Degradome Assisted MicroRNA Prediction in Plants

Salma Yousef A. Alzahrani

School of Computing Sciences

University of East Anglia

This dissertation is submitted for the degree of

Doctor of Philosophy

August 2022

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

To the loving memory of my dear father who inspired me to pursue a PhD

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Vincent Moulton, and my co-supervisors, Professor Tamas Dalmay and Dr. Leighton Folkes, for the invaluable advice, guidance and support that they have given me throughout the course of my studies. I would also like to thank the Saudi Cultural Bureau for the financial support they have provided over the years of my studies.

Special thanks to my children, Hashim and Kadi, and my husband, Mohammad, who gave me strength and joy in the hard moments and reasons to smile at the end of the day. Also, special thanks to my mom and my sister who never let me alone, in spite of the distance. Finally, I would like to thank my friends and my colleagues at UEA for their continued support.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Salma Yousef A. Alzahrani

August 2022

Publications

Salma Alzahrani, Christopher Applegate, David Swarbreck, Tamas Dalmay, Leighton Folkes, and Vincent Moulton. "Degradome Assisted Plant MicroRNA Prediction under Alternative Annotation Criteria." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, early access, Sep. 24, 2021. doi: 10.1109/TCBB.2021.3115023.

Abstract

Ribonucleic acid (RNA) is a polymeric molecule essential in various biological processes. In the past two decades, extensive research effort has been devoted to short non-coding regulatory RNAs called small RNAs (sRNAs). In particular, micro RNAs (miRNAs), 20–22 nucleotide in length, have emerged as an important class of gene regulators. In plants, miRNAs function at post-transcriptional level by suppressing the translation of their target messenger RNAs (mRNAs) through cleavage and degradation, leading to their participation in larger regulatory networks. In recent years, developments in next generation sequencing (NGS) technologies have enabled the large-scale sequencing of sRNAs and cleaved mRNA fragments, called the degradome. Consequently, multiple computational methods have been developed for the identification of miRNAs and their targets.

The advance in regulatory miRNA discoveries relies on understanding their biogenesis and function. Recently, a newly updated plant miRNA biogenesis criteria has been reported, which benefited in identifying more validated miRNAs compared to the old criteria. The new criteria bring the possibility of recommending a further update to the miRNA annotation rules. Moreover, the function of miRNAs is interpreted through their targets that could be determined and validated using degradome. The interactions between miRNAs and their target mRNAs contribute to biological regulatory networks.

In this thesis, we demonstrate a degradome-assisted approach that employs a hill-climbing algorithm to explore miRNAs with extreme biogenesis features in a

controlled manner. We apply this approach on *Arabidopsis thaliana*, evaluate its performance using differential expression analysis, and identify a potentially novel miRNA that has been previously missed by the existing miRNA prediction tools. The approach is presented within PAREfirst tool. Furthermore, we present PAREnet tool that utilises a degradome analysis tool to assist the simplifying, construction, and visualisation of sRNA-mediated regulatory networks on a genome-wide scale. Analysing the constructed simplified sRNA-mRNA network shows the possibility of unraveling the implications of sRNA-mediated regulation in biological processes.

In conclusion, the research focuses on identifying miRNAs, particularly condition specific miRNAs, with unique biogenesis, predicting their targets using degradome analysis, and presenting their interactions by constructing simplified sRNA-mRNA networks with retrievable biological reality. Through these efforts, the study could contributes towards enhancing our understanding of the biogenesis and function of plant miRNA, and the complexity of genes networks in plants.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate ‘take down’ action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Nucleic acids are bio-molecules that play major roles in all cells and viruses. The main two classes of nucleic acid are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) [192]. RNA is synthesised by the DNA in a process called transcription to produce coding RNA, messenger RNA (mRNA), that is responsible for protein synthesis through a process called translation, or non-coding RNA such as small RNA. Small RNAs (sRNAs) are short, non-coding RNA molecules that have been found to regulate the expression of genes which are known to be involved in many diverse plant biological processes such as growth and development, environmental adaptation, disease resistance, and stress response [17, 44, 89]. They mainly regulate the expression of plant genes at the post-transcriptional level by targeting mRNA molecules and silencing them through cleavage and degradation [18, 56, 117, 179]. The sRNAs in turn could then initiate further production of sRNAs to form cascades and networks of sRNA-mRNA interactions [124]. Recent advances in next-generation sequencing (NGS) technologies have made it possible to sequence sRNA datasets on a genome-wide scale from a variety of organisms, tissues, conditions, and developmental stages [58]. In addition, NGS have been used to capture sRNA mediated cleavage fragments, that are resulted from cleaving mRNAs,

using a high-throughput sequencing technique called degradome sequencing, which can then be used to identify functional sRNAs. [72].

Micro RNAs (miRNAs) are probably the most understood class of sRNA and typically have a sequence length in the range of 20-24 nucleotides (nt). They are derived from longer, single stranded hairpin-like structure called precursor. Figure 1.1 illustrates the secondary structure, the mature miRNA, and the miRNA* that is derived from the opposing arm of the hairpin. The understanding of miRNAs and their functions has been a subject undergoing intense study for the last decade [214]. Exciting progress has been made on the biogenesis and functions of miRNA, including a recent study that suggested new criteria for plant miRNA annotation [20], and various studies that identified regulatory cascades that have large effect on plant development, and responses to environmental stress conditions [22, 138, 151]. The work that we have carried out and presented in this thesis is primarily focused on exploring a wider range of computational miRNA prediction parameters in a controlled manner, which allows us to capture miRNAs that were not detected before. Moreover, we present a method that enable the identification and elucidation of potential miRNA cascades represented within a visualised sRNA-mediated regulatory network. Before we proceed, we give a brief overview of each chapter in this thesis.

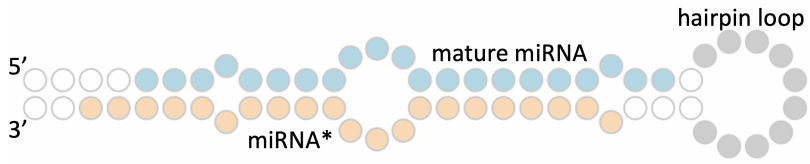


Figure 1.1 An example of miRNA hairpin-like secondary structure.

Chapter 2. In this chapter, we provide biological background information related to RNA silencing in plants. We focus on the most studied sRNA class, miRNAs, along with their biogenesis and functions in plant cells. We also give a brief description of other common sRNA classes, which have similar biogenesis to

miRNA. We then introduce high-throughput sequencing techniques that were used to retrieve sequencing data from different organisms. Finally, we introduce some RNA annotation databases that have greatly facilitated RNA studies.

Chapter 3. In this chapter, we provide a computational background on topics that are important and relevant to later chapters of this thesis. We present an overview on general bioinformatics tools that are necessary to this work. This is followed by a description of the most recent and commonly used computational miRNA prediction tools. In addition, we present a review of the available degradome-supported tools that are used to identify sRNA targets on a genome-wide scale using configurable targeting rules.

Chapter 4. In this chapter, we investigate the effect of applying the newly updated plant miRNA annotation criteria [20], and a more permissive criteria on miRNA prediction in *Arabidopsis thaliana* (*A. thaliana*) using existing miRNA prediction tools. In particular, we use an algorithm that was initially developed by Dr. Christopher Applegate to explore permissive miRNA parameters for miRCat, and we update it so it can be used to produce miRCat2 parameters. We also develop a new approach to miRNA prediction which is assisted by the functional information extracted from the analysis of degradome sequencing. We then demonstrate the improved performance of degradome-assisted miRNA prediction approach compared to the traditional prediction method. We evaluate the approach by applying sRNA differential expression analysis. Moreover, we observe how the miRNA predictions fit under the different criteria and show a potential novel miRNA that have been missed within *Arabidopsis thaliana*. We then present novel mature miRNA and miRNA* along with their predicted precursor. Finally, we introduce a new software tool, called PAREfirst, that can be used to perform degradome-assisted miRNA prediction using configurable parameters for miRCat2 and PAREsnip2.

Chapter 5. In this chapter, we introduce PAREnet, a tool for constructing sRNA-mRNA networks using degradome sequencing data. This tool was developed

to combine sRNA target predictions from PAREsnip2 with related information containing valid miRNA annotations. The output of this tool assist the user when visualising sRNA-mediated regulatory networks though the network visualisation software, Cytoscape. We then compare the *Arabidopsis thaliana* sRNA-mRNA network produced using degradome data alone, with the network produced by applying further filtration methods. Then, we analyse the network components and elucidate the sRNA-mRNA modules presented within the network. We then present how the miRNA candidates from Chapter 4 fit within the network. Finally, we perform transcript coverage analysis to determine if it can provide an element of validation to the target predictions.

Chapter 6. In this final chapter, we present suggestions of future extensions and improvements to this work. We conclude the chapter by presenting the overall conclusions of the work presented in the thesis and its implications for sRNA research.

The overall aim of this project is to explore sRNAome to identify miRNAs with extreme biogenesis and enhance the identification of meaningful sRNA-mRNA networks by incorporating flexible miRNA annotation criteria and utilizing the degradome for validation. The objectives include developing a novel combination method that integrates miRNA predictions using miRCat2 with degradome analysis using PAREsnip2, allowing for the identification of miRNAs with extreme biogenesis while minimizing the rate of false positive predictions that resulted from relaxing the miRNA annotation criteria. Additionally, we employ degradome analysis to simplify the wide-scale complex regulatory networks by reducing false positives in sRNA target predictions, leading to the construction of a simplified sRNA-mRNA network. The constructed network structural features will be analyzed to determine their relevance to other biological networks, and the interactions within the networks will be investigated for their biological significance.

Chapter 2

Small RNA Background

2.1 Summary

This chapter includes an introduction to the RNA biology relevant to the work presented within this thesis. First, it starts with a description of RNA silencing, as it is fundamental to this work, and includes a summary of the biogenesis and functional roles of different types of small RNAs. We then describe an introduction to sequencing techniques and production of sequencing data samples. Finally, we provide a brief account of the currently available small RNA annotation databases.

2.2 DNA and RNA

Nucleic acids are one of the major molecules that are essential for living beings. One of their major roles is gene expression which is a process that uses information from a gene to synthesise functional gene product that enables it to produce protein or non-coding RNA. One main class of nucleic acid is DNA, or deoxyribonucleic acid, which is a long molecule that contains genetic information for the development, functioning, and growth of all organisms. It is composed of two strands of nucleotide coiled around each other forming a double-helix structure. The DNA strands are

made up of four nucleotide bases: cytosine (C), guanine (G), adenine (A), and thymine (T), and are bound together according to complementary Watson-Crick base pairing rules forming C-G and A-T base pairs. As a result of the base pairing rules, DNA can replicate using one strand as a pattern to create a copy of the genetic material. The strands run in opposite directions of each other, these directions are represented by five-prime (5') end and three-prime (3') end carbons. The DNA can be duplicated using one strand and an enzyme called DNA polymerase [192].

Another type of nucleic acid is RNA, or ribonucleic acid. Unlike DNA, RNA is single stranded that is folded onto itself to form secondary structures depending on its required function. It also substitutes the nucleic base uracil (U) with the base thymine (T) (see Figure 2.1). RNA molecules are synthesised from DNA in a process called transcription. The resulting product of transcription is either coding RNA, also known as messenger RNA (mRNA), or non-coding RNA, such as transfer RNA (tRNA), ribosomal RNA (rRNA), and microRNA. The mRNA serves as a pattern for protein synthesis in a process called translation where mRNA is decoded into specific amino acids, tRNA delivers the amino acids to the ribosome, and then rRNA link amino acids together to help decode the information in mRNA into proteins. Basically, non-coding RNAs are not translated into coded proteins. Beside translation, non-coding RNA is involved in RNA processing and other gene regulation roles [57]. According to the length of RNA strands, RNA contains two types: long RNAs and small RNAs [55]. Long RNA strands are longer than 200 nt and mainly include long non-coding RNA (lncRNA) and mRNA. Small RNAs (sRNAs) mainly include microRNA, small interfering RNA (siRNA), small nucleolar RNA (snoRNAs), Piwi-interacting RNA (piwiRNA) [55].

The prediction of RNA secondary structure is an important step towards determining the function of RNA [53]. Several methods of the secondary structure prediction are based upon finding the folding structure with the minimum free energy (MFE) using dynamic programming algorithms [128]. In general, the free energy is

the amount of energy that is released upon folding an RNA sequence. Accordingly, the minimum free structure of nucleotides sequence is predicted by calculating the lowest possible energy needed to fold that structure. Moreover, the folding structure motifs such as: mismatches, hairpins, bulges, and loops are considered to give more understanding and enhance the prediction of the secondary structure [127].

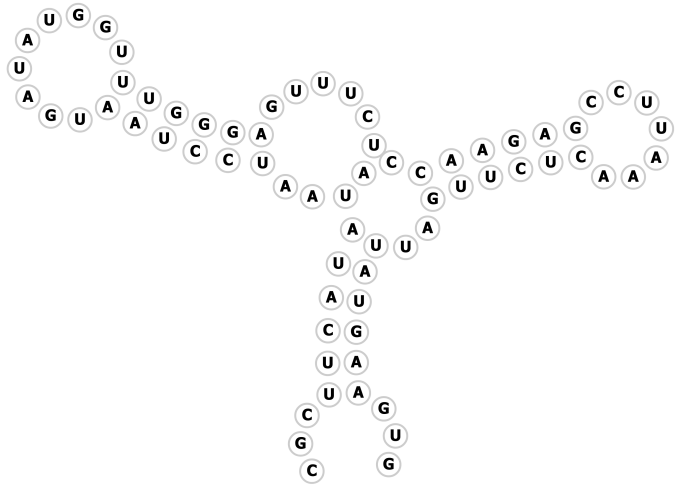


Figure 2.1 An example of an RNA secondary structure. The RNA secondary structure is a stem-loop structure that is formed by the complementary pairing of nucleotide bases (stem) and the non-paring of bases (loop) [129].

2.3 RNA silencing

RNA silencing, or RNA interference, is a mechanism where gene expression is negatively regulated by non-coding RNAs and, it is evolutionary conserved in most of eukaryotes [33]. In 1990, RNA silencing was first described in a plant model, where extra copies of flower pigmentation genes were introduced, which led to endogenous RNA and transgene suppression [139, 182]. There are at least three RNA silencing pathways in plants, the silencing of the mRNA by microRNA, the cytoplasmic siRNA silencing, and the DNA methylation [25]. All of the three pathways result in forming a double-stranded RNA (dsRNA), and this is called the

initiation step. The dsRNA is then cleaved by enzyme Dicer that has RNaseIII and produce 20-26-nt non-coding sRNAs [28, 33].

Moreover, in the effector step, the double-stranded sRNA duplex is then incorporated into an ribonucleoprotein complex, the RNA-induced silencing complex (RISC), by binding to a member of the Argonaute (AGO) protein family [128]. One strand of the sRNA duplex guides the RISC to its target and slices its complementary mRNA, while the rest of the duplex is removed [203]. After the targeting process, the complex can silence the target mRNA either by cleavage and degradation, or translational repression [40]. However, in plants, target degradation is more common due to the high complementarity between sRNAs and mRNAs in plant tissues [128]. The AGO protein that is bound with sRNA usually slices the mRNA between the tenth and eleventh position of the sRNA [117].

2.4 Small RNAs

sRNAs are an important class of non-coding RNAs. sRNAs play important roles in posttranscriptional gene regulation via target messenger RNA (mRNA) degradation or translation inhibition [109, 205], including antiviral defence, developmental timing, and genome adjustment [39, 62, 136]. The double-stranded sRNAs are excised from longer dsRNA by Dicer enzymes. Following that, one of the sRNA's strands is attached to an AGO protein forming a part of RISC complex molecule, which then targets a transcript. There are several classes of sRNAs that have roles in gene expression, such as microRNA, small interference RNA, natural antisense transcript siRNAs (nat-siRNAs) [32], and piwiRNAs [15]. The first two are the well known types and they have similar roles in RNA regulation. However, they are differentiated by their biogenesis, the miRNAs are extracted from a single strand hairpin, while siRNAs are derived from dsRNA [73] (see Figure 2.2).

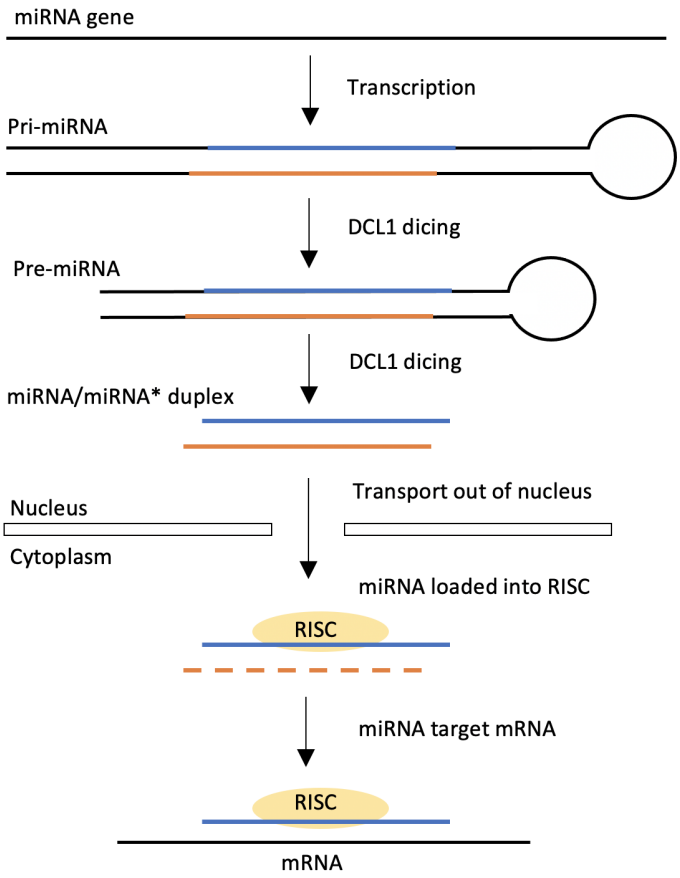


Figure 2.2 An overview of miRNA biogenesis and function in plants.

2.4.1 Micro RNAs

MicroRNAs (miRNAs) are about 20-24 nt small non-coding RNAs that play an important role in the gene expression regulating networks [24]. Recent studies have provided an explosive amount of information on miRNA regulation involvement in various biological processes, including organ development, phase transitions, and stress responses [37, 88, 209]. They were first discovered in 1993 within the developmental timing pathway in *C. elegans*, where instead of the production of the protein, two sRNAs were extracted from the developmental timing regulator. The shorter one with 22 nt, targeted another gene in the developmental timing pathway causing reduction in its protein level [107]. Again, another miRNA was found in *C.*

elegans, which was also discovered in fly and human genes. This discovery led to the investigation of the regulatory role of the miRNAs [153].

The first discovery of the existence of miRNAs in plants was in *Arabidopsis*. The approach that was used to identify miRNAs was the isolation through cloning of sRNAs from RNA biological samples [117]. This approach mostly identified miRNAs that are conserved in several tissues or plants, or with high expression levels [199]. As shown in Figure 2.2, the biogenesis of miRNAs in plants begins with the miRNA encoding gene being transcribed into primary miRNA (pri-miRNA), a long single stranded, by RNA polymerase II [108]. The pri-miRNA is processed by DCL1, a Dicer-like protein, and folds into a stem-loop structure called precursor miRNA (pre-miRNA) [176]. The Dicer cleaves the hairpin loop and generates a double stranded duplex miRNA/miRNA* with two-nucleotide 3' overhangs. The miRNA is the mature sequence, and miRNA* is its complementary, however, the pairing between miRNA and miRNA* is imperfect [154]. Plant miRNA precursors contain structural features that are important for Dicer recognition and precise processing [20, 29, 49]. These biogenesis features include the length of the precursor sequence, mismatches and bulges within the miRNA/miRNA* duplex, and the size of the hairpin loop (see Figure 2.3). After DCL1-mediated processing, the duplex is transported to the cytoplasm where one strand of the miRNA/miRNA* duplex, called guide miRNA, is incorporated with the Argonaute (AGO) protein and loaded into an RNA induced silencing complex (RISC), and the other strand is degraded [145, 184]. The miRNA then guides the RISC complex to target the mRNAs and prevent their translation through cleavage or translation repression [25].

2.4.2 miRNA-like RNAs

Recent work has shown a sRNA species of phased and half-phased miRNA-like RNAs are aligned to miRNA precursors beside the canonical miRNAs [206]. Most

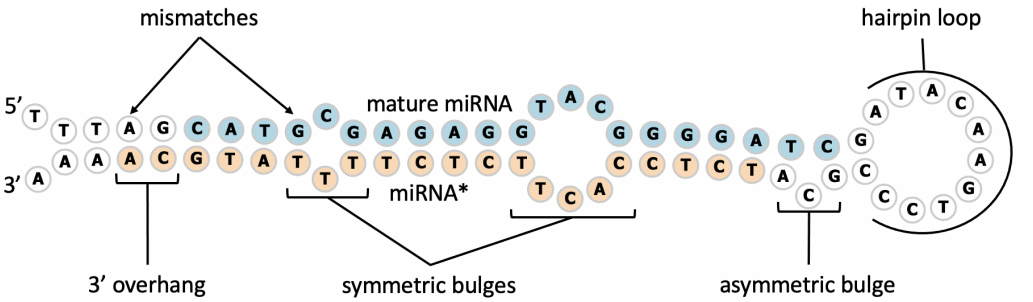


Figure 2.3 An example of miRNA hairpin-structure precursor that shows biogenesis features.

of these miRNA-like RNAs form pairing duplexes with high abundance similar to the miRNA/miRNA* duplexes [211]. It is also found that they have the same biogenesis pathway and functional roles as the known miRNAs. These miRNA-like RNAs were reported earlier in some plant and animal studies, but they were not examined as they were initially considered to be byproducts of Dicer activities. Moreover, some of these miRNA-like RNAs have been found to be real miRNAs as observed with miR477a and miR319b.2 [166].

2.4.3 Small interfering RNA

Small interfering RNAs (siRNAs) are 20-26 nt in length [32]. Their biogenesis is similar to miRNAs except that siRNAs originate from multiple sources and derived from double-stranded RNA sequences [45]. Both siRNAs and miRNAs biogenesis involve Dicer enzyme, which cleaves their precursors into small fragments that are incorporated into RNA-induced silencing complexes (RISCs) to regulate the gene expression. Four types of siRNAs are known in plants. First, trans-acting siRNA (tasiRNA) class, which its precursor derived from microRNA-mediated cleavage of tasiRNA-generating loci (TAS genes). The second type comprises the natural antisense transcript siRNA (nat-siRNA) which normally arises from cis-natural antisense transcripts [32]. The third type of siRNAs includes long siRNA (lsiRNA),

which are normally 30 to 40 nucleotides in length. The last class includes the heterochromatic siRNAs (hc-siRNAs), this class normally arise from transposon and repeat regions of the genome [18].

2.4.4 Other small RNAs

There are several further sRNA classes that will not concern us in this thesis [15, 18, 40]. These classes include:

- **Small nucleolar RNAs (snoRNAs)**: these guide the modifications of RNAs and they are highly involved in RNA nucleotide modification.
- **Small nuclear ribonucleic acids (snRNAs)**: these are involved in transcription, splicing, and formation of precursor mRNAs.
- **Piwi-interacting RNAs (piRNAs)**: these interact with Piwi proteins and silence genes. Recent studies suggest that they protect the genome from invasive transposable elements.

2.5 Next generation sequencing

DNA sequencing is a technology that enables the collection of the specific order of the four nucleotides: cytosine (C), guanine (G), adenine (A), and thymine (T), within the DNA strand [80]. Rapid sequencing has accelerated the research and discovery in a wide variety of research applications as in biological and medical fields [165]. Sequencing technology evolved massively when it became able to process complete DNA genomes from different species. The beginning of the DNA sequencing was in 1953 when Watson and Crick discovered the 3-dimensional structure of DNA [193]. Following that, researchers were able to produce RNA sequences, however, the RNA sequencing techniques failed in sequencing the long DNA molecule [86].

The first known DNA sequencing, or the first-generation sequencing methods, were presented in the 1970's [14], when Maxam and Gilbert developed chemical cleavage method [130], in addition to the chain termination method that was presented by Sanger, Nicklen, and Coulson [160]. These two techniques succeeded in producing complete DNA sequences. The Sanger method was widely adopted in improving the sequencing techniques, due to its reliability and simplicity [147].

In the 2000's, there was focused development on new sequencing techniques that contributed high-throughput sequencing technologies [14]. These techniques are capable of performing a massive amount of sequencing in parallel. This rapid technology is usually called next-generation sequencing (NGS) [126]. There are several platforms for the NGS, and below, we briefly describe some of these technologies:

- **Roche/454 sequencing**, which was presented in 2004 as the first NGS platform to achieve commercial introduction. This technology can generate relatively long reads and uses a picotiter plate to increase sequencing throughput. It applies pyrosequencing approach, which detects the release of pyrophosphate when additional nucleotides are added to a complementary strand of DNA being synthesized from a template sequence, allowing the sequence of the DNA fragment to be determined [141]. The Roche/454 sequencing technology is also called sequencing by synthesis.

- **Illumina/Solexa sequencing technology**, which uses sequencing by synthesis approach. It was released in 2006 and is currently the most common NGS system. The Illumina sequencing technique involves fragmenting the libraries randomly and attaching adaptors to both ends of each fragment. The fragments are then amplified through a process called bridge amplification to form clusters on a flow cell. Next, the sequencing process detects light signals emitted from the addition of a single nucleotide, where computer algorithms translate the signals into a nucleotide sequence. The first Illumina sequencing

machine was able to produce very short reads, roughly 35 nt in length. More recently, the output of the Illumina sequencing machines was improved to increase the read lengths to roughly 100 base pair in length [155]. Although Illumina has different sequencer models for short and long read sequencing, it is best known for short read sequencing. In this thesis, we focus exclusively on data obtained from Illumina sequencing techniques.

- **Oxford Nanopore Technology (ONT)**, which enables direct and real-time sequencing of long DNA or RNA molecules. It is distinguished from previous sequencing approaches, in that it directly detects the nucleotides without active DNA synthesis. It works by monitoring changes to an electrical current as individual nucleotides are passed through a protein nanopore. As the molecule passes through the pore, the unique electrical signal is recorded and used to identify the specific DNA or RNA sequence. This technology allows for a whole-genome sequencing with fast speed and cost-effective performance when compared to other sequencing methods. This long read sequencing technology allows for longer read lengths than other sequencing methods, with some reads extending over tens of thousands of base pair. However, ONT sequencing can have higher error rates with long read sequencing compared to short read sequencing [122].

There are more recent NGS platforms that were not described here, such as: Applied Biosystems sequencing by oligonucleotide ligation and detection (SOLiD) [185], Heliscope sequencer [178], and Pacific Biosciences SMRT technology [132]. Overall, the choice of sequencing method will depend on the specific research question and available resources. In the case of sRNA sequencing, short read length sequencing methods provide more comprehensive and accurate data than long read sequencing. Although, short read sequencing methods suffer from a relatively long sequencing time, the recent sequencer models, such as the latest Illumina models

(NextSeq 1000 and NextSeq 2000), reduce the run time while maintaining the accuracy and quality of the produced data [96, 84].

2.6 Annotation databases

There are various databases that are used for sRNA annotation such as miRBase [97], Rfam [75, 91], PMRD [212], and PMiREN [78]. Here, we describe the ones that are most important to this work:

- **miRBase repository**: is an online repository for all validated miRNAs with their assigned annotations. It was first established in 2002 as MiRNA Registry service [74, 76]. The primary aim of miRBase is assigning consistent names for novel miRNA sequences prior to their publications. In the nomenclature scheme, the name of a miRNA sequence should contain a species initial prefix and numeric suffix that is assigned sequentially, e.g., for *Arabidopsis thaliana*, the miRNA named ath-mir-166 [77]. Moreover, miRBase is searchable and facilitates bulk download of the published and annotated miRNAs. The miRNA annotations are human-readable and computer-parsable. It also provides a link to the evidence and resources that support the miRNA annotations [98]. Another function is the miRBase Target, a new database for predicted target genes for the miRNAs, although, not all validated miRNAs have predicted targets [76] due to a variety of challenges in the miRNA target prediction methods. Although computational methods have been developed to predict miRNA targets, these predictions are not always accurate due to different factors, including the complexity of miRNA-mRNA interactions and the presence of multiple potential target sites for each miRNA. Another factor is the tissue-specific and species-specific nature of miRNA regulation. Some miRNAs may only be expressed in specific tissues or organism, and their targets may be correspondingly restricted. This can make it difficult to predict targets

for miRNAs that have not been extensively studied in a particular tissue or organism.

Since its establishment, miRBase has rapidly and continuously grown, mainly due to the sRNA NGS technologies. The annotated miRNA loci increased from 15,172 loci in 142 species (release 16, 2010) to 24,521 loci in 206 species (release 20, 2013) [99], and later in 2014, the loci in release 21 increased to 28,645 entries, and lastly, to 38,589 entries in release 22 (v22), 2019 [97]. With the rapid growth in miRBase database, the quality of the miRNA sequences must be maintained. Hence, a new system was developed where it uses NGS sequencing datasets to annotate the levels of confidence for the miRNAs by assessing the pattern of reads that map to the locus of each miRNA annotation. The user can download miRNA sequences after filtering based on their confidence levels. miRBase is commonly used as a guide for researchers to predict novel miRNAs and assess the performance of miRNA prediction tools. Baohong Zhang et al. [207] investigated the miRNAs in miRBase and they observed that the mature sequences range from 19 to 24 nt in length, where the vast majority were 22-23 nt, however, 1.5% of miRNAs were outside that range [208].

- **Rfam**: is a database containing a collection of ncRNA families represented by multiple sequence alignments, covariance models, and other structured RNA elements, which include secondary and tertiary structures of RNA molecules, stem-loop or hairpin structures, hairpin loops, internal loops, and hairpin bulges. These structured RNA elements are crucial for the function of ncRNAs and they can facilitate the classification and identification of ncRNA classes in Rfam database. The current release is Rfam 14.7 containing 4,069 ncRNA families. Rfam is collaborating with miRBase to provide a comprehensive collection of miRNA families, where 1,246 miRNA families were created and

updated in Rfam. The ncRNA sequences within Rfam can be used to filter out known ncRNAs before identifying novel miRNAs.

2.7 Discussion

In this chapter, we provided an overview of DNA and RNA, and introduced RNA silencing pathways and secondary structures. We then discussed miRNAs, together with their biogenesis features and functioning mechanisms and introduced some other common sRNA classes. We then gave a brief background of next generation sequencing technologies. Finally, we described one of the most popular miRNA annotation databases.

Chapter 3

Computational Background

3.1 Summary

This chapter includes an introduction to the bioinformatics techniques relevant to the work presented within this thesis. We begin with a description of several tools and algorithms for processing DNA and RNA sequences. We then overview the miRNA detection methods that we use in this thesis. Finally, we provide a brief overview of sRNA target validation and introduce some sRNA target prediction tools that are important for this thesis.

3.2 Bioinformatics tools

We first briefly describe the bioinformatics tools that we shall use later in this thesis:

- **PaTMaN**: Pattern Matching in Nucleotide databases, is an alignment tool that identifies all occurrences for a short sequence within a genome-sized reference. To begin, the algorithm constructs a tree of all the query sequences. Each short read is included in the tree as a path starting from the root node and ending at a leaf node that carries the identifier for the corresponding query sequence. It

then searches the tree for matches between the reads and the reference, taking into account any mismatches or gaps in the alignment [148].

- **Bowtie2**: is an alignment tool for aligning short sequencing reads to a large genome reference. It first generates an index of the reference sequences using a Burrows-Wheeler transform, which creates a compressed representation of the reference genome. The short reads are then aligned to the index using a series of alignment steps, taking into account any mismatches or gaps in the alignment. The indexing technique used in Bowtie is the key to its speed and memory efficiency. Bowtie2 is an improved version that can process longer reads faster and more sensitive than Bowtie [103, 104].

- **SAMtools**: is a suite of programs that is widely used for processing and analysis NGS datasets. It also used for file format conversion and other file manipulation methods. It consists of three repositories: SAMtools, BCFtools, and HTSlib [51].

- **Genome browsers**: a genome browser is a graphical interface to display the graphical information of a biological database for genomic data. Among the best known are the Ensembl Genome Browser [82], UCSC Genome Browser [106], and NCBI Genome Data Viewer [195].

- **Basic Local Alignment Search Tool (Blast)**: is the most widely used tool for biological sequences comparing and searching [9].

- **Vienna RNA Package**: provides RNA secondary structure related computational tools. There are several tools that are frequently used for miRNA detection. RNAfold computes a minimal MFE secondary structure for an RNA sequence. In addition, RNALfold calculates all locally stable secondary structures of a long RNA sequence with a maximal base pair span. It is a practical way of scanning very large genomes for short RNA structures. Moreover,

RNAplot can be used to generate secondary structure graph of an input RNA sequence. the input format is produced by RNAfold program. Forna is one of the Vienna Web Services, it is a web interface that is used to visualise RNA secondary structure [120].

3.3 Computational miRNA prediction

Due to the importance of miRNA in gene expression regulation, its detection has become a major research area over the last decade [194]. The beginning of identifying the miRNAs in plants was in *Arabidopsis* using a powerful strategy, where the sRNAs from the biological samples were isolated and cloned [101, 117]. However, this strategy was not efficient for identifying miRNAs with low expression levels, neither for miRNAs that were not present in different tissues [199]. The emergence of NGS technologies produced a massive amount of data with high speed and low cost [165]. The hairpin structure of a miRNA precursor along with the biogenesis features that are shown in Figure 2.3 are key components of miRNA identification algorithms. Many computational tools were developed to detect miRNAs at transcriptome level, especially the ones with low-abundance [110]. However, these computational methods still suffer from high false positives, and many functional miRNAs are missed in the prediction [92]. A review of miRNA prediction tools can be found in [143]. We now describe two tools that are used in this thesis.

3.3.1 miRDeep

miRDeep was introduced in 2008 as one of the first miRNA prediction tools for animals NGS datasets [64]. miRDeep algorithm's starts by mapping the sequenced sRNAs to the genome and discards reads that map to other types of non-coding RNAs, such as tRNA and rRNA. The remaining reads are then used to identify the

secondary structure of potential miRNA precursors. Next, the algorithm examines the hairpin structure and the aligning of sRNA sequences for each potential precursor. This is performed as follow: for each precursor, the algorithm identifies the positions and the abundance of the sequences corresponding to a mature miRNA, a miRNA*, and a loop sequence within the hairpin, where the statistics of the reads positions and abundance within the precursor is referred to as the signature. The mature miRNA is defined as the sequence with the highest abundance, as it is sequenced more frequently , in the sRNA libraries, than other sequences within the precursor. The miRNA* is defined as the sequence aligned to the opposite arm of the hairpin from the mature miRNA with 2-nt 3' overhangs. The loop region is defined as the sequence between the mature miRNA and the miRNA*. After identifying the miRNA-loop-miRNA* hairpin structure, miRDeep ensure its reliability by requiring at least 14 nt pairings between the mature and star miRNAs. The potential precursors that did not pass these filters are considered inconsistent with miRNA biogenesis and are discarded. After that, it applies a probabilistic scoring on the candidate miRNA precursors by computing a probabilistic score for the frequencies of reads, positioning in correspondence with Dicer processing, and other features that contribute to the score. The results of miRDeep is scored potential miRNAs with their precursors, beside an estimation of the false positive rate of the results. An improved version, miRDeep2 [65], added more features and packages, and it provides an option to annotate known miRNAs if miRBase files for mature miRNAs and their precursors was provided. It also uses RNAfold to generate secondary structures of the miRNAs by calculating the minimum free energy (MFE). In addition, it gives the option for the user to compute the MFE of miRNA precursor using RANDfold [31]. There are other versions of miRDeep: miRDeep* [12], miRDeep-P [201], miRPlant [13], and miRDeep-P2 [100].

miRDeep* is an improved version of miRDeep and has a graphical user interface that integrates third party computational tools, such as genome alignment and RNA

secondary structure prediction, into a Java library. miRDeep-P was developed to identify plant miRNA by modifying miRDeep algorithm to adapt miRNA biogenesis in plants. Similar to miRDeep*, miRPlant was developed to extend miRDeep-P by providing a user-friendly interface that does not require any pre-installed computational tools. miRPlant dynamically plots miRNA hairpin structure with small reads for identified novel miRNAs. miRDeep-P2 (miRDP2) is an updated version of miRDeep-P, which was improved by employing a new filtering strategy and overhauling the algorithm. miRDP2 was shown to have better speed when tested on miRNA transcriptomes in plants with increasing genome sizes that included *Arabidopsis thaliana*, *Oryza sativa* (rice), *Solanum lycopersicum* (tomato), *Zea mays* (maize), and *Triticum aestivum* (wheat). By incorporating the newly updated plant miRNA annotation criteria [20] and developing a new scoring system, the accuracy of miRDP2 outperformed other programs [100].

3.3.2 miRCat and miRCat2

In 2012, Stocks et al. [170] introduced the UEA small RNA Workbench suite of tools that analyse and visualise NGS sRNA data. The suite was a successor to the UEA Small RNA Workbench web-based toolkit that was launched in 2008 [137]. The Java-based Workbench tools are interactive and user-friendly and were developed to provide more features comparing to the toolkit. One of the analysis tools is miRCat, which is a tool that uses plants or animals NGS datasets to predict miRNA precursor. miRCat searches for miRNA loci on the genome based on the locus criteria that were defined in previous study [184]. In brief, miRCat attempts to identify the loci that have two peaks of sRNA reads that map to one strand of the gene. The mature miRNA and miRNA* are expected to have distinct expression levels, resulting in two peaks in the sRNA read distribution on the same genomic strand. miRCat uses PatMaN [148] to map the sRNAs to the genome, in addition to RNAfold [120] and

RANDFOLD [31] programs, which investigate the secondary structure for each locus strand.

Unfortunately, most of the implemented prediction methods, including miRCat, suffer from high false positive rate. Therefore, a new approach called miRCat2 was introduced [143]. Although it is based on its predecessor, it incorporates several new features and improvements to increase the accuracy of miRNA identification. Some of the key differences between miRCat and miRCat2 is that the latter is optimised for large-scale genome analysis. For predicting secondary structures of candidate miRNA precursors, miRCat2 incorporates RNALfold algorithm [120], a modified version of RNAfold. It also incorporates a selection of miRDeep2 features, such as the scoring system for candidate precursors, that were briefly described in the previous subsection. As an input, miRCat2 requires at least one FASTA file containing sRNA sequences , in addition to a genome file for mapping the sequences with PatMaN. The tool generates results as a table for miRNA prediction candidates with their details. It has a user-friendly interface that allows the user to interact with the results by doing further analysis or visualizing the secondary structure for each candidate. Moreover, the user can export the candidates table, the analysis reports, and the secondary structure results.

3.4 sRNA target prediction and validation

sRNAs are involved in post-transcriptional gene regulation by binding and silencing specific mRNAs causing their degradation or inhibiting their translation. By identifying sRNA targets, researchers can gain insights of the sRNA-mRNA complex regulatory networks and how interactions between sRNAs and mRNAs contribute to the regulation of various biological processes such as cells development, organisms growth, and resistance to disease or stress. In plants, there is a high sequence complementarity between sRNAs and mRNA sequences [157]. Consequently, com-

putational tools were developed to predict plant sRNA target by using techniques that search for complementarity between a sRNA sequence and a potential target sequence [168, 204]. The majority of these tools use stringent targeting rules inferred from experimental observations. These rules are implemented within a position dependent scoring system based on the number of mismatches and target-bulged bases within the duplex. To the best of our knowledge, there are two sets of targeting rules for plants. The first set of rules was performed by Allen et al. [6] on a set of 94 validated sRNA target duplexes in *A. thaliana*. The second set of rules was performed by Fahlgren and Carrington on 155 validated target duplexes using a similar approach to the former rules, yet, a mismatch or G:U wobble at position 10 or 11 of the sRNA is permitted [59]. The difference between the two sets of rules is that the Fahlgren and Carrington rules permit a mismatch or G:U wobble at position 10 or 11 of the sRNA. Several computational tools and web servers are available for predicting plant sRNA targets including: psRNATarget [50], TargetFinder [60], and TAPIR [30]. However, these tools suffer from the rate of false positive predictions, thus, further experimental validation is required.

As mentioned above, sRNA targets a mRNA transcript and silences it at a post-transcriptional level through cleavage and degradation. The cleavage usually occurs between positions 10 and 11 of the sRNA [117]. The examination of mRNA cleavage fragments is an important step for sRNA targets validation. One method that is used to validate the putative targets of sRNAs is 5' rapid amplification of cDNA ends (RACE), which is used to identify cleavage products for a particular mRNA [66]. This method is time consuming as it needs to be performed for every predicted cleavage site on each mRNA. The advances of NGS techniques led to developing a new approach called Parallel Analysis of RNA Ends (PARE) protocol, or degradome sequencing, that is used to identify mRNA degradation products on a genome-wide scale [72]. The protocol is a modified 5'RACE combined with high-throughput sequencing methods. After mRNA cleavage, the downstream sequence

of the cleavage site, unlike the upstream sequence, does not degrades. The remaining mRNA sequences are not capped with an altered nucleotide known as a 5' cap. The PARE protocol selectively clones the 5' uncapped mRNA fragments, and following that, these fragments are subject to deep sequencing. The mRNA degraded fragments obtained using this method are called the degradome. When aligned to mRNA transcript, degradome sequences can provide evidence for a sRNA-mRNA interaction by showing clear peaks at the cleavage site that is corresponding to the targeting site of a sRNA. As a consequence, computational tools have been developed to use sRNA and degradome datasets to identify the interactions between sRNA and mRNA sequences. CleaveLand [2] was the first tool to analyse degradome data, other common tools in order of their first appearance, are SeqTar [213], PAREsnip [63], sPARTA [90], and PAREsnip2 [177]. We now review the degradome analysis tools that are important to this thesis.

3.4.1 CleaveLand

CleaveLand [2] was the first tool developed specifically to analyse degradome data, and it has been successfully used to identify sRNA targets in a variety of organisms. The first stage of the CleaveLand algorithm is to align the degradome data to the reference transcriptome using Bowtie. The alignments between mRNA and degradome fragments are processed to quantify the strength of the cleavage signal at each alignment site using a category system. Specifically, a category system is defined as follows where Category 0 interaction have the highest confidence:

- **Category 0:** There is more than one read at the cleavage site, the abundance is the maximum on the transcript, and there is only one maximum.
- **Category 1:** There is more than one read at the cleavage site, the abundance is the maximum on the transcript, and there is more than one maximum, i.e. there are two or more targeting signals with the same abundance.
- **Category 2:** There is more than one read at the cleavage site, the abundance is above the average fragment abundance but less than the maximum.
- **Category 3:** There is more than one read at the cleavage site, the abundance is less or equal to the average fragment abundance.
- **Category 4:** There is only one read at the cleavage site.

After that, CleaveLand reports potential target sites of a given sRNA with a Perl script called GStar that is a wrapper and parser for RNAplex [175]. Next, the resulted potential target sites are combined with those from the degradome read alignment stage to identify any matches opposite the position 10 of the sRNA. The selected sRNA-mRNA target interactions are then processed by calculating a p-value, and interactions that pass the p-value filter are reported.

3.4.2 PAREsnip

PAREsnip was introduced as the first tool that was able to perform a degradome analysis for the enormous sRNA datasets in a reasonable time [63]. It was introduced as a part of the UEA sRNA Workbench toolkit mentioned above [170]. For performing degradome analysis on a given organism, PAREsnip takes the following input files: sRNAome (sRNA dataset), degradome,

and a reference transcriptome, where the transcriptome contains the set of all cDNA sequences including coding sequences of mRNA and non-coding untranslated

regions (UTRs). A transcriptome library can be constructed using the transcriptomics technique, RNA sequencing (RNA-Seq) [102, 191]. Moreover, PAREsnip accepts a reference genome as an optional input. The tool uses PatMaN tool for sequence alignment if genome file is provided, where it only keeps the sRNA sequences that map to the genome. It also filters out the sequences that do not match to a user-configurable parameters such as sequence abundance and sequence complexity.

The first step of the PAREsnip pipeline is categorizing the potential cleavage positions based on the degraded fragments abundance on each site on the transcript. The five categories system that were implemented in CleaveLand was adopted for doing this. The PAREsnip algorithm constructs a 4-way search tree structure since there are four alphabets for the sRNA sequences (A, T, C, and G). This data structure enables PAREsnip to map the sequences faster and improved its computation rate. To search the encoded tree, the tool uses Allen et al. targeting rules [6]. The output file from PAREsnip is a list of a detailed interactions between sRNAs and their target mRNAs.

3.4.3 PAREsnip2

PAREsnip2 employs a search algorithm and sequence encoding technique to process the genome-wide scale datasets [177]. It shows a vast reduction in computation time and resource requirement compared to the previously implemented degradome analysis tools. The tool accepts the following input files: one or more sRNAome replicates, one or more degradome replicates, transcriptome (FASTA or GFF3 format), and a genome (optional unless using GFF3 as transcriptome). PAREsnip2 provides user-configurable targeting rules where the user can choose between the default Allen et al. rules or the Fahlgren and Carrington rules described above, this feature enables the users to search for non-canonical targets that would be missed by the existing targeting rules [34, 90]. Prior to the analysis, the tool provides optional

filtering techniques that can be used to discard the following sequences: sequences with ambiguous bases, low complexity sequences, sequences that are not conserved among replicates, and sRNA sequences that do not align to the genome.

A core part of PAREsnip2 algorithm is the binary encoding of sequence input into a number system by representing each nucleotide base (see Chapter 2) using two bits of computer memory. This process reduces the computation time and memory required to perform the analysis. The next stage of PAREsnip2 algorithm is the generation of target-sequence candidates by aligning the degradome fragments to the transcriptome using the binary encoding. The generated target candidates are then sorted into a 5-category system that is similar to the system defined in CleaveLand. A three-stage candidate filtering technique was developed to reduce the target-sequence candidates. Each target candidate that pass the three-stage filtering method is then aligned to the sRNA sequence by using the pre-chosen targeting rules. Once a potential target has been identified, two optional filtering methods, a minimum free energy (MFE) ratio filter and a *p*-value filter, can be performed to improve the confidence level of each target prediction.

The results of PAREsnip2 are provided in comma separated value (CSV) format where they include information about the sRNA-mRNA interaction, such as the category of the interaction and the cleavage position. It is also possible to produce target plots (t-plots) from PAREsnip2 results using the T-plot tool contained within the UEA Small RNA Workbench. A t-plot shows the degradation activity for a transcript where the cleavage site could be highlighted. The x axis gives nt positions along the transcript. The y axis gives the abundance of cleavage fragments.

3.5 Discussion

In this chapter, we have introduced several bioinformatics tools that are related to miRNA prediction methods. We discussed various computational tools used to identify different classes of sRNA from NGS data. We then introduced methods for predicting plant sRNA targets. Also, we introduced PARE, an NGS technique that identifies the mRNA cleavage products, degradome. Finally, we discussed several tools that apply the degradome analysis to predict sRNAs targets. In the next chapter, we will present a new method for identifying miRNAs using the degradome analysis, which will use some of these tools.

Chapter 4

Degradome Assisted miRNA Prediction

4.1 Summary

In this chapter, we present a new approach for miRNA detection based on the functional information provided by degradome. This work [11] was published in IEEE/ACM Transactions on Computational Biology and Bioinformatics and this chapter is an adapted version of that article. The parameter-search algorithm was initially developed by CA for miRCat, and SA modified it to comply with miRcat2. The parameter-search experiments were carried out by SA for miRCat2, and by CA for miRCat. CA, DS, TD, LF and VM analysed and interpreted the data produced using miRCat and PAREsnip. SA, TD, LF and VM analysed and interpreted the data produced using miRCat2 and PAREsnip2. SA took the lead in writing the chapter with input and critical feedback from LF and VM. In this chapter, we start by introducing the shortcomings of the current miRNA prediction methods, followed by recently suggested miRNA annotation criteria. We then investigate the effect of using more permissive parameters for miRNA prediction, in particular, we develop an algorithm to explore and evaluate different parameter combinations. Moreover,

we present a new combination approach to miRNA prediction, which uses the functional information extracted from a genome-wide degradome-assisted sRNA target analysis.

4.2 Background

As mentioned in Chapter 2, miRNAs are a class of non-coding sRNA that typically have a sequence length in the range of 20 to 24 nt [24]. The defining feature of a miRNA is the precise excision of a duplex from an RNA hairpin precursor structure by a Dicer like-enzyme. The duplex contains both a mature miRNA and a miRNA* sequences with a 2 nt overhang at the 3‘ ends. However, a pairing between a mature miRNA and a miRNA* within a duplex is often imperfect, including variation in the number of nt mismatches, bulges, and the number of nt within a bulge [154, 207]. The full-length precursor of a miRNA also exhibits variation in features such as its stem-loop folding composition as well as its length [49]. Such variability of features within both a miRNA stem-loop precursor and a mature miRNA duplex in plants can present a challenge for the accurate computational annotation of miRNAs within a genome-wide sRNA profile and in particular the correct attribution of sRNAs to the class of miRNAs [133].

In plants, it is typical for a high degree of complementarity between the sRNA and its target mRNA, often resulting in its translational silencing through cleavage and degradation [44, 117]. As discussed in Chapter 3, in recent years, NGS technologies have been used to capture an organism’s entire sRNAome profile in a single experiment on a genome-wide scale. Such a profile contains many classes of sRNA which are grouped based on their biogenesis and function.

4.2.1 Shortcomings of current miRNA prediction tools

The methods used within most tools for predicting plant miRNAs from the sRNAome, such as miRPlant, miRCat2, and miRDeep-P as discussed in Chapter 3, use dated annotation criteria from 2008 which employ a set of suggested miRNA biogenesis features. These features are comprised of a set of stringent criteria that have been used to model a miRNA. However, these criteria were published over a decade ago [133] and do not describe, or account for, a growing number of validated miRNAs that follow a model composed of a less stringent criteria. Therefore, the currently available tools risk discarding bona fide miRNAs. Moreover, in spite of using a stringent biogenesis model underpinning their prediction algorithms, some of these tools still tend to generate a large number of false positive predictions [76].

4.2.2 miRNA annotation criteria

Recently a new set of miRNA annotation criteria has been reported [20], suggesting that more flexibility is required in several of the criterion of the miRNA annotation model in order to identify validated miRNAs that were previously missed using the former criteria. The newly suggested model also applies some restrictions on the length of the miRNA, miRNA* and precursor. It also requires biological replication of the sRNA profile and suggests that further experimental validation beyond NGS of the sRNA profile is not required. In addition, the authors have suggested that their updates could contribute to the reduction of false positives. Even so, a more flexible choice of a less stringent set of parameters, e.g. allowing more mismatches within a duplex and increasing the size of gaps within a duplex, is likely to result in capturing more validated and verifiable miRNAs, and therefore to increase the total number of miRNA predictions overall.

We hypothesised that using more flexible miRNA annotation criteria could identify miRNAs with extreme biogenesis that would be missed using the current

criteria. However, allowing this flexibility in computational miRNA prediction methods may run the risk of increasing the rate at which false positives are predicted. To overcome this, we proposed utilising the degradome, which can be used to validate sRNA targets (see Chapter 3), in order to minimize the identification of false positives in the miRNA predictions. The degradome is useful to identify miRNA mediated cleavage [144], and the defined category system within degradome analysis tools, which were discussed in Chapter 3, ranks the confidence level of miRNA-mRNA interactions, and hence, it could support the miRNA identification.

In order to test our hypothesis, we introduced a novel combination method to predict miRNAs using miRCat2 by allowing some flexibility in the miRNA annotation criteria while utilising targeting information obtained from degradome analysis using PAREsnip2. In particular, our new approach using degradome data helps in the sRNA annotation effort in several ways. Firstly, by conceptually reducing the number of sRNA candidates to those that are potentially functional and cleavage capable. Secondly, the use of less-stringent miRNA secondary structure prediction parameters for miRNA candidates within the functional sRNA subset becomes feasible when modulating by their function. And thirdly, the predicted miRNA mediated cleavage signal and biogenesis information can be examined simultaneously to derive a final consensus miRNA candidate set that can be computationally filtered and ranked by confidence information for further experimental validation. Below we shall demonstrate that even though a greater number of candidate miRNAs tend to be generated with more flexible parameters, our combination method is able to reduce this number by employing degradome information. Our combination approach is made freely available in user-friendly software called 'PAREfirst' [11] that can be downloaded from: https://github.com/sRNAworkbenchuea/UEA_sRNA_Workbench.

4.3 Methodology

We begin by describing our approach to investigate the effect of allowing more permissive parameters on miRNA prediction. This approach is novel in that it explores the effect of more permissive parameters on miRNA prediction and degradome analysis. The method uses degradome analysis for miRNA prediction, which is a departure from the conventional use of degradome analysis for target prediction. While PAREsnip2 has been previously used for degradome analysis to identify sRNA targets, this method utilizes functional information obtained from degradome analysis to computationally filter and rank candidate miRNAs. Our approach uses the miRCat2 [143] tool for miRNA prediction and the PAREsnip2 [177] tool for degradome analysis. Recall that, in brief, miRCat2 is a miRNA prediction method that uses an entropy-based approach to detect miRNAs within a genome. As inputs, the method requires a reference genome and sRNAome. The method first identifies potential miRNA candidates based on sRNA abundance and then applies a number of filters such as mapping locus, size class distribution and miRNA-like alignment patterns on the candidates before calculating their miRNA secondary structures. The method outputs miRNA predictions in a tabular format and was selected for its improved accuracy when compared to similar tools.

PAREsnip2 is a degradome analysis method that can be used to identify sRNA targets. As discussed in Chapter 3, PAREsnip2 requires sRNAome, degradome, and transcriptome. The method first performs several optional quality filtering steps on the input sequences. The method's algorithm then encodes input sequences into a decimal number which is then used to make exact match sequence alignments and subsequently identify potential sRNA-target pairs. The tabular output contains information on the sRNAs and their potential target sites along with abundance and degradome assisted confidence metrics. PAREsnip2 was selected because it

shows a more efficient performance in term of computation time and resources, such as memory usage, compared to previously developed target prediction tools. The tool enables users to perform degradome analysis on large-scale datasets using configurable targeting rules. Both of miRCat2 and PAREsnip2 are implemented in the UEA sRNA workbench, and have the advantage that they can be easily configured and have been shown to perform comparatively well compared with other tools [169].

First, we implemented a parameter-search algorithm that is described in detail in the next subsection to produce a collection of roughly 150 exploratory parameter sets (denoted EPS), see Appendix A Table 1 for the complete list of the EPS. We then produced an updated miRCat2 parameter set (denoted UPS) based on the new criteria presented in [20]. For comparison, we present the main criteria for EPS and UPS in Table 4.1, together with the default miRCat2 parameter set (denoted DPS). Then, for each wild-type sRNA sample we obtained three sets of miRNA predictions using the miRCat2 tool with the DPS, UPS, and EPS.

Next, we performed a target analysis with PAREsnip2 and used its outcome to control the false positive miRNA predictions that could result from relaxing the biogenesis parameters without losing the majority of the validated miRNAs. The analysis was performed using the wild-type sRNA replicates, degradome replicates, the transcriptome, and the genome. In addition, we used Fahlgren and Carrington [59] targeting rules, allowed categories 0-3, disabled MFE and p-value filters, sRNA length from 18-25 nt, and disabled the core region multiplier. These more permissive parameters were used to capture validated sRNA-target interactions that would have been missed using the default settings [35].

Moreover, we used Dicer-like1 enzyme (DCL1) mutant sRNA data sets to validate the predicted functional miRNAs. In particular, we performed a differential expression (DE) analysis using DESeq2 [121] within iDEP9 [71] between the three wild-type sRNA replicates and the three DCL1-mutant sRNA replicates. The predicted miRNAs that were two-fold down-regulated in the DCL1-mutant were con-

miRCat2 parameter	DPS value	UPS value	EPS ranges
min_length (a)	20	20	18, 19, 20
max_length (b)	23	24	23, 24, 25, 26
min_fold_len (c)	45	40	40, 45
max_fold_len (d)	250	300	250, 300, . . . , 400
max_amfe (e)	-22	-22	-32, -27, . . . , -2
Complex (f)	0.90	0.90	0.50, 0.60, . . . , 0.90
clear_cut_perc (g)	0.92	0.90	0.52, 0.62, . . . , 0.92
gaps_mirna (h)	4	5	4, 5, . . . , 8
no_loop (i)	3	3	3, 4, . . . , 7
Repeats (j)	25	25	25, 30, . . . , 40
p-val (k)	0.05	0.05	0.05
RANDfold (l)	false	false	false
complex_loop (m)	true	true	true

Table 4.1 The miRCat2 parameters for DPS, UPS, and the ranges for EPS parameters. Parameters are labelled as follows: (a) minimum length of miRNA, (b) the maximum length of miRNA, (c) minimum length of precursor, (d) maximum length of precursor, (e) maximum value for the adjusted MFE for a miRNA precursor, (f) complexity of sequence, (g) percent of incident reads that should fall between the same start and end positions as the miRNA, (h) maximum number of consecutive gaps on the precursor on the miRNA location, (i) Maximum number of bulges in the loop area of the precursor, (j) maximum number of times a sRNA can map to a genome, (k) RANDfold computation, (l) threshold for the RANDfold value, and (m) if a precursor with multiple loops between miRNA and miRNA* is allowed.

sidered as enriched candidate miRNAs. Additionally, we performed a similar DE analysis between the wild-type replicates and DCL4-mutant triplicates. We show the results of DE analysis for WT_DCL1 in Appendix A Table 2, and for WT_DCL4 in Appendix A Table 3.

To further investigate the enriched miRNA candidates, we aligned the candidates to all the plant species miRNAs that were retrieved from miRBase using PatMaN [148], allowing one mismatch to allow for isomiRs. We also discarded the candidates that align to other RNA classes such as tRNAs, rRNAs, snoRNAs and snRNAs that were described in Chapter 2.

4.3.1 Generating permissive miRCat2 parameter sets

To generate exploratory parameter sets (EPS), we designed an iterative local parameter-search tool that uses a hill-climbing algorithm [158] to explore more permissive EPS combinations. More specifically, ranges were set for each configurable miRCat2 parameter (Table 4.1), which were provided to the algorithm as a technique to reduce the size of our parameter-space. The algorithm starts with a random selection of parameters within the given ranges. For each iteration, a check is made to each neighbouring EPS in which each parameter value is incremented or decremented. Predictions using miRCat2 were then made for each of the neighbouring EPS and the algorithm chooses a new EPS based on the score function described below. This process is repeated until there is no further improvement on the score and the highest scoring EPS is retained, a diagram illustrating this series of steps is presented in Figure 4.1. The algorithm was performed 100 times, using a randomly selected starting EPS for each run, of which we selected the highest scoring 50 EPS. The search was applied on each wild-type sRNA replicate, and we combined the top scoring 50 EPS to generate our final collection of 150 EPS that are listed in Appendix A Table 1.

The score of each EPS was calculated based on three sets: the set of miRCat2 predictions using the EPS being evaluated denoted by m, the set of predicted functional sRNAs from PAREsnip2 denoted by p, and the set of validated miRNAs from miRBase denoted by mb. For the purpose of the parameter search method, we obtained the set p for each replicate from PAREsnip2 using Allen targeting rules [6] and the default parameters with the exclusion of weak cleavage signals summarised as PAREsnip2 categories 2, 3 and 4 interactions to generate high confidence results. The score that we used is given by:

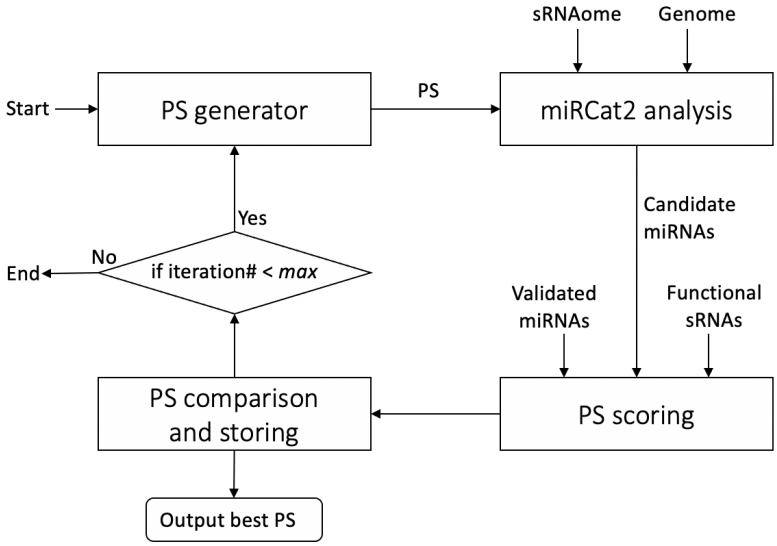


Figure 4.1 An overview of the parameter-search algorithm used to explore more permissive miRCat2 parameter sets (PSs). Solid rectangles represent processes, arrowed lines represent inputs and data flow, and round rectangles represent output.

$$score(EPS) = a + b + (1.5 * a/(a+b)) + (c/(c+d)) \tag{4.1}$$

where $a = \mid m \cap p \cap mb \mid$, $b = \mid m \cap p \cap mb' \mid$, $c = \mid m \cap p' \cap mb \mid$, *and* $d = \mid m \cap p' \cap mb' \mid$, where $'$ denotes the set complement.

In particular, the score in 4.1 was mainly calculated based on the intersection between the predictions of *m* and *p*, in addition to the ratio of *mb* in that intersection multiplied by 1.5. The factor 1.5 was chosen randomly as we found that the multiplication by a number greater than 1 would emphasise the score of EPS with an improved number of validated functional miRNA predictions in *a*. We also considered the non-functional validated predictions in *m*, hence, we added the ratio of non-functional predicted *mb* to the total non-functional predictions in *m*.

Retrieving the less stringent parameters using the parameter-search method requires running miRCat2 multiple times. However, the miRCat2 version that

was available at the time we started using the search method was depending on the database module within the UEA sRNA Workbench. Running this version multiple times on large sRNA datasets was imposing considerable computation time and memory resource constraints. Therefore, we contributed to the inclusion of a standalone version of miRCat2 that does not require the database module and performs the analysis within reasonable time frame compared to the main miRCat2 version. Moreover, we enabled additional features in miRCat2, such as including miRBase annotation in the results. The standalone miRCat2 version has been incorporated into the UEA sRNA Workbench and can only be executed through the command-line interface (CLI), allowing it to be integrated in other bioinformatics pipelines. The source code of the standalone miRCat2 version can be found in: https://github.com/sRNAworkbenchuea/UEA_sRNA_Workbench.

4.3.2 PAREfirst implementation

PAREfirst is a user-friendly, cross-platform (Windows, Linux and MacOS) tool that has been implemented in Java programming language (version 8) and is incorporated into the UEA sRNA Workbench [169]. This enables the users to utilize the existing pre-processing tools (e.g. Filter tool) of the Workbench to perform the PAREfirst analysis. The workflows within the Workbench are implemented via following the Model View Controller (MVC) framework and they can be connected together to form a workflow, which allows the flow of data between them and the ability to fully configure the workflows prior to runtime. Workflows in the Workbench are initiated by either the Database or FileManager module. The Database module creates a database on disk to store data during analysis. The FileManager module, on the other hand, serves as the input module that stores paths for input files. Both modules share a common GUI interface. Figure 4.2 presents a pipeline for PAREfirst, which combines miRCat2 and PAREsnip2 workflows into a tool within the UEA sRNA Workbench.

The user can perform a highly configurable analysis in a JavaScript graphical user interface (GUI) that produces an easily interpreted list of predicted miRNAs along with a visual representation of the prediction secondary structure and confidence metrics. Further validation of PAREfirst predictions can be achieved through several techniques including: sRNA-seq replication [20], laboratory experiments, and DE analysis [135].

PAREfirst accepts as an input the following files in FASTA format: a sRNA dataset file in redundant or non-redundant format, a degradome file in redundant format, a transcriptome file, and a genome file. The sRNA and degradome library files must be pre-processed, if they were not in the required format, to have the redundant or non-redundant format, and the adapters trimmed, this can be done using the filter ad adapter removal tools within the UEA sRNA Workbench [170]. Additionally, the user can configure the parameters for both PAREsnip2 and miRCat2 before starting the analysis. The tool performs the PARE analysis first using PAREsnip2 to produce the functional sRNAs that are stored by the database module in the Workbench. After that, it generates the miRNA predictions with miRCat2, which are also stored by the database module. The tool then retrieve all the stored predictions and combines the results to gain a set of functional miRNA predictions. Importantly, this method allows the use of less stringent rules for miRCat2, since the outcome of miRCat2 is controlled by the PARE analysis results.

As an output, PAREfirst exports and displays a table within a GUI interface containing information for the predicted functional miRNAs (Figure 4.3). In addition, the user is able to visualise and export target-plots (t-plots) [72] that are useful to distinguish true miRNA-mediated transcript cleavage sites from background noise, and the secondary structures for the predicted miRNA precursor using RNAplot [120] to visualise the hairpins. The computational efficiency of PAREsnip2 and miRCat2 allows PAREfirst to run on a desktop computer. Moreover, PAREfirst can be performed through command-line, which allows it to be used with other

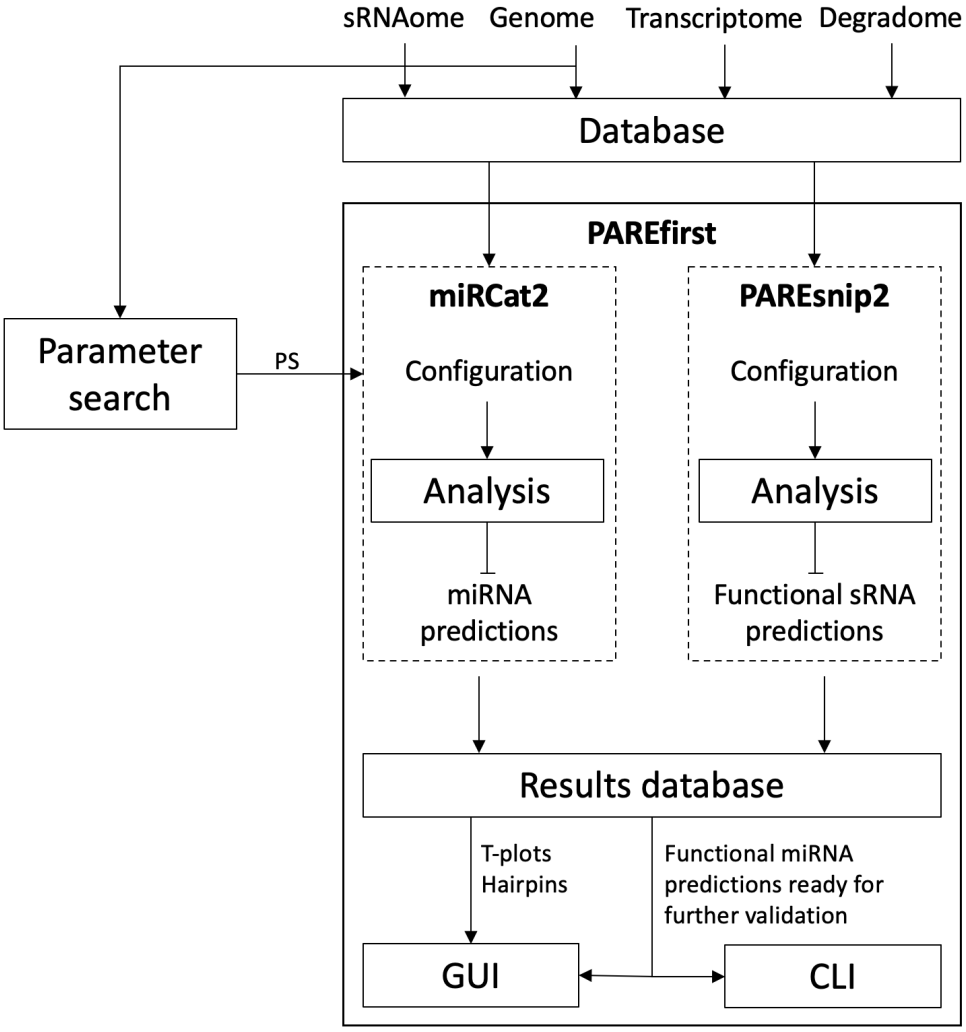


Figure 4.2 Schematic of the PAREfirst workflow used to perform a large-scale investigation of miRNAs and their targets evidenced through the degradome, along with parameter-search algorithm that provides less stringent parameter set (PS) for miRCat2 analysis. Solid rectangles represent processes, arrowed lines represent inputs, data flow, and output. The modules within PAREfirst are enclosed within dotted lines.

bioinformatics workflows. The PAREfirst source code, tutorial data, and manual files can be found on: https://github.com/sRNAworkbenchuea/UEA_sRNA_Workbench. The availability of the UEA sRNA Workbench on Github enables the community to to make modifications to current tools and create new ones that can be easily integrated into the existing framework. PAREfirst documentation can be found in Appendix A File 1.

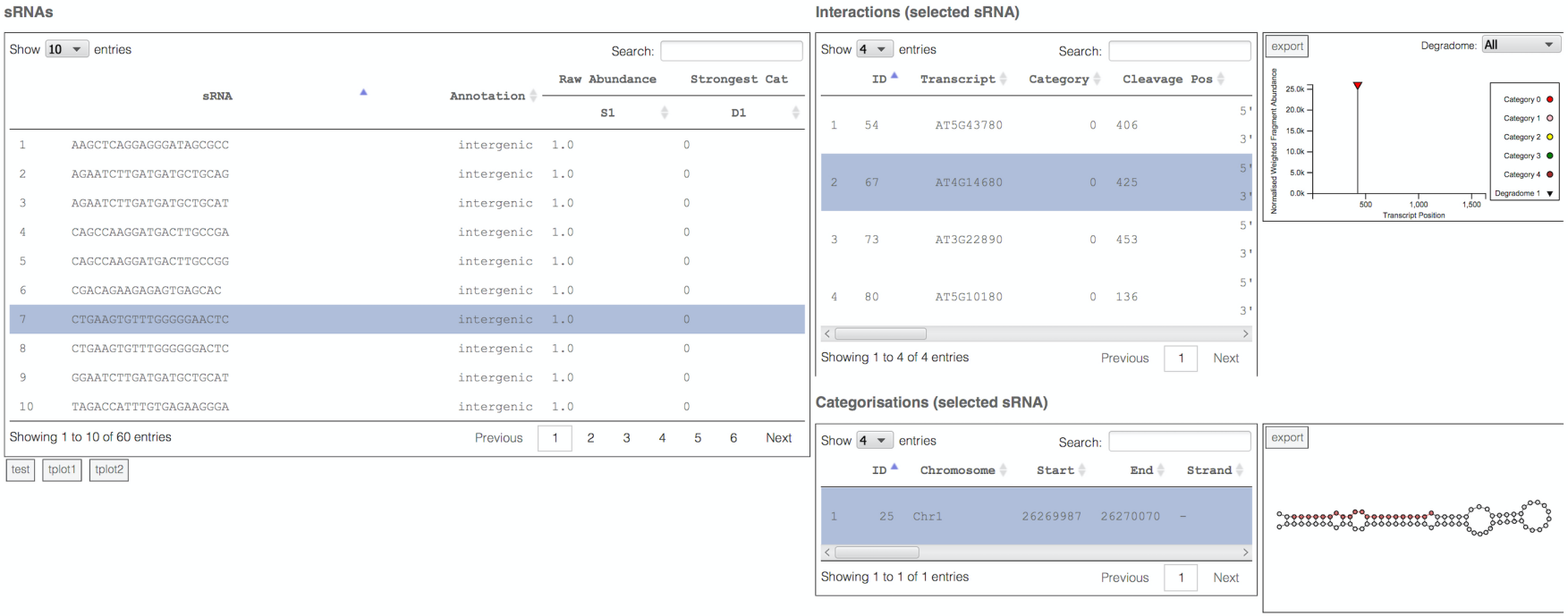


Figure 4.3 A screenshot of the results generated from PAREfirst within the UEA sRNA Workbench GUI.

To assess the performance of PAREfirst, we ran it on our data sets and benchmarked the results against other commonly used miRNA detection tools, miRCat2 and miRDeep-P2 [100]. To produce PAREfirst predictions, we used the EPS for miRCat2 analysis and the permissive PAREsnip2 parameters that were described above. The miRNA predictions for each of the other tools were obtained using the UPS for miRCat2 and the default miRDeep-P2 parameters (described in the user manual) for miRDeep-P2. We aligned the predictions to the mature miRNAs from miRBase to identify the validated miRNAs. Also, we excluded the miRNA candidates that aligned to other non-coding RNA classes.

4.3.3 Data sets

The organism that we considered for this study is *A. thaliana*. We used three wild-type and three DCL1-mutant *A. thaliana* sRNA biological replicates that are publicly available (GSE90771) [143], we called the wild-type sRNAome WTA, WTB, and WTC, and the DCL1 mutant samples: DCL1A, DCL1B, and DCL1C. Table 4.2 gives a summary of the three wild-type replicates and the three mutant replicates. Additionally, we used three *A. thaliana* DCL4-mutant datasets that were obtained from GEO (GSM4061704, GSM4061705 and GSM4061706) [87]. The PARE analysis was performed using the corresponding degradome for each wild-type replicate that are also available on GEO (GSE113958) [177]. For evaluation, we used the 326 unique mature miRNAs, which are excised from 426 precursors for *A. thaliana* from miRBase registry (v22) [76], and for the sake of this thesis, we refer to them as the validated miRNAs. The reference *A. thaliana* genome TAIR10 [174] was used, in addition to the transcriptome TAIR10 cDNA 20110103 representative gene model updated [27]. The DCL mutant raw data was processed using tools within the UEA sRNA Workbench.

Datasets	Redundant sequences	Unique sequences
WTA	6 698 044	1 362 057
WTB	4 514 396	1 107 617
WTC	5 173 806	1 121 816
DCL1A	12 296 993	3 042 828
DCL1B	11 234 476	2 548 906
DCL1C	15 347 268	3 132 360

Table 4.2 Total number of sRNAs in wild-type (WT) and Dicer (DCL1) mutant *A. thaliana* biological replicates.

We trimmed the adaptor sequences using the adapter trimming tool. Next, we aligned the sequences to the genome (TAIR10) with no mismatches allowed and discarded sequences with ambiguous bases using the Workbench Filter tool.

We further investigated the applicability of our method on other plants. Here we followed a similar approach for the data preparation and the target analysis parameters, however, we excluded the DE analysis due to the lack of DCL1-mutant data. The investigated species were the commonly studied tomato, *Solanum lycopersicum* (*S. lycopersicum*), and rice, *Oryza sativa* (*O. sativa*). For *S. lycopersicum* analysis, we used the publicly available sRNA data sets from GEO [180] (leaf GSM803579, flower GSM803580, and fruit GSM803581), and performed the target analysis with the corresponding tissue degradome data from a different study [119] (leaf GSM553688, flower GSM553689, and fruit GSM553690). The reference genome (SL3.0) and transcriptome (ITAG3.0) were downloaded from the Sol Genomics Network [61]. For *O. sativa* data, we used four sRNAome libraries [167] (Indica rice seedling and panicle, GSM562942 and GSM562943, Japonica rice seedling and panicle, GSM562946 and GSM562947). Also, we used two degradome libraries from another study [197] (seedling GSM455938 and panicle GSM455938), where we performed the target analysis on one tissue of the degradome data with the two corresponding tissue sRNA data sets. The reference genome and transcriptome were obtained from the Rice Annotation Project Database [93, 159]. The annotation for

both *S. lycopersicum* and *O. sativa* were performed using all plant mature miRNAs from the miRBase registry (v22).

4.4 Results

4.4.1 Comparison of miRCat2 analysis using different parameter sets

We hypothesised that using more flexible miRNA criteria would improve the prediction of valid and novel miRNAs. To test our hypothesis, we applied our method on the well-studied genome *A. thaliana*, as the objective of this analysis is to prove this concept, rather than to present a method that predict more valid miRNAs.

To compare the effect of applying more flexible alternative annotation criteria to miRNA prediction, we ran miRCat2 on wild-type sRNA data sets: WTA, WTB, and WTC, using the three parameter sets: DPS, UPS and EPS. For the purpose of improved confidence in predicted miRNAs, those having fewer than 10 reads were discarded from further analysis. We also filtered read counts by excluding isomiRs (sequences that are one or two nt shorter or longer than the canonical mature miRNAs), thus providing a clear quantification of the mature miRNA for each prediction. In addition, we considered conservation of the mature miRNA sequence across the three biological replicates in an attempt to provide a higher degree of confidence based upon multiple observation of the sequence [63]. For this thesis, we define a miRNA as conserved if it was expressed in at least two out of the three wild-type replicates. Furthermore, we used the validated miRNAs from miRBase as a reference to evaluate the results, even though we acknowledge its limitations with regards to the quality of miRNA annotations [20, 161]. We split the candidates into two groups: C and P; the reason for this categorization is that multiple miRNAs and miRNA-like RNAs can originate from one miRNA precursor [206, 211].

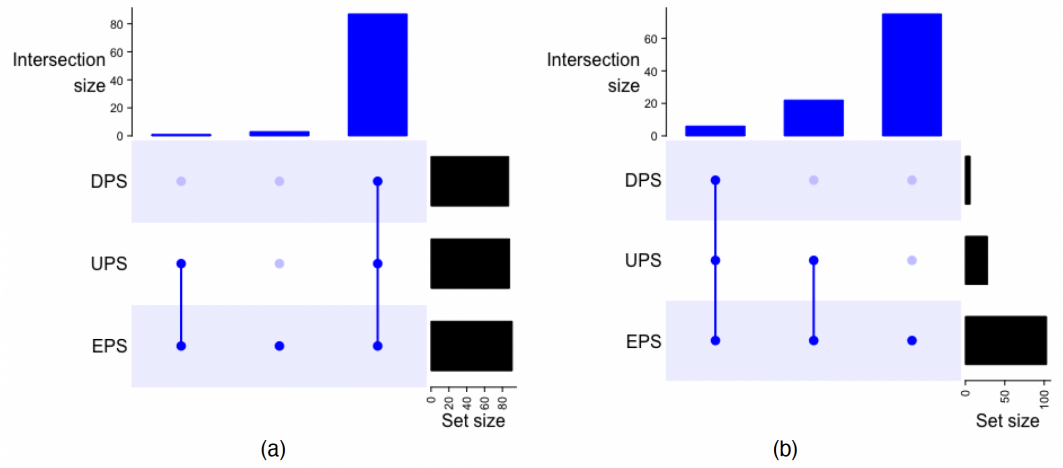


Figure 4.4 UpSets plots show the number of (a) validated miRNAs and (b) candidate miRNAs that are shared between miRCat2 predictions using DPS, UPS, and EPS parameter sets.

In Figure 4.4, we present the number of miRNAs that are shared between the miRCat2 predictions using DPS, UPS and EPS. We also present the number of miRCat2 predictions using the three parameter sets in Table 4.3(a). The table includes the number of all validated miRNAs within a sRNA replicate as well as the validated miRNAs predicted by miRCat2 using the three parameter sets. Comparing the results produced using each of the parameter sets, we observe that the UPS succeeded in predicting one or two more of the validated miRNAs in each replicate, and only predicted several new candidates when compared to DPS. Hence, it is likely that the false positive rate is still low and the performance of UPS is overall sufficient. Yet, using the EPS with miRCat2 performed slightly better in predicting more of the validated miRNAs. In particular, four more validated miRNAs that are conserved were predicted compared to DPS. However, it generated a high number of new candidate predictions that may include a number of false positives. To that end, increasing the flexibility increases the false positives in comparison to the updated criteria.

	Filter	Replicate	AV	DPS			UPS			EPS		
				V	C	P	V	C	P	V	C	P
(a)	None (all predictions)	WTA	132	87	8	33	89	29	33	92	129	43
		WTB	126	85	6	28	87	25	31	89	97	39
		WTC	127	90	6	26	91	27	28	95	106	37
		Conserved	136	87	6	31	88	28	33	91	103	42
(b)	PAREsnip2 filter	WTA	127	85	5	31	87	13	31	90	61	40
		WTB	109	74	4	25	75	9	28	77	35	33
		WTC	83	55	3	18	55	4	19	59	34	26
		Conserved	121	78	3	29	78	9	31	81	39	38
(c)	PAREsnip2 and DCL1 filters	WTA	91	63	3	27	64	8	26	67	33	34
		WTB	78	55	2	22	56	6	24	58	26	29
		WTC	59	41	2	18	41	3	19	45	17	26
		Conserved	85	58	1	26	58	6	27	61	23	34
(d)	Only DCL1 filter	WTA	95	65	5	29	66	17	28	69	77	37
		WTB	92	63	3	25	65	16	27	67	57	34
		WTC	93	67	3	25	68	17	26	72	57	34
		Conserved	97	65	3	28	66	18	29	69	63	37
(e)	PAREsnip2 and DCL4 filters	WTA	79	54	4	19	56	21	20	56	44	26
		WTB	79	46	4	17	57	19	21	48	28	23
		WTC	79	39	3	11	57	21	18	41	17	15
		Conserved	79	49	3	17	56	21	21	50	27	24
(f)	Only DCL4 filter	WTA	77	55	6	20	56	21	20	57	92	28
		WTB	66	55	6	19	57	19	21	58	75	28
		WTC	55	56	6	16	57	21	18	59	78	27
		Conserved	70	55	6	19	56	21	21	57	78	28

Table 4.3 The number of validated miRNAs that were found in our sRNA data sets, miRCat2 predictions using default parameters (DPS), updated parameters (UPS), and exploratory parameters (EPS) from the parameter-search method. AV: all validated miRNAs within a sRNA replicate, V: validated miRNAs predicted by miRCat2, C: candidate miRNAs predicted by miRCat2 and do not map to miRBase validated precursor loci, P: candidate miRNAs predicted by miRCat2 that map to a validated precursor but do not map to the canonical miRNA site. The conservation level used is between two or three replicates. All validated and candidate miRNAs have a read count above 10 reads.

We now present the results of applying the PAREsnip2 filter on the miRCat2 predictions. Although using less-stringent parameters for predicting miRNAs and their

precursors can introduce an increase in false positive predictions, we only consider the intersection between PAREsnip2 and miRCat2 predictions. Table 4.3(b) presents the number of miRNA predictions that are involved in a mRNA targeting interaction. It appears to be that the PAREnip2 filter kept a similar number of the validated miRNAs in each of the miRCat2 results across the three parameter sets, as true miRNAs are more likely to have a target. On the other hand, it discarded one-third or more of the non-validated candidates predicted by DPS and UPS. The functional filter reduced the majority of the EPS candidates, and upon manual inspection, we found that these candidates had secondary structures that were grossly inconsistent with miRNA biogenesis, hence, we consider this group to contain the highest number of potential false positive miRNA candidates. Accordingly, a function-first approach using degradome-assisted functional-filtering shows promising results, where we reduced the miRNA candidates to a list of 39 potential conserved functional miRNAs that can be carried forward for further investigation.

To validate the functional miRNA predictions, we performed a DE analysis between the wild-type and DCL1-mutant samples. Since the DCL1 has an important role in the miRNA biogenesis pathway in *A. thaliana*, knocking down its activity causes reduction in the expression of the miRNAs [95]. The outcome of applying the DCL1 validation filter was a set of predicted functional miRNAs that are enriched in the wild-type samples. The validation step discarded a number of the functional miRNAs and the results are shown in Table 4.3(c). The outcome shows a further reduction in the candidate predictions, where these remaining candidates could have a higher degree of confidence. Full details for all functional miRNA candidates and the conserved enriched functional miRNA candidates are found in Appendix A, Tables 4 and 5.

We also present the results for applying the DCL1 validation on all predictions in Table 4.3(d). It seems that several validated miRNAs were not down-regulated in the mutant samples. This could be because of a DCL1-independent pathway, and, in

some cases in *A. thaliana*, miRNAs are sometimes processed by a different Dicer family member such as DCL4 [25]. To investigate this hypothesis, we applied the DE analysis between wild-type and DCL4-mutants. The outcome of applying the DCL4 validation filter on the predicted functional miRNAs are presented in Table 4.3(e) and (f). The comparison between the up- and the down-regulation that occurred in WT_DCL1 and in WT_DCL4 is shown in Figure 4.5, which shows an overlap of seven validated miRNAs that were Down_WT_DCL1 and Up_WT_DCL4, and this could indicate a major involvement of DCL4 in some miRNA biogenesis pathways.

Furthermore, we investigated whether there is a need for applying more flexible criteria on other plant species, and assessed the applicability of the functional analysis filter on those species. To do that, we selected two plants genomes, including tomato and rice , based on the availability of a set of validated miRNAs and the availability of degradome datasets for each species. In Table 4.4, we present the number of predictions produced using both DPS and UPS, and the effect of applying the functional analysis filter on these predictions. As for *S. lycopersicum* results, the numbers of mature miRNAs that are present within the datasets were low, which explains the low numbers of miRCat2 predicted validated miRNAs. Also, we observed that the number of validated miRNA predictions did not increase in UPS compared to DPS. Hence, *S. lycopersicum* miRNA might benefit from applying more flexibility in miRCat2 parameters, which could improve the prediction of validated miRNAs, as observed in *A. thaliana* predictions. The application of functional analysis filter kept around half of the validated miRNAs across all replicates. On the other hand, majority of candidate miRNAs that were produced using both DPS and UPS, were filtered. Similarly to *A. thaliana*, the functional filter narrowed down the number of miRNA candidates, which supports the applicability of this filter on tomato datasets, however, further validation of these remaining candidates is required. Likewise, in order to enhance the prediction of validated miRNAs, further flexibility in the tomato miRNA biogenesis might be needed.

	Filter	Replicate	AV	DPS			UPS		
				V	C	P	V	C	P
S.lycopersicum	None (all predictions)	GSM803579	6	1	452	3	1	487	3
		GSM803580	7	2	403	3	2	428	3
		GSM803581	8	2	413	3	2	442	3
		Conserved	13	3	478	6	3	513	6
	PAREsnip2 filter	GSM803579	3	1	139	1	1	150	1
		GSM803580	2	1	100	1	1	105	1
		GSM803581	3	1	82	2	1	89	2
		Conserved	4	1	82	2	1	86	2
O.sativa	None (all predictions)	GSM562942	215	56	8	17	57	13	17
		GSM562943	202	33	119	5	35	183	5
		GSM562946	236	48	23	19	48	31	19
		GSM562947	265	58	165	10	60	248	11
		Conserved	375	92	9	27	92	14	27
	PAREsnip2 filter	GSM562942	47	29	3	6	29	8	6
		GSM562943	70	27	82	2	28	127	2
		GSM562946	12	3	4	2	3	4	2
		GSM562947	93	42	116	7	43	174	8
		Conserved	27	18	2	1	18	4	1

Table 4.4 The number of validated miRNAs that were found in *S. lycopersicum* and *O. sativa* sRNA Data Sets, miRCat2 predictions using default parameters (DPS), and updated parameters (UPS). AV: all validated miRNAs within a sRNA replicate, V: validated miRNAs predicted by miRCat2, C: candidate miRNAs predicted by miRCat2 and do not map to miRBase validated precursor loci, P: candidate miRNAs predicted by miRCat2 that map to a validated precursor but do not map to the canonical miRNA site. The conservation level used is between two, three, or four replicates. All validated and candidate miRNAs have a read count above 10 reads.

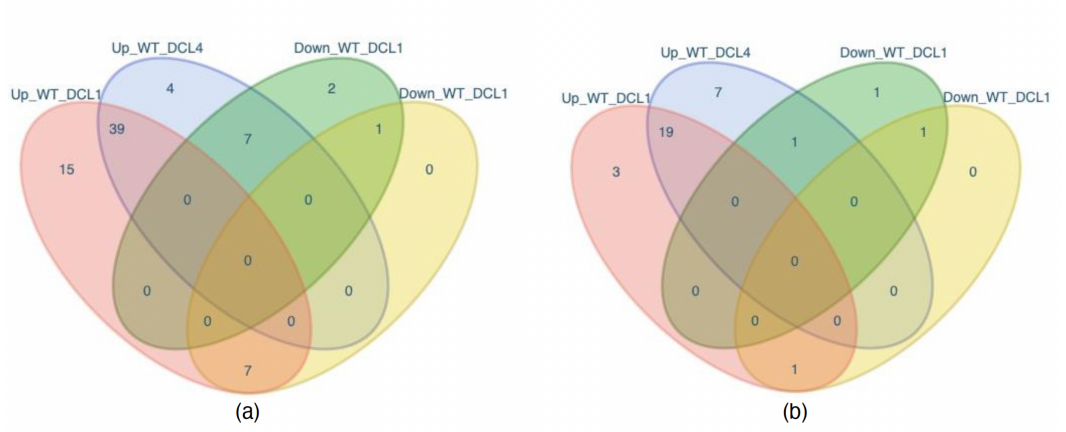


Figure 4.5 Schematic of the overlap between the number of the enriched miRNA predictions in wild-type vs DCL1-mutant, and in wild-type verses DCL4-mutant is shown in Venn diagrams for (a) the validated miRNAs and (b) the candidate miRNAs.

In *O. sativa*, there is a slight increment in the number of predicted validated miRNAs in the UPS results compared to the DPS, majority of the validated miRNAs that are present in the sRNA datasets were not reported with both DPS and UPS, thus, as observed in *A. thaliana* and *S. lycopersicum*, miRNA prediction in *O. sativa* may benefit from exploring more flexible biogenesis rules. The results also show that the functional filter excluded a minority of the predicted validated miRNAs in the seedling samples, with a reduction of 18% in GSM562943 and 27% in GSM562947. Yet, high proportion of the validated miRNA were excluded in panicle samples GSM562942 (48%) and GSM562947 (93%). Additionally, the filter excluded a majority of the candidates across all datasets. Correspondingly, these results suggest that our method might be applicable to other plant species, however, further investigation and validation are required.

4.4.2 Investigation of the miRNA annotation criteria

Further investigation on *A. thaliana* was applied to find whether the predicted secondary structures produced by miRCat2 with DPS, UPS, and EPS fit under the

2008 criteria, 2018 criteria, both criteria, or do not fit any of them. where the category of both criteria does not intersect with the 2008 or the 2018 criteria. The results are presented in Table 4.5. In this table we refer to the category where the precursor does not fit any criteria as 'Undefined'. In the following, we only consider miRNA precursors rather than the unique mature miRNA sequences, since the annotation is based on both the miRNA/miRNA* duplex and precursor structure features. We observed that most of the miRNAs that were predicted exclusively from any single replicate do not fit any criteria, hence, they were discarded since the conservation between replicates supports the confidence of the miRNA. Therefore, Table 4.5 only includes the grouping of the precursors that were predicted by miRCat2 in two or three sRNA replicates.

We observed that a few validated miRNAs do not fit any of the criteria, and a similar case was addressed by Axtell and Meyers [20] where some entries in miRBase may need to be revised. In addition, we looked into the validated miRNA precursor structures that are shown in miRBase and we observed that some of them have 1-nt overhangs at the 3' ends of the miRNA/miRNA* duplex instead of 2-nt as the annotation criteria suggested. Interestingly, even with the 2018 criteria included via the UPS for miRCat2, some of the predictions are still categorised as 'Undefined' criteria. As explained before, Table 4.5 also shows that the EPS predicts a higher number of all prediction candidate precursors than the DPS and UPS, and it seems that the majority of these candidates fit under the 'Undefined' category, and these candidates were considered to contain a high number of the false positives. Using the degradome assisted sRNA targets as a filter has discarded the majority of these false positives, while keeping most of the validated miRNA precursors.

	Conserved predictions	Validated miRNAs				Candidate miRNAs			
		2008 criteria	Both criteria	2018 criteria	Undefined	2008 criteria	Both criteria	2018 criteria	Undefined
DPS	All predictions	0	45	21	61	0	7	9	28
	+ PAREsnip2 filter	0	37	15	58	0	7	8	23
	+ DCL1 filter	0	31	10	41	0	6	8	17
UPS	All predictions	0	45	19	64	0	7	10	53
	+ PAREsnip2 filter	0	37	13	60	0	7	9	31
	+ DCL1 filter	0	31	10	41	0	6	9	22
EPS	All predictions	1	45	21	65	1	7	9	174
	+ PAREsnip2 filter	0	38	15	61	1	7	8	80
	+ DCL1 filter	0	31	12	43	1	6	8	52

Table 4.5 Categorization of the DPS, UPS, and EPS predictions based on the miRNA annotation criteria. The columns represent the validated and candidate miRNAs that fit the 2008, 2018, both criteria, or do not fit any criteria (Undefined), and the rows represent the filter layers that we applied on the miRNAs. The counts in the column of both criteria do not overlap with the 2008 or the 2018 criteria columns.

Furthermore, the majority of DPS predictions were kept during this step, where these predictions provide confidence through their strict miRNA features. As with the PAREsnip2 filter, applying the DE analysis validation, the DCL1 filter, also excluded the majority of the EPS candidate miRNAs that fit under the 'Undefined' category, and only excluded a minority from the rest of the results.

4.4.3 New miRNA and miRNA* candidates

We carried out an investigation of the miRNA candidates that are involved in a mRNA-target interaction and enriched in the wild-type samples. Before doing so, we excluded miRNAs that map to mature miRNAs, mature miRNA isomiRs, and the other RNA classes. As a result, we identified a potential novel miRNA with its miRNA* that map to one unannotated locus in the genome. To check if these candidates have already been annotated in other species, we performed a local alignment on all plant mature miRNAs from miRBase, and we found that the miRNA and miRNA* are not present in other genomes. Additionally, we aligned the enriched miRNA candidates to sequence databases using BLASTN algorithm within BLAST web interface, and we found no match with other plant species. The mature candidate was predicted along with its miRNA* and hairpin structure using UPS and EPS for miRCat2. The secondary structure of the candidate precursor with the highlighted miRNA/miRNA* duplex is shown in Figure 4.6(a). In addition, the majority of reads in the precursor align to the miRNA/miRNA* duplex as shown in Figure 4.6(b). This candidate precursor falls under the 2018 miRNA annotation criteria.

The abundance of the mature miRNA appears to be low (less than 100 reads) compared to most of the known miRNAs in our samples. Additionally, the mature abundance is double the miRNA*, which is a requirement for a well formed 'bona fide' miRNA duplex [20]. As shown in the WT_DCL1 DE analysis results in Appendix A Table 2, both the mature and the star sequence are differentially expressed

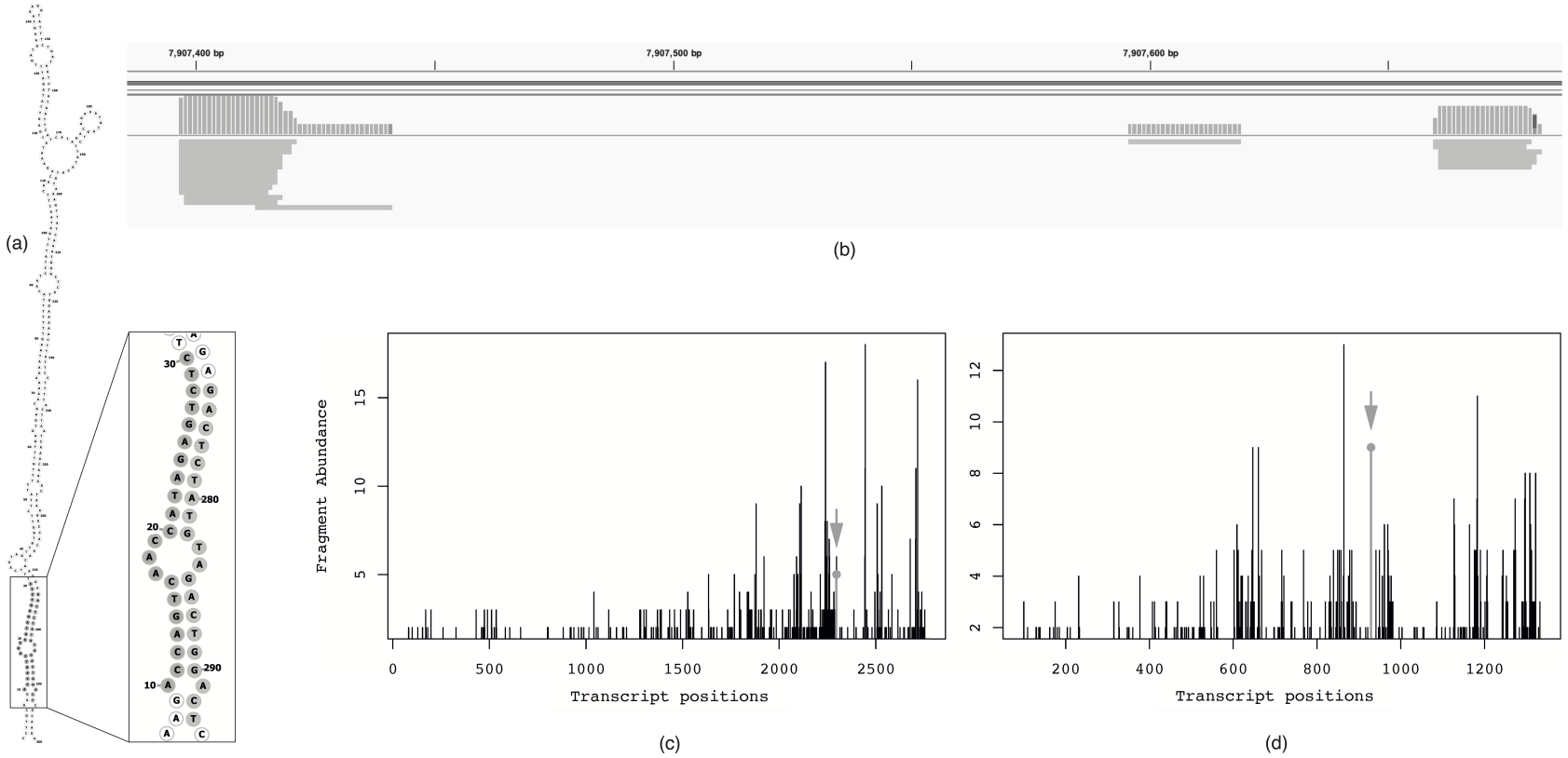


Figure 4.6 The novel miRNA candidate hairpin precursor with its mature miRNA and miRNA* sequences. The coordinates of the precursor locus within *A. thaliana* genome, Chromosome 4 is 7907388-7907687(+). (a) secondary structure of the precursor where the miRNA arises from the left arm of the hairpin and the miRNA* rises from the right arm [94], (b) a coverage plot of the precursor locus where miRNA alignments are presented on the left side and miRNA* alignments are presented on the right side of the plot, (c) t-plot of mature miRNA target mRNA AT4G00340, and (d) t-plot of miRNA* target mRNA AT3G61790 where arrows point at the cleavage positions.

with Log_2(fold-change) of 4.48 and adjusted p-value < 0.05. We present the t-plots for the most confident predicted mRNA-target interactions for the mature miRNA in Figure 4.6(c) and the miRNA* in Figure 4.6(d). It seems that the mature sequence does not show a strong signal in its t-plot compared to the miRNA*, which shows a higher peak. According to PAREsnip2 results, the mature miRNA is predicted to be involved in targeting interactions of category 2 with three different genes, while the miRNA* have six different targeting interactions of category 2. We discussed in detail potential functions of the mature miRNA and the miRNA* in Chapter 5. Also, details about the candidate precursor and the miRNA/miRNA* target interactions can be found in Appendix A, Table 5. Additionally, we explored alternative four *A. thaliana* data sets (flower GSM707678, leaf GSM707679, root GSM707680, and seedling GSM707681), and the potential novel miRNA and its star sequence were present in flower, leaf, and seedling samples. We performed miRCat2 analysis on these three data sets using UPS and the potential novel miRNA secondary structure was predicted with its miRNA/miRNA* duplex. The miRCat2 results for these new samples are presented in Appendix A, Table 6.

Furthermore, there are a number of miRNA candidates that derived from validated precursors in miRBase. These miRNAs are predicted with their complementary miRNA* (Appendix A Table 7), however, these miRNA* sequences are not registered in miRBase. Moreover, there are other studies that acknowledge these miRNA* [131, 190].

4.5 PAREfirst benchmarking

To investigate how PAREfirst compares to the existing traditional miRNA prediction tools, we ran our data sets through PAREfirst, miRCat2 and miRDeep-P2. We present the output of each tool in Appendix A Tables 8, 9 and 10, respectively.

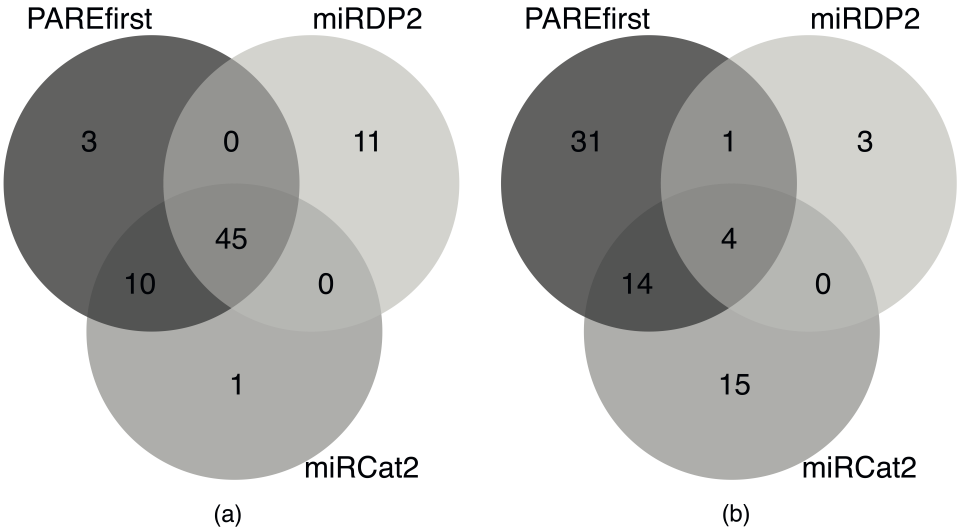


Figure 4.7 Venn diagram showing the intersection between the number of (a) validated miRNAs, and (b) candidate miRNAs that were predicted by PAREfirst, miRDeep-P2 (miRDP2), and miRCat2 in the three sRNA replicates.

Figure 4.7 shows the overlap of the number of validated and candidate predictions among the three tools, where these predictions are conserved between two or three out of three replicates. Figure 4.7(a) indicates that PAREfirst captures the majority of the other tool's validated miRNA predictions, in addition to other three miRNAs that were not identified using the updated miRNA annotation criteria. Figure 4.7(b) also shows that PAREfirst allows the identification of candidate miRNAs that as observed in this study, do not necessarily conform with the standard model.

4.6 Discussion

Axtell and Meyers [20] argued that a change in miRNA annotation criteria is necessary since some validated miRNAs are missed by the former criteria and some novel miRNAs might be missed too. Accordingly, there is a need to update the parameters used within the existing prediction tools, or design new tools that incorporate less stringent rules. In this chapter, we sought to investigate the effect of applying further flexibility to the miRNA annotation rules within a controlled method that

is assisted by functional analysis. By systematically exploring different parameter sets, it is clear that flexible parameters have an impact on miRNA prediction, and we should keep the balance between predicting miRNAs with novel features and an overestimation of the miRNA profile. We showed that using the updated and the less stringent criteria increases the capture of validated miRNAs while keeping a similar number of potential candidates through filtering with the degradome analysis. We applied our method to several publicly available *A. thaliana* sRNA data sets and we were able to identify a potentially novel miRNA candidate that has been previously missed by tools that are dependent on outdated miRNA-rule sets.

A. thaliana is a well-studied genome, and we can expect that the miRNA profile is well characterised, yet, we have been able to identify a potential novel miRNA. Applying our approach on less well annotated plant genomes could capture not only the miRNAs that are easily identified through existing methods but also the miRNAs that would otherwise be missed due to their extreme biogenesis characteristics. With this in mind, the variance within the parameters that identify novel miRNAs in *A. thaliana* may not be the most suitable in all cases, and we hypothesise that improvements in annotation results could be obtained from investigating species-specific parameter sets, however, due to time limitation, we were unable to pursue this hypothesis. To this end, we have provided some software to enable researchers to take this forward in the model species of their choice.

The degradome analysis is an NGS approach to identify miRNA mediated cleavage [144], and the defined category system within degradome analysis tools ranks the confidence level of miRNA-mRNA interactions. A degradome sequencing experiment offers many advantages when compared to low-throughput methods typically used for miRNA target validation, such as 5' RACE [117]. The advantages are not only present in time-cost savings, but also the global nature of the degradome profile being captured in a single experiment can reveal multiple miRNA-target interactions, useful for building miRNA-mediated gene regulatory networks [72, 188]. In

addition, when compared to traditional sequence similarity approaches for target prediction [204], using a degradome assisted miRNA target prediction approach provides valuable quantitative confidence values based on experimental evidence [2, 63, 177, 213]. Our results show that most of the enriched known and candidate miRNAs are predicted to be functional, which suggests that the degradome analysis provides useful supporting evidence for identifying functional miRNA candidates without using further validation steps. Be that as it may, a degradome assisted approach is somewhat limited by its dependence on the expression and tandem capture of the miRNA and miRNA mediated cleavage signal within the sRNAome and degradome data sets. The expression of many miRNAs and their targets are localized both temporally and spatially, specific to factors such as tissue, growth, and environment. However, testing for condition specific miRNA candidates and their targets is a common goal in sequencing experiments that investigate within and for such factors [54, 123, 151, 210]. Also, generating the degradome data required by our method can be challenging and is not necessarily straightforward [112]. However, new and optimized degradome protocols are regularly becoming available [38, 112, 113]. As NGS techniques become more accessible, we envisage more degradome libraries will become available, enabling the use of our approach with more varieties of species.

Prior NGS and miRNA computational prediction methods, biologist would have to assess possibly hundreds of genuine miRNAs, which was not possible without considerable time and resource constraints. Therefore, the development of bioinformatics methods has became essential to identify a selection of potential novel miRNA that could undergo further experimental validation. Experimental validation is the most direct and reliable method to assess the accuracy of computational predictions of genuine miRNAs. RNA gel blotting, such as northern and western blotting, has been widely used in molecular biology research to study gene expression in a biological sample, yet, it has become less common due to the continuous

advancement in NGS and high-throughput sRNA-seq techniques. Besides, Axtell and Meyers [20] argued that using sRNA-seq techniques to validate miRNA prediction in plants should alternate experimental validation, as RNA gel blotting accumulate sRNA expression but does not distinguish between miRNAs from other sRNA classes.

Another method to assess computational prediction is the use of mutant data to perform differential expression analysis. The differential expression is a strong indicator that the analysed sequences are interacting with the mutated genes. In particular, if the predicted miRNAs are significantly down-regulated in mutant data, it can provide evidence that the computational prediction is biologically relevant. In our results, we performed differential expression using Dicer-mutant datasets to further validate the predictions obtained from our method.

Furthermore, detection of the same novel miRNA in multiple biological samples could present evidence of it being a true miRNA. Replicates that come from the same condition, such as species, tissue, or developmental stage, could provide evidence for the expressed miRNAs to be specific to that condition [118]. We demonstrated in our results how the candidates mature miRNA/miRNA* were conserved across multiple *A. thaliana* sRNA datasets, which suggest that these candidate could be species-specific miRNAs. A further miRNA computational prediction assessment method is the benchmarking against other computational tools in the area by using a set of validated miRNAs to compare the performance of the different tools. The problem with benchmarking our method directly against other tools is that they use different parameters and they are not designed to be used in the way that we propose, i.e. find miRNAs that do not necessarily conform with the standard miRNA model, making it difficult to compare them directly and fairly with our approach. However, we have presented some results from miRDeep-P2 and miRCat2 to give some indication on how they compare.

The identification of biogenesis and function of a miRNA is important to understand its role within biological pathways and networks. Researchers and miRNA databases, such as miRBase, are enabling the provision of a reliable set of functional information that will enhance the advancement of the microRNA research field. In particular, miRBase is being improved not only by providing miRNA annotation entries, but by also including the functional information of these miRNAs [97]. Our functional approach moves toward the aim of identifying miRNAs and their target mRNAs, and we hope that it will have an impact on enriching the literature of miRNA functions.

In conclusion, our degradome-assisted method for miRNA prediction appears to provide broader predictions for plant miRNAs in a controlled manner. We have implemented it in PAREfirst, a freely available software that can be used to predict functional miRNAs. As more sequenced genomes of different species become available, we hope that our tool will play an important role in the understanding of biology and evolution through the annotation of novel miRNAs and their functions.

Chapter 5

sRNA Network Construction Using Degradome

5.1 Summary

Recent research provides evidence for the presence of large complex sRNA-mRNA networks that are involved in biological regulation. These networks can be predicted using computational sRNA target prediction tools. However, with previous computational methods, the predicted networks were difficult to interpret due to the rate of false positives. We hypothesised that with the advance in NGS technologies and computational methods, we could construct more informative sRNA-mRNA networks on a genome-wide scale. In the previous chapter, we introduced an miRNA prediction method to identify functional miRNAs using the degradome to support the findings. In this chapter, we utilise the degradome to generate easier to interpret sRNA networks which aims to seek more confident interactions.

We start by introducing some background about networks and biological networks. Next, we describe our newly implemented tool, PAREnet, that we used to construct sRNA networks. After that, we generate the sRNA networks for *Arabidopsis thaliana* using the interactions that are retrieved from PAREsnip2 results, and

we visualise the network using Cytoscape. We then compare different networks based on the used PAREsnip2 parameters, using strict and less strict parameters. We then assess the regulatory contribution of the sRNA-regulated network that was constructed by the strict parameters by looking closely into the individual sub-networks. Finally, we determine the contribution of the miRNA/miRNA* candidates (predicted in Chapter 4) to the constructed regulatory network.

5.2 Background

As we have seen, sRNAs are abundant molecules that carry out a variety of gene regulation functions within the cell, giving them a strong influence on the mRNA profile [26]. This influence results from sRNA molecules, mostly miRNAs, targeting and silencing multiple mRNAs, which in turn produce further sRNAs forming what tends to be cascades and networks of interactions between sRNAs and mRNAs [43, 124, 202]. The first step in these cascades requires an RNA-dependent RNA polymerase (RDR) to convert the targeted RNA into long, double-stranded RNA [83]. There are three cases when miRNAs could initiate the cascades: the miRNA duplex structure is asymmetrical [125], the miRNA is 20-24 nt in length [42], or if there are two target sites for miRNA within the mRNA [19]. Once the cascade is initiated, a high proportion of 21 nt siRNAs are generated which associate with AGO proteins [43].

Recent studies support the existence of sRNA networks of interactions in plants, and show that these networks are involved in the regulation of the biological pathways at the level of transcription. In particular, it has been found and verified in *A. thaliana* that an sRNA network is initiated by miR173 cleavage of the TAS genes TAS1 and TAS2 [43]. This cleavage is then followed by the production of tasiRNAs, which in turn target a group of pentatricopeptide repeat (PPR) genes that are involved in RNA processing [6, 162]. A network of miR173-tasiRNAs-PPR/TPR interactions is

shown in Figure 5.1. Moreover, another regulatory cascade is initiated by miR390 targeting TAS3, then followed by producing tasiRNAs that are involved in auxin response regulation, plant growth, and leaf morphology [3, 68]. It was also found that miR398 participates in stress adaption by regulating the expression of superoxide dismutase (SOD) enzyme related genes [172, 173].

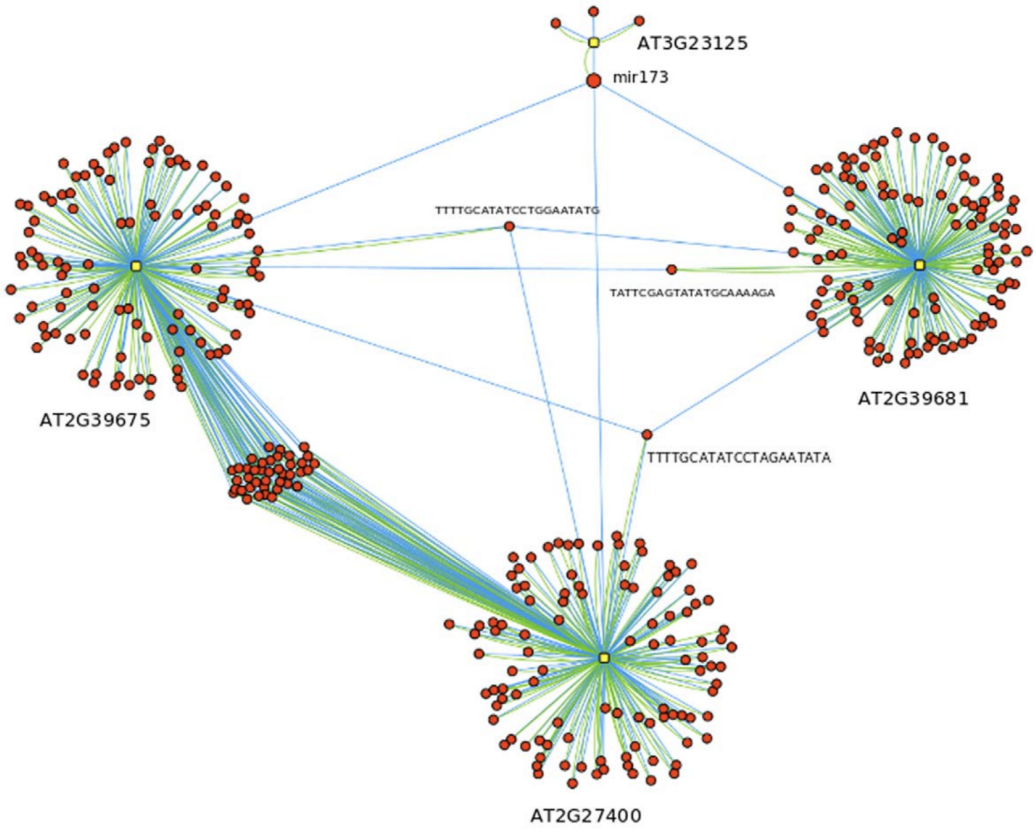


Figure 5.1 Visualisation of the miR173-tasiRNAs-PPR/TPR network. Yellow squares are mRNAs, red circles are sRNAs. Blue edges are sRNA to mRNA target, the green edge is RNA to sRNA source. The large red circle is miR173. Figure from MacLean *et al.* [124].

Networks are often described in terms of a mathematical entity called graph. It contains nodes and edges, and in the case of biological networks, the interacting molecules (e.g. sRNAs and mRNAs) are the nodes and the interactions between them the edges. The edges between those nodes are directed when the interactions only follow one direction such as sRNA-mRNA targeting interactions, however,

edges could be undirected when the link follows either way as in protein-protein interactions [171]. The term that describes the number of edges that are linked to a node, whether they come in or out, is called the degree of the node. The distribution of node degree is a commonly used feature for network analysis [21, 215].

Complex real-world networks, such as internet, protein interaction networks, and social networks, have power-law scale-free degree distribution. Also, there are mathematical features that are important to characterise such networks including high clustering coefficient, and assortativity or disassortativity. Random networks are used to model these complex networks in order to help understand their characteristics, as they largely un-clustered and have distinct characteristics from real-world networks. Generating a random network starts with a set of nodes where the connections between nodes are chosen randomly based on a probability distribution.

Biological networks, including sRNA networks, have mathematical graph characteristics that are more similar to real-world networks than random networks [124] (see also [10]). They have the characteristic of the power-law degree distribution where the majority of nodes have low degree and very few nodes, called hubs, have the highest range of node degree [4]. Real-world networks show behaviours of assortativity or disassortativity [140, 152]. In assortative networks, nodes with high degree tend to link to nodes with similar degree, while in disassortative networks, high degree nodes link to a greater number of low degree nodes. Biological networks tend to show a disassortative pattern. Moreover, real-world networks are highly clustered and have relatively short path lengths between nodes. These characteristics are common in biological networks, and they contribute to network resilience to failure [21, 146].

Recent studies support the evidence of the existence of regulatory sRNA networks [124, 70, 111], the associated sRNA experiments provide hints for the richness and complexity of the networks. It also became apparent that to identify changes in the

plant sRNA response to stress, a network of interactions between sRNAs and mRNA targets was required in order to serve as reference for mapping changes [183].

The investigation of sRNA regulatory networks relies on identifying the interactions between sRNAs and their target mRNAs. Those interactions have often been predicted by using sequence similarities methods, such as psRNATarget [50] and TAPIR [30], between sRNAs and and their target transcripts. An important downside of these methods is the rate of false positives, which leads to enlarged and complicated networks and makes them difficult to interpret [124]. As we have seen in Chapter 3 , the advance in NGS knowledge and computational methods, together with degradome sequencing, has made it possible to identify the interactions more reliably, and therefore, helped to clarify the sRNA regulatory network [2, 115, 177]. Moreover, current sRNA target prediction tools could enable the construction of sRNA networks on a genome-wide scale. However, the constructed networks tend to be large and complex, and hence, difficult to determine whether they reflect biological reality. One way to simplify the regulatory networks is to reduce the rate of false positives in the predictions, this can be achieved through adjusting the tools configuration in order to produce predictions with high confidence levels.

We hypothesised that using computational methods for degradome analysis to obtain a set of high confident predicted sRNA-mRNA interactions, could allow for the construction of a simplified sRNA-mRNA network, and thus, the possibility to retrieve meaningful biology from such a network. In this chapter, we present a tool, PAREnet, that we developed within the UEA sRNA Workbench that allows us to explore this hypothesis by enabling the construction of sRNA networks systematically from interactions predicted using the output from PAREsnip2 software. The constructed network characteristics are then analysed to test weather they are relative to other biological networks. Also, the interactions in the networks are investigated to assess if they provide biological meaning.

5.3 Methods

In this section, we introduce our sRNA-mRNA network construction method that uses PAREsnip2 to perform target prediction and output sRNA-mRNA interactions. We designed it to post process PAREsnip2 results and output a table of interactions that can be used as an input for Cytoscape [164]. We chose PAREsnip2 for this method because it had shown to generate more confident interactions using the degradome [177], and hence, less false positives interactions, which is an important factor to construct simpler networks. Additionally, PAREsnip2 shows improvement in computation performance, i.e. required computation time and memory usage, when compared to previously developed degradome analysis tools.

5.3.1 Network construction

We implemented a freely available and open source software tool, called PAREnet, for sRNA-mRNA network construction using the Java programming language (version 8). The cross-platform tool is incorporated into the UEA sRNA Workbench and it can only be performed through the command-line interface, allowing PAREnet to be used in other bioinformatics pipelines. PAREnet integration into the Workbench provides the ability to utilise other incorporated helper tools for PAREnet analysis, such as PatMaN aligment tool. Contrary to PAREfirst (see 4.3.2), PAREnet uses the FileManager module to store the absolute path of the input files in order to reduce memory usage during runtime. The source code for PAREnet can be found on: https://github.com/sRNAworkbenchuea/UEA_sRNA_Workbench, and also can be found in Appendix B File 1.

PAREnet accepts as an input the following mandatory files: sRNAome in redundant FASTA format, degradome in redundant FASTA format, and transcriptome in FASTA format. Alternatively, the user can provide PAREsnip2 results file and

transcriptome. Also, the user has the option to provide the following files: a genome in FASTA format, mature miRNAs from miRBase in FASTA format [99], and experimentally validated interactions from miRTarBase [85]. Although, miRBase file is optional, it is required only when providing the validated interactions. When providing a genome, sRNAs are aligned using PaTMaN [148] and the sequences that do not align to the genome are discarded prior the analysis. For the target prediction analysis, the user can configure the parameters for PAREsnip2 as they require. The stringency of the parameters determines the complexity of networks and as a consequence the ability to extract meaningful biology from them.

The tool performs the target prediction using PAREsnip2 to produce predicted functional sRNAs with their mRNA targets, we refer to the interactions between sRNA and mRNAs by target interactions. The interactions are stored in Hash Table data structure, which can speed up the search and access to its elements and provides better synchronization than other data structures. Next, the tools parses PAREsnip2 output to extract the functional sRNA reads that are then searched against the RFam database [69, 75] for tRNA, rRNA, and snoRNA-derived sequences using PaTMaN with exact match, and the aligning reads are discarded. Moreover, the identification of sRNA source gene is required to uncover the regulatory cascades of sRNAs, thus, the tool aligns the filtered functional sRNAs, using PatMaN with exact match, to the set of transcripts that were predicted to have cleavage sites for these sRNAs, we refer to the interactions between sRNAs and their source genes by source interactions. After that, if known miRNAs are provided, the known miRNAs and their isomiRs are annotated by aligning the unique sRNA reads to the known miRNA reads allowing up to two mismatches. Then the tool updates the miRNA annotation details for the aligned reads. Moreover, if validated interactions from miRTarBase are provided, the tool parses the validated interactions, and then it checks the stored interactions for the validated interactions within the predicted results and update their details. The tool also extracts the gene ontology terms of each transcript given within the PAREsnip2

results to be used as node labels. The information obtained from PAREsnip2 results (e.g. interaction category), known miRNAs, validated interactions, source genes, and transcripts' annotations are then used to construct network visual components such as labels, colors, and shapes of nodes and edges. For example, the interaction category and the type of interaction (target or source) specify the edge color, and the predicted known miRNA nodes are labeled with miRID and have distinguished color from other sRNAs. An overview of PAREnet workflow is shown in Figure 5.2.

The output of PAREnet is provided in comma-separated value (CSV) format that can be viewed in any CSV file viewer. The CSV file can be used as an input for Cytoscape in order to build sRNA-mRNA networks. The table includes information about the sRNA-mRNA interactions such as genes annotation, type of interaction (target or source), targeting categories, the known miRNAs annotation, and the validated interactions. In Cytoscape, the user can customise and highlight the nodes and edges within the constructed network using Style features [164].

5.3.2 Datasets

We carried out genome-wide degradome analysis and creation of sRNA networks were using three *A. thaliana* sRNA replicates (WTA, WTB, and WTC), which were previously published by Dalmya's lab at UEA (GEO accession number GSE90771) [143], and the corresponding degradome datasets (DegA, DegB, and DegC), which are also available on GEO (GSE113958). The transcriptome used in our analysis was obtained from the *Arabidopsis* Information Resource (TAIR) containing the cDNA for the updated representative gene model, in addition to the reference genome TAIR10 [174]. For the purpose of miRNA annotation, we used *A. thaliana* mature miRNA sequences obtained from miRBase (v22) [76]. To validate the predicted interactions, we used a set of experimentally validated *A. thaliana* interactions that we obtained from miRTarBase [85].

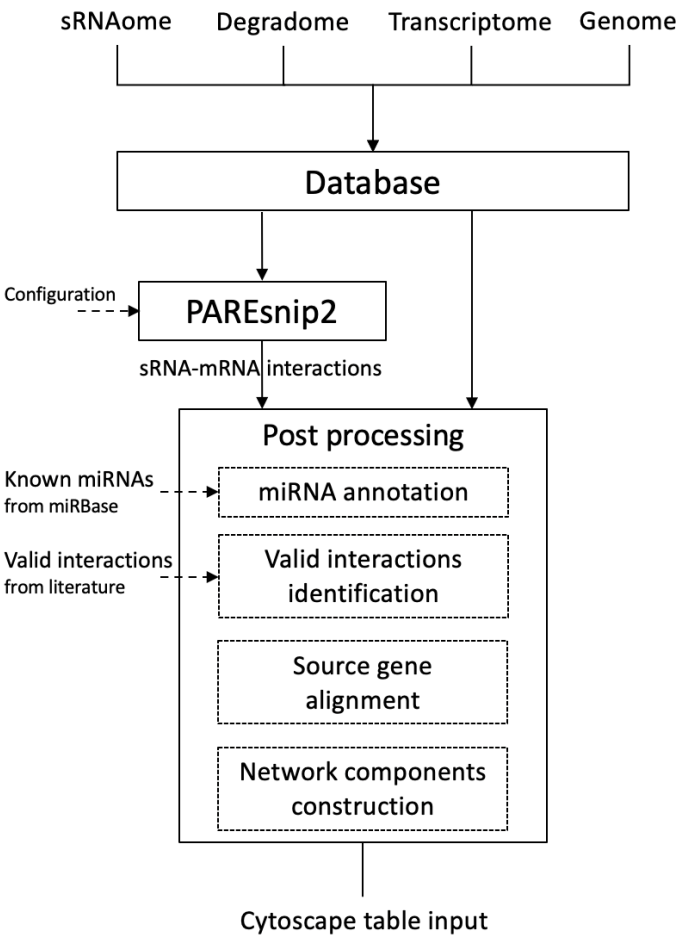


Figure 5.2 Schematic of the PAREnet workflow to construct sRNA-mRNA networks using degradome analysis. Solid rectangles represent processes, dotted rectangles represent sub-processes that are dependant on input data, solid arrows represent inputs and data flow, dashed arrows represent optional inputs, and lines represent output.

5.3.3 Network visualization and analysis

The constructed networks visualization and structural analysis were performed using Cytoscape (version 3.9.1) [164]. Statistical analysis for degree distribution and assortativity were carried out using R statistical package (version 4.3.0) [149] and NetworkX (version 3.1) [79].

5.4 Results

PAREsnip2 allows configurable parameters in order to determine the confidence level of predicted targeting interactions. We expected that using less strict parameters could produce a high number of false positives, and thus, lead to constructing a large complex network that are difficult to elucidate. On the other hand, using strict configuration could discard a proportion of false positives, hence, a simpler and more informative network could be constructed. To compare the effects of varying PAREsnip2 filters and category system on the sRNA networks construction, we first performed the degradome analysis through PAREsnip2 by using a less strict configuration. In particular, we used Carrington targeting rules [39], allowed categories 0-3, disabled core region multiplier, disabled *p*-value filter, and disabled MFE filter. We present the number of predicted interactions, along with number of sRNAs and genes that are involved in them, using the less strict parameters in Table 5.1 . We visualised the interactions using Cytoscape (version 3.9.1), which generated a large and complex network for each replicate. In Figure 5.3, we show the generated network of replicate WTA_vs_DegA, similar visualised networks were generated for the other replicates.

	Less strict PS2 parameters			**Strict PS2 parameters**		
Replicates	**sRNAs**	**Genes**	**Interactions**	**sRNAs**	**Genes**	**Interactions**
WTA_DegA	27235	194636	194636	16301	8530	20932
WTB_DegB	17078	12760	100830	9710	6072	12821
WTC_DegC	20231	14055	171223	14847	8650	21533

Table 5.1 Summary of the number of the target analysis predictions that were produced using PAREsnip2 (PS2) on three *A. thaliana* sRNA replicates against their corresponding degradome replicates.

Further analysis was performed using more strict PAREsnip2 configuration in order to produce a higher rate of high confidence interactions. In particular, we used

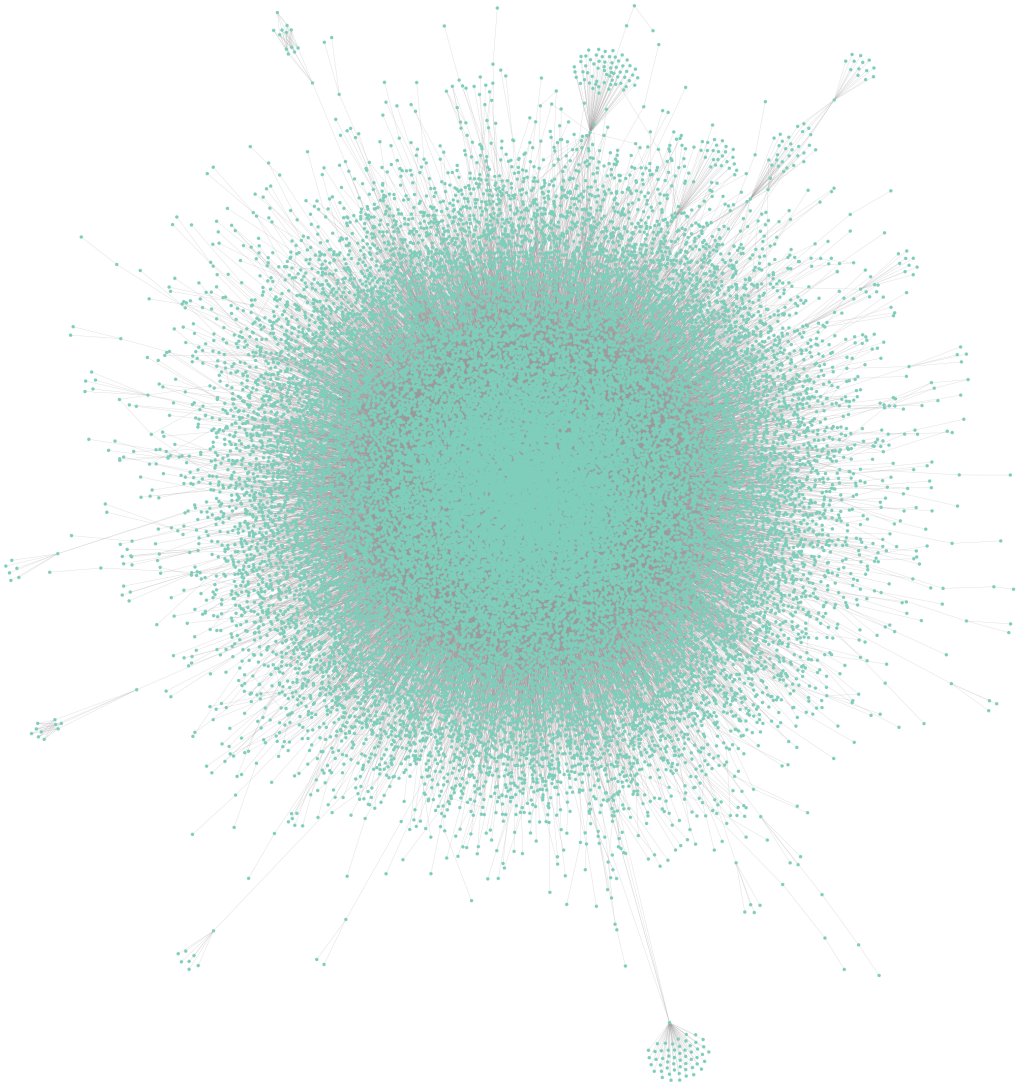


Figure 5.3 A visual construction of large and complex network of sRNA-mRNA interactions that were produced by PAREsnip2 using less strict parameters.

Carrington targeting rules, the default PAREsnip2 parameters, and only allowed interactions with the strongest signal on the transcripts, i.e. categories 0-2. We disabled the core region score multiplier as this step contributes to the identification of more experimentally validated interactions [177]. The numbers of predicted interactions, and sRNAs/genes corresponding to these interactions are presented in Table 5.1, which shows a reduction in the number of interactions in each replicate compared to the less strict configuration. For network construction and in-depth

assessment of the network, we used the set of interactions obtained from PAREsnip2 using the strict parameters. Specifically, we only included the interactions that are conserved among the three replicates, in order to construct a network with more confident interactions. The set of conserved interactions (see Appendix B Table 1) were then utilised by PAREnet to generate a sRNA-mRNA network that is presented in Figure 5.5.

Moreover, we evaluated and compared between the results of each parameters set, the less strict and the strict parameters. We used the set of experimentally validated interactions that we obtained from miRTarBase [85] to construct three validation classes. First, the true positives (TP) that consists of the predicted interactions with experimental validation. Second, the false positives (FP), which is the set of predicted interactions that has no current experimental validation. Third, the class of positive (P) data that included the total experimentally validated interactions present in the dataset. We provide the sensitivity and precision for each set of parameters, where sensitivity is calculated as TP/P, which is the proportion of predicted validated interactions, and precision is calculated as TP/(TP+FP), which is the proportion of predicted validated interactions among the total number of reported interactions. When evaluating the performance of each parameters set, we did not use specificity as a performance metric because it is challenging to accurately determine the class of true negatives. We show the differences in sensitivity and precision between the two sets in Table 5.2. The table shows that the strict parameters set provides increased precision compared to the less strict parameters set, whilst also maintaining sensitivity on most datasets. Over all datasets, the PAREsnip2 less strict parameters with a mean sensitivity of 49% versus 48.33% for the strict parameters. The mean precision for the less strict parameters was 95.67% versus 98.33% for the strict parameters.

A more detailed analysis of the network structure allows us to study the presented sRNA-regulated network and identify features that could be compared to related

Dataset	P	LS. V	LS. NV	S. V	S. NV	LS. TPR	LS. PPV	S. TPR	S. PPV
WTA	367	184	11	171	4	50%	94%	47%	98%
WTB	364	175	6	170	2	48%	97%	47%	99%
WTC	337	165	7	174	4	49%	96%	51%	98%

Table 5.2 Comparison of sensitivity and specificity between the less strict and the strict PAREsnip2 parameters on the *A. thaliana* datasets. P: positives, LS.: less strict parameters, S.: strict parameters, V: validated, NV: non-validated, TPR: true positive rate or sensitivity, and PPV: positive predictive value or precision.

complex biological networks. To better understand the properties of the regulatory networks presented here, structural features were analysed using the Cytoscape plugin, NetworkAnalyzer [16]. The results in Table 5.3 illustrate the network structural features for the networks constructed using the less strict parameters and the network constructed using the strict parameters. However, for simplicity, we shall refer to the former network as complex network, and to the latter as simplified network, for the rest of this analysis. The complex network consisted of 158 components, or sub-networks, while the simplified network consisted of 488 components. The average number of neighbours decreased from 9.072 in the complex network, to 3.487 in the simplified network. The path length in this analysis increased from 4.740 in the complex network, to 4.812 in the simplified network, which shows consistence with other complex biology where the path length is usually around 6. Contrary to biological networks, the clustering coefficient in both, complex and simplified networks, was extremely low. Network density, represents the proportion of the edges in the network to the all possible edges, was very low in both networks, consistent with results from other studies [105], however, the density in the simplified network was greater than the complex network.

Furthermore, we observed a significant relationship between node degree and the frequency of nodes of that degree (Figure 5.4). In totality, the heavy-tailed total degree distribution for both for in and out degrees, indicated that a small number of

Structural feature	Less strict	Strict and conserved
Number of nodes	42809	1655
Number of edges (target interactions)	194636	1546
Connected components (sub-networks)	158	488
Average number of neighbours	9.072	3.487
Path length	4.740	4.812
Clustering coefficient	0	0
Network density	0	0.015

Table 5.3 Summary statistics generated by Cytoscape Network Analyser, of the network produced using PAREsnip2 with the less strict parameters on one of the *A. thaliana* (WTA vs DegA), and the network produced using PAREsnip2 with the strict parameters and conservation approach (interactions that predicted in all three replicates) on three *A. thaliana* sRNA replicates against their corresponding degradome replicates.

high-degree nodes, while the majority have low degree. Due to the directed nature of sRNA networks, degree distribution for sRNAs and mRNAs were analysed separately, which also showed a heavy-tailed distribution reflecting a power-law distribution. Additionally, we observed a dissortativity pattern where high-degree nodes tend to connect with low-degree nodes, as demonstrated by the negative correlation observed in Figure 5.4. These characteristics indicate the presence of very few and highly connected nodes, hubs, which provide networks a higher robustness towards random disruption [4]. The observed heavy-tailed node degree distribution and dissortativity characteristics are phenomenon found in various biological networks. These results are consistent with previous studies in *A. thaliana* [124, 183].

We now focus on the simplified network for an in-depth assessment of the underlying regulatory interactions. As mentioned above, the simplified network consisted of 488 sub-networks, where the largest sub-network composed of 230 nodes and the remainder sub-networks are composed of less than 66 nodes, which implies uneven distribution of the number of nodes per sub-network. The sub-networks were ordered based on the number of nodes involved within them. On

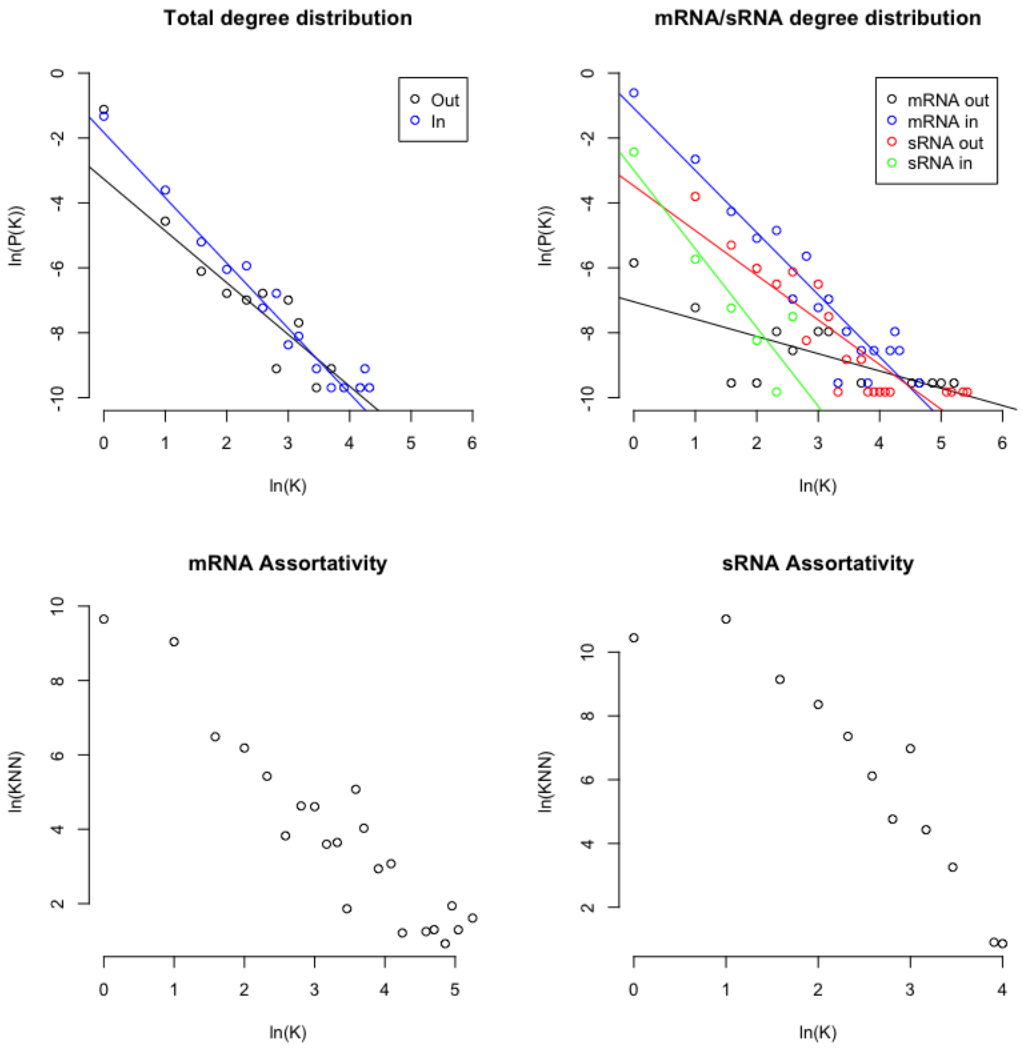


Figure 5.4 Degree distribution and assortativity in the simplified *A. thaliana* sRNA-mRNA network. The top row shows the degree distribution for all nodes in the network (left) and the individual nodes, i.e. sRNAs and mRNAs (right). The bottom row shows the assortativity for mRNAs and sRNAs respectively. K: node degree, p(K): the number of nodes with degree K divided by total nodes, and KNN: the average degree of the nearest neighbour for nodes with degree K.

the upper half appear the sub-networks with higher number of nodes, and higher number of validated interactions. On the other hand, the sub-networks on the lower half have less than three nodes and lower number of validated interactions. We also looked more closely at some of these sub-networks to investigate whether they were constructed randomly or if they had biological significance i.e they contain interactions that contribute to biological pathways. To do so, we selected the sub-

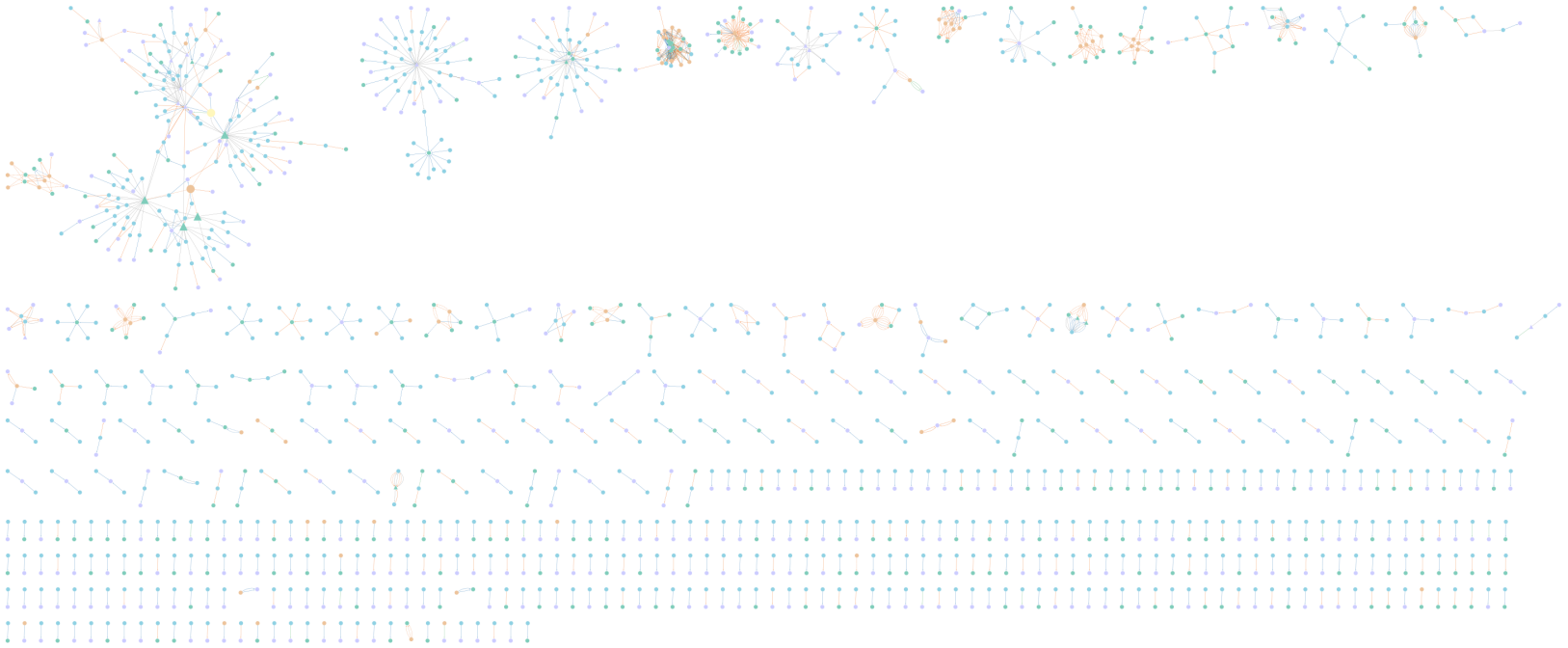


Figure 5.5 A representation of the sRNA-regulated network that was constructed from conserved interactions that were predicted by performing PAREsnip2 analysis with strict parameters on three *A. thaliana* replicates. The network was visualised using Cytoscape. Blue circles are sRNAs, and orange circles are validated miRNAs. Green and purple circles represent annotated and un-annotated target genes, respectively. Green and purple triangles represent annotated and un-annotated source genes, respectively. Grey edges are source interactions. Orange, green, and blue solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions.

networks that are composed of 20 or more nodes, and involved known miRNAs and validated edges (interactions), and then manually identified if they contained previously published biological pathways. For clarification, we removed isoforms of sRNAs from the selected sub-networks below. This step was carried out through closer inspection into the cluster of sRNAs that target the exact gene or group of genes, discarding the sRNAs that have an exact match to a shorter sequence within the cluster. We first selected the largest sub-network that appears on the top-left of Figure 5.5. This sub-network was uncovered when using PAREsnip2 and conservation filtering methods and it is shown in Figure 5.6. It involves miRNAs and sRNAs that contribute to the previously described TAS network that was verified in *A. thaliana* [43]. The degradome analysis supports that miR173 cleaves primary transcripts of the TAS1 family (TAS1A, TAS1B, and TAS1C) and TAS2 genes, which then leads to the cleavage of AT5G16640 and AT1G63080, pentatricopeptide repeat (PPR) transcripts, and AT1G63130, tetratricopeptide repeat (TPR) transcript, by TAS2-derived tasiR2140. Recent studies suggest that PPRs are loci for PPR-derived secondary siRNAs and they participate in post-transcriptional regulation of chloroplast and mitochondrial genes, this role might be a consequence of their importance to the sRNA network [81, 189]. Our sub-network also shows further 18 TAS2-derived sRNAs target PPR and TPR transcripts leading to PPR/TPR-derived sRNAs directed against PPR and TPR gene transcripts. Most of the targeted PPR members were validated as targets of sRNAs in the computational analysis and experimental validation that was performed in [7, 157]. This TAS2 cascade suggests that these sRNAs could be new potential tasi-RNAs and siRNAs. Moreover, within the same sub-network (Figure 5.6), we also find previously validated interactions including miRNAs targeting the PPR and TPR transcripts that are involved in the TAS2 cascade, these interactions are mediated by both forms of miR161 (miR161.1 and miR161.2) and miR400. There are also potential new interactions involving two members of the miR158 family, miR158a miR158b, targeting a PPR (AT1G62860)

and a TPR (AT3G15130) genes. The miR158-AT1G62860 categoty-2 interaction was previously predicted using psRNATarget [50] and was validated by modified 5' RLM-RACE in [181]. Finally, the sub-network shows three of the miR159 family members (miR159a, miR159b, and miR159c) target MYB33 and MYB65 genes, where these interactions were validated in [8]. miR159b was also predicted to target TCP24, which is a validated target for the miR319 family that is closely related in sequence with the miR159 family [138]. Also, in [163], they confirmed that several TCP genes in *A. thaliana* were regulated by miR319, and in our results, we have TCP2 and TCP24 cleaved by three members of miR319 family (miR319a, miR319b, and miR319c). The network interactions that we obtained in this work are consistent with the results published in [183]. For simplicity, we have omitted isomiRs and duplicate interactions from Figure 5.6, the detailed interactions of miR173/TAS cascade are shown in Appendix B Figure 1.

A visualisation of another large sub-network, presented in Figure 5.7, shows 10 targets that were identified for the highly similar miRNAs, miR156 and miR157. These targets are genes from the SQUAMOSA promoter binding protein-like (SPL) genes, which represent a family of transcription factors that are defined by a plant-specific DNA-binding domain. They have important regulatory roles in plant development, growth, and stress responses. The interactions between miR156/miR157 and 11 (out of 17) SPL genes have been verified in previous studies [67, 157]. Also, we predict that miR391 targets four of the SPL genes and to the best of our knowledge, these interactions were not identified in other studies, hence, these could be potentially new targets for miR391 as two of them show high confidence level (Category-0). miR391 was validated to target TAS3 in *A. thaliana* [6, 60], PRS3 (AT1G10700) in *A. thaliana*, but both genes could not be identified as targets for miR391 in this work. Furthermore, the sub-network in Figure 5.7 involves miR396, encoded by miR396a and miR396b, targeting six growth-regulating factor (GRF) genes and the basic helix–loop–helix transcription factor bHLH74 (AT1G10120).

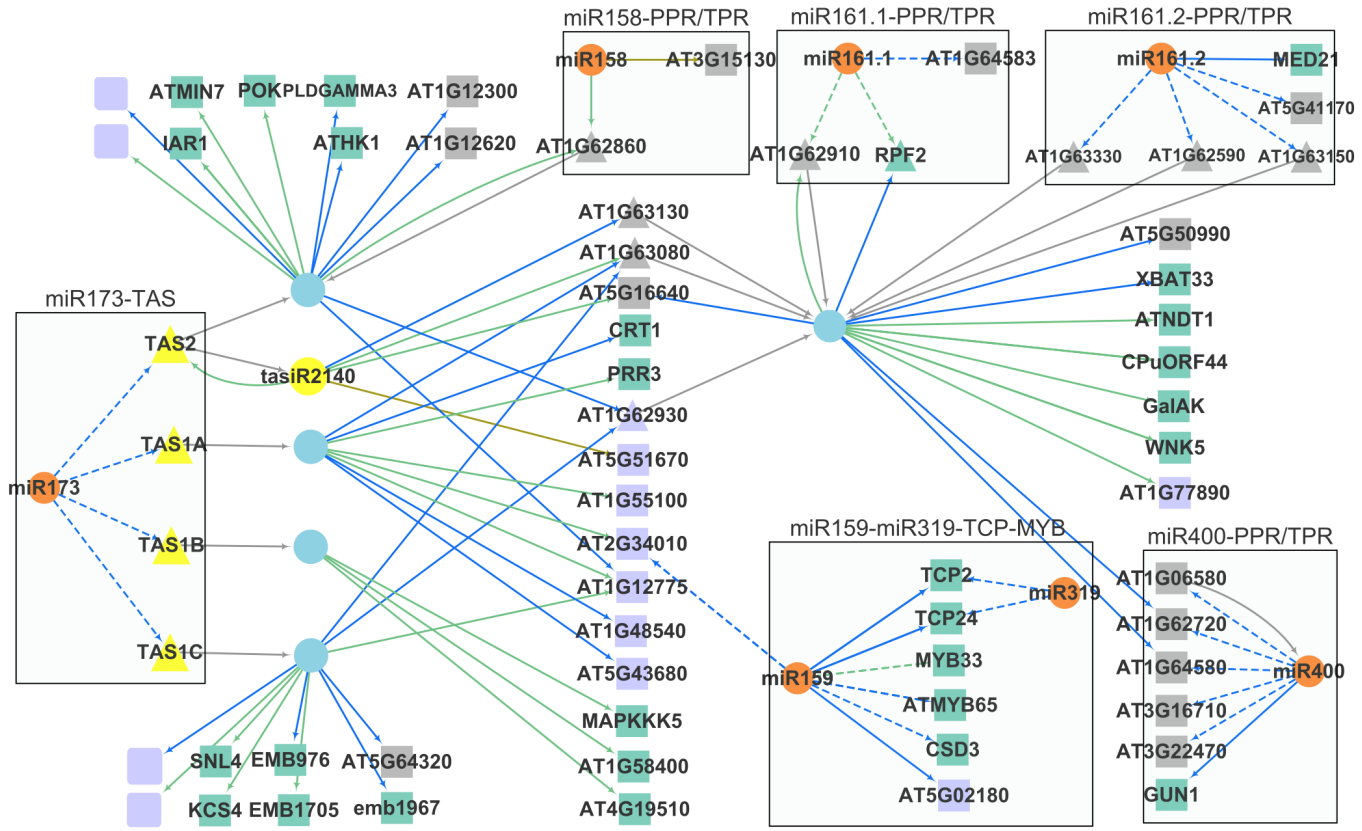


Figure 5.6 A visual representation of miR173/TAS regulatory network, the largest hub viewed in Figure 5.5. The other validated interactions and validated miRNAs that are present within the miR173/TAS sub-network, were grouped within the boxes. Orange circles are validated miRNAs, blue circles are grouped sRNAs, and the yellow circle is ta-siR2140. Yellow, grey, green, and purple rectangles/triangles represent TAS (Trans-acting small interfering RNAs), PPR/TPR (Pentatricopeptide/Tetratricopeptide repeat-like superfamily), annotated genes, and genes with no previous annotations, respectively. Rectangles are targeted genes, and triangles are targeted and source genes. Unlabelled purple rectangles are grouped un-annotated genes. Blue, brown, and green solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions. Details for all of the nodes and interactions can be found in Appendix B Figure 1 and Appendix B Table 1.

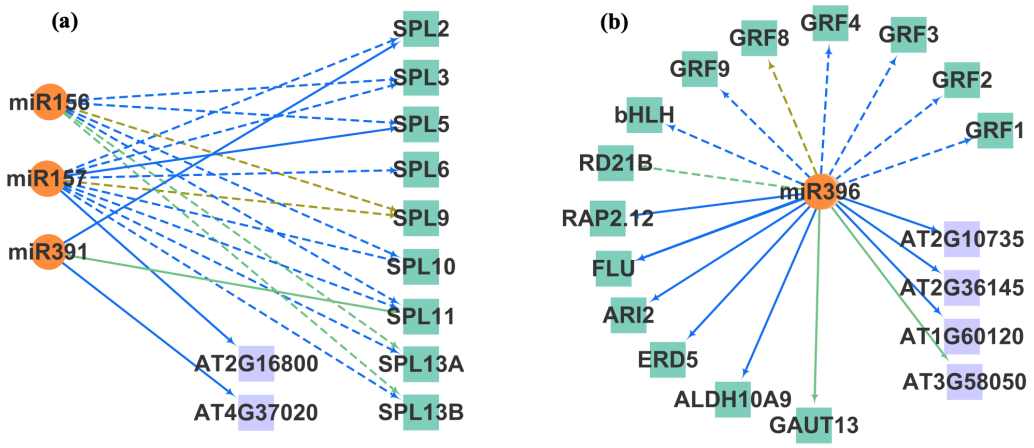


Figure 5.7 A visual representation of two sub-networks presented in Figure 5.5 and represent regulatory networks that involve: (a) validated mediated-interactions of miR156 and miR157 (controls proper development of lateral organs), and (b) validated mediated-interactions of miR396 (controls leaf development). Orange circles are validated miRNAs. Green, and purple rectangles/triangles represent annotated genes and genes with no previous annotations, respectively. Rectangles represent targeted genes. Blue, brown, and green solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions. Details for all of the nodes and interactions can be found in Appendix B Figures 2 and 3 and Appendix B Table 1.

In *A. thaliana*, miR396 plays an important role in regulating cell proliferation activity during leaf development by repression of the GRF genes, and regulating root growth by repression the expression of the bHLH74 gene [22, 52, 89]. Seven out of nine GRF genes (GRF1-4 and GRF7-9), and bHLH74 were validated as targets for miR396 in *A. thaliana*. Two other potential miR396 targets were AT5G43060, that encodes ESPONSIVE TO DEHYDRATION 21B (RD21B, also referred to by MMG4.7), and FLUORESCENT IN BLUE LIGHT (FLU), these targets were previously validated in [46, 200]. However, the miR396-FLU interaction does not appear as validated in Figure 5.7 due to the lack of this interaction from the list of experimentally validated interactions described in the previous section. Moreover, the potential interaction between miR396 and GAUT13 (AT3G01040), encodes a putative galacturonosyltransferase, was previously characterised in [115]. Furthermore,

there remain new five potential miR396 category-0 interactions and an interaction of category-2, which to the best of our knowledge have not previously identified.

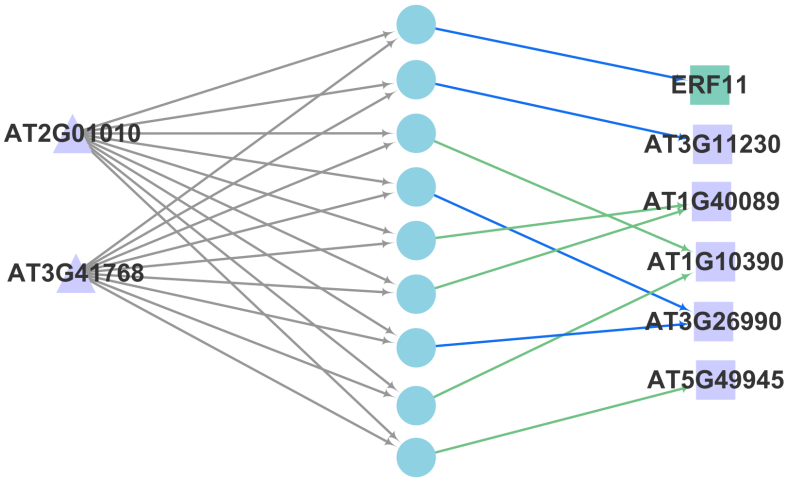
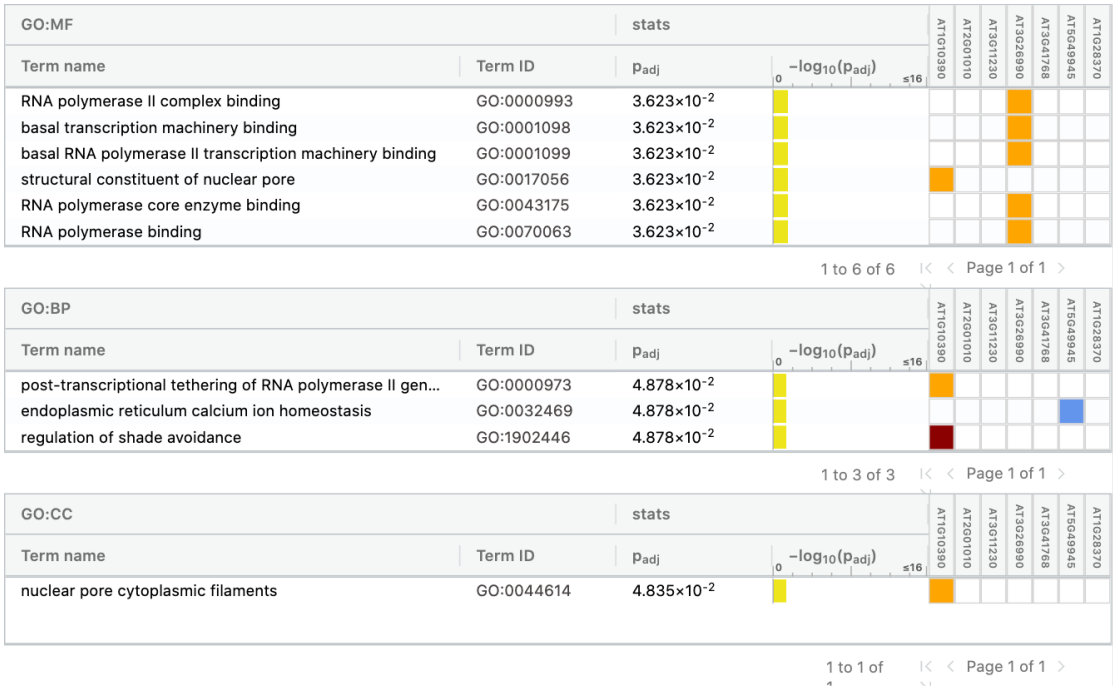


Figure 5.8 A visual representation of a sub-network that does not involve validated miRNAs nor validated interactions. Blue circles are sRNAs. Green, and purple rectangles/triangles represent annotated genes and genes with no previous annotations, respectively. Rectangles represent targeted genes and triangles are targeted and source genes. Blue, brown, and green solid edges are target interactions of categories 0, 1, and 2, respectively.

To further assess the regulatory contribution of the resulting sRNA mediated networks, further analysis can be performed on a sub-network that does not involve validated miRNAs or validated interactions (Figure 5.8). We performed functional enrichment analysis using g:profiler [150] for the gene group within the sub-network. In particular, we input the list of candidate genes that are shown in Figure 5.8 into g:GOSt tool (available on g:profiler web server) to determine if they map to known functional information sources and detects statistically significantly enriched biological processes and used the Benjamini–Hochberg False Discovery Rate as a correction method with significance threshold of 0.05. The enrichment analysis results are shown in Figure 5.9, which could indicate potential biological functions for the candidate genes. The candidate genes could then be investigated further and selected for experimental validation.



GO:MF

Term name	Term ID	stats p_{adj}	stats $-\log_{10}(p_{adj})$ (0 … ≤16)	AT1G10390	AT2G01010	AT3G11230	AT3G26990	AT3G41768	AT5G49945	AT1G28370
RNA polymerase II complex binding	GO:0000993	3.623×10^{-2}								
basal transcription machinery binding	GO:0001098	3.623×10^{-2}								
basal RNA polymerase II transcription machinery binding	GO:0001099	3.623×10^{-2}								
structural constituent of nuclear pore	GO:0017056	3.623×10^{-2}								
RNA polymerase core enzyme binding	GO:0043175	3.623×10^{-2}								
RNA polymerase binding	GO:0070063	3.623×10^{-2}								

1 to 6 of 6 Page 1 of 1

GO:BP

Term name	Term ID	stats p_{adj}	stats $-\log_{10}(p_{adj})$ (0 … ≤16)	AT1G10390	AT2G01010	AT3G11230	AT3G26990	AT3G41768	AT5G49945	AT1G28370
post-transcriptional tethering of RNA polymerase II gen...	GO:0000973	4.878×10^{-2}								
endoplasmic reticulum calcium ion homeostasis	GO:0032469	4.878×10^{-2}								
regulation of shade avoidance	GO:1902446	4.878×10^{-2}								

1 to 3 of 3 Page 1 of 1

GO:CC

Term name	Term ID	stats p_{adj}	stats $-\log_{10}(p_{adj})$ (0 … ≤16)	AT1G10390	AT2G01010	AT3G11230	AT3G26990	AT3G41768	AT5G49945	AT1G28370
nuclear pore cytoplasmic filaments	GO:0044614	4.835×10^{-2}								

1 to 1 of Page 1 of 1

Figure 5.9 Gene Ontology (GO) enrichment analysis performed by g:profiler [150] on the gene candidates involved in the sub-network that is shown in Figure 5.8.

In Chapter 4, we identified miRNA and miRNA* candidates and their predicted targets that were conserved among multiple *A. thaliana* biological replicates. The candidates were predicted by PAREsnip2 to have multiple Category 3 targeting signals. We investigated whether these candidates contribute to regulatory networks. The candidates' interactions were predicted using the less strict parameters that are described above. Through close inspection into the complex network in Figure 5.3, we extracted the candidates nodes manually from the network along with their second neighbour nodes. We also extracted the first neighbours of the gene nodes, the other sRNA nodes that target the same genes as the candidates. The extracted nodes and edges are disconnected from the main network and are shown in Figure 5.10. The mature miRNA shows two target sites of category-3 signal on two different genes. The first target is a novel chloroplast protein, SAFEGUARD 1 (SAFE1), which was shown to suppress singlet oxygen-induced stress responses in *A. thaliana* [187]. There are no studies suggest miRNA-mediated regulation of

SAFE1 protein, however, recent studies are exploring the link between miRNA biogenesis, particularly miRNA expression, and oxygen signaling pathways, which are suppressed by SAFE1 [23, 134]. The other target, encodes a receptor-like protein kinase (RLK4) that is expressed in roots [186]. In [198], they concluded that miRNAs mediate the regulation of leaf senescence in maize. On the contrary, the miRNA* have category-3 interactions with five genes, two of which are annotated. The first target gene is SYTF, a member of synaptotagmin-like (SYT) genes family that have been identified in *A. thaliana* [36, 47]. To the best of our knowledge, no studies were conducted for the interactions between sRNAs and SYT genes in *A. thaliana*. The other annotated gene is the ribosomal protein (RP) S6, which is involved in regulating growth processes in *A. thaliana* as suggested in previous studies [41, 48]. RPS6 has another potential category-3 interaction involving miR162, although the interactions between sRNAs and RPs have not been studied in plants, several studies in mammals suggested that RP targeted by miRNAs [5, 142, 156].

Finally, we investigated the t-plots that we produced for transcript genes within the network and we observed that transcripts with high confidence target interactions (category-0) tend to have lower degradome coverage than transcripts with lower confidence interactions (category-2) (see Figure 5.12). Consequently, we hypothesise that there could be a relation between the transcript coverage, represented by the number of degraded fragments that map to a transcript, and interaction confidence. We further pursued this observation by comparing the transcript degradation coverage with the number of interactions reported (Figure 5.11) and we found that using the PAREsnip2 p-value and MFE filters reduces the number of interactions while retaining experimentally validated interactions. We also found that using interaction conservation between PAREsnip2 analysis of biological replicates further reduce the number of hits per transcript while retaining most of validated interactions. This approach could provide a further validation layer for the degradome analysis results.

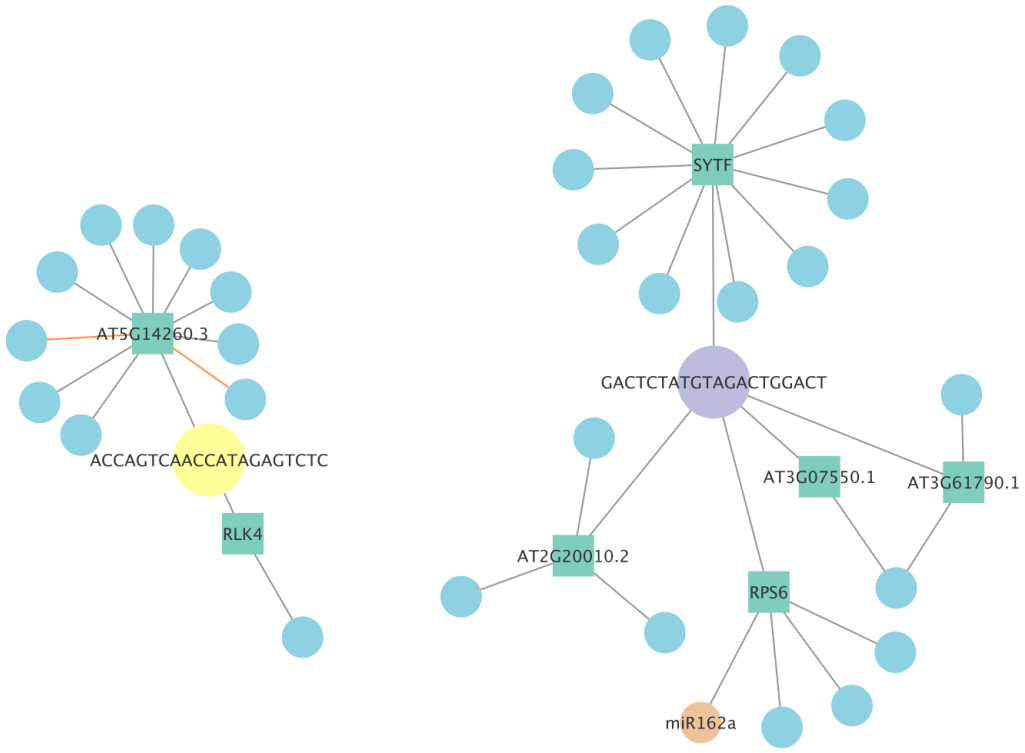


Figure 5.10 A visual representation of the partial network that involved the predicted miRNA/miRNA* candidates. Large yellow circle is the candidate miRNA, large purple circle is the candidate miRNA*, orange circle is validated miRNA, and blue circles are sRNAs. Green squares are targeted genes. Orange and grey solid edges are target interactions of categories 0 and 3, respectively.

5.5 Discussion

Prior to the introduction of PARE analysis [1, 72], sequence complementarity between the sRNA and the mRNA was used to identify cleavage sites within the transcripts. The NGS techniques has advanced tremendously over the last decade and using the degradome allowed us to obtain a higher level of confidence in the predicted interaction. Although the degradome provide an element of validation to the target prediction, using the degradome alone was not enough to reveal the biological functions of the massive cloud of sRNA regulatory interactions. It is increasingly recognized that the need of additional methods of filtration is necessary to construct networks that are easier to interpret [183].

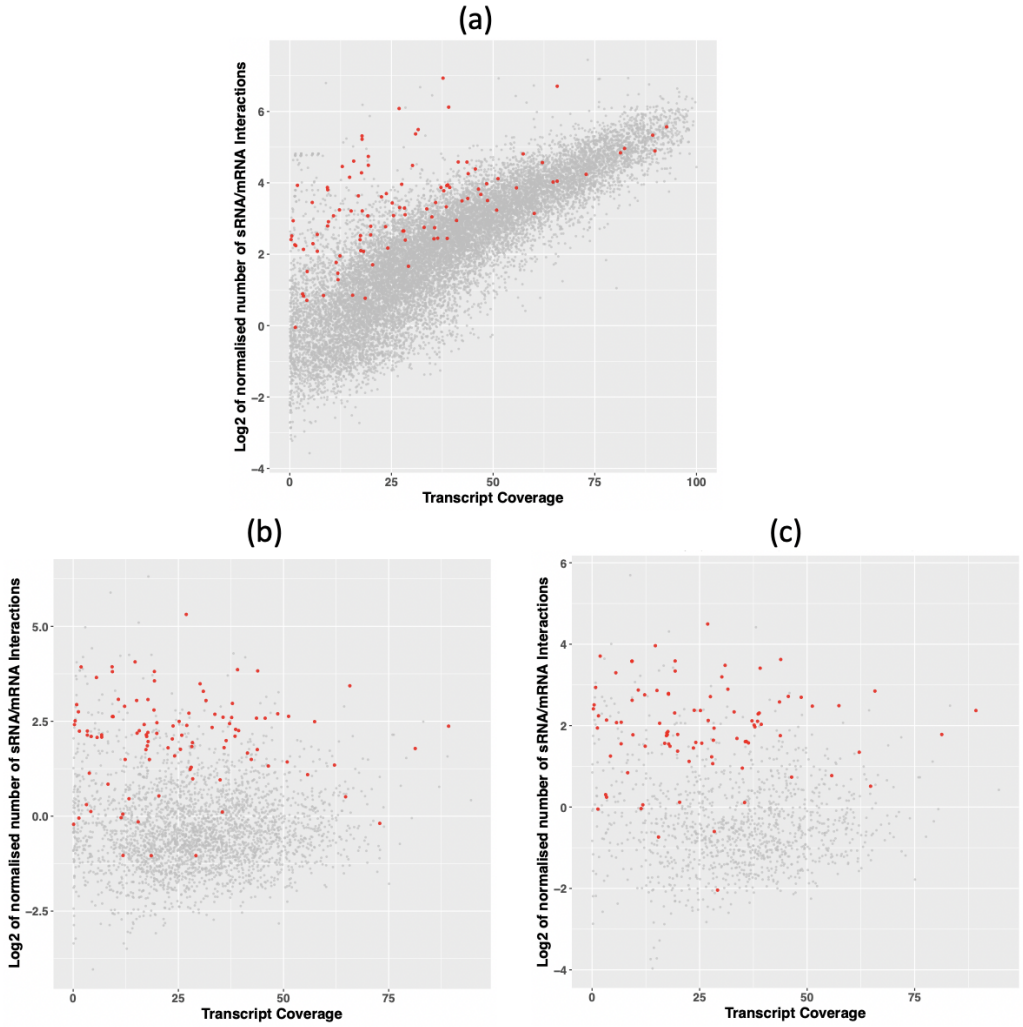


Figure 5.11 Transcript/degradome coverage analysis. Plots show a progressive reduction in the number of retained interactions from PAREsnip2 analysis after using filtering techniques and replicate conservation. a) Shows analysis without the use of p-value and MFE filtering or conservation. b) Shows analysis with p-value and MFE filtering and no conservation. c) Shows analysis with p-value and MFE filters and conservation of interactions obtained from two degradome analyses of *A.thaliana* biological replicates. For all plots, data points represent transcripts. A red circle data point contains a validated sRNA/mRNA interaction.

To investigate this, we performed degradome analysis using PAREsnip2 to compare the output of applying less strict parameters with the output of applying strict parameters. The less strict parameters produce interactions with weak cleavage signals on the transcripts, i.e. category-3, in addition to applying filtration steps provided by PAREsnip2, such as MFE and p-value filters. On the other hand, the

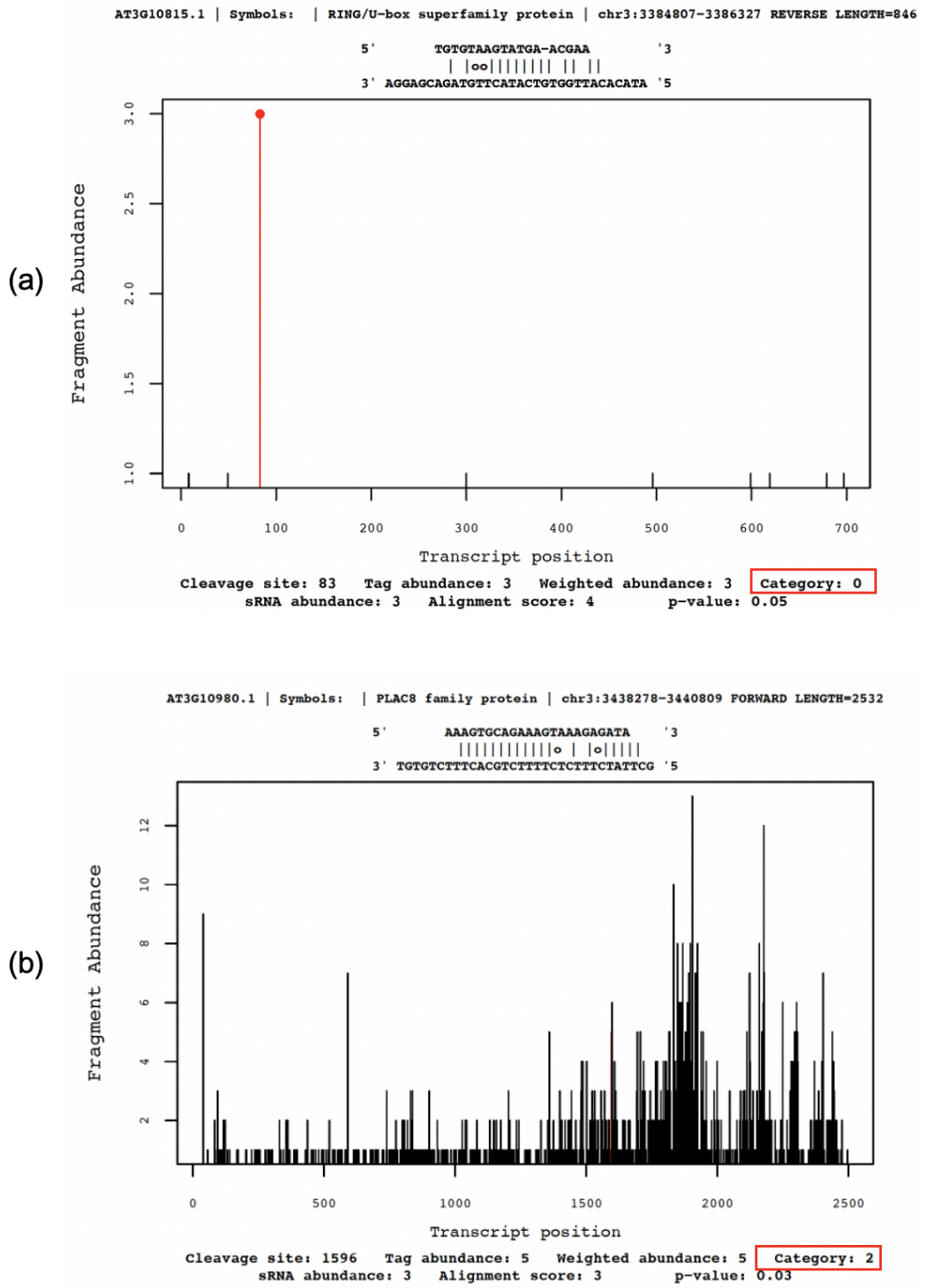


Figure 5.12 Example transcript t-plots showing (a) low transcript coverage, and (b) high transcript coverage.

strict parameters generate interactions with strong cleavage signals on the transcript, i.e. categories 0,1 and 2, with the application of MFE and p-value filters. We then

evaluated the performance of each of the parameter sets and calculated the sensitivity and precision for all the predictions using *A. thaliana* datasets. We presented how the sensitivity of the output was maintained for the less strict and strict parameters across all the datasets, while the precision increased in the interactions set that was resulted from using the strict parameters. To that end, using degradome data and applying PAREsnip2 filters results in a higher precision output that could facilitate a more detailed overview of regulatory interactions and an in-depth assessment of the underlying sRNA regulatory networks.

The complex network constructed from the sRNA-mRNA interactions, that were produced using the less strict parameters, was large, complex and difficult to elucidate. We speculate that the cause of this complexity is due to the higher proportion of putative false positive predictions, i.e. the Category-3 interactions consist of both genuine cleavage sites and random degradation, which is evidenced by the low abundance of the degradation signals on the transcript. Conversely, the set of interactions that we obtained from using the strict parameters allowed us to construct a simpler sRNA-mediated regulatory network with higher precision of predicted interactions, which could help to understand and elucidate the biological information within the network. Accordingly, as observed in our results, PAREsnip2 filtration methods and a conservation approach facilitated the identification of previously validated sRNAs regulatory cascade, such as TAS, PPR, GRF, and SPL networks, that were hidden within the unfiltered complex network.

The representation of sRNA-mediated regulatory networks enables the examination of its mathematical structural properties. Analysing the topology and structural characteristics of such networks has uncovered intriguing similarities among diverse biological systems. In the results, we compared between the complex and the simplified *A. thaliana* networks using the structural features. We observed that the number of sub-networks in the simplified network was increased when compared to the complex network, this could be due to the reduced number of predicted interactions

in the simplified network, which led to disconnecting the 'hairball' component that was present in the complex network. Also, there was a reduction of the average number of neighbours in the simplified network, this might also be related to the reduction of predicted interactions, which led to the reduction in the nodes degree. Although the clustering coefficient in other biological networks tends to be higher than the random networks, the sRNA networks presented in the results showed an extremely low value. Perhaps one reason for this may be to do with the nature of interactions between the nodes, i.e. an sRNA node interacts with multiple gene nodes (neighbours), however, the neighbour nodes have no interactions between them. To that end, some of the proprieties of the simplified sRNA-mRNA network may not be relative to other biological networks, yet, the network tends to show a real biological entity and that its interactions were not assigned randomly.

Degradome sequencing enabled the development of a variety of computational methods to identify sRNA-mRNA interactions on a genome-wide scale. In this work, we chose PAREsnip2 as it showed improved prediction performance when compared to other computational methods in the area, including: CleaveLand4, sPARTA, and PAREsnip. The accuracy of the output of a sRNA target prediction tool can be assessed through benchmarking against other tools' predictions. One way to benchmark different prediction tools is using a set of experimentally validated miRNA-mRNA interactions to perform comparative approach based on metrics, such as sensitivity, specificity, precision, and accuracy. Correspondingly, computational performance benchmarking can lead to the development of more accurate and reliable sRNA-mRNA prediction methods, which could help to advance our understanding of sRNA-mediated regulatory networks.

Furthermore, the inclusion of differential expression analysis can be used to assess the performance of a prediction method. It may also provide further indication into the function of specific sRNAs when combined with degradome analysis. For instance, if a sRNA is determined to be up-regulated in a given sample, a targeted

mRNA for this sRNA is identified using degradome analysis, and the mRNA is determined be down-regulated, this may support the prediction of sRNA target as sRNAs act as negative regulators for the genes.

The wide scale studies of sRNA regulatory networks are limited due to the challenges posed by the high false positive rate in the sRNA target predictions produced by the computational tools. PAREnet, utilise degradome analysis to provide an approach into generating more confident sRNA-mRNA and easier to elucidate networks. Beside *A. thaliana*, the tool could be used to construct networks for various plant genomes, such as rice, tomato, potato, maize, and wheat. Due to time constraint and the limited availability of degradome datasets for some plant genomes, network analysis was not performed on these species datasets. However, with the advance in NGS, degradome datasets are becoming more available, and thus, future work will focus on more comprehensive analysis using exploration and comparative approach on more plant datasets networks. In conclusion, our approach enables the identification of potentially new interactions that might be of interest to investigate and experimentally validate, and it may open new directions of research towards sRNA mediated regulation of mRNAs in plants.

Chapter 6

Future Work and Conclusion

6.1 Summary

In this thesis, we have provided an introduction into sRNA biology and RNA silencing in plants. We then presented an overview of the computational methods used for analysing sRNA, including the important class of miRNA, and degradome. We developed a new tool that is based on a new approach for miRNA classification that combines biogenesis and functional criteria. We also introduced a tool that enable the visualisation and analysis of sRNA-meditated regulatory networks using degradome. In this chapter, we shall discuss some possible extensions to this work.

6.2 Future work

6.2.1 Micro RNA prediction

The miRCat2 parameter search algorithm that was described in Chapter 4 was used to explore less stringent miRCat2 parameters to detect miRNAs with extreme biogenesis characteristics in *A. thaliana*, and we succeeded in identifying candidate novel miRNAs. Due to their low abundance (compared to known miRNAs), we assumed

these miRNA candidates are species-specific, and the non-conserved miRNAs are typically less expressed than conserved miRNAs. Also, some miRNAs are condition-specific that are specifically expressed in a particular developmental stage, tissue, or stress response condition [54, 123, 151, 210]. We hypothesise that these miRNAs could have extreme biogenesis features, such as high number of mismatches and bulges within miRNA/miRNA* duplex, particularly in less-studied genomes. Further parameter search experiments, such as those detailed in Chapter 4, could be applied to investigate if our algorithm behaves differently on different genomes, tissues, or stress conditions. Candidate miRNAs might be detected and they could be analysed as with the methods described in Chapter 4.

Furthermore, the intense study of miRNAs has led to a steady increase in available miRNA repositories that archive miRNA biogenesis and functional information. miRbase is one of the most common repositories and it was used to annotated miRNAs in this work. A recent miRNA repository was introduced, PmiREN, which accepts miRNA entries based on the newly suggested miRNA biogenesis criteria. PmiREN presents 16 more miRNAs in *A. thaliana* than miRBase, which is a high number considering the well studied genome model. However, not all miRNA entries in miRBase were present in PmiREN due to their filtering criteria. Accordingly, the parameter search algorithm in Chapter 4 could be performed using the PmiREN miRNAs, or even combine miRNAs from both repositories.

6.2.2 sRNA-mediated regulatory networks

- **Network construction for other species**: *A. thaliana* is a heavily studied plant, therefore, the majority of the components in the constructed network included previously described interactions. Further work could be performed on less annotated plant genomes that have sequenced degradome data, such as rice and tomatoes [119, 197]. It would be interesting to see if the constructed network

of a less annotated genome would have similar structural features to the network in Chapter 5 (e.g. the composition of sub-networks). If unknown sub-networks were consistent between multiple networks from different replicates, regulatory functions could be suggested for the un-annotated sRNAs/genes involved, thus, the sub-networks could be investigated further and some could be selected for experimental validation.

- **Network construction for samples under stress condition**: The complex regulatory networks of sRNAs in response to stress need to be elucidated. Therefore, our work could also be extended by constructing networks for control and stress treated samples and applying GO and KEGG enrichment analysis. In [70, 114, 196], they performed a comprehensive integrated analysis to provide better insights into the regulatory network components associated with stress conditions. The samples that were analysed in the studies could be analysed through PAREnet in order to visualise their networks, which could provide a different approach into comparing between the samples by addressing the differences in the networks components.

- **Annotation of other sRNA classes**: For our analysis in Chapter 5, we focused exclusively on miRNA annotations within the predicted interactions. Thus, we ignored investigating other important sRNA classes, such as phasiRNA and tasiRNA, that have been identified as components of the reguolatory networks associated with biological processes [1, 116, 183]. Future work could involve the annotation of other sRNA classes, which in turn would provide a more complete picture of the sRNA networks.

6.2.3 UEA sRNA Workbench

The UEA sRNA Workbench source code is available on GitHub, enabling the bioinformatics community to make their contribution to the software package. Here

we suggest some future amendments that could be implemented to improve the performance of the tools that we have developed to make most of the Workbench package:

- **Potential improvements to PAREfirst**: Currently, PAREfirst analysis can only be performed on one replicate at a time. Using multiple datasets as input for PAREfirst and applying conservation approach between these datasets would enhance the confidence level within the predicted results. Moreover, PAREfirst enables the prediction of miRNAs with extreme biogenesis due to its degradome-assisted approach that allow to explore less strict miRNA biogenesis in a controllable manner. Therefore, PAREfirst could be improved to allow it to identify potential miRNA-like miRNAs that are derived from known miRNA precursors, and provide this information within the results.

- **Potential improvements to PAREnet**: Similar to PAREfirst, PAREnet only accepts one sRNA and one degradome datasets, or one PAREsnip2 results file at a time to generate the network interactions. Enabling the input of multiple datasets of a single genome will help provide additional confidence to the predicted interactions. Moreover, we observed in Chapter 5 that the network involved sRNA nodes that have multiple isoforms and filtering the isoforms will make the network components simpler. Therefore, a new step could be implemented to filter the isoforms. The suggested feature could be implemented to filter the set of sRNAs that have the same target or set of targets and map them to the genome. If the reads aligned to one locus in the genome, one read with the longer sequence is selected and the other are discarded. A further addition that could be implemented into the tool would be the annotation of other sRNA classes such as tasiRNAs. Other technical features that we could improve include allowing PAREnet to be accessible for many users by enabling input files from other degradome-assisted target

prediction methods, such as CleaveLand and sPARTA. Additionally, a GUI version of PAREnet could be developed and Cytoscape.js, an open-source graph library, could be implemented as a plug-in to visualise the networks in the GUI interface where PAREnet could provide a JSON file as an output that could be used for visualising networks via Cytoscape.js. Finally, t-plot viewer feature could be implemented within the GUI version of PAREnet where users could view t-plots of the genes within the visualised network.

- **Parameter-search method**: The less-strict miRNA parameters that were produced for Chapter 4 analysis were chosen based on *A. thaliana* datasets and miRBase v22 entries. As the understanding of miRNA is continuously evolving and changing, the parameter-search method (described in Chapter 4) could be implemented into the UEA sRNA Workbench as a new feature that enables the users to extract new miRNA features based on their datasets and updated set of validated miRNA entries.

6.3 Thesis conclusion

With the development and advancement of NGS technology, a wide range of species, tissues and conditions sequencing datasets have become available for degradome analysis. For this reason, broad scale studies of sRNA mediated regulation in less-studied genomes, other than the model organism *A. thaliana*, are becoming possible. Also, the study of sRNAs and their regulatory roles within the biological networks became more accessible due to the availability of a variety of bioinformatics techniques that can process the enormous sequencing data with lower resources. However, the scope of these studies is limited due to challenges posed by the high false positive rate of bioinformatics predictions of sRNA activities.

A significant body of research is dedicated to understanding the complexity of miRNA biogenesis and mechanism of functioning, to which new discoveries are still being added. The expression of many miRNAs and their targets are localized both temporally and spatially, specific to factors such as species, tissue, growth, and stress conditions, yet, testing for species-specific or condition-specific miRNAs is a challenge and a common goal in this field. We showed here that we could systematically explore a wider range of miRNA biogenesis features and define less strict parameters that would have an impact on miRNA prediction. However, we should keep the balance between predicting miRNAs with novel features and predicting overestimated miRNAs. This balance was achieved by utilising degradome data, which added a further element of validation into the miRNA prediction.

The development of bioinformatics tools enabled the processing of large-scale biological datasets. Evaluating these tools and validating their predictions are crucial steps in assessing the accuracy of their predictions and determining whether they reflect biological reality. There are several methods to validate and assess the predictions of bioinformatics tools, including experimental validation, differential expression analysis, and benchmarking against other computational methods. In the case of miRNA prediction, experimental validation, such as RNA gel blotting, is the most direct and reliable method to assess the accuracy of miRNA predictions. On the other hand, experimental validation for sRNA target predictions is carried out through 5' RACE. By confirming the computational predictions using experimental techniques, we can gain insights into the underlying mechanisms of miRNA-mediated gene regulation and improve our understanding of biological systems. Another performance assessment method is the differential expression analysis between wild-type and mutant datasets. To validate miRNA predictions the fold change is calculated to determine if predicted miRNAs have increased expression in the wild-type data when compared to the expression in the mutant dataset. Similarity, if the sRNA, within a sRNA-mRNA interaction, is significantly up-regulated in

a sRNA dataset, and the predicted target of that particular sRNA is significantly down-regulated in gene expression data, it may provide evidence that the prediction method is biologically relevant. Further performance assessment method, is the benchmarking against other computational methods, which is a useful approach to assess the accuracy of a bioinformatics tool. Comparing the performance of different tools can help identify the strengths and weaknesses of each method and provide insight into how well they can generate predictions. One way to benchmark the different tools is to use an experimentally validated data, i.e. a set of validated miRNAs (can be obtained from miRBase) and a set of experimentally validated sRNA-mRNA interactions (can be obtained from miRTarBase). The performance of these methods can then be compared based on metrics such as sensitivity, specificity, precision, and accuracy.

Our work moves toward the aim of identifying miRNAs, their target mRNAs, and their contribution to the regulatory biological networks. We developed an exploratory approach of identifying underestimated miRNAs, and the construction of a genome-level network of interactions between sRNAs and transcripts, which allowed a visualization of high-level interactions of sRNA regulatory cascades. We hope that the methods described in this thesis will have an impact on enriching the literature of miRNA biogenesis and function and enable us to get closer to understand the complexity of genes networks in plants. This could have applications to areas that have potentially important implications for agriculture such as improving crop production and resistance to plant stress conditions.

Appendix A

Some of the tables referenced within Chapter 4 contain a large number of rows, such as predicted miRNAs, that are not practical to include within this thesis. However, for completeness, a brief description of each table is provided below, and the actual data can be found on the Computer Society Digital Library at the following url: https://doi.org/10.1109/TCBB.2021.3115023.

Appendix A File 1 PAREfirst tool documentation.

Appendix A Table 1 contains the 150 parameter sets (denoted as EPS) that were obtained from the parameter search method.

Appendix A Table 2 contains the differential expression analysis results between wild-type and DCL1-mutants sRNA libraries using DESeq2 within iDep9.0.

Appendix A Table 3 contains the differential expression analysis results between wild-type and DCL4-mutants sRNA libraries using DESeq2 within iDep9.0.

Appendix A Table 4 contains the functional miRNA with hairpin predictions. The hairpins in this table were predicted by miRCat2 using EPS.

Appendix A Table 5 contains the enriched predicted functional miRNAs with hairpins that were predicted in at least two out of three wild-type replicates. The miRNA hairpins were predicted by miRCat2 using EPS. Also, included are the PAREsnip2 targeting interaction predictions for the novel miRNA and miRNA* candidates.

Appendix A Table 6 contains the output of running miRCat2 with the UPS parameters on the flower sample GSM707678, leaf sample GSM707679, root sample GSM707680, and the seedling sample GSM707681.

Appendix A Table 7 contains the suggested annotation for miRNA* sequences in miRBase.

Appendix A Table 8 contains the output of running PAREfirst with the EPS parameters on wild-type replicates.

Appendix A Table 9 contains the output of running miRCat2 with the UPS parameters on wild-type replicates.

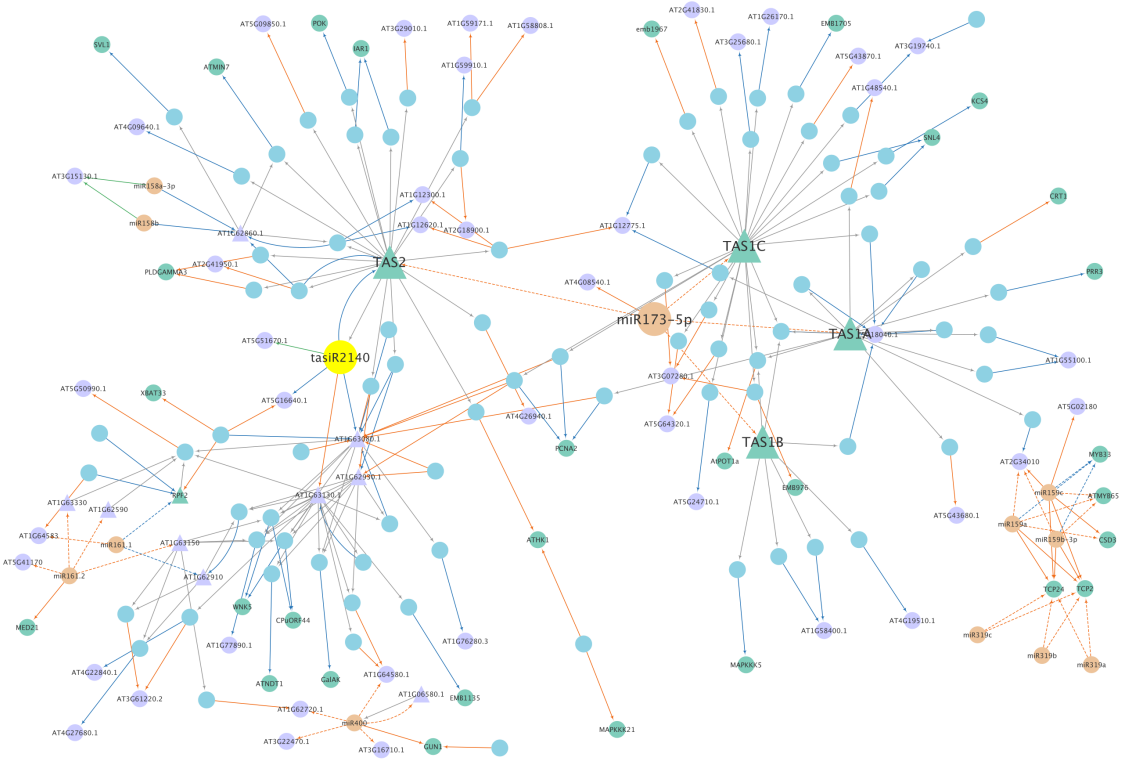
Appendix A Table 10 contains the output of running miRDeep-P2 with the default settings on wild-type replicates.

Appendix B

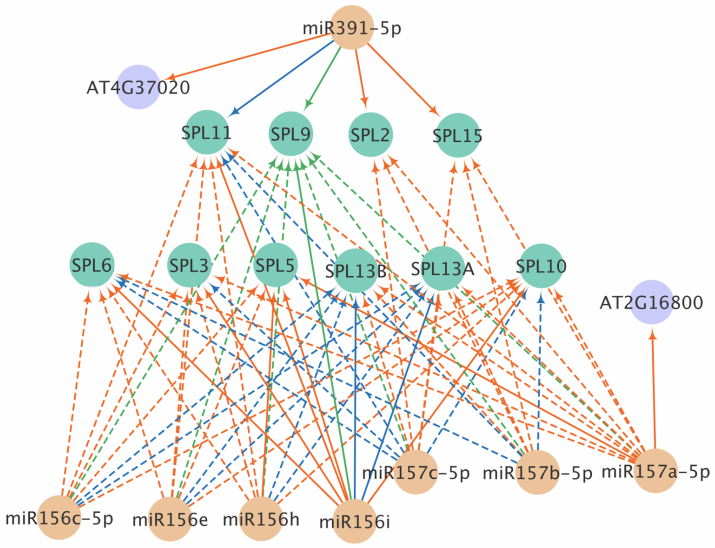
The tables referenced within Chapter 5 contains a large number of rows, representing the sRNA-mRNA interactions, that are not practical to include within this thesis. However, for completeness, a brief description of each table is provided below and the actual data is provided as supplementary information included with the thesis.

Appendix B File 1 Source code (in Java) for the implemented PAREnet tool, which is a part of the UEA sRNA Workbench.

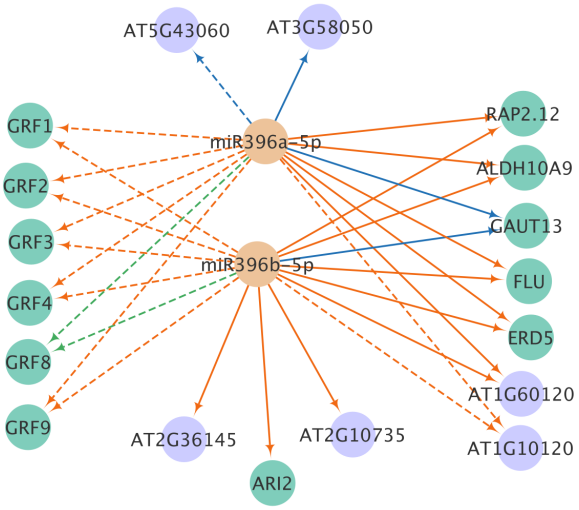
Appendix B Table 1 The results from PAREnet analysis on three *A. thaliana* wild-type replicates using strict PAREsnip2 parameters.



Appendix B Figure 1 A visual representation of detailed miR173/TAS regulatory network, the largest hub viewed in Figure 5.5. Blue circles are sRNAs, orange circles are validated miRNAs, the large orange circle is miR173, and the large yellow circle is ta-siR2140. Green and purple circles represent annotated and un-annotated target genes, respectively. Green and purple triangles represent annotated and un-annotated source genes, respectively. Large green triangles are TAS1A, TAS1B, TAS1C, and TAS2. Grey edges are source interactions. Orange, green, and blue solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions.



Appendix B Figure 2 A visual representation of the detailed sub-network presented in Figure 5.7(a) and represent a regulatory network that involves validated mediated-interactions of miR156 and miR157 (controls proper development of lateral organs). Blue circles are sRNAs, and orange circles are validated miRNAs. Green and purple circles represent annotated and un-annotated target genes, respectively. Green and purple triangles represent annotated and un-annotated source genes, respectively. Grey edges are source interactions. Orange, green, and blue solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions.



Appendix B Figure 3 A visual representation of a sub-network presented in Figure 5.7(b) and represent a regulatory network that involves validated mediated-interactions of miR396 (controls leaf development). Blue circles are sRNAs, and orange circles are validated miRNAs. Green and purple circles represent annotated and un-annotated target genes, respectively. Green and purple triangles represent annotated and un-annotated source genes, respectively. Grey edges are source interactions. Orange, green, and blue solid edges are target interactions of categories 0, 1, and 2, respectively. Dashed edges are validated target interactions.

References

[1] Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the *Arabidopsis* Degradome. *Current Biology*, 18(10):758–762.

[2] Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics (Oxford, England)*, 25(1):130–131.

[3] Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouché, N., Gasciolli, V., and Vaucheret, H. (2006). DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Current biology: CB*, 16(9):927–932.

[4] Albert, R. (2005). Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957.

[5] Alkhatabi, H. A., McLornan, D. P., Kulasekararaj, A. G., Malik, F., Seidl, T., Darling, D., Gaken, J., and Mufti, G. J. (2016). RPL27A is a target of miR-595 and may contribute to the myelodysplastic phenotype through ribosomal dysgenesis. *Oncotarget*, 7(30):47875.

[6] Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants. *Cell*, 121(2):207–221.

[7] Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H., Spatafora, J. W., and Carrington, J. C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nature Genetics*, 36(12):1282–1290.

[8] Allen, R. S., Li, J., Alonso-Peral, M. M., White, R. G., Gubler, F., and Millar, A. A. (2010). MicroR159 regulation of most conserved targets in *Arabidopsis* has negligible phenotypic effects. *Silence*, 1(1):1–18.

[9] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[10] Alzahrani, S. (2015). Generating and Comparing Small-RNA Networks in Plants. Master's thesis.

[11] Alzahrani, S., Applegate, C., Swarbreck, D., Dalmay, T., Folkes, L., and Moulton, V. (2021). Degradome Assisted Plant MicroRNA Prediction under Alternative Annotation Criteria. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP.

[12] An, J., Lai, J., Lehman, M. L., and Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research*, 41(2):727–737.

[13] An, J., Lai, J., Sajjanhar, A., Lehman, M. L., and Nelson, C. C. (2014). miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics*, 15(1):275.

[14] Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New biotechnology*, 25(4):195–203.

[15] Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007). The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. *Science*, 318(5851):761–764.

[16] Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

[17] Aukerman, M. J. and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *The Plant Cell*, 15(11):2730–2741.

[18] Axtell, M. J. (2013). Classification and Comparison of Small RNAs from Plants. *Annual Review of Plant Biology*, 64(1):137–159.

[19] Axtell, M. J., Jan, C., Rajagopalan, R., and Bartel, D. P. (2006). A two-hit trigger for siRNA biogenesis in plants. *Cell*, 127(3):565–577.

[20] Axtell, M. J. and Meyers, B. C. (2018). Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *The Plant Cell*, 30(2):272–284.

[21] Bansal, S., Khandelwal, S., and Meyers, L. A. (2009). Exploring biological network structure with clustered random networks. *BMC bioinformatics*, 10(1):1–15.

[22] Bao, M., Bian, H., Zha, Y., Li, F., Sun, Y., Bai, B., Chen, Z., Wang, J., Zhu, M., and Han, N. (2014). miR396a-mediated basic helix–loop–helix transcription factor bHLH74 repression acts as a regulator for root growth in *Arabidopsis* seedlings. *Plant and Cell Physiology*, 55(7):1343–1353.

[23] Barczak-Brzyżek, A., Brzyżek, G., Koter, M., Siedlecka, E., Gawroński, P., and Filipecki, M. (2022). Plastid retrograde regulation of miRNA expression in response to light stress. *BMC plant biology*, 22(1):1–15.

[24] Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297.

[25] Baulcombe, D. (2004). RNA silencing in plants. *Nature*, 431(7006):356–363.

[26] Baulcombe, D. (2006). Short Silencing RNA: The Dark Matter of Genetics? *Cold Spring Harbor Symposia on Quantitative Biology*, 71(0):13–20.

[27] Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis (New York, N.Y.: 2000)*, 53(8):474–485.

[28] Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366.

[29] Bologna, N. G., Schapire, A. L., Zhai, J., Chorostecki, U., Boisbouvier, J., Meyers, B. C., and Palatnik, J. F. (2013). Multiple rna recognition patterns during microrna biogenesis in plants. *Genome research*, 23(10):1675–1689.

[30] Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*, 26(12):1566–1568.

[31] Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics (Oxford, England)*, 20(17):2911–2917.

[32] Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in *Arabidopsis*. *Cell*, 123(7):1279–1291.

[33] Brodersen, P. and Voinnet, O. (2006). The diversity of RNA silencing pathways in plants. *Trends in genetics: TIG*, 22(5):268–280.

[34] Brousse, C., Liu, Q., Beauclair, L., Deremetz, A., Axtell, M. J., and Bouche, N. (2014a). A non-canonical plant microRNA target site. *Nucleic acids research*, 42(8):5270–5279.

[35] Brousse, C., Liu, Q., Beauclair, L., Deremetz, A., Axtell, M. J., and Bouché, N. (2014b). A non-canonical plant microRNA target site. *Nucleic Acids Research*, 42(8):5270–5279.

[36] Cabanillas, D. G., Jiang, J., Movahed, N., Germain, H., Yamaji, Y., Zheng, H., and Laliberté, J.-F. (2018). Turnip mosaic virus uses the SNARE protein VTI11 in an unconventional route for replication vesicle trafficking. *The Plant Cell*, 30(10):2594–2615.

[37] Candar-Cakir, B., Arican, E., and Zhang, B. (2016). Small RNA and degradome deep sequencing reveals drought-and tissue-specific microRNAs and their important roles in drought-sensitive and drought-tolerant tomato genotypes. *Plant biotechnology journal*, 14(8):1727–1746.

[38] Carpentier, M.-C., Bousquet-Antonelli, C., and Merret, R. (2021). Fast and Efficient 5'P Degradome Library Preparation for Analysis of Co-Translational Decay in *Arabidopsis*. *Plants (Basel, Switzerland)*, 10(3).

[39] Carrington, J. C. and Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science*, 301(5631):336–338.

[40] Chapman, E. J. and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nature Reviews. Genetics*, 8(11):884–896.

[41] Chen, G.-H., Liu, M.-J., Xiong, Y., Sheen, J., and Wu, S.-H. (2018). TOR and RPS6 transmit light signals to enhance protein translation in deetiolating *Arabidopsis* seedlings. *Proceedings of the National Academy of Sciences*, 115(50):12823–12828.

[42] Chen, H.-M., Chen, L.-T., Patel, K., Li, Y.-H., Baulcombe, D. C., and Wu, S.-H. (2010). 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proceedings of the National Academy of Sciences*, 107(34):15269–15274.

[43] Chen, H.-M., Li, Y.-H., and Wu, S.-H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, 104(9):3318–3323.

[44] Chen, X. (2004). A MicroRNA as a Translational Repressor of APETALA2 in *Arabidopsis* Flower Development. *Science*, 303(5666):2022–2025.

[45] Chen, X. (2005). MicroRNA biogenesis and function in plants. *FEBS letters*, 579(26):5923–5931.

[46] Chorostecki, U., Crosa, V. A., Lodeyro, A. F., Bologna, N. G., Martin, A. P., Carrillo, N., Schommer, C., and Palatnik, J. F. (2012). Identification of new microRNA-regulated genes by conserved targeting in plant species. *Nucleic Acids Research*, 40(18):8893–8904.

[47] Craxton, M. (2004). Synaptotagmin gene content of the sequenced genomes. *BMC genomics*, 5(1):1–14.

[48] Creff, A., Sormani, R., and Desnos, T. (2010). The two *Arabidopsis* RPS6 genes, encoding for cytoplasmic ribosomal proteins S6, are functionally equivalent. *Plant molecular biology*, 73(4):533–546.

[49] Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *The Plant Cell*, 23(2):431–442.

[50] Dai, X. and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic acids research*, 39(suppl_2):W155–W159.

[51] Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

[52] Debernardi, J. M., Rodriguez, R. E., Mecchia, M. A., and Palatnik, J. F. (2012). Functional specialization of the plant miR396 regulatory network through distinct microRNA–target interactions. *PLoS genetics*, 8(1):e1002419.

[53] Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301.

[54] Djami-Tchatchou, A. T., Sanan-Mishra, N., Ntushelo, K., and Dubery, I. A. (2017). Functional Roles of microRNAs in Agronomically Important Plants-Potential as Targets for Crop Improvement and Protection. *Frontiers in Plant Science*, 8:378.

[55] Dozmorov, M. G., Giles, C. B., Koelsch, K. A., and Wren, J. D. (2013). Systematic classification of non-coding RNAs by epigenomic similarity. *BMC Bioinformatics*, 14(S14):S2.

[56] Eamens, A. L., Smith, N. A., Curtin, S. J., Wang, M.-B., and Waterhouse, P. M. (2009). The *Arabidopsis thaliana* double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA (New York, N.Y.)*, 15(12):2219–2235.

[57] Eddy, S. R. (2001). Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

[58] Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.

[59] Fahlgren, N. and Carrington, J. C. (2010). miRNA Target Prediction in Plants. In *Plant MicroRNAs*, pages 51–57. Humana Press.

[60] Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., Law, T. F., Grant, S. R., Dangl, J. L., and Carrington, J. C. (2007). High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of MIRNA genes. *PloS One*, 2(2):e219.

[61] Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., and Mueller, L. A. (2015). The Sol Genomics Network (SGN)–from genotype to phenotype to breeding. *Nucleic Acids Research*, 43:D1036–1041.

[62] Finnegan, E. J. and Matzke, M. A. (2003). The small RNA world. *Journal of cell science*, 116(23):4689–4693.

[63] Folkes, L., Moxon, S., Woolfenden, H. C., Stocks, M. B., Szittya, G., Dalmay, T., and Moulton, V. (2012). PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Research*, 40(13):e103.

[64] Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, 26(4):407–415.

[65] Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52.

[66] Frohman, M. A. et al. (1990). RACE: rapid amplification of cDNA ends. *PCR protocols: A guide to methods and applications*, 28.

[67] Gandikota, M., Birkenbihl, R. P., Höhmann, S., Cardon, G. H., Saedler, H., and Huijser, P. (2007). The miRNA156/157 recognition element in the 3 UTR of the *Arabidopsis* SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. *The Plant Journal*, 49(4):683–693.

[68] Garcia, D., Collier, S. A., Byrne, M. E., and Martienssen, R. A. (2006). Specification of leaf polarity in *Arabidopsis* via the trans-acting siRNA pathway. *Current biology: CB*, 16(9):933–938.

[69] Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., et al. (2009). Rfam: updates to the RNA families database. *Nucleic acids research*, 37(suppl_1):D136–D140.

[70] Garg, V., Khan, A. W., Kudapa, H., Kale, S. M., Chitikineni, A., Qiwei, S., Sharma, M., Li, C., Zhang, B., Xin, L., Kishor, P. K., and Varshney, R. K. (2019). Integrated transcriptome, small RNA and degradome sequencing approaches provide insights into *Ascochyta* blight resistance in chickpea. *Plant Biotechnology Journal*, 17(5):914–931.

[71] Ge, S. X., Son, E. W., and Yao, R. (2018). iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, 19(1):534.

[72] German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., Paoli, E. D., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. *Nature Biotechnology*, 26(8):941–946.

[73] Ghildiyal, M. and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108.

[74] Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Research*, 32:D109–111.

[75] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441.

[76] Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl_1):D140–D144.

[77] Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl_1):D154–158.

[78] Guo, Z., Kuang, Z., Wang, Y., Zhao, Y., Tao, Y., Cheng, C., Yang, J., Lu, X., Hao, C., Wang, T., Cao, X., Wei, J., Li, L., and Yang, X. (2020). PmiREN: a comprehensive encyclopedia of plant miRNAs. *Nucleic Acids Research*, 48(D1):D1114–D1121.

[79] Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[80] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8.

[81] Hou, Y., Zhai, Y., Feng, L., Karimi, H. Z., Rutter, B. D., Zeng, L., Choi, D. S., Zhang, B., Gu, W., Chen, X., et al. (2019). A *Phytophthora* Effector Suppresses Trans-Kingdom RNAi to Promote Disease Susceptibility. *Cell host & microbe*, 25(1):153–165.

[82] Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., Garcia Giron, C., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Gonzalez Martinez, J., Marugán, J. C., Maurel, T., McMahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh, D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Salam, A., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., De Silva, N., Flint, B., Frankish, A., Hunt, S. E., IIsley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R., and Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891.

[83] Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., Kasschau, K. D., and Carrington, J. C. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway

in *Arabidopsis* reveals dependency on miRNA-and tasiRNA-directed targeting. *The Plant Cell*, 19(3):926–942.

[84] Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811.

[85] Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., Tang, Y., Chen, Y.-G., Jin, C.-N., Yu, Y., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic acids research*, 48(D1):D148–D154.

[86] Hutchison, C. A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237.

[87] Jiang, N., Gutierrez-Diaz, A., Mukundi, E., Lee, Y. S., Meyers, B. C., Otegui, M. S., and Grotewold, E. (2020). Synergy between the anthocyanin and RDR6/SGS3/DCL4 siRNA pathways expose hidden features of *Arabidopsis* carbon metabolism. *Nature Communications*, 11(1):2456.

[88] Jin, Q.-Y., Peng, H.-Z., Lin, E.-P., Li, N., Huang, D.-N., Xu, Y.-L., Hua, X.-Q., Wang, K.-H., and Zhu, T.-J. (2016). Identification and characterization of differentially expressed miRNAs between bamboo shoot and rhizome shoot. *Journal of Plant Biology*, 59(4):322–335.

[89] Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular Cell*, 14(6):787–799.

[90] Kakrana, A., Hammond, R., Patel, P., Nakano, M., and Meyers, B. C. (2014). sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Research*, 42(18):e139.

[91] Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R., Bateman, A., and Petrov, A. I. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1):D192–D200.

[92] Kang, W. and Friedländer, M. R. (2015). Computational Prediction of miRNA Genes from Small RNA Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3:2–7.

[93] Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., and Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, 6(1):4.

[94] Kerpedjiev, P., Hammer, S., and Hofacker, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379.

[95] Khraiwesh, B., Arif, M. A., Seumel, G. I., Ossowski, S., Weigel, D., Reski, R., and Frank, W. (2010). Transcriptional Control of Gene Expression by MicroRNAs. *Cell*, 140(1):111–122.

[96] Kono, N. and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. *Development, growth & differentiation*, 61(5):316–326.

[97] Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162.

[98] Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39:D152–157.

[99] Kozomara, A. and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42:D68–73.

[100] Kuang, Z., Wang, Y., Li, L., and Yang, X. (2019). miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics*, 35(14):2521–2522.

[101] Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543):853–858.

[102] Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210.

[103] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

[104] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

[105] Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology*, 4(1):213.

[106] Lee, B. T., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A., Lee, C., Muthuraman, P., Nassar, L., Nguy, B., Pereira, T., Perez, G., Raney, B., Rosenbloom, K., Schmelter, D., Speir, M., Wick, B., Zweig, A., Haussler, D., Kuhn, R., Haeussler, M., and Kent, W. (2022). The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*, 50(D1):D1115–D1122.

[107] Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

[108] Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060.

[109] Li, C. and Zhang, B. (2016). Micrornas in control of plant development. *Journal of cellular physiology*, 231(2):303–313.

[110] Li, F., Pignatta, D., Bendix, C., Brunkard, J. O., Cohn, M. M., Tung, J., Sun, H., Kumar, P., and Baker, B. (2012). MicroRNA regulation of plant innate immune receptors. *Proceedings of the National Academy of Sciences*, 109(5):1790–1795.

[111] Li, Y., Nie, T., Zhang, M., Zhang, X., Shahzad, K., Guo, L., Qi, T., Tang, H., Wang, H., Qiao, X., Feng, J., Lin, Z., Wu, J., and Xing, C. (2022). Integrated analysis of small RNA, transcriptome and degradome sequencing reveals that micro-RNAs regulate anther development in CMS cotton. *Industrial Crops and Products*, 176:114422.

[112] Li, Y.-F., Zhao, M., Wang, M., Guo, J., Wang, L., Ji, J., Qiu, Z., Zheng, Y., and Sunkar, R. (2019). An improved method of constructing degradome library suitable for sequencing using Illumina platform. *Plant Methods*, 15:134.

[113] Lin, S.-S., Chen, Y., and Lu, M.-Y. J. (2019). Degradome Sequencing in Plants. *Methods in Molecular Biology (Clifton, N.J.)*, 1932:197–213.

[114] Liu, H., Able, A. J., and Able, J. A. (2020a). Integrated analysis of small rna, transcriptome, and degradome sequencing reveals the water-deficit and heat stress response network in durum wheat. *International journal of molecular sciences*, 21(17):6017.

[115] Liu, J., Liu, X., Zhang, S., Liang, S., Luan, W., and Ma, X. (2021). TarDB: an online database for plant miRNA targets and miRNA-triggered phased siRNAs. *BMC Genomics*, 22(1):348.

[116] Liu, Y., Teng, C., Xia, R., and Meyers, B. C. (2020b). PhasiRNAs in plants: their biogenesis, genic sources, and roles in stress responses, development, and reproduction. *Plant Cell*, 32(10):3059–3080.

[117] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science (New York, N.Y.)*, 297(5589):2053–2056.

[118] Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., Hatzimichael, E., Kirino, Y., Honda, S., Lally, M., et al. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate-and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences*, 112(10):E1106–E1115.

[119] Lopez-Gomollon, S., Mohorianu, I., Szittya, G., Moulton, V., and Dalmay, T. (2012). Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. *Planta*, 236(6):1875–1887.

[120] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

[121] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

[122] Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 14(5):265–279.

[123] Ma, X., Zhang, X., Zhao, K., Li, F., Li, K., Ning, L., He, J., Xin, Z., and Yin, D. (2018). Small RNA and Degradome Deep Sequencing Reveals the Roles of microRNAs in Seed Expansion in Peanut (Arachis hypogaea L.). *Frontiers in Plant Science*, 9:349.

[124] MacLean, D., Elina, N., Havecker, E. R., Heimstaedt, S. B., Studholme, D. J., and Baulcombe, D. C. (2010). Evidence for large complex networks of plant short silencing RNAs. *PloS One*, 5(3):e9901.

[125] Manavella, P. A., Koenig, D., and Weigel, D. (2012). Plant secondary sirna production determined by microrna-duplex structure. *Proceedings of the National Academy of Sciences*, 109(7):2461–2466.

[126] Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

[127] Mathews, D. H. (2006). Revolutions in RNA secondary structure prediction. *Journal of Molecular Biology*, 359(3):526–532.

[128] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292.

[129] Mathews, D. H. and Turner, D. H. (2006). Prediction of rna secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–278.

[130] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564.

[131] May, P., Liao, W., Wu, Y., Shuai, B., Richard McCombie, W., Zhang, M. Q., and Liu, Q. A. (2013). The effects of carbon dioxide and temperature on microRNA expression in *Arabidopsis* development. *Nature Communications*, 4:2145.

[132] McCarthy, A. (2010). Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chemistry & Biology*, 17(7):675–676.

[133] Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., Cao, X., Carrington, J. C., Chen, X., Green, P. J., Griffiths-Jones, S., Jacobsen, S. E., Mallory, A. C., Martienssen, R. A., Poethig, R. S., Qi, Y., Vaucheret, H., Voinnet, O., Watanabe, Y., Weigel, D., and Zhu, J.-K. (2008). Criteria for Annotation of Plant MicroRNAs. *The Plant Cell*, 20(12):3186–3190.

[134] Mielecki, J., Gawroński, P., and Karpiński, S. (2020). Retrograde signaling: understanding the communication between organelles. *International journal of molecular sciences*, 21(17):6173.

[135] Mohorianu, I. and Moulton, V. (2010). Revealing biological information using data structuring and automated learning. *Recent Patents on DNA & Gene Sequences (Discontinued)*, 4(3):181–191.

[136] Molnár, A., Csorba, T., Lakatos, L., Várallyay, É., Lacomme, C., and Burgyán, J. (2005). Plant virus-derived small interfering rnas originate predominantly from highly structured single-stranded viral rnas. *Journal of virology*, 79(12):7812–7818.

[137] Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics (Oxford, England)*, 24(19):2252–2253.

[138] Nag, A., King, S., and Jack, T. (2009). miR319a targeting of TCP4 is critical for petal growth and development in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, 106(52):22534–22539.

[139] Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *The Plant Cell*, 2(4):279–289.

[140] Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

[141] Nyrén, P. and Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, 151(2):504–509.

[142] Ørom, U. A., Nielsen, F. C., and Lund, A. H. (2008). MicroRNA-10a binds the 5 UTR of ribosomal protein mRNAs and enhances their translation. *Molecular cell*, 30(4):460–471.

[143] Paicu, C., Mohorianu, I., Stocks, M., Xu, P., Coince, A., Billmeier, M., Dalmay, T., Moulton, V., and Moxon, S. (2017). miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics*, 33(16):2446–2454.

[144] Pantaleo, V., Szittya, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., and Burgyan, J. (2010). Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. *The Plant Journal: For Cell and Molecular Biology*, 62(6):960–976.

[145] Park, M. Y., Wu, G., Gonzalez-Sulser, A., Vaucheret, H., and Poethig, R. S. (2005). Nuclear processing and export of microRNAs in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3691–3696.

[146] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData mining*, 4(1):1–27.

[147] Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, 93(2):105–111.

[148] Prüfer, K., Stenzel, U., Dannemann, M., Green, R. E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531.

[149] R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[150] Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198.

[151] Ravichandran, S., Ragupathy, R., Edwards, T., Domaratzki, M., and Cloutier, S. (2019). MicroRNA-guided regulation of heat stress response in wheat. *BMC genomics*, 20(1):488.

[152] Redner, S. (2008). Teasing out the missing links. *Nature*, 453(7191):47–48.

[153] Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906.

[154] Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. *Genes & Development*, 16(13):1616–1626.

[155] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597.

[156] Reza, A. M. M. T. and Yuan, Y.-G. (2021). micrornas mediated regulation of the ribosomal proteins and its consequences on the global translation of proteins. *Cells*, 10(1):110.

[157] Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, 110(4):513–520.

[158] Russell, S. J. and Norvig, P. (2016). *Artificial Intelligence : A Modern Approach*. Malaysia; Pearson Education Limited,.

[159] Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.-c., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., Matsumoto, T., Sasaki, T., and Itoh, T. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant & Cell Physiology*, 54(2):e6.

[160] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

[161] Saçar, M. D., Hamzeiy, H., and Allmer, J. (2013). Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? *Journal of Integrative Bioinformatics*, 10(2):215.

[162] Schmitz-Linneweber, C. and Small, I. (2008). Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science*, 13(12):663–670.

[163] Schommer, C., Debernardi, J. M., Bresso, E. G., Rodriguez, R. E., and Palatnik, J. F. (2014). Repression of cell proliferation by miR319-regulated TCP4. *Molecular plant*, 7(10):1533–1544.

[164] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

[165] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

[166] Sobkowiak, L., Jarmolowski, A., Karlowski, W., and Szweykowska-Kulinska, Z. (2012). Non-canonical processing of *Arabidopsis* pri-miR319a/b/c generates additional microRNAs to target one RAP2. 12 mRNA isoform. *Frontiers in plant science*, 3:46.

[167] Song, X., Li, P., Zhai, J., Zhou, M., Ma, L., Liu, B., Jeong, D.-H., Nakano, M., Cao, S., Liu, C., Chu, C., Wang, X.-J., Green, P. J., Meyers, B. C., and Cao, X. (2012). Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. *The Plant Journal: For Cell and Molecular Biology*, 69(3):462–474.

[168] Srivastava, P. K., Moturu, T. R., Pandey, P., Baldwin, I. T., and Pandey, S. P. (2014). A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC genomics*, 15(1):1–15.

[169] Stocks, M. B., Mohorianu, I., Beckers, M., Paicu, C., Moxon, S., Thody, J., Dalmay, T., and Moulton, V. (2018). The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. *Bioinformatics (Oxford, England)*, 34(19):3382–3384.

[170] Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T., and Moulton, V. (2012). The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics (Oxford, England)*, 28(15):2059–2061.

[171] Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

[172] Sunkar, R., Kapoor, A., and Zhu, J.-K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in *Arabidopsis* is mediated by downregulation of miR398 and important for oxidative stress tolerance. *The Plant Cell*, 18(8):2051–2065.

[173] Suzuki, T., Ikeda, S., Kasai, A., Taneda, A., Fujibayashi, M., Sugawara, K., Okuta, M., Maeda, H., and Sano, T. (2019). Rnai-mediated down-regulation of dicer-like 2 and 4 changes the response of 'moneymaker'tomato to potato spindle tuber viroid infection from tolerance to lethal systemic necrosis, accompanied by up-regulation of mir398, 398a-3p and production of excessive amount of reactive oxygen species. *Viruses*, 11(4):344.

[174] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, 36:D1009–1014.

[175] Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics (Oxford, England)*, 24(22):2657–2663.

[176] Tang, G., Reinhart, B. J., Bartel, D. P., and Zamore, P. D. (2003). A biochemical framework for RNA silencing in plants. *Genes & Development*, 17(1):49–63.

[177] Thody, J., Folkes, L., Medina-Calzada, Z., Xu, P., Dalmay, T., and Moulton, V. (2018). PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Research*, 46(17):8730–8739.

[178] Thompson, J. F. and Steinmann, K. E. (2010). Single molecule sequencing with a HeliScope genetic analysis system. *Current protocols in molecular biology*, 92(1):7–10.

[179] Tomari, Y., Du, T., Haley, B., Schwarz, D. S., Bennett, R., Cook, H. A., Koppetsch, B. S., Theurkauf, W. E., and Zamore, P. D. (2004). RISC assembly defects in the Drosophila RNAi mutant armitage. *Cell*, 116(6):831–841.

[180] Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641.

[181] Tripathi, A. M., Singh, R., Singh, A., Verma, A. K., Mishra, P., Narayan, S., Shirke, P. A., and Roy, S. (2021). Abolished miR158 activity leads to 21-nucleotide tertiary phasiRNA biogenesis that targets NHX2 in *Arabidopsis thaliana*. *bioRxiv*.

[182] van der Krol, A. R., Mur, L. A., de Lange, P., Mol, J. N., and Stuitje, A. R. (1990). Inhibition of flower pigmentation by antisense CHS genes: promoter and minimal sequence requirements for the antisense effect. *Plant Molecular Biology*, 14(4):457–466.

[183] Vargas-Asencio, J. A. and Perry, K. L. (2020). A small RNA-mediated regulatory network in *Arabidopsis thaliana* demonstrates connectivity between phasiRNA regulatory modules and extensive co-regulation of transcription by miRNAs and phasiRNAs. *Frontiers in plant science*, 10:1710.

[184] Vaucheret, H., Vazquez, F., Crété, P., and Bartel, D. P. (2004). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes & Development*, 18(10):1187–1197.

[185] Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4):641–658.

[186] Walker, J. C. (1993). Receptor-like protein kinase genes of *Arabidopsis thaliana*. *The Plant Journal*, 3(3):451–456.

[187] Wang, L., Leister, D., Guan, L., Zheng, Y., Schneider, K., Lehmann, M., Apel, K., and Kleine, T. (2020). The *Arabidopsis* SAFEGUARD1 suppresses singlet oxygen-induced stress responses by protecting grana margins. *Proceedings of the National Academy of Sciences*, 117(12):6918–6927.

[188] Wang, R., Yang, Z., Fei, Y., Feng, J., Zhu, H., Huang, F., Zhang, H., and Huang, J. (2019). Construction and analysis of degradome-dependent microRNA regulatory networks in soybean. *BMC genomics*, 20(1):534.

[189] Wang, X., An, Y., Xu, P., and Xiao, J. (2021). Functioning of ppr proteins in organelle rna metabolism and chloroplast biogenesis. *Frontiers in plant science*, 12:627501.

[190] Wang, X., Wang, Y., Dou, Y., Chen, L., Wang, J., Jiang, N., Guo, C., Yao, Q., Wang, C., Liu, L., Yu, B., Zheng, B., Chekanova, J. A., Ma, J., and Ren, G. (2018). Degradation of unmethylated miRNA/miRNA*s by a DEDDy-type 3 to 5 exoribonuclease Atrimmer 2 in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, 115(28):E6659–E6667.

[191] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

[192] Watson, J. D. and Crick, F. H. C. (1953a). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

[193] Watson, J. D. and Crick, F. H. C. (1953b). The Structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 18:123–131.

[194] Williamson, V., Kim, A., Xie, B., McMichael, G. O., Gao, Y., and Vladimirov, V. (2013). Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Briefings in Bioinformatics*, 14(1):36–45.

[195] Wolfsberg, T. G. (2010). Using the NCBI Map Viewer to Browse Genomic Sequence Data. *Current Protocols in Bioinformatics*, 29(1):1.5.1–1.5.25.

[196] Wu, C., Li, X., Guo, S., and Wong, S.-M. (2016a). Analyses of RNA-Seq and sRNA-Seq data reveal a complex network of anti-viral defense in TCV-infected *Arabidopsis thaliana*. *Scientific reports*, 6(1):1–12.

[197] Wu, L., Zhang, Q., Zhou, H., Ni, F., Wu, X., and Qi, Y. (2009). Rice MicroRNA effector complexes and targets. *The Plant Cell*, 21(11):3421–3435.

[198] Wu, X., Ding, D., Shi, C., Xue, Y., Zhang, Z., Tang, G., and Tang, J. (2016b). microRNA-dependent gene regulatory networks in maize leaf senescence. *BMC plant biology*, 16(1):1–15.

[199] Xie, Z., Kasschau, K. D., and Carrington, J. C. (2003). Negative feedback regulation of Dicer-Like1 in *Arabidopsis* by microRNA-guided mRNA degradation. *Current biology: CB*, 13(9):784–789.

[200] Yang, C.-Y., Huang, Y.-H., Lin, C.-P., Lin, Y.-Y., Hsu, H.-C., Wang, C.-N., Liu, L.-Y. D., Shen, B.-N., and Lin, S.-S. (2015). MicroRNA396-targeted SHORT VEGETATIVE PHASE is required to repress flowering and is related to the development of abnormal flower symptoms by the phyllody symptoms1 effector. *Plant Physiology*, 168(4):1702–1716.

[201] Yang, X. and Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics (Oxford, England)*, 27(18):2614–2615.

[202] Yoshikawa, M., Peragine, A., Park, M. Y., and Poethig, R. S. (2005). A pathway for the biogenesis of trans-acting siRNAs in *Arabidopsis*. *Genes & Development*, 19(18):2164–2175.

[203] Yu, H. and Kumar, P. P. (2003). Post-transcriptional gene silencing in plants by RNA. *Plant Cell Reports*, 22(3):167–174.

[204] Yue, D., Liu, H., and Huang, Y. (2009). Survey of Computational Algorithms for MicroRNA Target Prediction. *Current Genomics*, 10(7):478–492.

[205] Zamore, P. D. and Haley, B. (2005). Ribo-gnome: the big world of small RNAs. *Science*, 309(5740):1519–1524.

[206] Zeng, C., Xia, J., Chen, X., Zhou, Y., Peng, M., and Zhang, W. (2017). MicroRNA-like RNAs from the same miRNA precursors play a role in cassava chilling responses. *Scientific Reports*, 7(1):1–9.

[207] Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006). Conservation and divergence of plant microRNA genes. *The Plant Journal*, 46(2):243–259.

[208] Zhang, B., Stellwag, E. J., and Pan, X. (2009). Large-scale genome analysis reveals unique features of microRNAs. *Gene*, 443(1-2):100–109.

[209] Zhang, H., Hu, J., Qian, Q., Chen, H., Jin, J., and Ding, Y. (2016). Small RNA profiles of the rice PTGMS line *Wuxiang S* reveal miRNAs involved in fertility transition. *Frontiers in Plant Science*, 7:514.

[210] Zhang, J., Lin, Y., Wu, F., Zhang, Y., Cheng, L., Huang, M., and Tong, Z. (2021). Profiling of microRNAs and their targets in roots and shoots reveals a potential miRNA-mediated interaction network in response to phosphate deficiency in the forestry tree betula luminifera. *Frontiers in genetics*, 12:552454.

[211] Zhang, W., Gao, S., Zhou, X., Xia, J., Chellappan, P., Zhou, X., Zhang, X., and Jin, H. (2010a). Multiple distinct small RNAs originate from the same microRNA precursors. *Genome Biology*, 11(8):1–81.

[212] Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y., and Su, Z. (2010b). PMRD: plant microRNA database. *Nucleic Acids Research*, 38(suppl–1):D806–813.

[213] Zheng, Y., Li, Y.-F., Sunkar, R., and Zhang, W. (2012). SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Research*, 40(4):e28.

[214] Zhu, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., Sun, Z., and Wu, J. (2010). mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Research*, 38(suppl_2):W392–W397.

[215] Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & development*, 21(9):1010–1024.