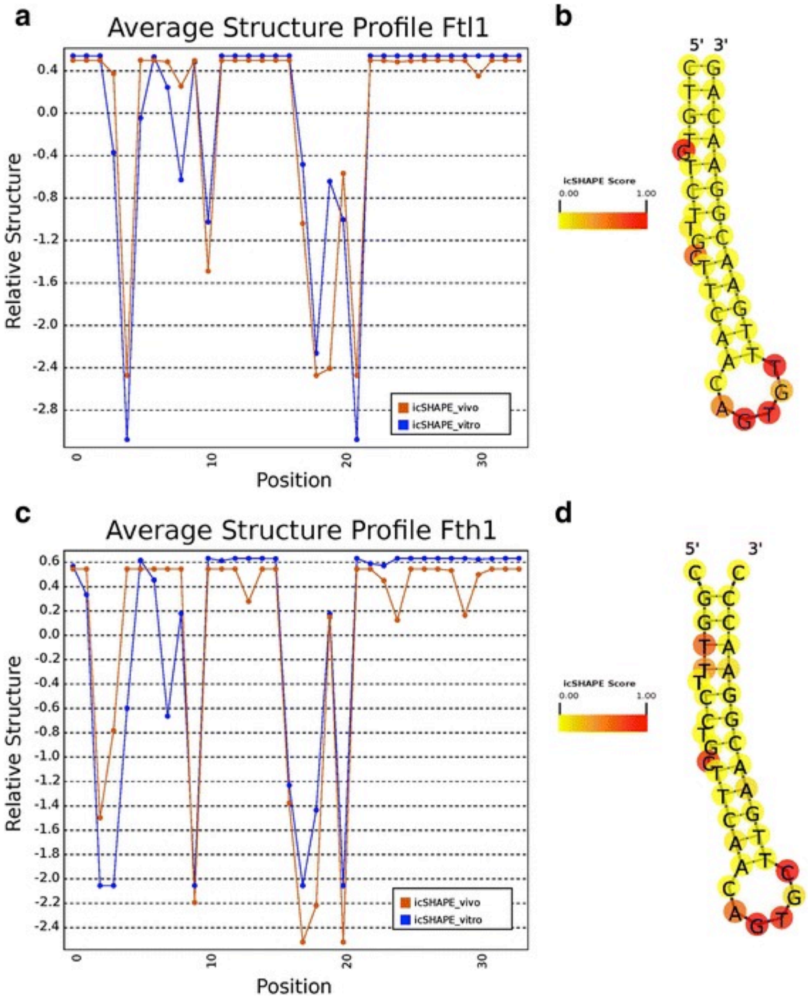


Figure 2.3: icSHAPE score profiles for the iron response element (IRE) hairpins of murine Ftl1 (a) and Fth1 (c) visualized using Structure Surfer's standardized data output. The Ftl1 IRE is located at position 76 to 110 in the transcript and, in the genome, is located on chromosome seven from position 45459777 to 45459811 on the non-reference strand. Fth1's IRE is located at position 83 to 117 in transcript variant 1. In the genome its coordinates are from 9982728 to 9982762 on the reference strand of chromosome 19. In vitro reactivity scores from the database are superimposed on in silico predicted structures for Ftl1 (b) and Fth1 (d) using SAVoR [4]. Red indicates positions with higher reactivity, showing evidence of low secondary structure. Positions colored in yellow have lower reactivity and are more likely paired

In order to develop our visualization of RNA secondary structure scores, we inspected a well-characterized class of highly structured elements, the iron response element (IRE). IREs are short stem-loops that act as binding sites for the RNA-binding protein (RBP) IRE-BP. They are found within the 5' untranslated regions (UTRs) of several mRNAs including two that encode the heavy and light chains of Ferritin in mouse, Fth1 and Ftl1, respectively. We visualized these two specific IREs using the database's icSHAPE structure scores (Figures 2.3A and C). In both structure score profiles, we see a 5 nucleotide stretch of low structure scores indicating an unpaired region. Indeed, each of these corresponds to the position of the unstructured loop region of the IRE. Also as expected, the structured stem region of the IRE has comparatively high scores. The 5' and 3' ends of the feature, which are not predicted to participate in the stem, have intermediate scores (Figures 2.3A and C).

In both score profiles, there are several single nucleotide positions along the stem region with

sharply low structure scores. We used the RNA annotation tool SAVoR[112] to superimpose icSHAPE reactivity scores onto RNAfold structures for the two loops (Figures 2.3B and D). Because Ftl1 is on the negative strand with respect to the genome, its score were reversed in order before they were superimposed. Strikingly, the two most reactive positions outside of the loop region in Ftl1 correspond to single nucleotide bulges in the stem at positions 5 and 11. This is not as clear in Fth1. While bulges in the predicted structure do generally correspond to peaks, as in the highly reactive bulge at position 11, there are two highly reactive nucleotides at positions 4 and 5, which are predicted to be paired. This provides evidence for the importance of structure probing techniques to define regions that differ in structure in vivo and in silico. One explanation for the differences is that in silico techniques do not always generate true RNA secondary structure. This may be a limitation of the algorithm used or it may be the result of nucleotide modification affecting structure in a way not reflected by the input sequence. For icSHAPE in particular, very high structure scores sometimes represent bases that are highly constrained, but in a way that makes them more rather than less reactive. More detailed experiments are needed to understand the exact source of disagreements between annotated structures and icSHAPE scores. Structure Surfer allows such differences to be detected easily and visually.

The human homologs of the mouse IRE features have no scores in any of the four human data sets, which illustrates a key issue to consider when dealing with RNA secondary structure. Structure measurement depends on RNA expression, sequencing depth, and technique specific biases. Many regions of potential interest have no scores or low score density. Fortunately, it is still possible to interrogate regions with low score density to detect overarching structure trends using a data aggregation approach.

Structure Surfer's interface provides such an approach by allowing users to input multiple regions aggregated into a single bed file and find the average structure score for all of the incorporated data sets across this collection of regions. This is useful for investigating repeated structural patterns across functionally related regions. For example, it has been noted that there are local decreases in RNA secondary structure at the start and stop of the coding sequence (CDS)[38,40,103,113]. To test Structure Surfer's aggregation mode, we queried the database with a set of sites containing every annotated CDS start codon in the human genome centered in a window of 9 nucleotides up- and downstream of these elements. Similarly, another file was entered using every CDS stop codon and their 9 nt up- and downstream surrounding sequences. When averaged across all of the input features, every human data set shows a dip in secondary structure around both the CDS start and stop codons (Figures 2.4A and B, respectively).

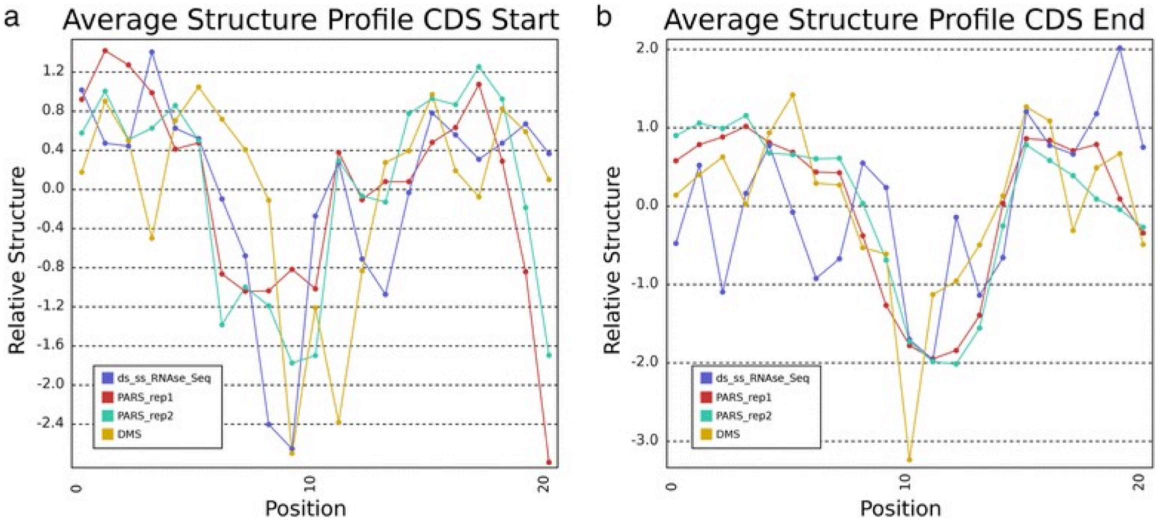


Figure 2.4: a-b Structure scores from the Structure Surfer database aggregated across all annotated human start codons (a) and stop codons (b). The three nucleotides of the start (ATG) and stop codons (e.g. TAA) occupy nucleotide positions 9–11 on each of the plots respectively. The score at each position is calculated as the average score across all nucleotides at that position relative to the codon. At both starts and stops, we note a dip in secondary structure consistent across experiments indicating that these positions, on average, have lower structure than the nucleotides surrounding them

Individual CDS start and stop sites may have very low score densities, but taken together, their average scores indicate broad agreement between the data sets and agreement with this previously described structural trend in numerous eukaryotic organisms[38,40,103,113]. This example shows how Structure Surfer can be used to reveal trends in RNA secondary structure across biologically related regions.

Example use case: RNA-binding protein interaction motifs

As an example application of Structure Surfer, we also used it to query the structural patterns at and around RBP interacting motif sites. Many RBPs bind their target transcripts according to sequence specificity, however it is likely that the structural environment around these sequences is also important. A recent high-throughput study applied the RNAcompete protocol to identify sequence motifs for 244 RBPs across multiple organisms[11]. We selected the RBPs from human and mouse that were interrogated by this study, and scanned both genomes for matches to RNAcompete-derived motifs. For each selected RBP, we computed an average structure score across all matching sites (all data from these analyses can be downloaded from http://tesla.pcbi.upenn.edu/structuresurfer/RBP_motif_structure.pdf).

We found several examples of RBPs whose predicted binding sites show a consensus structural environment across experiments. For example, motif matches for cytoplasmic polyadenine (polyA) binding protein-5 (PABPC5) show a strong unstructured trend when structure scores of all sites are averaged (Figure 2.5A). We observe the same result when we search for PABPC5 sites in the mouse genome and average their icSHAPE scores (Figure 2.5B). The opposite trend is found for motifs recognized by SNRPA, a component of the splicing machinery. All experiments report a local peak in structure at SNRPA motif sites in both human and mouse (Figures 2.5C and D).

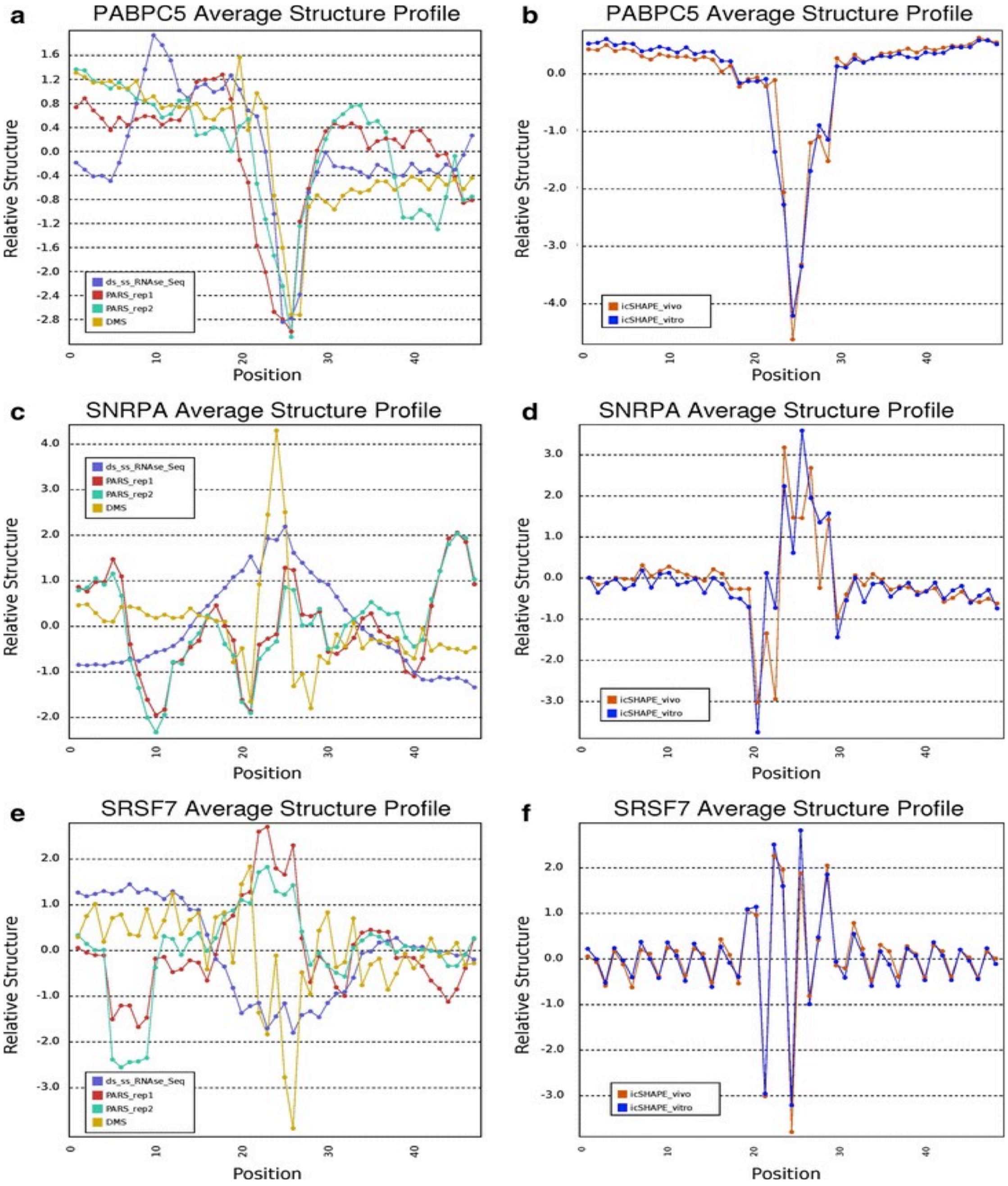


Figure 2.5: Examples of structure score aggregation using the data from Structure Surfer across RBP motif match sites for three RBPs, PABPC5 (a-b), SNRPA (c-d), and SRSF7 (e-f). Human structure scores are aggregated at match sites in the human exome (a, c, and e), and mouse

scores are aggregated at match sites in the mouse exome (b, d, and f). In all examples, the RBP interacting motif sequence is a heptamer occupying nucleotide positions 21–27. The score at each position is calculated as the average score across all nucleotides at that position relative to the RBP motif. PABPC5 shows a consistent dip in secondary indicating that sites matching its motif have, on average, less secondary structure than surrounding nucleotides. The SNRPA motif shows the opposite trend. Specifically, the average structure scores at sites containing this motif are higher than the surrounding nucleotides indicating that these sites tend to be double stranded. Sites for SRSF7 show a more complex pattern in which the different experiments do not form a consensus. PARS demonstrates evidence for a peak in average secondary structure at SRSF7 motifs, while ds/ssRNA-seq and DMS display evidence for a dip in average secondary structure. The icSHAPE experiments both show a region where some positions appear to be involved in base pairing and others appear unpaired

Unlike the examples above where we consistently find the same pattern across the different structure data sets, we also observe sites where there is not a consensus. For instance, the collection of predicted interaction sites of SRSF7 appear to be structured according to PARS, but unstructured according to DMS and ds/ssRNA-seq (Figure 2.5E). Interestingly, the icSHAPE experiments report an average structural environment with some highly reactive positions and some positions that appear protected (Figure 2.5F). One possible explanation for the icSHAPE result is that highly reactive sites compete for reagent with their slightly less reactive neighbors even if the entire region is unstructured. If this is the case it may also explain the difference in signal between the other methods. While it is difficult to interpret non-consensus sites, they may provide some insight into the types of features that are differentially detectable between the four methods.

2.4 Summary

Structure Surfer is a database of RNA secondary structure information compiled from six different experiments across four distinct methods from human and mouse. The web interface allows users to visualize secondary structure patterns at any genomic region of interest. For instance, we visualized a known feature type, the IREs of murine Ferritin heavy and light chain mRNAs, and revealed a pattern of structure scores that match the in silico RNAfold-predicted secondary structure for these elements. When the scores provided by the structure probing methods are sparse, we find that a data aggregation approach reveals broad overall structural trends in a collection of transcript regions (i.e. the area around all transcript start codons). Therefore, we have also implemented a data aggregation option in the web interface to interrogate files containing a collection of such regions. Using this interface, we demonstrate the ability to visualize a known structural trend, specifically the dips in secondary structure at translation start and stop sites. Also using aggregation, we see intriguing patterns of secondary structure at predicted binding sites of specific RBPs. However, these are only two of the many possible use cases of Structure Surfer. Specifically, we hypothesize that there will be structural patterns corresponding to nuances in splicing, translation, and many other important processes.

Authors Contributions

NDB, IMS, and BDG, conceived the study and designed the experiments. IMS performed the experiments. NDB, HK, LSW, and BDG analyzed the data and set up the Structure Surfer database and visualization system. DMC and NDB developed the web interface. NDB and BDG wrote the paper with assistance from all authors. The authors have read and approved the manuscript for publication.

CHAPTER 3. High throughput probing of Arabidopsis microRNA precursors

ABSTRACT

microRNAs (miRNAs) are small non-coding RNAs that function in post-transcriptional gene regulation. In plants, they are transcribed by RNA Polymerase II from their own genes as longer RNAs with regions of self-complementarity that fold into stem loop structures, which are called primary miRNA transcripts. The stem region of these stem loops contains the microRNA sequence and is recognized and subsequently cleaved by the RNase III DICER-LIKE1, thereby releasing the mature miRNA. Thus, the secondary structure of the hairpin is important for this cleavage step. Secondary structures can be determined in silico based on nucleotide complementarity and there are useful computational tools, however the computed structures must be validated. We used high

throughput enzymatic probing to validate annotated secondary structures of miRNA precursor transcripts.

3.1 Introduction

In plants and animals, microRNAs (miRNAs) play an important role in post-transcriptional gene regulation. These short non-coding RNAs bind to target transcripts using nucleotide complementarity and negatively regulate them, either by targeting them for cleavage or by inhibiting their translation.[114–117]

The biogenesis of miRNAs has multiple steps and some of them differ between animals and plants. Plant miRNAs, as well as many animal miRNAs, are initially transcribed by RNA Polymerase II (Pol II) into primary miRNAs (pri-miRNAs).[118,119] Like mRNAs and other PolI II transcribed RNAs, these transcripts are capped[119] and polyadenylated. Some have introns that are spliced out.[120,121] Each one has regions of imperfect self-complementarity that pair with each other and cause it to fold into a characteristic hairpin loop, a secondary structure that is critical for its processing. The mature miRNA is processed out of the hairpin's stem in a series of endonuclease cuts.

In animals, two distinct nucleases are required. Drosha makes two cuts at the base of the stem, removing the 5' and 3' ends of the transcript.[122] The resulting precursor miRNA (pre-miRNA) is then exported from the nucleus. In the cytosol, Dicer, another nuclease, makes one or more pairs of cuts on the opposite end, removing the loop and leaving the mature miRNA and its reverse complement.[123–126] A similar pattern of cuts occurs in plants, but they are performed by a single Dicer-like molecule (DCL1) and both steps take place in the nucleus.[117,120,127,128] Interestingly, in plants the two processing steps can happen in either order depending on the specific miRNA, meaning that some miRNAs are processed base-to-loop while others are processed loop-to-base.[69]

The sequence properties of miRNAs have made them an appealing topic for computational studies. In the genome, miRNAs appear as imperfect inverted repeats which makes it possible to discover them by searching for such patterns.[96,129,130] These putative miRNAs can then be validated experimentally. In addition to the miRNAs themselves, their regulatory targets have also been predicted *in silico* by taking advantage of the near perfect complementarity that plant miRNAs have with their target sites[96,97,100] as well as empirically through miRNA overexpression.[131]

Algorithms have been developed to predict the secondary structure of RNAs using their sequences by determining the minimum free energy (MFE) structure.[132–134] This method considers possible secondary structures based on nucleotide complementarity and chooses the most likely conformation by calculating the free energy of the unpaired nucleotides, favoring solutions with more nucleotide pairing.

While miRNAs themselves are short and unstructured, both pri-miRNAs and pre-miRNAs have secondary structures important for their processing. Many such structures are curated in miRBase, the primary source for miRNA annotation.[135–138] This database contains many high confidence entries, however new ones are deposited frequently and some have yet to be validated[29]. For some entries, the annotated structure has been experimentally determined however in many cases it is calculated by MFE.

Chemical and enzymatic probing have revealed differences between several predicted human pre-miRNAs structures and their experimentally determined structures.[139] This finding suggests that thermodynamic optimality is not sufficient to fully explain human pre-miRNA folding. In this study we use high-throughput enzymatic probing to characterize the secondary structure at miRBase annotated miRNAs. We investigate whether plant miRNAs, which are different from mammalian miRNAs in their biogenesis, arise from precursors that are also folded in ways that do not match their predicted structures. Additionally we propose a method for updating structure annotations using high throughput structure-probing data.

3.2 Results

We obtained RNA Sequencing (RNA-Seq) reads from two nuclease probing libraries. These libraries were prepared by Shawn Foley in the Brian Gregory's lab as controls for previously published experiment.[113] In each one, total RNA from the purified nuclei of *Arabidopsis* seedlings had been treated with a nuclease sensitive to RNA secondary structure. One sample was treated with RNase ONE, which preferentially digests single-stranded RNA. The other was treated with RNase V1, which preferentially digests double stranded RNA. We used the normalized read counts from the two libraries to calculate a structure score for each nucleotide in the genome (see methods). The resulting scores represent the relative degree of base pairing at each position in a given RNA molecule. Higher scores indicate stronger evidence of pairing while lower scores indicate a more unpaired region of the RNA. We used these scores to inspect miRNAs taken from miRBase.

The current miRBase annotation (v21) contains entries for 325 known and predicted *Arabidopsis* pri-miRNA transcripts. The predictions take into account the known or likely position of the mature miRNA within the stem-loops, but for many transcripts the exact positions of the 3' and 5' ends of the RNA are unclear. Of these annotated pri-miRNAs, 168 were detectable in our RNA-Seq libraries. We further narrowed our focus to transcripts that had calculable structure scores for at least half of their nucleotides resulting in a list of 88. We used the structure scores of these transcripts to fit a cubic spline for smoothing.

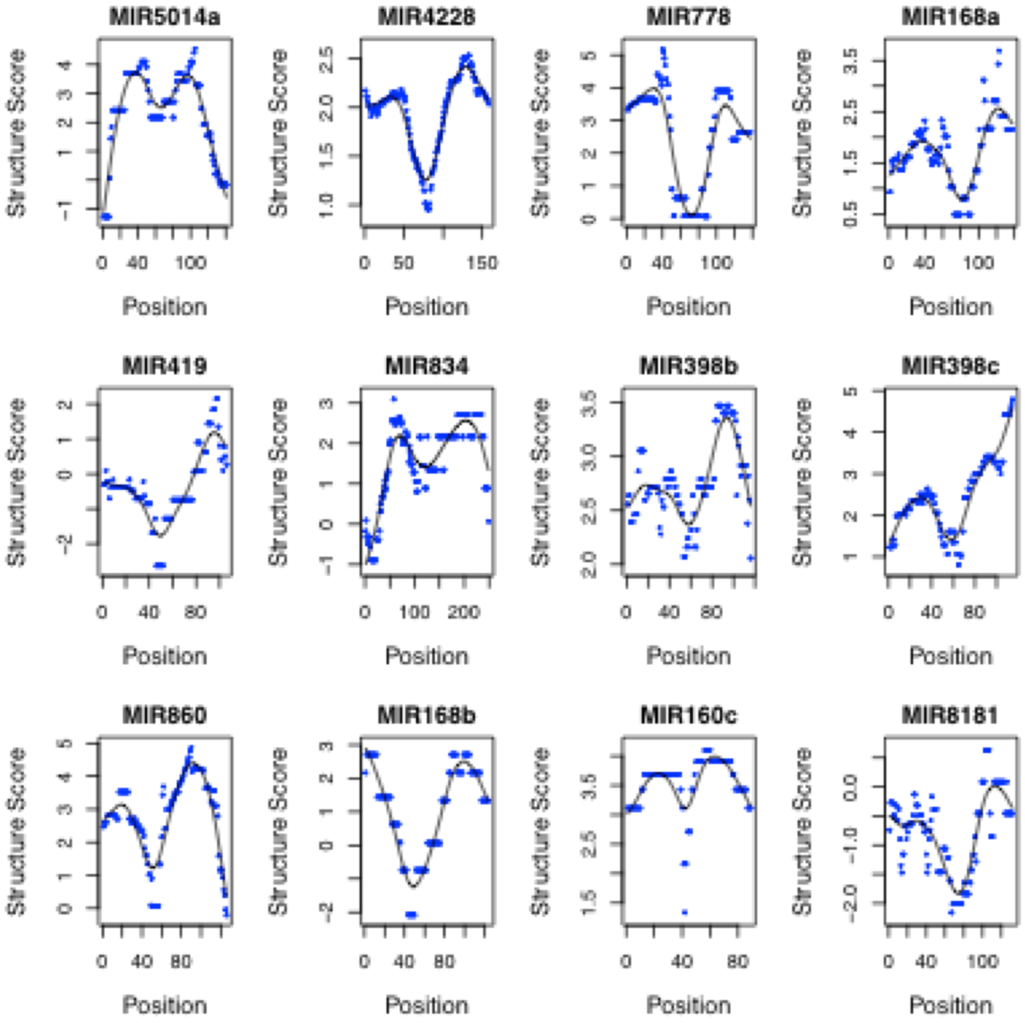


Figure 3.1: Example structural profiles of annotated miRNAs. Each position is given a structure score with higher scores indicating more evidence of a nucleotide being paired and lower scores indicating more evidence of it being unpaired. A cubic spline is fit to each profile revealing a characteristic pattern of an unstructured region near the center flanked by highly structure regions.

The structure profiles of many of these genes follow a characteristic pattern (Figure 3.1). They contain a central local minimum flanked by two local maxima indicating a region of low secondary structure in between two regions of high secondary structure. The simplest interpretation of this pattern is that it reflects the pri-miRNA's hairpin loop. The 5' structured region is paired with the 3' structured region forming the stem and the unstructured region between them is the loop. This pattern is evident even in profiles with positions that could not be scored due to low sequence depth (Figure 3.2). In such profiles we interpolated missing scores using the spline fit.

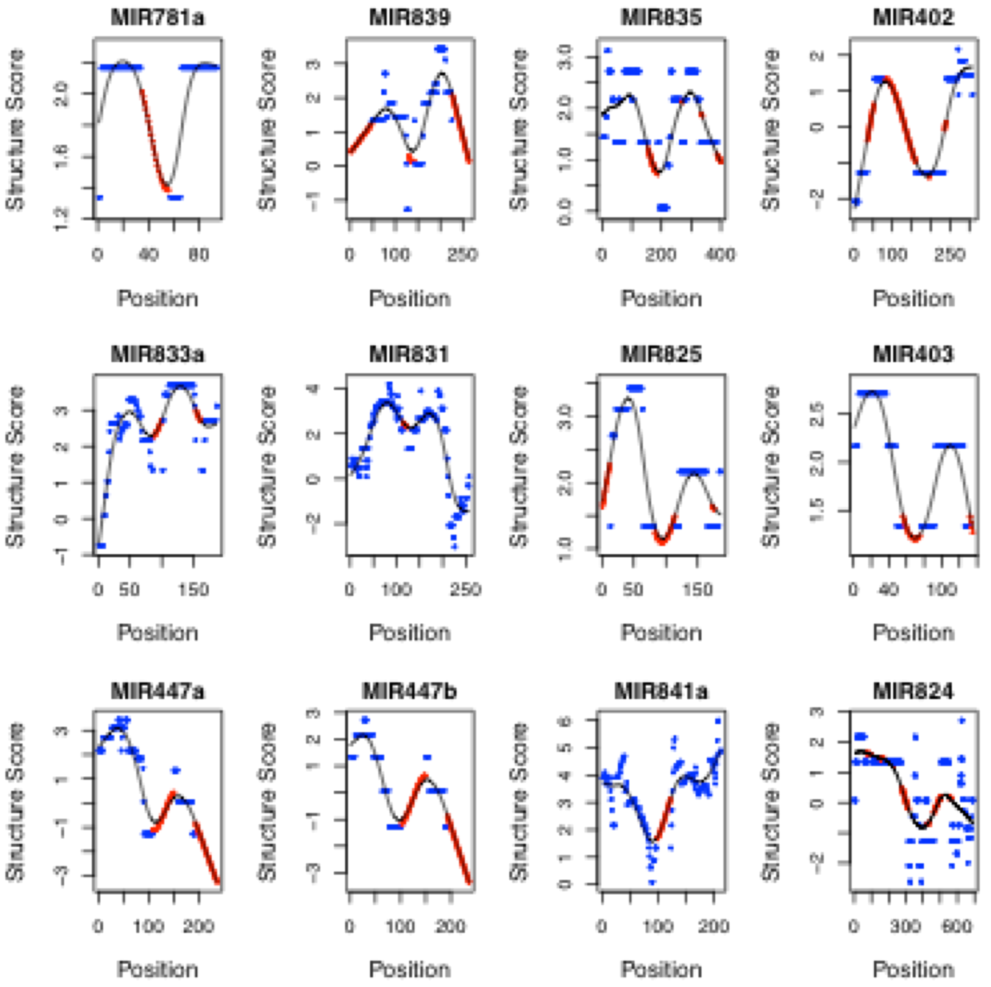
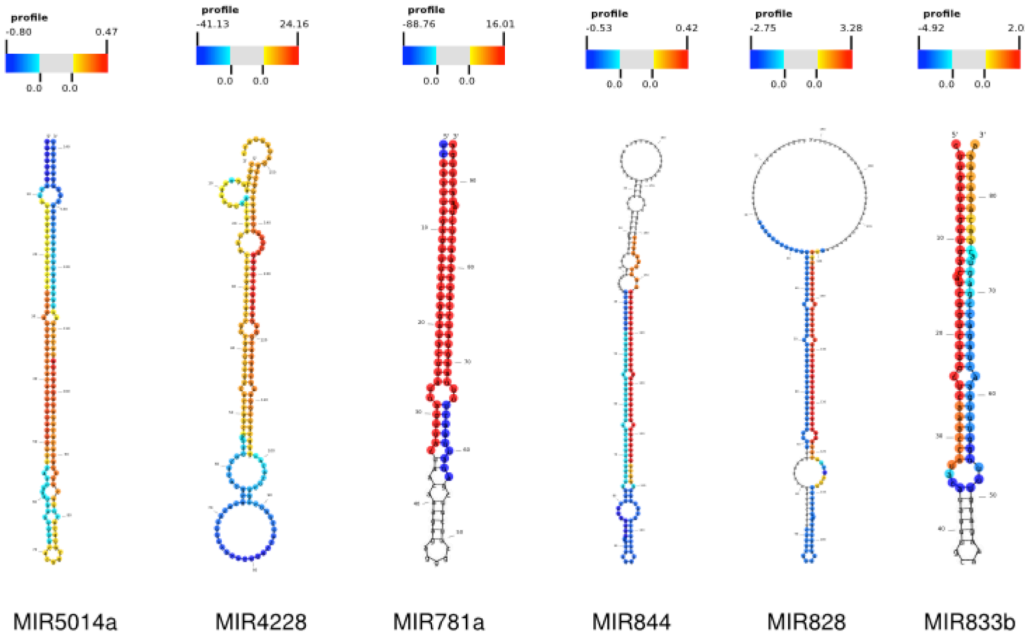


Figure 3.2: Examples of structure profiles with incomplete data. Despite having positions that could not be scored due to sparse data, profiles reveal the same pattern of secondary structure as completely scored profiles. Missing scores (red) are interpolated according to a cubic spline.

We compared our structure scores to the annotated structures from miRBase (Figure 3.3). In many cases, we found that positions with low structure scores, which we would predict to be unpaired, were annotated as paired, especially in the loop. This is consistent with a previous study that found differences between the secondary structures of *in vitro* transcribed pri-miRNA and RNA secondary structures empirically determined by through and enzymatic and chemical probing.[139] Our data recapitulate this finding in high throughput and *in vivo.*



Drawing 3.3: Examples of annotated secondary structures with empirical ds/ssRNA-seq structure scores superimposed. Warm colors indicate high secondary structure, cool colors indicate low secondary structure.

3.3 DISCUSSION

Powerful computational tools have been developed for detecting miRNAs using genomic data, however validation of such predictions remains an important step in miRNA discovery. Small RNA sequencing (smRNA-Seq) provides evidence of true miRNA transcripts by demonstrating the expression of mature miRNAs at candidate loci.[71] MiRBase, the primary source for miRNA annotations, curates smRNA-Seq data sets alongside its other annotations for each miRNA when appropriate datasets are available. Clear smRNA-Seq signal at a given locus is convincing evidence that a smRNA is being produced there.

It is important to note that miRNAs are only one of several classes of smRNA produced by plants. They are defined, in part, by their biogenesis. Each miRNA begins as a Pol II transcript and folds into pri-miRNA with a distinct hairpin loop structure, which is then cleaved to produce a pre-miRNA. The pre-miRNA is, in turn, cleaved to produce the mature miRNA.

Because pri-miRNA, pre-miRNAs and mature miRNAs all exist simultaneously in the cell, the secondary structure of a sequence is not simply defined. A given genomic locus is associated with the both the structured precursor molecules and the unstructured mature miRNA. Crucially, the dataset we analyzed was prepared from purified nuclei, limiting the amount of cytosolic mature miRNA and enriching for the structured precursors.

We observed a pattern of RNA secondary structure consistent with the hairpin loops of pri-miRNAs and pre-miRNAs at multiple annotated miRNA loci providing evidence that the smRNAs produced at those loci are true miRNAs arising from pri-miRNA transcripts. We argue that high throughput enzymatic probing effectively validates proposed miRNA locus annotations.

In addition, we noticed differences between the structural patterns we observed and the annotated structures from miRBase. In particular we noted many nucleotides that were annotated to be involved in base pairing interactions but appeared to be unpaired in our data. We proposed a method for updating existing annotations using empirical secondary structure information. High throughput enzymatic probing, when combined with nuclear purification, is a powerful tool for validating miRNAs and honing their structural annotations at a genomic scale. It will be intriguing to see this combination of techniques applied to other systems. Emerging model organisms with less genomic miRNA annotation will likely benefit from more rapid and confident miRNA detection. It might also be applied to disease states in which miRNA misregulation plays an important role.

3.4 METHODS

Raw sequencing reads were obtained from experiment, performed by my colleagues in the Gregory lab, on RNA secondary structure in the *Arabidopsis* nucleus.[113] The data included one library prepared from RNA that had been treated with RNase ONE, an enzyme that digests single-stranded RNA, and a second library in which the RNA had been treated with RNase V1, which digests double stranded RNA.

All reads were trimmed using cutadapt.[109] Trimmed and untrimmed reads were combined and mapped to the genome (TAIR10) using tophat.[110] Reads that were too short to be trimmed and reads that could not be aligned were discarded. We allowed up to two mismatches per read and a maximum edit distance of two. Reads that mapped to more than five locations in the genome were discarded. Read coverage at each nucleotide was calculated using bedtools[111] and normalized to library size.

We calculated a structure score based on the read coverage at each nucleotide in the two libraries. Nucleotides that are comparatively more represented in the RNase ONE treated library than in the RNase V1 library are considered to be more likely to be in structured regions while the opposite pattern indicates a higher likelihood of being in an unstructured region. The structure score is calculated as the generalized log ratio of normalized RNase ONE coverage to normalized RNase V1 coverage at each position *i*.

$$S_i = glog\left(ONE_i / V1_i\right)$$

A higher structure score indicates a higher degree of secondary structure while a lower score indicates more evidence of unstructured RNA. The scores were loaded to a mySQL database and can be browsed and downloaded using our web tool, Structure Surfer (http://tesla.pcbi.upenn.edu/structuresurfer/).

A list of *Arabidopsis* miRNAs was obtained from miRBase along with their sequences, genomic coordinates and structural annotations.[135–138] The coordinates were used to query our database. All miRNAs having computable scores for at least half of their positions were selected for further analysis. The structural annotations from miRBase were converted into RNAFold[132] constraint syntax. The miRNA sequences were folded with RNAFold both with and without constraints.

For each miRNA, a cubic spline with seven degrees of freedom was fitted to the structure scores. At nucleotide positions with missing scores, the score was interpolated from the spline. The second derivative of each spline was then calculated in order to determine the concavity at each nucleotide. Regions of the miRNA that were concave down were considered likely to be involved in base pairing while concave up positions were considered to be likely unpaired.

This information was used to update the constraints parsed from the miRBase annotations. Each annotated base pairing interaction was considered individually. In cases where both nucleotides involved in the pair occurred in concave up regions, the structure constraints were updated to reflect that those nucleotides were unpaired. Regions in which only one nucleotide in a pair occurred in a concave up region were considered ambiguous and were allowed to fold according to MFE.

CHAPTER 4. Genome-wide mapping of uncapped and cleaved transcripts (GMUCT)

This chapter was adapted from Willmann, M. R., Berkowitz, N. D. & Gregory, B. D. Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes--GMUCT 2.0. *Methods San Diego Calif* 67, 64–73 (2014).

Abstract

The advent of high-throughput sequencing has led to an explosion of studies into the diversity, expression, processing, and lifespan of RNAs. Recently, three different high-throughput sequencing-based methods have been developed to specifically study RNAs that are in the process of being degraded. All three methods—genome-wide mapping of uncapped and cleaved transcripts (GMUCT), parallel analysis of RNA ends (PARE), and degradome sequencing—take advantage of the fact that Illumina sequencing libraries use T4 RNA ligase 1 to ligate an adapter to the 5′ end of RNAs that have a free 5′-monophosphate. This condition for T4 RNA ligase 1 substrates means that mature mRNAs are not substrates of the enzyme because they have a 5′-cap moiety. As a result, these sequencing libraries are specifically made up of clones of decapped or degrading mRNAs resulting from 5′-to-3′ or nonsense-mediated decay (NMD) and the 3′ fragment of cleaved microRNA (miRNA) and small interfering RNA (siRNA) target RNAs. Here, we present a massively streamlined protocol for GMUCT that takes 2–3 days, can be initiated with as little as 5 μg of starting total RNA, and involves only one gel size-selection step. We show that the resulting datasets are similar to those produced using the previous GMUCT and PARE protocols. In total, our results suggest that this method will be the preferable approach for future studies of RNA degradation intermediates and small RNA-mediated cleavage in eukaryotic transcriptomes.

4.1. Introduction

I briefly discussed the relationship between RNA secondary structure, smRNAs and RNA cleavage. Specifically, many smRNAs have structured precursors, which are cleaved to produce smRNAs. Additionally, many regulatory smRNAs including members of the miRNA and siRNA classes regulate their targets by cleaving them. In this chapter I will discuss a technique developed by my colleagues that is capable of detecting cleavage events on a transcriptome-wide scale.

The key principle behind this is the observation that cleavage of a capped, polyadenylated RNA results in two products, a capped RNA with a 3' hydroxyl at the cleavage site and a polyadenylated RNA with a 5' monophosphate. The later molecule can be directly ligated to a 5' sequencing adapter. Intact transcripts are protected from such ligation by their 5' caps resulting a sequencing libraries which represent only the transcripts which have been cleaved or are actively undergoing degradation.

Recently, three different high-throughput sequencing based methods (genome-wide mapping of uncapped and cleaved transcripts (GMUCT), parallel analysis of RNA ends (PARE), and degradome sequencing) have been developed to study the degradation of mRNAs by exploiting this chemistry[72,140,141]. The pathways that degrade mRNAs are highly conserved in eukaryotes. Most turnover of normal, functional mRNAs occurs by 5′-to-3′ degradation, 3′-to-5′ decay, or RNA silencing[142–146]. Aberrant mRNAs are degraded by one of three pathways—nonsense-mediated mRNA decay (NMD), non-stop decay (NSD), and no-go decay (NGD)[147–150].

In total, the GMUCT, PARE, and degradome approaches are useful for studying 5′-to-3′ decay, small RNA-mediated target cleavage, and NMD, but not 3′-to-5′ decay, NSD, or NGD because a key enzyme (T4 RNA ligase 1) used to construct the library of molecules to be sequenced requires an available 5′-monophosphate on every RNA to be cloned. This requirement also means that mature, functional mRNAs are not substrates of these high-throughput methods because they have a 5′-cap. As a result, these sequencing libraries are specifically made up of clones of decapped or degrading mRNAs from 5′-to-3′ exoribonucleases, the 3′ fragment of cleaved miRNA and siRNA targets, and NMD decaying mRNAs.

Here, we present a streamlined method for GMUCT that we call GMUCT 2.0. The new method reduces the time necessary to make a GMUCT library from 5–6 to 2–3 days and decreases the amount of starting total RNA from 50 μg to 5 μg[141,151]. We use this protocol to make libraries of degrading and cleaved RNAs isolated from plant tissue and human cell lines. We also compare the data generated from this improved, more time-efficient method to data obtained using the original GMUCT[141], PARE[140], and degradome sequencing[72] approaches to show that the data produced is similar. Thus, we have significantly improved a methodology that will be widely useful for future studies of mRNA turnover in all eukaryotic organisms.

4.2. GMUCT 2.0 protocol

The new method for making GMUCT sequencing libraries is summarized in Figure 1, where it is also compared to the original GMUCT method[141,151].

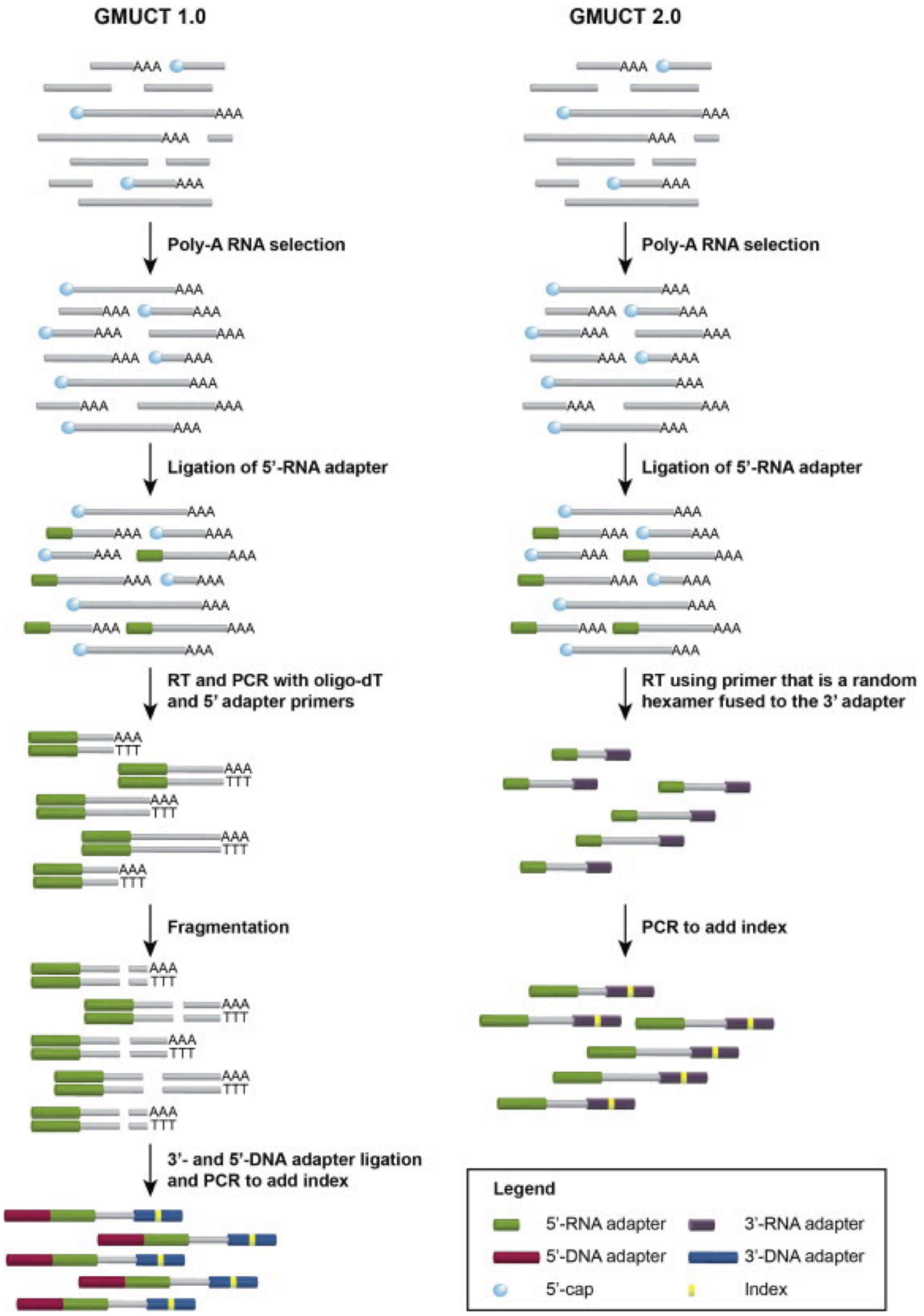

GMUCT 1.0
GMUCT 2.0
AAA
Poly-A RNA selection
Ligation of 5'-RNA adapter
RT and PCR with oligo-dT and 5' adapter primers
RT using primer that is a random hexamer fused to the 3' adapter
TTT
Fragmentation
PCR to add index
3'- and 5'-DNA adapter ligation and PCR to add index
Legend
5'-RNA adapter
3'-RNA adapter
5'-DNA adapter
3'-DNA adapter
5'-cap
Index

Figure 4.1: Comparison of the original GMUCT (GMUCT 1.0) and the new streamlined GMUCT 2.0. Both methods begin the same way, with the selection of poly-A RNA, ligation of the 5'-RNA

adapter, and reverse transcription, but the type of reverse transcription is different in each case. The original GMUCT protocol performs a traditional reverse transcription with an oligo-dT primer followed by PCR using oligo-dT and 5' adapter primers. The double-stranded DNA is then fragmented, both 3' and 5' DNA adapters are ligated, and the library is amplified, adding indices in the process. In GMUCT 2.0, the reverse transcription is performed using a primer that has the 3'-adapter sequence on the 5' end and a random hexamer on the 3' end, allowing for the adapter to be added during reverse transcription. The library is amplified and indices are added by PCR.

4.3 Protocol

4.3.1 Experimental techniques

For the complete experimental method please refer to Willmann et al.[152] Briefly, library preparation can be accomplished in three days, starting with RNA purification from a tissue of interest. We typically purify our RNA using the Qiagen miRNeasy Mini Kit. To verify that the starting total RNA is of high quality, the concentration, purity, and degradation should be examined using Agilent’s BioAnalyzer or a Nanodrop spectrophotometer and then running an agarose gel.

We purify polyadenylated RNA using the Dynabeads mRNA DIRECT kit. Many other companies sell poly-A selection kits, and these are also likely to work. We then ligate on the 5' sequencing adapter, capturing only RNA molecules with free 5' monophosphate groups. We then perform one or two more rounds of poly-A selection, this time to purify the ligation products away from any unligated 5' adapters. This step is important for reducing adapter-adapter clones.

On the second day of the protocol, we reverse transcribe the adapter ligated RNAs with primer containing a random hexamer sequence and the sequence of the 3' sequencing adapter. It's this priming step that chiefly differentiates GMUCT 2.0 from other techniques that sequence cleaved and degrading transcripts. It allows us to avoid fragmentation and 3' adapter ligation, steps which take time and reduce efficiency. Following this step we amplify the adapter ligated molecules using PCR.

On the final day, we size select the amplicons on polyacrylamide gel. The samples are then cut and extracted from the gel and precipitated. After resuspension, they are ready to be quantified and sequenced.

4.3.2 Computational analysis

Using this updated protocol, we constructed 2 replicate GMUCT libraries using RNA from mixed stage flower buds of the Arabidopsis thaliana accession Columbia (Col-0) and three human cell lines—the cervical cancer cell line HeLa, human embryonic kidney 293T (HEK293T) cell line, and the human chronic myelogenous leukemia cell line K562. After receiving the Illumina sequencing data, the reads were trimmed to remove the adapter sequences, the abundance of each unique read was determined, and unique reads were mapped to the Arabidopsis and human genomes, respectively. The vast majority of the inserts were 50 nt or longer (data not shown). Because (1) the GMUCT protocol clones only pieces of RNA that have a free 5′-monosphosphate, which are produced as a result of many of the degradation processes in the cell, and (2) the exact position of the 3′ end is not important because it is randomly determined based on where the random hexamer primer binds to the RNA, the important information from each cloned insert is the genomic location of the nucleotide at the 5′ end of each read. As a result, all mapped sequencing reads that start with the same 5′ nucleotide are summed together to calculate the number of cleavage/degradation events that occurred at that position of the RNA molecule in the tissue or cell line of interest.

Using this information, we plotted the summed distribution of GMUCT 2.0 reads across all protein-coding transcripts for the libraries from the *Arabidopsis* flower buds and the three human cell lines to obtain an idea of the frequency of decay products in different parts of mRNAs (Figures 3A and D). In all libraries of both species, there were significantly more reads at the 3′ end of mRNAs compared to the 5′ end. This is similar to the result we obtained by reanalyzing the *Arabidopsis* flower bud data obtained using the original GMUCT[141] and PARE[140] protocols (Figures 3B and C, respectively). This same distribution of mRNA degradation products has previously been reported for degradome sequencing libraries[72]. In total, these results indicate that the highly improved GMUCT 2.0 protocol captures the expected collection of turnover events in protein-coding mRNAs of both plants and animals.

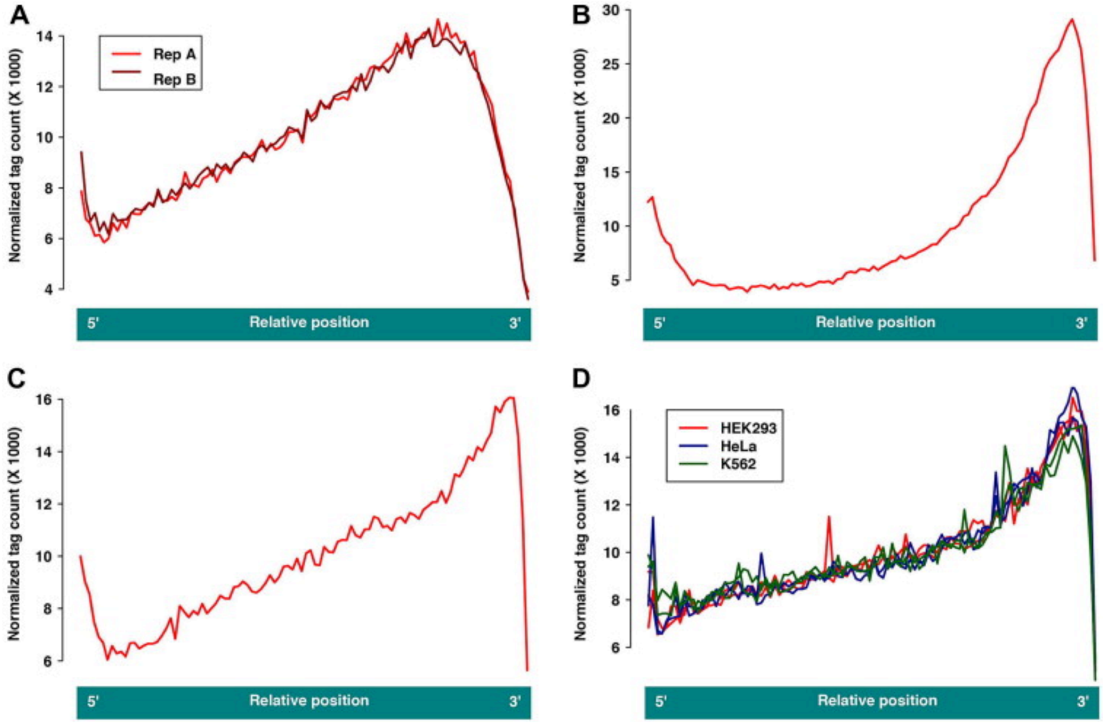


Figure 4.2: Number of GMUCT reads for eukaryotic transcripts is positively correlated with their abundance. For every detectable Arabidopsis transcript, its overall abundance as determined by mRNA-seq (x-axis) is plotted against its total cleavage profile as determined using GMUCT 2.0 (y-axis) for one Arabidopsis Col-0 flower bud replicate (A) and the other (B). r values denote the Pearson correlation between the two datasets.

We also found that the total abundance of GMUCT 2.0 reads varied greatly across all protein-coding transcripts. However, in general the abundance of GMUCT reads for a transcript was positively correlated with the overall abundance of that mRNA calculated using mRNA sequencing (mRNA-seq) data (Figure 4). This result reveals that increased levels of mRNA degradation events will be interrogated for RNAs that are more abundant and vice versa. These results are consistent with what was previously observed using the degradome sequencing approach[72], and suggest that the number of normal molecules decaying 5′-to-3′ and targeted by NMD for a given protein-coding transcript are proportional to its abundance.

One of the most promising uses of GMUCT, and other sequencing methods that specifically clone degrading mRNAs, is for validating and identifying miRNA and siRNA cleavage sites on target mRNAs. We looked at the average abundance of GMUCT reads with 5′ ends at miRNA target cleavage sites compared to the 100 bp up- or downstream for both *Arabidopsis* and humans. In both of the GMUCT 2.0 datasets for *Arabidopsis* flower buds there was a clear peak of read 5′ ends precisely at the predicted miRNA cleavage sites (Figure 2A). Similar results were obtained when reanalyzing the original *Arabidopsis* GMUCT[141] and PARE datasets[140] (Figure 2B and C, respectively). Conversely, none of the libraries made from the human cell lines showed a peak in abundance at predicted miRNA cleavage sites (Figure 2D).

This result is consistent with the current understanding of the differences between plant and mammalian miRNA activity. While miRNA-directed target cleavage and translational inhibition are known to occur in both plants and animals, cleavage is thought to be more common in plants and translational inhibition more common in animals. This difference in silencing mechanism has been attributed to the disparity in complementarity between target RNAs and miRNAs in plants versus mammals. One confounding factor is that the highly complementary miRNAs found in plants are more unambiguous to predict than their less complementary mammalian counterparts. Our metatranscript approach exhaustively summarizes predicted mammalian targeting relationships and finds no signal. It is premature to infer absence from a negative result, however these data are reassuring about the consistency of our method compared to other approaches. In total, these results also validate the usefulness of data obtained using the GMUCT 2.0 protocol for future studies of smRNA-directed cleavage events.

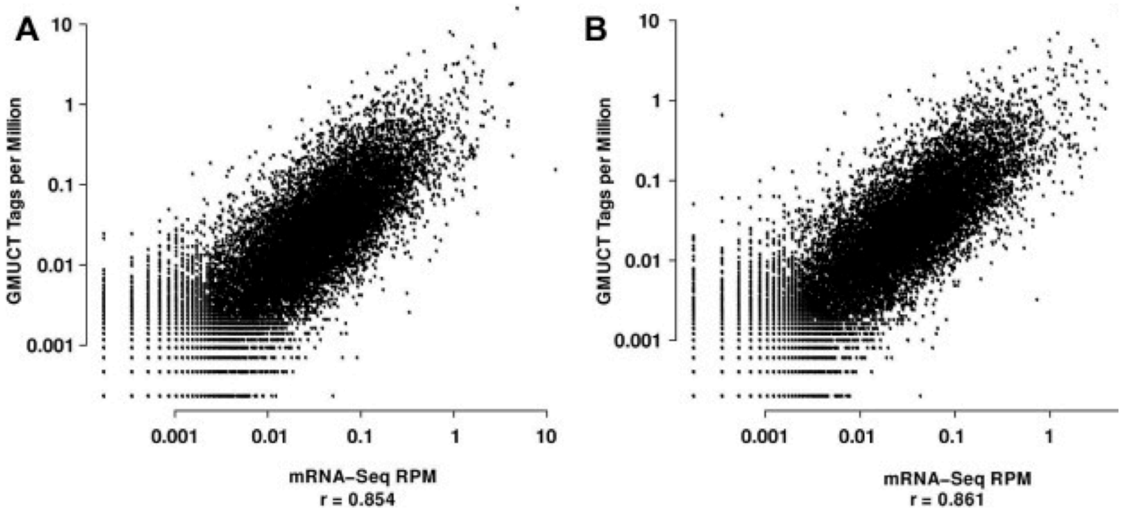


Figure 4.3: Number of GMUCT reads for eukaryotic transcripts is positively correlated with their abundance. For every detectable Arabidopsis transcript, its overall abundance as determined by mRNA-seq (x-axis) is plotted against its total cleavage profile as determined using GMUCT 2.0 (y-axis) for one Arabidopsis Col-0 flower bud replicate (A) and the other (B). r values denote the Pearson correlation between the two datasets.

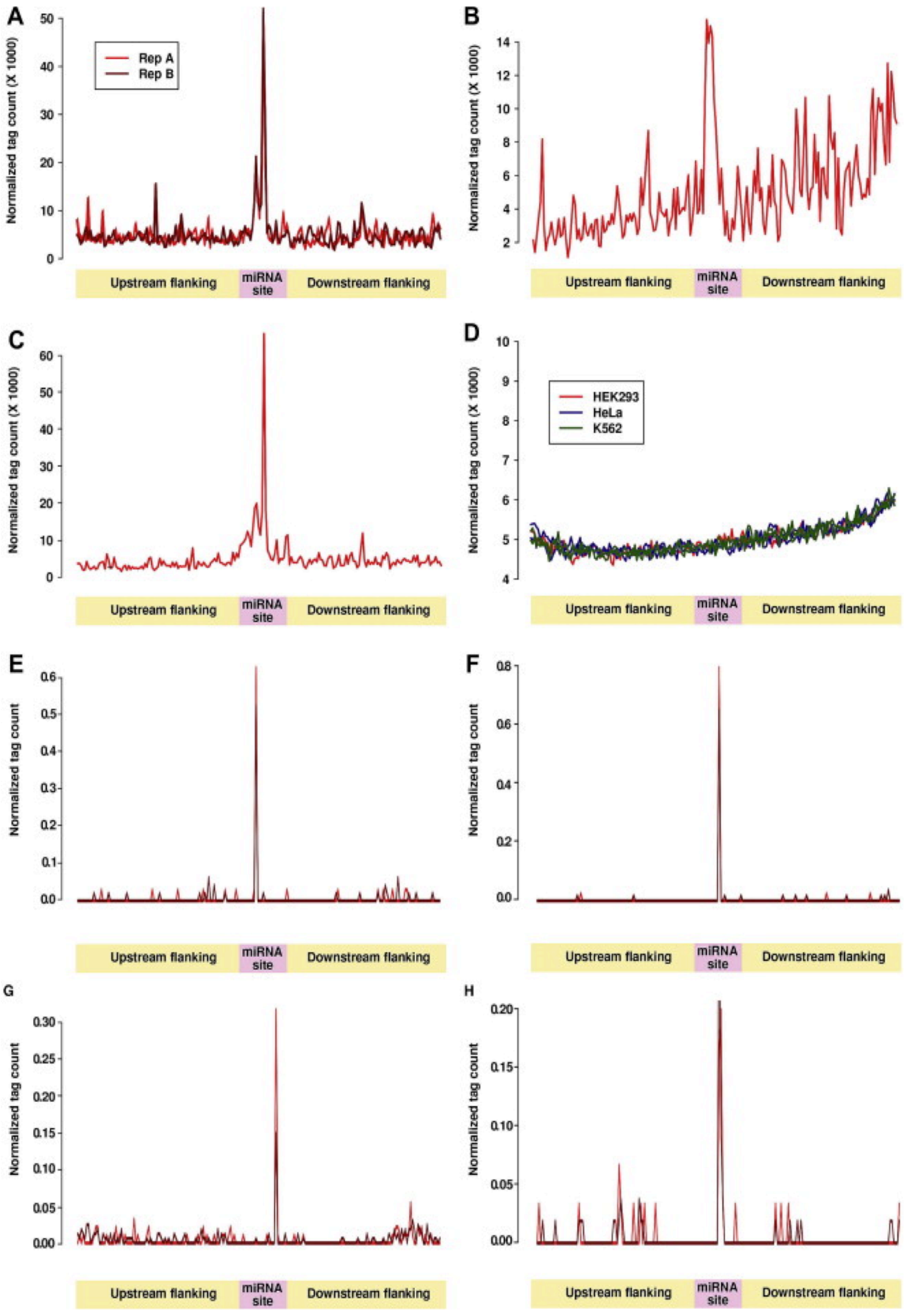

A
Normalized tag count (X 1000)
0
10
20
30
40
50
Rep A
Rep B
Upstream flanking
miRNA site
Downstream flanking
B
Normalized tag count (X 1000)
2
4
6
8
10
12
14
Upstream flanking
miRNA site
Downstream flanking
C
Normalized tag count (X 1000)
0
10
20
30
40
50
60
Upstream flanking
miRNA site
Downstream flanking
D
Normalized tag count (X 1000)
4
5
6
7
8
9
10
HEK293
HeLa
K562
Upstream flanking
miRNA site
Downstream flanking
E
Normalized tag count
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Upstream flanking
miRNA site
Downstream flanking
F
Normalized tag count
0.0
0.2
0.4
0.6
0.8
Upstream flanking
miRNA site
Downstream flanking
G
Normalized tag count
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Upstream flanking
miRNA site
Downstream flanking
H
Normalized tag count
0.00
0.05
0.10
0.15
0.20
Upstream flanking
miRNA site
Downstream flanking

Figure 4.4: GMUCT reads are highly abundant at the miRNA target cleavage site in Arabidopsis, but not in human cell lines. Each graph shows the average number of GMUCT read 5' ends at the collection of miRNA target cleavage sites and the 100 nt up- and downstream. This analysis was done by summing the number of reads whose 5' ends map at each nucleotide in the window for each miRNA target transcript, renormalizing all of the reads for each transcript in this window such that the total across a window was 1, adding the normalized total from all of the transcripts for each nucleotide, plotting the result, and scaled to one million counts. The miRNA target site cleavage profiles are plotted for the two Arabidopsis Col-0 GMUCT 2.0 libraries from this paper (A), the Arabidopsis Col-0 GMUCT 1.0 library from (1) (B), the Arabidopsis Col-0 PARE library from (3) (C), and the six human cell line GMUCT 2.0 libraries from this paper (D). There are two GMUCT 2.0 replicates for HEK293 (red lines), HeLa (blue lines), and K562 (green lines) cells. The list of miRNA targets used for the Arabidopsis analysis was a compilation of targets validated by the PARE method (3). The list is found online at http://www.mpss.udel.edu/at_pare/. The list of miRNA targets used for the human analysis is composed of predicted human targets found at http://www.microrna.org[23]. The specific miRNA target site cleavage profiles for the two Arabidopsis Col-0 GMUCT 2.0 libraries from this paper are provided for the specific miRNA targets AT4G37740.1 (E), AT1G10120.1 (F), AT2G16600.2 (G), and AT1G27340.1 (H).

4.4 Conclusions

The development of high-throughput sequencing and the increased interest in understanding the different levels of gene expression regulation has resulted in many new genomic approaches for analyzing very specific populations of mRNAs. GMUCT, PARE, and degradome sequencing were all developed to study mRNA degradation resulting in free 5′-monophosphates on decaying mRNAs, including 5′-to-3′ exoribonucleolytic decay, RNA silencing directed target RNA cleavage, and non-sense mediated decay of aberrant mRNAs.

Here, we present an improved GMUCT protocol that produces results similar to the longer original method and to the two other high-throughput sequencing-based methods (PARE and degradome sequencing) for studying mRNA degradation but requires less time and less starting total RNA. Additionally, the distribution of sequencing reads across transcripts, the correlation between GMUCT reads and the steady state abundance of mRNAs measured by mRNA sequencing (mRNA-seq) libraries, and miRNA target cleavage in plants are similar for the different methods. Thus, our improved methodology is more time efficient, and still captures the desired information.

As expected and in contrast to the analysis of plant GMUCT data, analyzing the average abundance of GMUCT products at human miRNA target sites did not provide obvious evidence of widespread cleavage. These results are consistent with most previous data concerning animal miRNA silencing mechanisms. It would be of interest, however, to further parse the GMUCT data from the human cell lines to test the hypothesis that it is the level of complementarity between a miRNA and its target that determines whether this smRNA will induce the cleavage of its target. Overall, the shorter time investment, increased simplicity, and lower required input of starting total RNA needed for the improved GMUCT method will likely lead to the increased use of this approach for future studies of mRNA degradation in all eukaryotic organisms.

Accession numbers

All GMUCT 2.0 and mRNA-seq sequencing data generated for this study from *Arabidopsis* Col-0 mixed stage flower buds (2 GMUCT replicate libraries and 2 mRNA-seq replicate libraries) and the three (HeLa, HEK293T, and K562) human cell lines (2 GMUCT replicates for each (6 libraries total)) were deposited in GEO under the accession GSE47121.

5 RNA Dependent RNA polymerases (RDR) and their substrates

5.1 INTRODUCTION

In general, eukaryotic RNA molecules are initially transcribed as single stranded transcripts by DNA-dependent RNA polymerases. However, it is becoming increasingly clear that double-stranded RNA (dsRNA) molecules have important roles in gene regulation and other cellular processes. Thus, it is not surprising that cells have multiple ways of creating double stranded structures. In a previous chapter, I discussed two examples of such structural elements, microRNA stem loops and iron response elements. Both form double stranded regions by folding into a hairpin shape using intra-molecular hydrogen bonds. Such intra-molecular bonds are also found in other important functional RNAs such as rRNAs, tRNAs, and ribozymes. In addition to structures formed by intra-molecular bonds, many eukaryotic organisms produce intermolecular dsRNA molecules using a class of enzymes known as RNA-dependent RNA polymerases (RDRs).

This class of enzymes uses an RNA molecule as a template to synthesize a reverse complementary RNA. A useful comparison can be made to reverse transcriptases, a class of enzymes widely used in various biological experiments, that synthesize complementary DNA (cDNA) using an RNA molecule as a template. RDRs behave analogously except they produce RNA instead of cDNA.

The functional roles of RDRs are diverse, but in general they are important players in the biosynthesis of small RNAs (smRNAs) that function in various RNA silencing pathways. Perhaps the most straightforward role for RDRs is in the production of small interfering RNAs (siRNAs), which once produced can target unwanted transcripts, such as transposons and viral RNAs in both *cis* and *trans*. More specifically, a RDR can use the unwanted transcript as a template and produce its reverse complement, making it double stranded. The resulting double stranded molecule can then be cleaved by a RNase III DICER enzyme into siRNAs. This cleavage has two effects: the unwanted transcript is destroyed and siRNAs complementary to it are produced. These complementary siRNAs can then act as specificity factors to guide repressive factors to other copies of the unwanted transcript by base pairing with them.

The mechanisms of suppression vary between RDRs as do the templates they prefer and the siRNAs they produce. For instance, some substrate RNAs require priming with a smRNA acting as a primer, while others can be directly used as templates without priming [154]. This diversity likely stems from the long evolutionary history of RDRs. RDRs are ancient genes [155], and it is thought that there were three RDR genes in the last common ancestor of animals, plants and fungi. Over time, each of these has given rise to a family of orthologs, the α, β, and γ RDRs.

Interestingly, most animals have lost their RDRs during their evolution. However, RDRs can be found in the nematode *C. elegans.* They are widely found in plants and fungi where they have developed a broad range of biological functions. In addition to suppressing unwanted exogenous RNAs, RDRs are also involved in regulating endogenous genes during development and stress [156–158]. Furthermore, it is likely that there are roles for RDRs that have not yet been discovered.

The model plant *Arabidopsis thaliana* is an appealing system for studying RDRs. Its genome contains three α RDRs and three γ RDRs, and the substrates and functions of these proteins have been studied to different degrees. For instance, the γ RDRs have not been assigned any function. However, they are evolutionarily conserved and they are transcribed in multiple tissues, suggesting that they have some function. The α RDRs (RDR1, 2 and 6) have each been functionally described and remain a topic of current research.

RDR1 and RDR6 are known to play an important role in defending plants against infection by several viruses[159–166]. RDR2 likely has a more limited antiviral role, but mutations in any of the three α RDRs result in increased susceptibility to viral infection[161]. RDRs defend against viruses through post-transcriptional gene silencing (PTGS).[162]

In antiviral PTGS, an RDR uses a viral transcript as a template to synthesize the complementary strand resulting in a perfectly complementary double-stranded RNA molecule. This double-stranded molecule is subsequently cleaved by an RNase III DICER/DICER-LIKE (DCL) family protein,[167] resulting in both turnover of the target transcript as well as production of virus-derived siRNAs (vsiRNAs). These vsiRNAs are then loaded into ARGONAUTE (AGO) family proteins. AGOs use these vsiRNAs as specificity factors to detect additional viral transcripts and cleave them, further increasing turnover of these pathogens.[159,168–170]

In addition to their function in turning over viral transcripts, RDRs also play a role in down-regulating endogenous transcripts in plants. Some of the most well studied examples of this type of regulation belong to the RDRs involved in making the class of siRNAs known as trans-acting siRNA (TAS), which are encoded by endogenous loci (*TAS* genes). siRNAs made from dsRNAs produced by RDRs using these gene transcripts as templates regulate other target RNAs in *trans*, analogous to miRNA-directed silencing. Also like miRNAs, RNAs from these loci are transcribed by Pol II, processed into smRNAs by a DCL protein, and these siRNAs are loaded into AGO1, which then uses them as specificity factors to target unrelated transcripts for silencing.[171] In contrast to vsiRNAs, they regulate transcripts other than their own precursors, and are thus referred to as trans acting siRNAs (ta-siRNAs). One important difference between miRNA and ta-siRNAs is in their requirements for biogenesis.

Specifically, *TAS* transcript processing begins when a primary RNA molecule produced by Pol II is targeted by a miRNA for AGO1-mediated cleavage.[172–174] The details of the targeting vary between family members and, as a group, they have important differences from other miRNA targeting events.[173–176]

After being cleaved, a *TAS* transcript is then stabilized by SUPPRESSOR OF GENE SILIENCING3 (SGS3) and recognized by RDR6.[174,177,178] RDR6 then synthesizes the reverse complement of the transcript, making it double stranded. This double stranded molecule is then cleaved into ta-siRNAs, primarily by DCL4, although DCL2 and DCL3 can also inefficiently produce smRNAs from these templates.[128] DCL4's cuts produce siRNAs in 21 nucleotide (nt) intervals starting from the initial miRNA cleavage site resulting in ta-siRNAs that are in 21 nt phase. Thus, the same set of ta-siRNAs is produced from each *TAS* dsRNA molecule.[172,175]

Some of the ta-siRNAs are loaded onto AGO1, which uses them to target unrelated RNAs, including the mRNAs encoding some members of the AUXIN RESPONSE FACTOR (ARF) family of transcription factors (e.g. ARF3).[179]. These ARF transcription factors are broadly important, including during the development of leaves[177,178,180,181] and flowers[178,182,183] and in phase change.[178]

In addition to targeted regulation of single genes such as the ARFs, some RDRs have a role in genome-wide regulation. For instance, RDR2 is particularly important in forming and maintaining regions of heterochromatin. Its pathway has some parallels to the ta-siRNA pathway that is directed by RDR6 but there are significant differences.

More specifically, RDR2 along with an SGS3-like cofactor RDM12,[184–186] uses a primary transcript as a template to produce a double-stranded RNA molecule which is then cleaved by a DCL, in this case DCL3,[187,188] into smRNAs. These heterochromatin-associated siRNAs (hsiRNAs) are mostly 24 nt in length[189] and they are the most abundant smRNAs in the cell.

They are, as with the 21 nt ta-siRNAS, loaded onto AGO proteins and used as specificity factors. However the AGOs loaded with hsiRNAs are AGO4, 6, and 9.[190,191] RISC complexes containing these AGOs are involved in formation and maintenance of DNA methylation and subsequent heterochromatin formation. Because they use hsiRNAs to identify their targets, this process is called RNA dependent DNA methylation (RdDM).[73,188,189,192,193]

Two other important components of the RdDM pathway are the plant specific RNA polymerases Pol IV[194] and Pol V. Pol IV produces the transcripts that RDR2 uses as templates.[194–197] These transcripts include transposons, telomeres, pericentromeric regions, and other repetitive elements.[189,194–197] The hsiRNAs produced from these transcripts, once loaded onto an AGO, guide methylation at similar repetitive loci including their own loci of origin.[190,191,198] Pol V also produces transcripts from these regions. However, these transcripts act as scaffolds that physically interact with hsiRNA-loaded AGOs, recruiting this complex to the site to be methylated.[199–202]

The detailed mechanisms described for RDR6's role in PTGS and RDR2's role in RdDM are two examples of critical roles played by RDRs in the cell. However RDRs likely have functions that are not yet described. For example, the γ RDRs (RDR3, 4, and 5) have no assigned roles despite the fact that they are evolutionarily conserved and they are detectably expressed in multiple tissues.

In order to better understand the roles of the RDRs in plant transcriptomes, I describe a genome-wide search for endogenous RDR substrates. To do this, we used RNA-seq to identify genomic loci that require RDRs for full transcription of both double-stranded RNAs and smRNAs. The list of well-characterized substrates for the a RDRs provides an opportunity to internally validate my analytical approaches and adds confidence when novel substrates are identified. I further characterize these loci by describing the smRNAs that they produce and investigating the likely targets of these smRNAs.

5.2 Methods

Library Preparation

Total RNA was purified from *Arabidopsis* tissue, either leaf or unopened flower bud. Sequencing libraries were prepared for smRNA-seq and dsRNA-seq as described in Zheng et al. 2010 and sequenced on an Illumina HiSeq using the 50 nt single end sequencing protocol as per manufacturer's instructions (Illumina Inc., San Diego, CA).

Sequencing Read Processing and Alignment

Some of the sequencing libraries were sequenced more than once to achieve greater sequencing depth. Multiple runs for each library were pooled by concatenating fastq files. For reads with small inserts, often the 3' sequencing adapter is included in the sequence. These adapter sequences were trimmed using cutadapt.[109] Trimmed and untrimmed reads were separately mapped to the *Arabidopsis* genome (TAIR10) using TopHat,[110,203] an aligner which takes into account splice junctions. Reads that could not be trimmed or mapped were discarded. We allowed up to two mismatches per read and a maximum edit distance of two. Because of the repetitive nature of some RDR substrates, we allowed up to 100 alignments per read.

Read Count Computation

The number of reads aligning to each transcript was counted using HTSeq.[204] This was done in a strand specific manner, separately counting reads that aligned to the sense and antisense sequences. The genome was partitioned into unbiased 500 nt bins. The number of reads falling into each bin was counted using bedtools.[111]

Differential Expression and Hit Calling

Differential expression between Col-0 and each of the mutant genotypes was computed using the R package edgeR[205] for all library types. Features were considered to be putative RDR substrates if they had a decrease in smRNA expression of at least 33% with a false discovery rate below 0.1 and at least a nominal decrease in dsRNA expression.

Size Classification of smRNA

Trimmed smRNA-seq reads were used as a proxy for intact smRNAs. For each genotype, the set of putative RDR substrate genomic bins was intersected with the smRNAs in Col-0. These were classified by length and 5' nucleotide.

Target Prediction and Analysis

The set of all smRNAs aligning to the bins described above, excluding those assigned only to RDR2, was converted to a list of unique sequences. These were used to query psRNATarget[100] for target sites among the set of *Arabidopsis* transcripts, excluding miRNAs. GMUCT data were obtained from Willmann et al., the study discussed in a previous chapter. GMUCT scores and PhastCons scores[206] for each base were intersected with 100 nt windows surrounding predicted targeting interactions.

5.3 Results

Internal Validation with TAS loci

In order to test the sensitivity of my differential expression quantification, I inspected a very well characterized family of RDR substrates, the *TAS* loci. These transcripts are targeted by miRNAs triggering recruitment of SGS3 and RDR6. RDR6 synthesizes the reverse complement of each *TAS* RNA making it a suitable substrate for DCL4, which cleaves it into 21 nt ta-siRNAs.[173–176,207] Therefore, I expected to see differential expression in both dsRNA and smRNA between Col-0 and *rdr6* and *sgs3* mutant plants (Figure 5.1).

In general, the TAS loci show strong differential expression of smRNA in both the *rdr6* and *sgs3* mutant backgrounds consistent with the roles of those two proteins in ta-siRNA production. In addition, I see significant differential expression on both the sense and the antisense strand for seven out of the ten *TAS* loci inspected in bud and six out of the ten in leaf. Antisense expression, and loss of antisense expression in the mutants, is an unambiguous signal of RDR activity as the antisense strand is not a Pol II transcript but a direct product of RDR activity.

TAS3a and *TAS3b* were not highly expressed enough in Col-0 to make differential expression calls. *TAS4* is differentially expressed in leaf, however, despite its negative log fold change in bud, it does not reach significance after correction for multiple testing.

The differential expression pattern in the dsRNA is consistent with the smRNA, but weaker due to lower expression in Col-0. This lower expression is likely due to the intermediate nature of these molecules. Depending on the kinetics of RDR6 and DCL3, they may exist in the cell very briefly, and therefore be rare in our dsRNA sequencing (dsRNA-seq) libraries.

This analysis shows that RDR6 substrates are detectable by smRNA-seq and suggests that substrates of other RDRs will be detectable as well. The dsRNA-seq experiments provide a second line of evidence, increasing my confidence that hits identified using these methods are true RDR substrates. Taken together with the observation of differential expression of both sense and antisense RNAs in both dsRNA-seq and smRNA-seq, it demonstrates that these loci are most likely differentially expressed because they are direct substrates of RDRs as opposed to being differentially expressed due to upstream regulation.

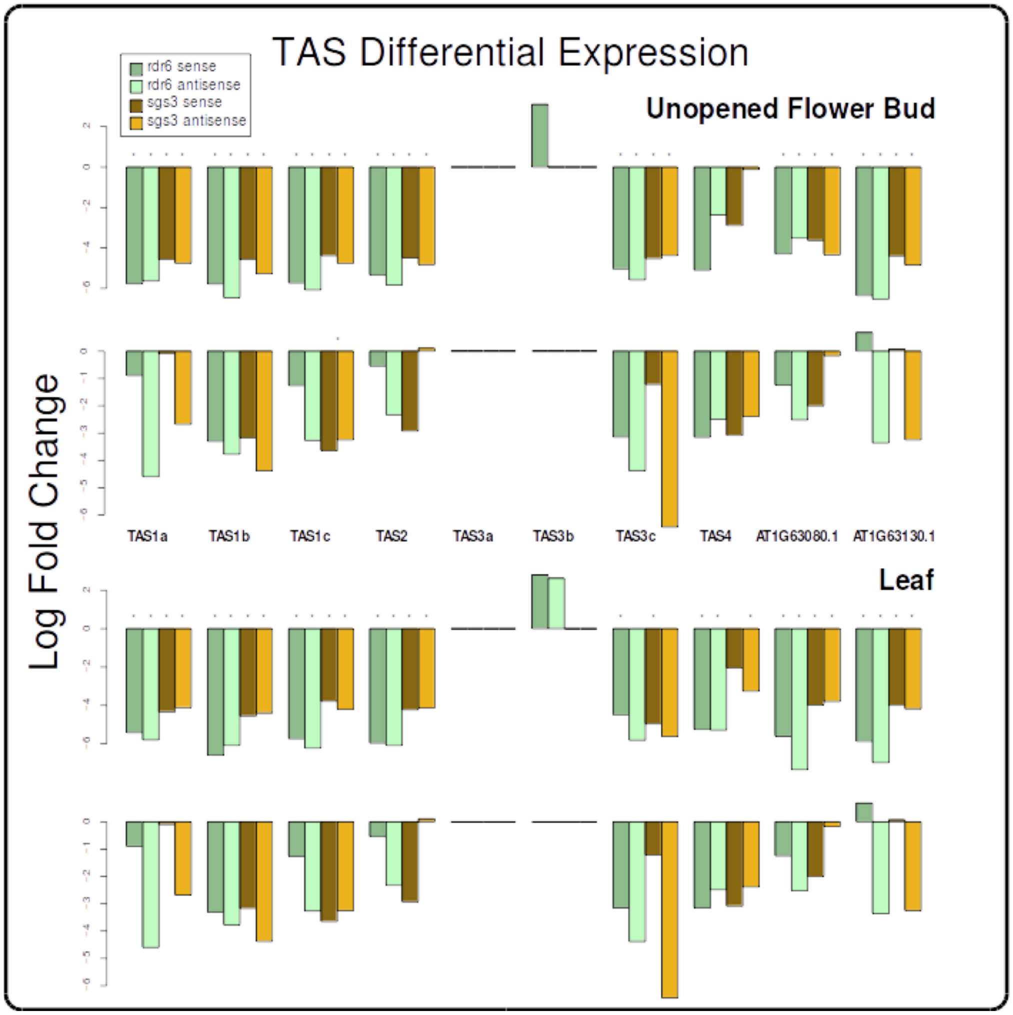


Figure 5.1: Differential expression of the TAS loci, a class of known RDR6 and SGS3 substrates. In both mutants, most of the TAS loci show decreases in expression of both double stranded RNA (dsRNA) and smRNA consistent with loss of RDR activity as well as loss of smRNAs requiring a double stranded precursor. This differential expression is detectable on both the sense and anti-sense strands, again consistent with absent RDR activity.

RDR2 Genomic Bins

I next viewed the differential expression of unbiased 500 nt genomic bins between Col-0 and an *rdr2* mutant plants in order to test my pipeline's sensitivity to large scale genomic trends. RDR2 is involved in RdDM and is known to silence many targets including telomeres, the pericentromeric regions, transposons, and other repetitive elements. Inspecting chromosome 1 (Figure 5.2) reveals differentially expressed bins in both smRNA-seq and dsRNA-seq tiling the chromosome.

Across the genome, 14,697 bins were identified as putative RDR2 substrates in unopened flower buds, with 3313 found on chromosome 1. Hits tile the entire chromosome and are most abundant in the pericentromeric regions. I don't observe enrichment for hits near the telomeres, possibly because the low complexity of these regions interferes with sequencing read alignment. Nearly half of the total putative substrates, 6204, overlapped with an annotated transposon, transposon gene, or transposon fragment. This ratio is consistent across chromosomes, for example 1217 such transposon related bins were detected on chromosome 1. These data suggest that, in addition to being able to detect single gene substrates like the templates used by RDR6, the combination of dsRNA-seq and smRNA-seq is also powerful when detecting genome-wide RDR substrate trends.

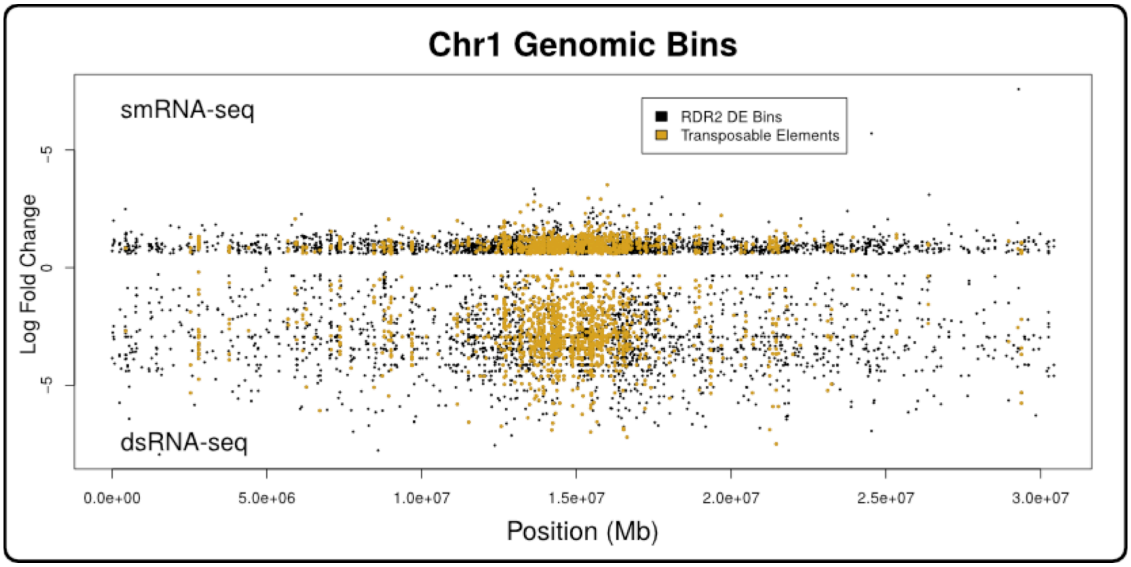
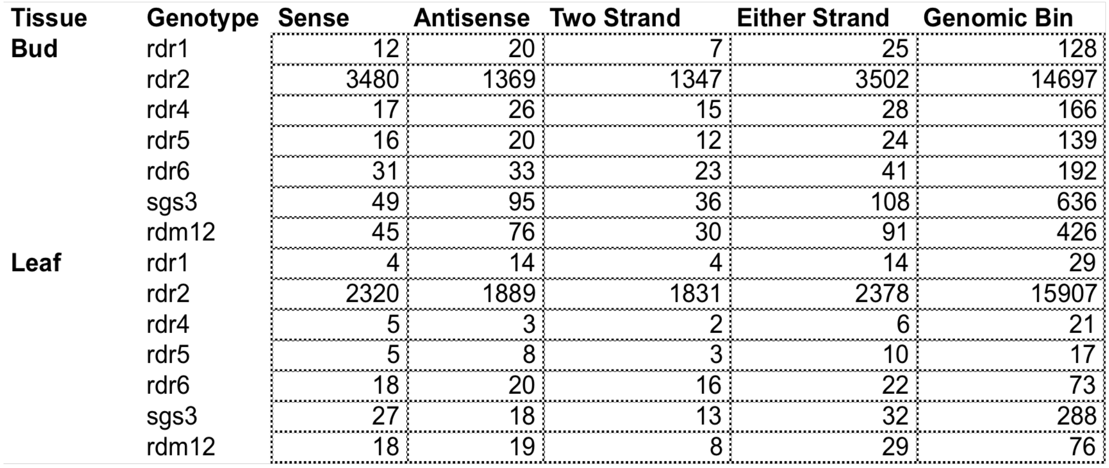


Figure 5.2: Differential expression of smRNA (up) and dsRNA (down) in rdr2 *within 500 nt genomic bins. Marked bins intersect with annotated transposable elements*

Total RDR Substrate Identification

I scanned the *Arabidopsis* genome for regions where wild type Col-0 plants robustly expressed both dsRNA and smRNA but lost expression of both in *rdr1, 2, 4, 5,* or *6*. I also looked for lost expression in mutants of the RDR2 cofactor *RDM12* and the RDR6 cofactor *SGS3*. I interpret such sites as putative RDR substrates. Their differential expression patterns are consistent with transcripts that depend on RDRs to synthesize their reverse compliments making them suitable substrates for cleavage into smRNAs by a member of the DCL family.

I considered two types of loci, TAIR10 annotated transcripts and unbiased 500 nt windows tiling the genome. This approach is designed to detect known genes that are substrates of RDRs as well as regions of the genome that have no annotated genes but are transcribed into RDR templates. I found putative targets for each of the proteins inspected (Table 1).



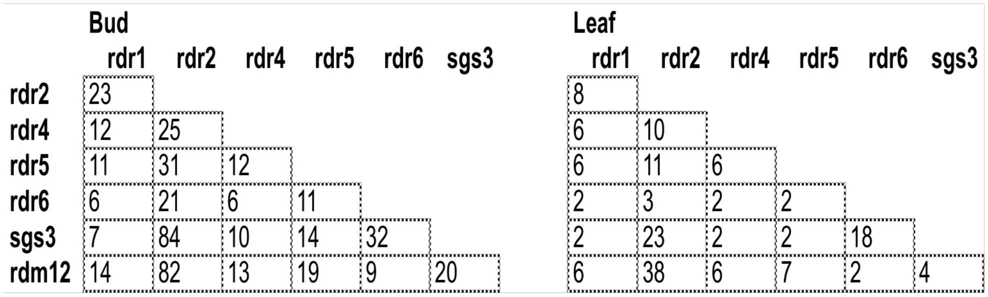
Tissue	Genotype	Sense	Antisense	Two Strand	Either Strand	Genomic Bin
Bud	rdr1	12	20	7	25	128
	rdr2	3480	1369	1347	3502	14697
	rdr4	17	26	15	28	166
	rdr5	16	20	12	24	139
	rdr6	31	33	23	41	192
	sgs3	49	95	36	108	636
	rdm12	45	76	30	91	426
Leaf	rdr1	4	14	4	14	29
	rdr2	2320	1889	1831	2378	15907
	rdr4	5	3	2	6	21
	rdr5	5	8	3	10	17
	rdr6	18	20	16	22	73
	sgs3	27	18	13	32	288
	rdm12	18	19	8	29	76

Table 5.1: Summary of putative RDR substrates identified in each of the transcripts across two tissue, unopened flower bud and leaf. For annotated transcripts, strand information is given. Significance calls can refer to sense or antisense or their intersection (two strand) or union (either strand). For 500nt genomic bins strand information was ignored.

Many of the transcripts identified showed differential expression of the antisense strand. This provides evidence that these transcripts are directly used as templates by RDRs. In some cases we identified both sense and antisense transcripts as differentially expressed hits.

By far the greatest number of putative RDR substrates was found for RDR2. This is consistent with the fact that hsiRNAs, which depend on RDR2 for synthesis, are the most abundant smRNAs in the cell. It also makes sense that RDR2, which is important for proper heterochromatin formation, would result in gross, genome-wide differences in transcription when mutated.

I also note that many putative RDR substrates appear differentially expressed in more than one mutant background (Table 2). These counts represent the number of transcript that that had sense or antisense hits, or both.



	Bud						Leaf					
	rdr1	rdr2	rdr4	rdr5	rdr6	sgs3	rdr1	rdr2	rdr4	rdr5	rdr6	sgs3
rdr2	23						8					
rdr4	12	25					6	10				
rdr5	11	31	12				6	11	6			
rdr6	6	21	6	11			2	3	2	2		
sgs3	7	84	10	14	32		2	23	2	2	18	
rdm12	14	82	13	19	9	20	6	38	6	7	2	4

Table 2: Shared hits between genotypes using the union of sense and antisense differential expression calls.

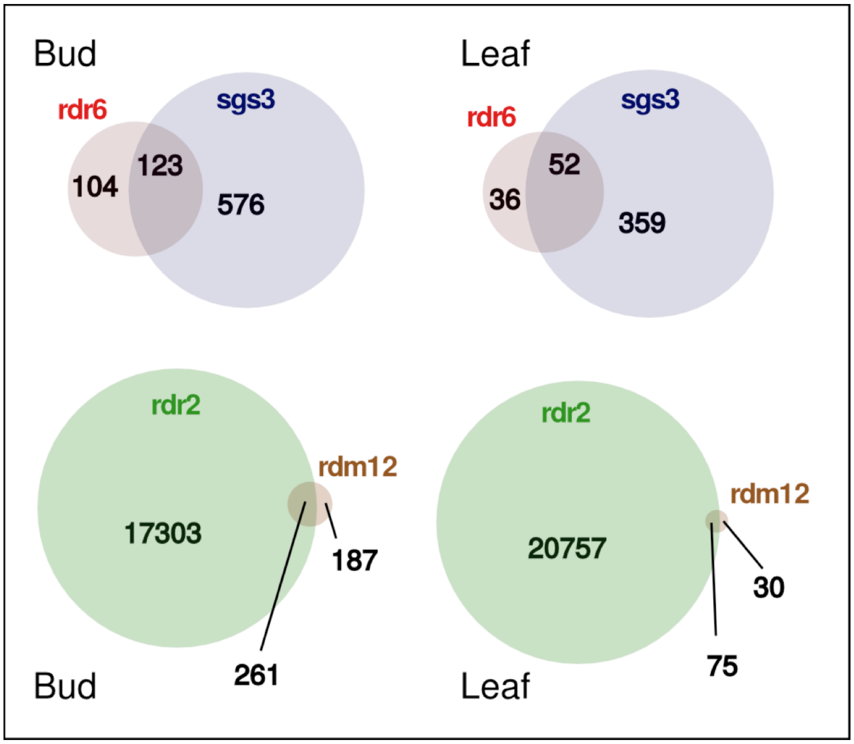


Figure 5.3: Shared genomic bin hits between RDRs and their known cofactors

Notably, RDR6 and its cofactor SGS3 share the bulk of their transcriptomic hits in both bud and leaf tissues. Similarly, RDM12 shares nearly all of its hits with RDR2. However, RDR2 has many more putative substrates, suggesting that it has robust activity even in the absence of its cofactor or that it requires RDM12 at only a subset of its substrates.

Both of these trends are repeated, but weaker when considering genomic bins (Figure 5.1). RDR6 and SGS3 show substantial overlap, but they are each associated with bins that were not identified in their cofactor. Similarly, RDM12 shares a majority of its bins with RDR2 but also has bins that are not shared.

The two γ RDRs included in this study, RDR4 and 5, both have putative substrates. I will discuss each substrate list in more detail. However, it is notable that they have several shared hits, consistent with their high degree of homology. These data show that the combination of smRNA-seq and dsRNA-seq is powerful enough to detect genome-wide effects of RDR activity and co-factor relationships.

RDR1

I found a small number of putative substrates for RDR1. The presence of putative endogenous targets for RDR1 is notable because RDR1 is generally considered to act on exogenous targets such as viruses and transgenes. The list includes several proteins that either lack functional descriptions or have very minimal functional information such as assignment to a putative gene family. While it is impossible to draw conclusions from a lack of characterization, one might speculate that these genes are ncRNAs, possibly even giving rise to functional siRNAs. There are several annotated ncRNAs on the list including snRNA-like genes as well as known protein coding genes. Interestingly, many of these hits were identified on the antisense strand lending evidence of RDR activity.

The lists of putative RDR4 and RDR5 substrates are approximately divided between annotated transposable elements and annotated genes, suggesting that they may have multiple roles. Like RDR1, the two γ RDRs investigated here are associated with several uncharacterized and poorly characterized transcripts. Again, it would be premature to conclude anything about the functions of these genes. However, their lack of functional characterization invites speculation that they do not encode proteins.

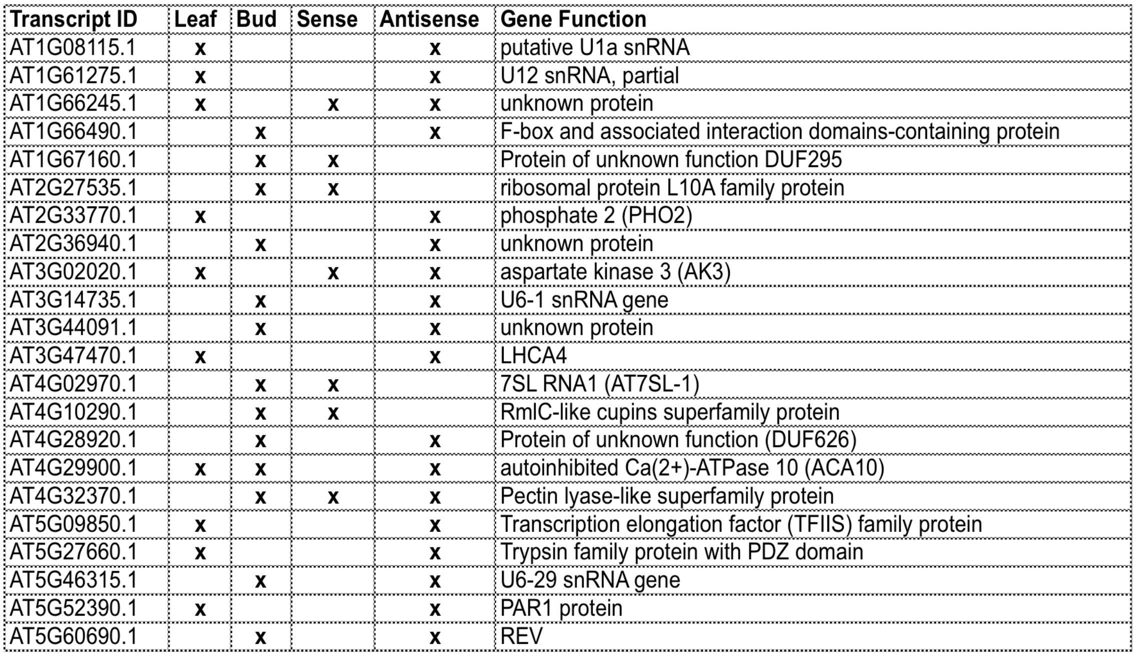
I visualized the smRNA profiles of the most highly expressed putative substrates in Col-0 (Figure 4). In general, I do not see a pattern consistent with phased siRNAs being progressively excised. Instead, I see one or two dominant siRNAs for each transcript. This may be due to the excision preferences of a DCL or the loading preferences of an AGO. The dominant siRNAs for a given transcript can come from either the sense or antisense strand or both. The overall composition of siRNAs originating from putative a RDR substrates is similar to the total smRNA population in Col-0 (Figure 5.5).

RDR4 and RDR5

The lists of putative RDR4 and RDR5 substrates are approximately divided between annotated transposable elements and annotated genes, suggesting that they may have multiple roles. Like RDR1, the two γ RDRs investigated here are associated with several

uncharacterized and poorly characterized transcripts. Again, it would be premature to conclude anything about the functions of these genes. However, their lack of functional characterization invites speculation that they do not encode proteins.

I visualized the smRNA profiles of the most highly expressed putative substrates in Col-0 (Figure 4). In general, I do not see a pattern consistent with phased siRNAs being progressively excised. Instead, I see one or two dominant siRNAs for each transcript. This may be due to the excision preferences of a DCL or the loading preferences of an AGO. The dominant siRNAs for a given transcript can come from either the sense or antisense strand or both. The overall composition of siRNAs originating from putative a RDR substrates is similar to the total smRNA population in Col-0 (Figure 5).



Transcript ID	Leaf	Bud	Sense	Antisense	Gene Function
AT1G08115.1	x			x	putative U1a snRNA
AT1G61275.1	x			x	U12 snRNA, partial
AT1G66245.1	x		x	x	unknown protein
AT1G66490.1		x		x	F-box and associated interaction domains-containing protein
AT1G67160.1		x	x		Protein of unknown function DUF295
AT2G27535.1		x	x		ribosomal protein L10A family protein
AT2G33770.1	x			x	phosphate 2 (PHO2)
AT2G36940.1		x		x	unknown protein
AT3G02020.1	x		x	x	aspartate kinase 3 (AK3)
AT3G14735.1		x		x	U6-1 snRNA gene
AT3G44091.1		x		x	unknown protein
AT3G47470.1	x			x	LHCA4
AT4G02970.1		x	x		7SL RNA1 (AT7SL-1)
AT4G10290.1		x	x		RmlC-like cupins superfamily protein
AT4G28920.1		x		x	Protein of unknown function (DUF626)
AT4G29900.1	x	x		x	autoinhibited Ca(2+)-ATPase 10 (ACA10)
AT4G32370.1		x	x	x	Pectin lyase-like superfamily protein
AT5G09850.1	x			x	Transcription elongation factor (TFIIS) family protein
AT5G27660.1	x			x	Trypsin family protein with PDZ domain
AT5G46315.1		x		x	U6-29 snRNA gene
AT5G52390.1	x			x	PAR1 protein
AT5G60690.1		x		x	REV

Table 5.3: Putative RDR1 substrates

Transcript ID	Leaf	Bud	Sense	Antisense	Gene Function
AT1G52315.1		x		x	Regulator of Vps4 activity in the MVB pathway protein
AT1G64035.1		x	x	x	pseudogene, putative serpin
AT1G66245.1	x		x		unknown protein
AT1G70840.1		x	x	x	MLP-like protein 31 (MLP31)
AT2G13540.1		x		x	ABA HYPERSENSITIVE 1 (ABH1)
AT2G15340.1		x		x	glycine-rich protein
AT2G16080.1	x			x	pseudogene, similar to OSJNBb0041J06.18 cultivar-group)}
AT2G27520.1		x	x	x	F-box and associated interaction domains-containing protein
AT2G27535.1		x	x		ribosomal protein L10A family protein
AT4G13992.1		x	x	x	Cysteine/Histidine-rich C1 domain family protein
AT4G32370.1		x	x	x	Pectin lyase-like superfamily protein
AT5G35605.1		x	x	x	tRNA-Arg (anticodon: CCT)

Table 5.4: Putative RDR4 substrates

Transcript ID	Leaf	Bud	Sense	Antisense	Gene Function
AT1G09026.1	x			x	unknown protein
AT1G26762.1		x		x	unknown protein
AT1G47389.1		x	x		unknown protein
AT1G59870.1		x	x		PENETRATION 3 (PEN3)
AT1G64035.1	x		x	x	pseudogene, putative serpin
AT1G66245.1	x		x	x	unknown protein
AT1G67160.1		x		x	Protein of unknown function
AT1G76040.2		x	x		calcium-dependent protein kinase 29 (CPK29)
AT2G02670.1		x	x	x	pseudogene, hypothetical protein
AT2G16580.1		x		x	SAUR-like auxin-responsive protein family
AT2G27535.1		x	x		ribosomal protein L10A family protein
AT3G03620.1		x		x	MATE efflux family protein
AT3G06125.2		x		x	Unknown gene
AT3G14210.1	x		x		epithiospecifier modifier 1 (ESM1)
AT3G17530.1		x	x		F-box and associated interaction domains-containing protein
AT3G50350.2		x	x		Protein of unknown function (DUF1685)
AT4G10290.1		x		x	RmlC-like cupins superfamily protein
AT4G29290.1		x	x		low-molecular-weight cysteine-rich 26 (LCR26)
AT5G11420.1	x			x	molecular function unknown
AT5G35605.1		x	x	x	tRNA-Arg (anticodon: CCT)
AT5G52390.1	x		x		PAR1 protein

Table 5.5: Putative RDR5 substrates

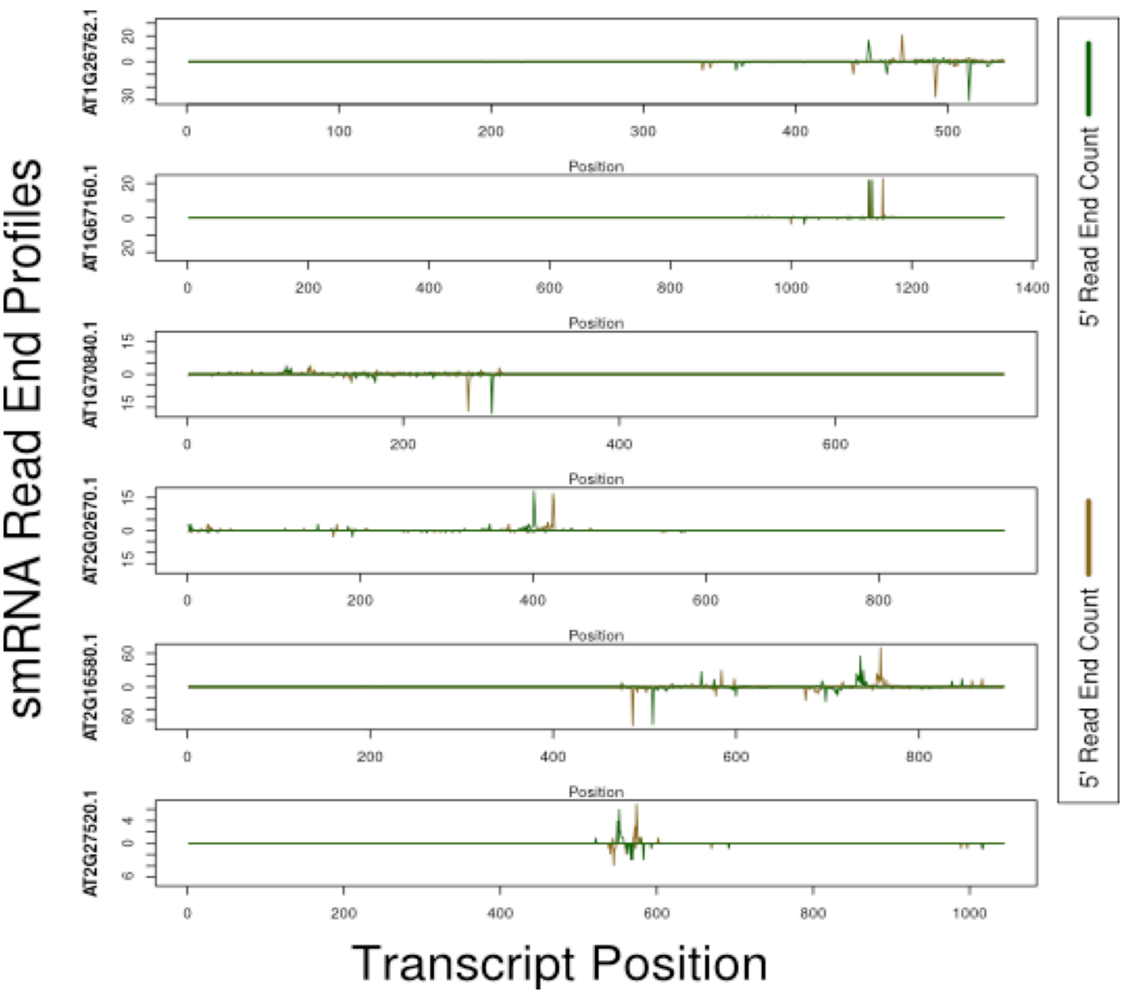


Figure 5.4: smRNA excision profiles of selected putative RDR4 and RDR5 substrates. The number of 5' and 3' read ends is plotted for each position. Peaks above the x axis indicate excisions from the sense strand, peaks below the axis indicate excisions from the antisense strand

RDR6 and SGS3

The putative substrates of RDR6 and SGS3 indicated by my detection method identified many of the previously known substrates of these proteins, providing internal validation of the hit lists. These include the *TAS* loci discussed, above as well as multiple members of the Pentatricopeptide Repeat (PPR) Superfamily. In addition, we see many novel substrates including several genes associated with auxin response, an important plant hormone. This relates to the *TAS3* genes because the ta-siRNAs produced from these transcripts regulate a family of transcription factors involved in response to auxin, the ARFs.[177,179,180,182]

Transcript ID	Leaf	Bud	Sense	Antisense	Gene Function
AT1G12300.1	x		x		Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G12460.1		x	x	x	Leucine-rich repeat protein kinase family protein
AT1G12820.1	x			x	auxin signaling F-box 3 (AFB3)
AT1G47389.1		x	x		unknown protein
AT1G50055.1	x		x	x	Trans-acting siRNA1b primary transcript (TAS1b). Regulated by miR173.
AT1G53490.1	x		x	x	RING/U-box superfamily protein
AT1G62860.1	x		x	x	pseudogene of pentatricopeptide (PPR) repeat-containing protein
AT1G62910.1	x		x	x	Pentatricopeptide repeat (PPR) superfamily protein
AT1G62914.1	x		x	x	pentatricopeptide (PPR) repeat-containing protein
AT1G62930.1	x		x	x	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G63080.1	x		x	x	Pentatricopeptide repeat (PPR) superfamily protein
AT1G63150.1	x		x	x	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G63230.1	x			x	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G64583.1	x		x	x	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G70080.1		x	x	x	Terpenoid cyclases/Protein prenyltransferases superfamily protein
AT2G16580.1		x	x	x	SAUR-like auxin-responsive protein family
AT2G27400.1	x		x	x	Trans-acting siRNA1a primary transcript (TAS1a). Regulated by miR173.
AT2G28350.1		x	x		auxin response factor 10 (ARF10)
AT2G31070.1		x		x	TCP domain protein 10 (TCP10)
AT2G39675.1		x	x	x	Trans-acting siRNA1c primary transcript (TAS1c)
AT2G39681.1	x		x	x	Trans-acting siRNA primary transcript
AT2G45160.1		x		x	HAIRY MERISTEM 1 (HAM1)
AT3G15730.1	x			x	phospholipase D alpha 1 (PLDALPHA1)
AT3G17510.1		x		x	CBL-interacting protein kinase 1 (CIPK1)
AT3G17530.1	x		x		F-box and associated interaction domains-containing protein
AT3G23690.1	x		x	x	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT3G26810.1		x	x	x	auxin signaling F-box 2 (AFB2)
AT3G50022.1		x		x	unknown protein
AT3G60630.1		x		x	HAIRY MERISTEM 2 (HAM2)
AT4G14610.1	x		x	x	pseudogene, disease resistance protein (CC-NBS-LRR class
AT4G21370.1		x	x		similar to ARK1
AT4G32370.1		x	x	x	Pectin lyase-like superfamily protein
AT5G16640.1	x	x	x	x	Pentatricopeptide repeat (PPR) superfamily protein
AT5G38850.1	x		x	x	Disease resistance protein (TIR-NBS-LRR class)
AT5G39370.1		x	x	x	Curculin-like (mannose-binding) lectin family protein
AT5G49615.1		x	x	x	trans-acting siRNA (tasi-RNA)
AT5G60690.1		x		x	REVOLUTA (REV)
AT5G63020.1	x		x	x	Disease resistance protein (CC-NBS-LRR class) family

Table 5.6: Putative RDR6 substrates

Like the previously discussed RDRs, RDR6 and its cofactor SGS3 are associated with unannotated and poorly annotated genes. Inspecting the siRNA excision profiles of some of the known substrates reveals some evidence of phasing (Figure 5.6), particularly of *TAS1A*, which shows multiple phased excision sites. Other examples show two or more phased sites, but like the putative substrates of RDR4 and RDR5, one or two siRNAs dominate the profile. The siRNA populations mapping to RDR6 and

Transcript ID	Leaf	Bud	Sense	Antisense	Gene Function
AT1G09260.1		x	x		Chaperone DnaJ-domain superfamily protein
AT1G10745.1		x	x		Maternally expressed gene (MEG) family protein
AT1G14225.1		x	x		unknown protein
AT1G26762.1		x	x	x	unknown protein
AT1G32225.1		x		x	unknown protein
AT1G35515.1		x	x		high response to osmotic stress 10 (HOS10)
AT1G45040.1		x	x		pseudogene, hypothetical protein
AT1G51150.1	x			x	DegP protease 6 (DegP6)
AT1G52315.1		x	x		Regulator of Vps4 activity in the MVB pathway protein
AT1G59680.1		x	x	x	embryo sac development arrest 1 (EDA1)
AT1G64035.1		x	x	x	pseudogene
AT1G66245.1	x		x	x	unknown protein
AT1G67105.1		x	x		other RNA
AT1G67240.1		x	x	x	Mutator-like transposase family
AT1G76040.2		x		x	calcium-dependent protein kinase 29 (CPK29)
AT2G04870.1		x	x		unknown protein
AT2G05335.1		x	x		SCR-like 15 (SCRL15)
AT2G13540.1		x	x	x	ABA HYPERSENSITIVE 1 (ABH1)
AT2G15340.1	x		x		glycine-rich protein
AT2G27535.1		x		x	ribosomal protein L10A family protein
AT3G04250.1	x		x		F-box associated ubiquitination effector family protein
AT3G04717.1		x	x	x	pseudogene
AT3G05755.1	x			x	tRNA-Pro (anticodon: CGG)
AT3G10900.1	x		x	x	Glycosyl hydrolase superfamily protein
AT3G19350.1		x	x	x	maternally expressed pab C-terminal (MPC)
AT3G22770.1	x			x	F-box associated ubiquitination effector family protein
AT3G27590.1		x	x		unknown protein
AT3G28899.1	x		x		unknown protein
AT3G44091.1		x	x		unknown protein
AT4G00960.1		x		x	Protein kinase superfamily protein
AT4G07965.1		x	x		unknown protein
AT4G10190.1	x		x	x	F-box and associated interaction domains-containing protein
AT4G10290.1		x		x	RmlC-like cupins superfamily protein
AT4G13075.1		x	x	x	RALF-like 30 (RALFL30)
AT4G28470.1		x	x		26S proteasome regulatory subunit S2 1B (RPN1B)
AT4G32370.1		x	x	x	Pectin lyase-like superfamily protein
AT5G35605.1		x	x	x	tRNA-Arg (anticodon: CCT)
AT5G42140.1		x		x	Regulator of chromosome condensation (RCC1) family
AT5G42265.1		x	x	x	unknown pseudogene
AT5G47300.1	x			x	unknown protein

Table 5.7: Putative SGS3 dependent substrates

SGS3 genomic bins also show enrichment for 21 nt siRNAs, consistent with the well-described ta-siRNA biogenesis pathway.

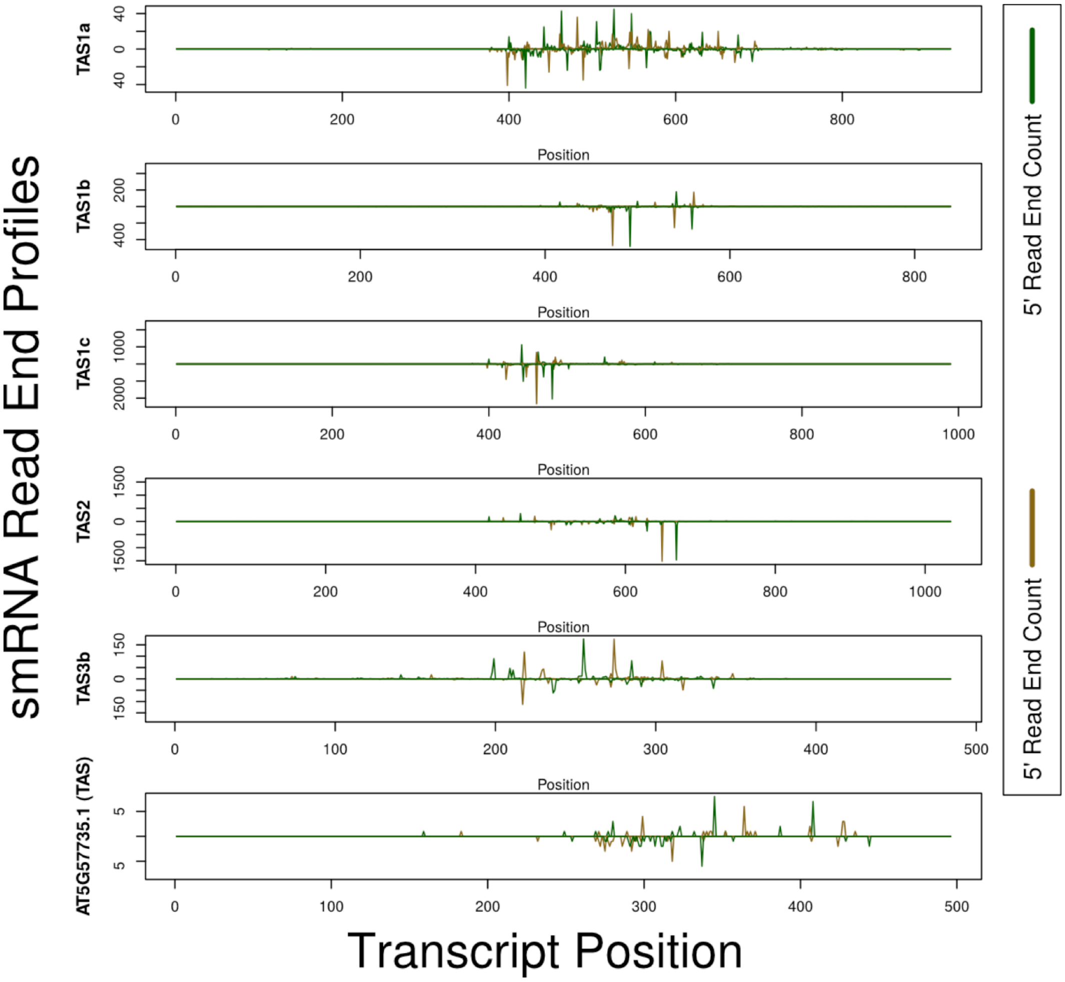


Figure 5.6: smRNA excision profiles for known RDR6/SGS3 substrates from the TAS family. The counts of 3' and 5' smRNA read ends are shown. Upward peaks are from the sense strand, downward peaks are antisense

RDR substrate derived siRNAs and their targets

Plant siRNAs bind their targets through nucleotide pairing along almost their entire length, so it is possible to predict targets of siRNAs by searching for sites with near perfect complementarity *in silico.* In order to predict targets for siRNAs that arise from RDR substrate loci, I intersected a list of the smRNAs expressed in Col-0 with each set of genomic bins identified as RDR substrates, excluding the very large set associated with RDR2.

These were then collapsed down to a set of unique putative siRNAs. This unique list was used to query the set of Arabidopsis transcripts, excluding miRNAs using the target search tool psRNATarget.[100] I identified 12,304 predicted targeting interactions for RDR-dependent siRNAs identified in our unopened flower bud analyses (Table 7).

Inspecting the list for known previously characterized targeting relationships. Among the targets, I found multiple members of the Pentatricopeptide Repeat (PPR) family as well as members of a family with similar annotation, the Tetratricopeptide Repeat (TPR) genes. I found 139 targeting interactions involving 68 distinct PPRs and 119 interactions involving 73 distinct TPRs. Many members of these families appear as both putative RDR substrates, and as targets of the siRNAs that require RDR activity suggesting a potential self-regulatory mechanism,[186] as previously described.

In addition, I found the known targets of *TAS3*-derived ta-siRNAs, ARF2, ARF3 and ARF4.[177,179,182,208,209] All three are predicted by psRNATarget to be targeted at multiple sites by multiple siRNAs, the majority of which were from *TAS3A* or *TAS3B*. To further characterize these target sites I analyzed cleavage data at the miRNA target sites.

Bud tissue was selected for this analysis because it could be integrated with a dataset from an experiment that can detect cleavage events. As discussed in a previous chapter, GMUCT is a sequencing based technique that takes advantage of the fact that RNA molecules which are cleaved or are being degraded have a free 5′ monophosphate. The protocol selectively clones such molecules into sequencing libraries.[210] When the resulting reads are mapped to transcripts, each detectable cleavage event appears as a pileup of reads with a common 5' end corresponding to the cleavage site.

The targeting interactions on the *ARF2* and *ARF3* transcripts fall into two regions on each transcript, with each site being targeted by multiple distinct siRNAs. This is possible because the sites overlap, but the siRNAs either have mismatches, are of slightly different lengths, or are have positions offset by a nucleotide or more. Similarly, psRNATarget predicts three target sites on *ARF4*, two of which match multiple *TAS3A* or *TAS3B* ta-siRNAs. One matches a region of the genome with no known transcripts. No cleavage signal was detected on *ARF2*, I inspected the GMUCT 2.0 signal for *ARF3* and *ARF4* (Figure 7).

Both replicates of GMUCT 2.0 show sharp peaks at the centers of both *ARF3* target sites indicating a cleavage event. The GMUCT 2.0 coverage was lower for *ARF4*. However, there are peaks at the centers of both sites. Evolutionary conservation data from a phastCons analysis was available for the *ARF3* sites, and shows that both are highly conserved across plant species.

Taken with the other findings in this study, the analysis of the ARF genes suggests that my pipeline is powerful enough to capture RDR substrates, identify the dominant smRNAs produced at those sites, predict targets for them and validate the targets that are cleaved, all on a genome-wide scale.

5.4 Discussion

In this study, I present a pipeline that uses RNA-seq technology to discover substrates of RDR, to characterize the smRNAs that they are processed into and to identify the targets of those smRNAs. I found putative substrates for five out of the six RDRs encoded in the *Arabidopsis* genome.

The selection of *Arabidopsis* was advantageous because it allowed several opportunities for internal validation that strengthened my confidence in the novel substrates. The ta-siRNAs and the *TAS* loci are among the best-characterized RDR substrates. The pipeline I show here was able to detect them using two lines of evidence, smRNA-seq and dsRNA-seq. In addition, it was able to predict targets of these siRNAs and demonstrate cleavage of those targets.

In addition, I observed a less well-studied class of RDR6 substrates, the PPR genes. This large family of genes is known to produce siRNAs, which target members of the same gene family. I was able to detect this family as both substrates and siRNA targets.

I was also able to detect genome-wide differences in the *rdr2* mutant. I noted that the mutant lost the ability to produce both dsRNAs, the direct product of RDR2, as well as the hsiRNAs that are cleaved from them. I was additionally able to distinguish the population of hsiRNAs from the ta-siRNAs produced by RDR6. For both RDR6 and RDR2 I observed common substrates that depend on both the RDR and a known cofactor, SGS3 and RDM12 respectively.

Taken together, these observations add confidence to the putative substrates detected for the previously uncharacterized γ RDRs. The substrates assigned to these genes include many transcripts, which also lack a known function, but can be shown to produce smRNAs. Some of these smRNAs are predicted to target other transcripts. While more data must be collected before hypothesizing about their function, these finding invite speculation about their possible similarities to hsiRNAs and ta-siRNAs.

In the future, it will be intriguing to functionally characterize these substrates as well as the substrates of RDR3. Of particular interest are the uncharacterized substrates of the γ RDRs as the function of these genes is still unknown. Because of the high homology of these genes, functional redundancy may limit the number of substrates that can be identified in a single mutant.

Applying the pipeline used here to an *rdr3*/*rdr4*/*rdr5* triple mutant will likely reveal more. To date, such a mutant has not been developed due to the difficulty of deleting three genes that exist as tandem repeats on the chromosome. However, with recent advances in CRISPR technology, development is very likely achievable. Ultimately it will be possible, when enough substrates are identified, to infer the criteria by which γ RDRs select their targets.

CHAPTER 6: Conclusions

Summary

In this dissertation, I discussed three intimately linked aspects of RNA biology: RNA secondary structure, RNA cleavage, and processing of small RNAs from their longer double-stranded precursors. First, I discussed a new database of RNA secondary structure data drawn from diverse high-throughout sequencing-based structure mapping techniques. I showed how browsing the database could reveal both regions of consensus among the techniques and transcript features where they differed. I argued that the regions of difference showed the limitations of interpreting any one technique, but that by integrating them I could be more confident about the structural properties of specific transcripts.

In browsing the database, I navigated to transcripts with known structures. In particular, I was able to see the hairpin structures of two murine Iron Response Elements (IREs). The database showed evidence of structure in the stem of both IREs and lack of structure in the loop, agreeing with the known structures of these elements. I also examined another class of hairpins, primary and precursor miRNA (pri-miRNA and pre-miRNA) transcripts. I considered them together for two reasons. One is that the stem and loop, the regions with well-defined structures, are often the same between pri- and pre-miRNA. The other, more practical reason is that many of the currently available annotation resources do not distinguish between them.

I observed miRNAs with low score density as well as structures that seemed internally inconsistent, for example, hairpins in which one strand of the stem appeared highly structured and the other strand appeared unstructured. I speculated that these results might be caused by a mixture of pri- pre- and mature miRNAs in the cell leading to structural ambiguity.

Fortunately, I had access to data from an experiment performed in the model plant *Arabidopsis thaliana* in which RNA had been purified from nuclei. Plant pri- and pre-miRNAs are processed in the nucleus and exported to the cytosol as mature miRNAs, so I reasoned that nuclear miRNAs would be enriched for the pri- and pre- forms. I computed structure scores from this experimentally determined data sets and loaded them into the database. In Chapter 2, I discussed the structure score patterns of miRNAs from these data. Briefly, I found scores consistent with stem loops in most of the miRNAs with high score coverage and I showed some examples. I found that, even when score coverage was incomplete, I could interpolate missing scores and still observe a stem loop pattern.

This recurring structural pattern is an important link between RNA secondary structure and smRNAs and it also points toward RNA cleavage. After all, miRNAs are processed into their mature form through cleavage events. Additionally, in plants, many miRNA are used to target transcripts for cleavage. Just as our lab has been interested in developing tools to study RNA structure genome-wide, we have also worked to develop a method for detecting cleavage and degradation using high-throughput sequencing. In this dissertation, I focused on the latest refinement of our technique, GMUCT 2.0.

I described the technique in detail in Chapter 3 and an even more complete protocol is given in Willmann et al, 2014. Briefly, GMUCT 2.0 exploits a useful fact of RNA chemistry. Specifically, transcripts that have been cleaved, or are undergoing 5' to 3' degradation, have a free monophosphate on their 5' ends. Contrast this with intact transcripts, which have a 7-methylguanylate cap in that position. GMUCT 2.0 creates sequencing libraries from the pool of molecules with free 5' monophosphates. The resulting libraries, when aligned, reveal transcripts undergoing degradation, as well as cleavage sites, including those induced by miRNAs incorporated into an RNA-induced silencing complex.

From my analyses, I found that GMUCT 2.0 expression results correlated highly with mRNA-seq expression for most *Arabidopsis* transcripts. This makes sense given that a transcript must be expressed in order to be degraded. It also opens up the possibility of normalizing GMUCT 2.0 expression to mRNA-seq expression to identify highly degraded transcripts. The profiles of transcripts in our GMUCT libraries had a 3' bias, possibly hinting at the kinetics of degradation.

I also inspected the targets of miRNAs. For both *Arabidopsis* and human tissue culture cells, I created a meta-transcript profile by normalizing and combining expression across a list of miRNA target sites. In *Arabidopsis*, I found the meta-transcript profile had a sharp peak in the center of the miRNA footprint consistent with miRNA induced cleavage. The meta-transcripts for the human cell lines didn't show a similar peak. Again this agrees with all currently available data that suggests mammalian miRNAs regulate their target transcripts by a mechanism that does not involve cleavage, but is focused on translation inhibition.

One useful consequence of GMUCT 2.0 target site peaks, at least in plants, is that they can be used to verify putative smRNA targeting interactions. This is one of several analyses that I integrated in Chapter 4. In that chapter, I discussed a class of polymerases that use RNA as a template for RNA synthesis (RDRs). The *Arabidopsis* genome encodes six such enzymes. The double-stranded RNAs that they produce act as precursors for at least two classes of characterized siRNAs.

I used dsRNA-seq to look for regions of the genome where double-stranded RNAs were being made in wild type plants but were reduced in *rdr* mutants. Similarly, I used smRNA-seq to look for smRNAs that were lost in mutants. I reasoned that regions that lost expression in both sequencing types, whether or not they were annotated genes, were likely to be RDR substrates. In this way, I compiled a list of putative targets for each known RDR in two tissues.

Many of the putative substrates I identified were known, especially for the most well studied RDRs, which gave me confidence that the analysis was powerful enough to detect previously unknown substrates. I found novel substrates for each of the RDRs, including two γ RDRs for which no substrates are currently known.

I was also interested in the targeting interactions of the siRNAs that rely on RDRs for their synthesis. I found predicted targets for the siRNAs associated with each of these proteins. Because some siRNAs are involved in cleavage of their targets, I integrated GMUCT 2.0 data into my analysis to identify which of the target sites were cleaved. I found many interesting examples, including some known cleavage targets as well as some novel ones.

The analysis of the RDR substrates brought together aspects of all of my other projects. It required an understanding of RNA-secondary structure, smRNAs, and siRNA-induced cleavage, as well as all the tools used for building RNA-seq based pipelines. It also challenged my understanding of plant biology and cell biology in general. Thus, it represents a microcosm of the concepts presented in this dissertation.

Future Directions

Each of the projects I discussed is still ongoing with intriguing possibilities for the future. The databases I described originally contained six mammalian data sets. An additional four were recently added for *Arabidopsis.* As techniques for inferring secondary structure continue to develop and become widespread, the available data will only grow.

One core improvement for the future is development of the scoring methods. Every technique has its own distinct score formula, but they share a key oversimplification. Each of them represents a per-base *ratio* of expression. However, none of them take into account the *magnitude* of expression. As a thought experiment, consider the following ratios where the numerator represents evidence that a given nucleotide is paired and the denominator is evidence that it is unpaired: 1/4 vs 1000/4000.

It's clear that the latter contains more information than the former, however under all of the current scoring methods the two would produce similar scores. Each of the current formulas is a reasonable approach at converting sequencing reads to interpretable secondary structure information, especially since the techniques are relatively new.

It would have been premature to make assumptions about the properties of the data. I discussed in Chapter 2 how strikingly different the score distributions are for the four techniques. However, as more data is generated it will be important to develop a concept of variance for the scores so that they can be weighted by confidence.

Even without making the score calculations more sophisticated, it's clear that the database is useful in its current form. One future direction is to simply browse it for features of interest. I found that it can visualize short hairpins and I found that miRNA precursors in particular have a characteristic pattern, especially when missing scores are interpolated. An appealing application of this is to encode this pattern formally so that pri- and pre-miRNAs can be detected algorithmically.

This would add an additional layer of data for miRNA detection software that currently detects putative miRNAs using sequence alone. It would also benefit users of miRNA databases such as miRBase who are unsure of miRNA entries supported only by *in silico* evidence. Currently there is only one dataset for one organism with structure scores from nuclei. However, as such experiments continue in the Gregory lab and in others, algorithm development for structure-based miRNA detection will become a real possibility.

Related to this is the determination of pri- and pre-miRNA structures. I discussed in Chapter 2 how the structures recorded in miRBase are often determined by *in silic*o folding algorithms, which use minimum free energy to maximize structure likelihood. It has been shown that, at least for some famous miRNAs, that these structures are not always perfectly correct. I suspect it will be possible to update *in silic*o structures using empirical scores from the database. I proposed a method for doing this. However more work needs to be done to validate and, most likely, refine these methods.

There is also additional work to do on the *Arabidopsis* RDR project. I showed how data integration can be used to track RDR substrates from their precursors to their biosynthesis to their eventual function as mature siRNAs by integrating dsRNA-seq, smRNA-seq, and GMUCT 2.0 data. As I've worked on various iterations of this project I've been interested in making it increasingly automated. In its current state, it still requires an expert user to run many of the steps.

As I, and others, continue to develop the pipeline it will evolve into a tool that can be used without computational training. This will enable researchers to probe for substrates of all RDRs across multiple tissues and organisms. Ultimately the goal is to determine what makes a transcript a suitable RDR substrate. Given the complexity of known mechanisms, this will likely require additional analytic steps along with careful experimental validation. Internal validation with the known substrates I explored here will continue to be a valuable metric for technical improvement.

BIBLIOGRAPHY

1. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).

2. Fenton, M. J. Review: transcriptional and post-transcriptional regulation of interleukin 1 gene expression. *Int. J. Immunopharmacol.* **14,** 401–411 (1992).

3. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9,** 102–114 (2008).

4. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582,** 1977–1986 (2008).

5. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84,** 291–323 (2015).

6. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17,** 407–421 (2016).

7. Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoform□specific translational regulation via alternative transcription start sites in mammals. *Mol. Syst. Biol.* **12,** 875 (2016).

8. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15,** 829–845 (2014).

9. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8,** 479–490 (2007).

10. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).

11. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–177 (2013).

12. Hentze, M. W. *et al.* A model for the structure and functions of iron-responsive elements. *Gene* **72,** 201–208 (1988).

13. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31,** 147–157 (1982).

14. Madhani, H. D. snRNA Catalysts in the Spliceosome's Ancient Core. *Cell* **155,** 1213–1215 (2013).

15. Madhani, H. D. & Guthrie, C. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71,** 803–817 (1992).

16. Kim, S. H. *et al.* Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185,** 435–440 (1974).

17. Yusupova, G. & Yusupov, M. High-resolution structure of the eukaryotic 80S ribosome. *Annu. Rev. Biochem.* **83,** 467–486 (2014).

18. Bhaskaran, H., Rodriguez-Hernandez, A. & Perona, J. J. Kinetics of tRNA folding monitored by aminoacylation. *RNA* **18,** 569–580 (2012).

19. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. *Science* **289,** 920–930 (2000).

20. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28,** 411–418 (2003).

21. Ramakrishnan, V. The Ribosome Emerges from a Black Box. *Cell* **159,** 979–984 (2014).

22. Nguyen, G. T. D. T., Scaife, M. A., Helliwell, K. E. & Smith, A. G. Role of riboswitches in gene regulation and their potential for algal biotechnology. *J. Phycol.* **52,** 320–328 (2016).

23. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111,** 747–756 (2002).

24. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29,** 11–17 (2004).

25. Winkler, W. C. Metabolic monitoring by bacterial mRNAs. *Arch. Microbiol.* **183,** 151–159 (2005).

26. Babitzke, P. Regulation of transcription attenuation and translation initiation by allosteric control of an RNA-binding protein: the Bacillus subtilis TRAP protein. *Curr. Opin. Microbiol.* **7,** 132–139 (2004).

27. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches Control Fundamental Biochemical Pathways in Bacillus subtilis and Other Bacteria. *Cell* **113,** 577–586 (2003).

28. Chen, G. & Yanofsky, C. Tandem Transcription and Translation Regulatory Sensing of Uncharged Tryptophan tRNA. *Science* **301,** 211–213 (2003).

29. Dambach, M. *et al.* The ubiquitous yybP-ykoY riboswitch is a manganese-responsive regulatory element. *Mol. Cell* **57,** 1099–1109 (2015).

30. Furukawa, K. *et al.* Bacterial riboswitches cooperatively bind Ni(2+) or Co(2+) ions and control expression of heavy metal transporters. *Mol. Cell* **57,** 1088–1098 (2015).

31. Salehi-Ashtiani, K., Lupták, A., Litovchick, A. & Szostak, J. W. A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene. *Science* **313,** 1788–1792 (2006).

32. Teixeira, A. *et al.* Autocatalytic RNA cleavage in the human β-globin pre-mRNA promotes transcription termination. *Nature* **432,** 526–530 (2004).

33. Ganot, P., Bortolin, M.-L. & Kiss, T. Site-Specific Pseudouridine Formation in Preribosomal RNA Is Guided by Small Nucleolar RNAs. *Cell* **89,** 799–809 (1997).

34. Kiss, T. Small Nucleolar RNAs: An Abundant Group of Noncoding RNAs with Diverse Cellular Functions. *Cell* **109,** 145–148 (2002).

35. Kiss-László, Z., Henry, Y., Bachellerie, J.-P., Caizergues-Ferrer, M. & Kiss, T. Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. *Cell* **85,** 1077–1088 (1996).

36. Kozak, M. Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol. Cell. Biol.* **9,** 5134–5142 (1989).

37. Svitkin, Y. V. *et al.* The requirement for eukaryotic initiation factor 4A (eIF4A) in translation is in direct proportion to the degree of mRNA 5′ secondary structure. *RNA* **7,** 382–394 (2001).

38. Li, F. *et al.* Global analysis of RNA secondary structure in two metazoans. *Cell Rep.* **1,** 69–82 (2012).

39. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).

40. Li, F. *et al.* Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. *Plant Cell* **24,** 4346–4359 (2012).

41. Robertus, J. D. *et al.* Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature* **250,** 546–551 (1974).

42. Selmer, M. *et al.* Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313,** 1935–1942 (2006).

43. Ban, N. *et al.* Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature* **400,** 841–847 (1999).

44. Berman, H. M. The Protein Data Bank: a historical perspective. *Acta Crystallogr. A* **64,** 88–95 (2008).

45. Lyngsø, R. B. & Pedersen, C. N. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **7,** 409–427 (2000).

46. Akutsu, T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.* **104,** 45–62 (2000).

47. McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29,** 1105–1119 (1990).

48. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148 (1981).

49. Rivas, E. & Eddy, S. R. A dynamic programming algorithm for RNA structure prediction including pseudoknots1. *J. Mol. Biol.* **285,** 2053–2068 (1999).

50. Liu, H., Xu, D., Shao, J. & Wang, Y. An RNA folding algorithm including pseudoknots based on dynamic weighted matching. *Comput. Biol. Chem.* **30,** 72–76 (2006).

51. Bompfünewerer, A. F. *et al.* Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.* **56,** 129–144 (2008).

52. Hofacker, I. L. & Stadler, P. F. Memory efficient folding algorithms for circular RNA secondary structures. *Bioinforma. Oxf. Engl.* **22,** 1172–1176 (2006).

53. Galil, Z. & Park, K. Parallel Algorithms for Dynamic Programming Recurrences with More Than O(1) Dependency. *J. Parallel Distrib. Comput.* **21,** 213–222 (1994).

54. Xia, F. & Jin, G. Fine-grained parallelism accelerating for RNA secondary structure prediction with pseudoknots based on FPGA. *J. Bioinform. Comput. Biol.* **12,** 1450008 (2014).

55. Tijerina, P., Mohr, S. & Russell, R. DMS Footprinting of Structured RNAs and RNA-Protein Complexes. *Nat. Protoc.* **2,** 2608–2623 (2007).

56. Stern, S., Moazed, D. & Noller, H. F. in (ed. Enzymology, B.-M. in) **164,** 481–489 (Academic Press, 1988).

57. McGinnis, J. L., Dunkle, J. A., Cate, J. H. D. & Weeks, K. M. The Mechanisms of RNA SHAPE Chemistry. *J. Am. Chem. Soc.* **134,** 6617–6624 (2012).

58. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (2005).

59. RajBhandary, U. L. & Chang, S. H. Studies on polynucleotides. LXXXII. Yeast phenylalanine transfer ribonucleic acid: partial digestion with ribonuclease T-1 and derivation of the total primary structure. *J. Biol. Chem.* **243,** 598–608 (1968).

60. Ehresmann, C. *et al.* Probing the structure of RNAs in solution. *Nucleic Acids Res.* **15,** 9109–9128 (1987).

61. Volkin, E. & Cohn, W. E. On the structure of ribonucleic acids. II. The products of ribonuclease action. *J. Biol. Chem.* **205,** 767–782 (1953).

62. Uchida, T., Arima, T. & Egami, F. Specificity of RNase U2. *J. Biochem. (Tokyo)* **67,** 91–102 (1970).

63. Desai, N. A. & Shankar, V. Single-strand-specific nucleases. *FEMS Microbiol. Rev.* **26,** 457–491 (2003).

64. Knapp, G. Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol.* **180,** 192–212 (1989).

65. Favorova, O. O., Fasiolo, F., Keith, G., Vassilenko, S. K. & Ebel, J. P. Partial digestion of tRNA--aminoacyl-tRNA synthetase complexes with cobra venom ribonuclease. *Biochemistry (Mosc.)* **20,** 1006–1011 (1981).

66. Lockard, R. E. & Kumar, A. Mapping tRNA structure in solution using double-strand-specific ribonuclease V1 from cobra venom. *Nucleic Acids Res.* **9,** 5125–5140 (1981).

67. Sand, M. The pathway of miRNA maturation. *Methods Mol. Biol. Clifton NJ* **1095,** 3–10 (2014).

68. Cai, Y., Yu, X., Hu, S. & Yu, J. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7,** 147–154 (2009).

69. Bologna, N. G., Mateos, J. L., Bresso, E. G. & Palatnik, J. F. A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *EMBO J.* **28,** 3646–3656 (2009).

70. Chapman, E. J. & Carrington, J. C. Specialization and evolution of endogenous small RNA pathways. *Nat. Rev. Genet.* **8,** 884–896 (2007).

71. Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription sta... - PubMed - NCBI. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21295131. (Accessed: 8th March 2016)

72. Addo-Quaye, C., Eshoo, T. W., Bartel, D. P. & Axtell, M. J. Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. *Curr. Biol.* **18,** 758–762 (2008).

73. Willmann, M. R., Endres, M. W., Cook, R. T. & Gregory, B. D. The Functions of RNA-Dependent RNA Polymerases in Arabidopsis. *Arab. Book Am. Soc. Plant Biol.* **9,** (2011).

74. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75,** 843–854 (1993).

75. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75,** 855–862 (1993).

76. Eulalio, A. *et al.* Deadenylation is a widespread effect of miRNA regulation. *RNA* **15,** 21–32 (2009).

77. Moss, E. G., Lee, R. C. & Ambros, V. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell* **88,** 637–646 (1997).

78. Olsen, P. H. & Ambros, V. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216,** 671–680 (1999).

79. Bagga, S. *et al.* Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122,** 553–563 (2005).

80. Wu, L., Fan, J. & Belasco, J. G. MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 4034–4039 (2006).

81. Giraldez, A. J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312,** 75–79 (2006).

82. Arribas-Hernández, L., Kielpinski, L. J. & Brodersen, P. mRNA decay of most Arabidopsis miRNA targets requires slicer activity of AGO1. *Plant Physiol.* (2016). doi:10.1104/pp.16.00231

83. Baldrich, P. & San Segundo, B. MicroRNAs in Rice Innate Immunity. *Rice N. Y. N* **9,** 6 (2016).

84. Han, Y., Zhang, B., Qin, X., Li, M. & Guo, Y. Investigation of a miRNA-Induced Gene Silencing Technique in Petunia Reveals Alterations in miR173 Precursor Processing and the Accumulation of Secondary siRNAs from Endogenous Genes. *PloS One* **10,** e0144909 (2015).

85. Branscheid, A. *et al.* SKI2 mediates degradation of RISC 5'-cleavage fragments and prevents secondary siRNA production from miRNA targets in Arabidopsis. *Nucleic Acids Res.* **43,** 10975–10988 (2015).

86. Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J. & Zhang, B.-T. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* **7,** 411 (2006).

87. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37,** 495–500 (2005).

88. Sætrom, O., Snøve, O. & Sætrom, P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA* **11,** 995–1003 (2005).

89. Enright, A. J. *et al.* MicroRNA targets in Drosophila. *Genome Biol.* **5,** R1 (2003).

90. Kiriakidou, M. *et al.* A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* **18,** 1165–1178 (2004).

91. Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA N. Y. N* **10,** 1507–1517 (2004).

92. Rusinov, V., Baev, V., Minkov, I. N. & Tabler, M. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res.* **33,** W696-700 (2005).

93. Lewis, B. P., Shih, I. -hun., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115,** 787–798 (2003).

94. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39,** 1278–1284 (2007).

95. Li, L., Xu, J., Yang, D., Tan, X. & Wang, H. Computational approaches for microRNA studies: a review. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **21,** 1–12 (2010).

96. Jones-Rhoades, M. W. & Bartel, D. P. Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA. *Mol. Cell* **14,** 787–799 (2004).

97. Rhoades, M. W. *et al.* Prediction of Plant MicroRNA Targets. *Cell* **110,** 513–520 (2002).

98. Xie, F. L. *et al.* Computational identification of novel microRNAs and targets in Brassica napus. *FEBS Lett.* **581,** 1464–1474 (2007).

99. Zhang, Y. miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res.* **33,** W701-704 (2005).

100. Dai, X. & Zhao, P. X. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* gkr319 (2011). doi:10.1093/nar/gkr319

101. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505,** 701–705 (2014).

102. Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519,** 486–490 (2015).

103. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).

104. Peattie, D. A. Direct chemical method for sequencing RNA. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1760–1764 (1979).

105. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 4679–4682 (1980).

106. Hector, R. D. *et al.* Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic Acids Res.* **42,** 12138–12154 (2014).

107. Mortimer, S. A. & Weeks, K. M. A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry. *J. Am. Chem. Soc.* **129,** 4144–4145 (2007).

108. Mortimer, S. A., Trapnell, C., Aviran, S., Pachter, L. & Lucks, J. B. SHAPE-Seq: High-Throughput RNA Structure Analysis. *Curr. Protoc. Chem. Biol.* **4,** 275–297 (2012).

109. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

110. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

111. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

112. Li, F. *et al.* SAVoR: a server for sequencing annotation and visualization of RNA structures. *Nucleic Acids Res.* **40,** W59-64 (2012).

113. Gosai, S. J. *et al.* Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the Arabidopsis Nucleus. *Mol. Cell* **57,** 376–388 (2015).

114. Baulcombe, D. RNA silencing in plants. *Nature* **431,** 356–363 (2004).

115. Jones-Rhoades, M. W., Bartel, D. P. & Bartel, B. MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS. *Annu. Rev. Plant Biol.* **57,** 19–53 (2006).

116. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431,** 343–349 (2004).

117. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. MicroRNAs in plants. *Genes Dev.* **16,** 1616–1626 (2002).

118. Lee, Y. *et al.* MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23,** 4051–4060 (2004).

119. Xie, Z. *et al.* Expression of Arabidopsis MIRNA Genes. *Plant Physiol.* **138,** 2145–2154 (2005).

120. Kurihara, Y. & Watanabe, Y. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12753–12758 (2004).

121. Yamaguchi, A. & Abe, M. Regulation of reproductive development by non-coding RNA in Arabidopsis: to flower or not to flower. *J. Plant Res.* **125,** 693–704 (2012).

122. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425,** 415–419 (2003).

123. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).

124. Knight, S. W. & Bass, B. L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293,** 2269–2271 (2001).

125. Ketting, R. F. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev.* **15,** 2654–2659 (2001).

126. Hutvágner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293,** 834–838 (2001).

127. Park, W., Li, J., Song, R., Messing, J. & Chen, X. CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA Metabolism in Arabidopsis thaliana. *Curr. Biol.* **12,** 1484–1495 (2002).

128. Gasciolli, V., Mallory, A. C., Bartel, D. P. & Vaucheret, H. Partially Redundant Functions of Arabidopsis DICER-like Enzymes and a Role for DCL4 in Producing trans-Acting siRNAs. *Curr. Biol.* **15,** 1494–1500 (2005).

129. Friedländer, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26,** 407–415 (2008).

130. Hendrix, D., Levine, M. & Shi, W. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.* **11,** R39 (2010).

131. Schwab, R. *et al.* Specific Effects of MicroRNAs on the Plant Transcriptome. *Dev. Cell* **8,** 517–527 (2005).

132. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **6,** 26 (2011).

133. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Res.* **36,** W70–W74 (2008).

134. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

135. Griffiths-Jones, S., Saini, H. K., Dongen, S. van & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36,** D154–D158 (2008).

136. Griffiths-Jones, S., Grocock, R. J., Dongen, S. van, Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34,** D140–D144 (2006).

137. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** D152–D157 (2011).

138. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–D73 (2014).

139. Krol, J. *et al.* Structural Features of MicroRNA (miRNA) Precursors and Their Relevance to miRNA Biogenesis and Small Interfering RNA/Short Hairpin RNA Design. *J. Biol. Chem.* **279,** 42230–42239 (2004).

140. German, M. A. *et al.* Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* **26,** 941–946 (2008).

141. Gregory, B. D. *et al.* A link between RNA metabolism and silencing affecting Arabidopsis development. *Dev. Cell* **14,** 854–866 (2008).

142. Meyer, S., Temme, C. & Wahle, E. Messenger RNA Turnover in Eukaryotes: Pathways and Enzymes. *Crit. Rev. Biochem. Mol. Biol.* **39,** 197–216 (2004).

143. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10,** 94–108 (2009).

144. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8,** 113–126 (2007).

145. Houseley, J. & Tollervey, D. The Many Pathways of RNA Degradation. *Cell* **136,** 763–776 (2009).

146. Liu, Q. & Paroo, Z. Biochemical Principles of Small RNA Pathways. *Annu. Rev. Biochem.* **79,** 295–319 (2010).

147. Belostotsky, D. A. & Sieburth, L. E. Kill the messenger: mRNA decay and plant development. *Curr. Opin. Plant Biol.* **12,** 96–102 (2009).

148. Kervestin, S. & Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nat. Rev. Mol. Cell Biol.* **13,** 700–712 (2012).

149. Inada, T. Quality control systems for aberrant mRNAs induced by aberrant translation elongation and termination. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1829,** 634–642 (2013).

150. Schweingruber, C., Rufener, S. C., Zünd, D., Yamashita, A. & Mühlemann, O. Nonsense-mediated mRNA decay — Mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1829,** 612–623 (2013).

151. Endres, M., Cook, R. & Gregory, B. in *MicroRNAs in Development* (ed. Dalmay, T.) 209–223 (Humana Press, 2011).

152. Willmann, M. R., Berkowitz, N. D. & Gregory, B. D. Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes--GMUCT 2.0. *Methods San Diego Calif* **67,** 64–73 (2014).

153. Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. The microRNA.org resource: Targets and expression. *Nucleic Acids Res.* **36,** D149–D153 (2008).

154. Devert, A. *et al.* Primer-Dependent and Primer-Independent Initiation of Double Stranded RNA Synthesis by Purified Arabidopsis RNA-Dependent RNA Polymerases RDR2 and RDR6. *PLoS ONE* **10,** (2015).

155. Zong, J., Yao, X., Yin, J., Zhang, D. & Ma, H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447,** 29–39 (2009).

156. Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R. & Zhu, J.-K. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell* **123,** 1279–1291 (2005).

157. Katiyar-Agarwal, S. *et al.* A pathogen-inducible endogenous siRNA in plant immunity. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 18002–18007 (2006).

158. Ron, M., Alandete Saez, M., Eshed Williams, L., Fletcher, J. C. & McCormick, S. Proper regulation of a sperm-specific cis-nat-siRNA is essential for double fertilization in Arabidopsis. *Genes Dev.* **24,** 1010–1021 (2010).

159. Boutet, S. *et al.* Arabidopsis HEN1: a genetic link between endogenous miRNA controlling development and siRNA controlling transgene silencing and virus resistance. *Curr. Biol. CB* **13,** 843–848 (2003).

160. Diaz-Pendon, J. A., Li, F., Li, W.-X. & Ding, S.-W. Suppression of antiviral silencing by cucumber mosaic virus 2b protein in Arabidopsis is associated with drastically

reduced accumulation of three classes of viral small interfering RNAs. *Plant Cell* **19,** 2053–2063 (2007).

161. Donaire, L. *et al.* Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. *J. Virol.* **82,** 5167–5177 (2008).

162. Mourrain, P. *et al.* Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101,** 533–542 (2000).

163. Muangsan, N., Beclin, C., Vaucheret, H. & Robertson, D. Geminivirus VIGS of endogenous genes requires SGS2/SDE1 and SGS3 and defines a new branch in the genetic pathway for silencing in plants. *Plant J. Cell Mol. Biol.* **38,** 1004–1014 (2004).

164. Pandey, S. P., Gaquerel, E., Gase, K. & Baldwin, I. T. RNA-directed RNA polymerase3 from Nicotiana attenuata is required for competitive growth in natural environments. *Plant Physiol.* **147,** 1212–1224 (2008).

165. Qi, X., Bao, F. S. & Xie, Z. Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. *PloS One* **4,** e4971 (2009).

166. Wang, X.-B. *et al.* RNAi-mediated viral immunity requires amplification of virus-derived siRNAs in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 484–489 (2010).

167. Garcia-Ruiz, H. *et al.* Arabidopsis RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during Turnip Mosaic Virus infection. *Plant Cell* **22,** 481–496 (2010).

168. Cuperus, J. T. *et al.* Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat. Struct. Mol. Biol.* **17,** 997–1003 (2010).

169. Morel, J.-B. *et al.* Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. *Plant Cell* **14,** 629–639 (2002).

170. Qu, F., Ye, X. & Morris, T. J. Arabidopsis DRB4, AGO1, AGO7, and RDR6 participate in a DCL4-initiated antiviral RNA silencing pathway negatively regulated by DCL1. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 14732–14737 (2008).

171. Vazquez, F. *et al.* Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol. Cell* **16,** 69–79 (2004).

172. Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121,** 207–221 (2005).

173. Montgomery, T. A. *et al.* AGO1-miR173 complex initiates phased siRNA formation in plants. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 20055–20062 (2008).

174. Yoshikawa, M., Peragine, A., Park, M. Y. & Poethig, R. S. A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. *Genes Dev.* **19,** 2164–2175 (2005).

175. Axtell, M. J., Jan, C., Rajagopalan, R. & Bartel, D. P. A two-hit trigger for siRNA biogenesis in plants. *Cell* **127,** 565–577 (2006).

176. Chen, H.-M. *et al.* 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 15269–15274 (2010).

177. Adenot, X. *et al.* DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Curr. Biol. CB* **16,** 927–932 (2006).

178. Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H. L. & Poethig, R. S. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev.* **18,** 2368–2379 (2004).

179. Fahlgren, N. *et al.* Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. *Curr. Biol. CB* **16,** 939–944 (2006).

180. Garcia, D., Collier, S. A., Byrne, M. E. & Martienssen, R. A. Specification of leaf polarity in Arabidopsis via the trans-acting siRNA pathway. *Curr. Biol. CB* **16,** 933–938 (2006).

181. Li, H. *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* **17,** 2157–2171 (2005).

182. Matsui, A. *et al.* tasiRNA-ARF Pathway Moderates Floral Architecture in Arabidopsis Plants Subjected to Drought Stress. *BioMed Res. Int.* **2014,** (2014).

183. Tantikanjana, T., Rizvi, N., Nasrallah, M. E. & Nasrallah, J. B. A Dual Role for the S-Locus Receptor Kinase in Self-Incompatibility and Pistil Development Revealed by an Arabidopsis rdr6 Mutation. *Plant Cell* **21,** 2642–2654 (2009).

184. Finke, A., Kuhlmann, M. & Mette, M. F. IDN2 has a role downstream of siRNA formation in RNA-directed DNA methylation. *Epigenetics* **7,** 950–960 (2012).

185. Zhang, C.-J. *et al.* IDN2 and its paralogs form a complex required for RNA-directed DNA methylation. *PLoS Genet.* **8,** e1002693 (2012).

186. Zheng, Q. *et al.* Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.* **6,** e1001141 (2010).

187. Lu, C. *et al.* MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res.* **16,** 1276–1288 (2006).

188. Xie, Z. *et al.* Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLoS Biol.* **2,** e104 (2004).

189. Chan, S. W.-L., Henderson, I. R. & Jacobsen, S. E. Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat. Rev. Genet.* **6,** 351–360 (2005).

190. Havecker, E. R. *et al.* The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. *Plant Cell* **22,** 321–334 (2010).

191. Zheng, X., Zhu, J., Kapoor, A. & Zhu, J.-K. Role of Arabidopsis AGO6 in siRNA accumulation, DNA methylation and transcriptional gene silencing. *EMBO J.* **26,** 1691–1701 (2007).

192. Qi, Y. *et al.* Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443,** 1008–1012 (2006).

193. Vrbsky, J. *et al.* siRNA-mediated methylation of Arabidopsis telomeres. *PLoS Genet.* **6,** e1000986 (2010).

194. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat. Genet.* **37,** 761–765 (2005).

195. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).

196. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).

197. Pontier, D. *et al.* Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev.* **19,** 2030–2040 (2005).

198. Olmedo-Monfil, V. *et al.* Control of female gamete formation by a small RNA pathway in Arabidopsis. *Nature* **464,** 628–632 (2010).

199. Jackel, J. N., Storer, J. M., Coursey, T. & Bisaro, D. M. Arabidopsis RNA polymerases IV and V are required to establish H3K9 methylation, but not cytosine methylation, on geminivirus chromatin. *J. Virol.* (2016). doi:10.1128/JVI.00656-16

200. Kanno, T. *et al.* Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. *Curr. Biol. CB* **14,** 801–805 (2004).

201. Wierzbicki, A. T., Haag, J. R. & Pikaard, C. S. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135,** 635–648 (2008).

202. Yang, D.-L. *et al.* Dicer-independent RNA-directed DNA methylation in Arabidopsis. *Cell Res.* **26,** 66–82 (2016).

203. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

204. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinforma. Oxf. Engl.* **31,** 166–169 (2015).

205. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26,** 139–140 (2010).

206. Li, F. *et al.* Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* **24,** 4346–4359 (2012).

207. Montgomery, T. A. *et al.* Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell* **133,** 128–141 (2008).

208. Dotto, M. C. *et al.* Genome-wide analysis of leafbladeless1-regulated and phased small RNAs underscores the importance of the TAS3 ta-siRNA pathway to maize development. *PLoS Genet.* **10,** e1004826 (2014).

209. Marin, E. *et al.* miR390, Arabidopsis TAS3 tasiRNAs, and their AUXIN RESPONSE FACTOR targets define an autoregulatory network quantitatively regulating lateral root growth. *Plant Cell* **22,** 1104–1117 (2010).

210. Willmann, M. R., Berkowitz, N. D. & Gregory, B. D. Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes--GMUCT 2.0. *Methods San Diego Calif* **67,** 64–73 (2014).