Computational analysis of small RNAs and the RNA degradome with application to plant water stress

Leighton Folkes

Supervisor: Prof. Vincent Moulton

Co-supervisor: Prof. Tamas Dalmay

A thesis submitted for the Degree of Doctor of Philosophy at the University of East Anglia

March 2014

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Acknowledgement

Thank you to my supervisor Professor Vincent Moulton and my co-supervisor Professor Tamas Dalmay for the invaluable advice, guidance and support that they have given me over the course of my studies. I would also like to thank Dr. Simon Moxon for his encouragement and inspirational words of wisdom. Thank you to the BBSRC for funding my PhD. Finally I would like to thank my wife Ruthie for being there for me throughout my journey of discovery.

Publications

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. Nucl. Acids Res., 10.1093/nar/gks277.

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, 28, 20592061.

Mapleson,D., Mohorianu,I., Pais,H., Stocks,M., Folkes,L. and Moulton,V. (2014) Processing Large-scale Small RNA Datasets in Silico. In Xu,J. (ed), Next Generation Sequencing: Current Technologies and Applications. Caister Academic Press, p. 150.

*Authors contributed equally to this work.

Statement of originality

I certify that this thesis, and the research to which it refers, are the product of my own work and that any ideas or quotations from the work of other people, published or otherwise are fully acknowledged.

Abstract

Water shortage is one of the most important environmental stress factors that affects plants, limiting crop yield in large areas worldwide. Plants can survive water stress by regulating gene expression at several levels. One of the recently discovered regulatory mechanisms involves small RNAs (sRNAs), which can regulate gene expression by targeting messenger RNAs (mRNAs) and directing endonucleolytic cleavage resulting in mRNA degradation. A snapshot of an mRNA degradation profile (degradome) can be captured through a new high-throughput technique called Parallel Analysis of RNA Ends (PARE) by using next generation sequencing technologies. In this thesis we describe a new user friendly degradome analysis software tool called PAREsnip that we have used for the rapid genome-wide discovery of sRNA/target interactions evidenced through the degradome. In addition to PAREsnip and based upon PAREsnip's rapid capability, we also present a new software tool for the construction, analysis and visualisation of sRNA regulatory interaction networks. The two new tools were used to analyse PARE datasets obtained from*Medicago truncatula* and *Arabidopsis*

thaliana. In particular, we have used PAREsnip for the high-throughput analysis of PARE data obtained from *Medicago* when subjected to dehydration and found several sRNA/mRNA interactions that are potentially responsive to water stress. We also present how we used our new network visualisation and analysis tool with PARE datasets obtained from *Arabidopsis* and discovered several novel sRNA regulatory interaction networks. In building tools and using them for this kind of analysis, we gain a better understanding of the processes and mechanisms involved in sRNA mediated gene regulation and how plants respond to water stress which could lead to new strategies in improving stress tolerance.

Contents

Chapter 1

Introduction

The work that we have carried out and presented in this thesis is primarily focused upon the development of new computational tools and algorithms that can be used for the analysis of small RNAs (sRNAs) and their targets. In particular, we develop a new tool that can be used for the high-throughput identification and validation of sRNA targets and a new tool that can be used for the generation, discovery and visualization of sRNA regulatory interaction networks. Through the application of these new tools to plant sRNA and degradome datasets, we have identified a number of novel sRNA/target interactions and interaction networks. In addition, we have identified biologically interesting sRNA/target interactions that are potentially involved in a plant's response to water stress. Below we give an overview of the contents of this thesis.

Chapter 2. In this chapter we provide some relevant biological background information on RNA silencing and sRNA biogenesis and function in plants. We provide an overview of the methods used by next generation sequencing technologies to obtain the data that our new tools analyse. We then go on to review several computational sRNA target prediction tools and sRNA target validation methods. Finally, with sequence alignment being a core operation of sRNA/target analysis, we review and benchmark several short read alignment tools. These background topics are important and relevant to later chapters of this thesis.

Chapter 3. In this chapter we describe the development of a new software tool along with its embedded novel algorithms that can be used to rapidly identify and validate sRNA/target interactions using libraries obtained from a high-throughput sequencing method called Parallel Analysis of RNA Ends (PARE). We describe the tools confidence measures such as its p-value calculations and use of mRNA degradation signals. We then go on to benchmark the software and compare the tool with a previous low-throughput approach. We use our new tool to analyse plant sRNA and degradome datasets and we demonstrate that conservation of sRNAs and mRNA cleavage signals that are found in multiple samples can be used to filter out background noise and confidently identify sRNA/target interactions. Through the use of the tool and using a our conservation methods, we identified over 4000 putative sRNA/target interactions. The idea for us-

ing multiple datasets from plants that are biological replicates to filter out background degradation and confidently identify interactions was jointly conceived by Dr. Simon Moxon and myself. The idea and design of the algorithms that are used to search for sRNA/mRNA interactions as well as the implementation of the software, experimental testing, refinement of the methods and the generation of results were my contribution to this work. Dr. Hugh Woolfenden and Dr. Mathew Stocks provided assistance for making the tool compatible with t-plot visualization tool called VisSR and the UEA sRNA Workbench.

Chapter 4. In this chapter we describe the application of our new software to datasets obtained from plants subjected to water stress. We begin with a brief background on the importance of understanding how plants respond to water stress and why this is important in relation to changes in our climate. We go on to explain the composition of the datasets that we used in our analysis and the methods that we used to conduct the analysis. The degradome datasets that we describe in this chapter were prepared by our collaborator Dr. Goyrgy Szzitya (UEA School of Biological Sciences). We continue this chapter by describing the results of the analysis using our new tool and describe the identification of a number of novel sRNA/mRNA interactions potentially involved in plant water stress response. In particular, two candidate interactions were selected for deeper investigation and we consider the genes and sRNAs that are potentially

involved within the water stress responsive interactions in more detail.

Chapter 5. Here we describe the development and use of a new software tool that we designed to identify, analyse and visualize sRNA regulatory interaction networks that are evidenced through the RNA degradome. We begin this chapter with a brief background on regulatory sRNA networks and continue by describing the methods employed by the tool for network construction and visualization. We use the tool to analyse over 4000 sRNA/target interactions in *Arabidopsis thaliana* that were described in Chapter 3. We identify a number of novel regulatory sRNA interaction networks. The initial concept of the tool originated from discussions with Dr. Simon Moxon and Professor Vincent Moulton. The development and implementation of the software along with experimental testing, refinement of the methods and the generation of results were my contribution to this work.

Chapter 6. In this final chapter we present a suggestion for how the work in this thesis could be continued. We provide a road-map for the design and use of a new tool that could be used for the prediction of novel miRNAs by using a function first approach. We begin by explaining a preliminary study on the prediction of miRNA-like sRNAs. We then go on to suggest a framework for developing a new tool that could be used to predict novel miRNAs that fall slightly outside the strict miRNA classification criteria, but have supporting functional evidence through the degradome. We

then suggest how data could be generated and analyzed by the tool to predict novel miRNAs, optimize the tool's parameter settings and validate the predictions. We end this chapter, and this thesis by presenting our overall conclusions.

Chapter 2

Background

2.1 Summary

In this chapter we present a whistle-stop tour of some of the key aspects of both the biology and computational methods involved in the work presented in this thesis. RNA silencing is the biological process at the core of this work and we begin with a description of the RNA silencing process and machinery. We then go on to describe the effectors of RNA silencing which are tiny RNA molecules called small RNAs (sRNAs) and we detail several classes of sRNAs. sRNAs are found in biological samples in great abundance and next generation sequencing (NGS) methods are used to obtain the readout of the RNA content from experimental samples in huge volumes. So, we give an overview of the technology and principles behind NGS.

Using the data (sRNA reads) obtained from NGS technology, a first step

in understanding a sRNA's function is to identify messenger RNAs (mRNA) that can be targeted by them. Because of the size of the data generated by NGS technology, computational predictions of which mRNAs are targeted by sRNAs are required. Therefore, we describe several popular sRNA target prediction tools. As the tools only provide predictions, experimental validation is subsequently required and we explain two of the current experimental sRNA target validation methods. At the core of all DNA/RNA sequence analysis are sequence alignment tools of which there are many that are freely available, each with their own advantages and disadvantages. As they are so important to the sRNA bioinformatician, we also briefly review several of the most popular sequence alignment tools that are relevant to this work.

2.2 RNA silencing

RNA silencing is a phenomenon that was independently discovered in animals and plants in the early 1990s. The core RNA silencing machinery is now known to be highly conserved between eukaryotic kingdoms, and the common feature of all RNA silencing pathways is the production of non-coding small RNAs (sRNAs), mostly in the size range of 20 to 25 nucleotides (nt). These sRNAs are excised from longer, double-stranded or hairpin RNA precursors by RNaseIII-type enzymes called Dicers [16] to

form a double stranded sRNA duplex. One strand of the initial sRNA duplex is recruited into a member of the Argonaute protein family, which can be part of a larger complex known as the RNA Induced Silencing Complex (RISC). The sRNA component confers sequence specificity to RISC by establishing Watson-Crick base pairs i.e. pairs in the form of guanine:cytosine (G:C) and adenine:uracil (A:U) hydrogen bonds [127], with potential target mRNA molecules. Having bound to its target, the complex can silence the target at the transcriptional or translational level by employing one of the following mechanisms: (i) cleavage and degradation, (ii) translational repression, (iii) DNA methylation and heterochromatin formation [24]. This highly versatile machinery plays important roles in gene regulation, defence against pathogens and genome maintenance [21],[72].

In plants, sRNA-mediated post-transcriptional gene regulation generally leads to messenger RNA (mRNA) cleavage and degradation due to a high degree of sequence complementarity between the sRNA and its mRNA target [7]. This cleavage is highly specific and the mRNA is "sliced" by an Argonaute protein between positions 10 and 11 of the bound sRNA [73]. Below we describe several of the major sRNA classes that function within the RNA silencing mechanism such as microRNAs (miRNAs) and small interfering RNAs (siRNAs). These sRNAs regulate other RNA molecules, in particular messenger RNAs (mRNA), and are important repressors of gene expression. For recent reviews of RNA silencing see [56],[113].

2.2.1 miRNAs

The first endogenous small non-coding RNA known as microRNA (miRNA) was discovered in 1993 within the nematode model organism *C. elegans* [63]. Since this initial discovery, miRNAs have also been characterized in plants and viruses. In plants, mature miRNAs typically have a sequence length of 21 or 22nt. Their biogenesis is a multi-step process which begins in the nucleus of a cell [12],[61] (see figure 2.1). A single stranded primary miRNA (pri-miRNA) is transcribed from a miRNA gene by an RNA polymerase II enzyme [61],[64]. The pri-miRNA is able to fold into an imperfect hairpin type structure and is processed by an RNaseIII, Dicer-Like 1 enzyme [120] resulting in a precursor miRNA (pre-miRNA). Further processing is carried out by DCL1 to leave a double stranded RNA (dsRNA) duplex comprised of the mature miRNA or guide strand annealed to its complementary sequence called the miRNA "star" (miRNA*) or passenger strand. The duplex exits the nucleus and enters the cytoplasm [94] where it is separated by a helicase enzyme [12]. The mature miRNA is recruited by an Argonaute protein (AGO) and loaded into an RNA induced silencing complex (RISC) [124]. The mature miRNA (guide strand) confers sequence specificity to RISC, which acts to negatively regulate target mRNAs. Depending upon the level of complementarity between miRNA/mRNA, the mRNA is silenced by either endonucleolytic cleavage or arrest of protein translation [14],[73].

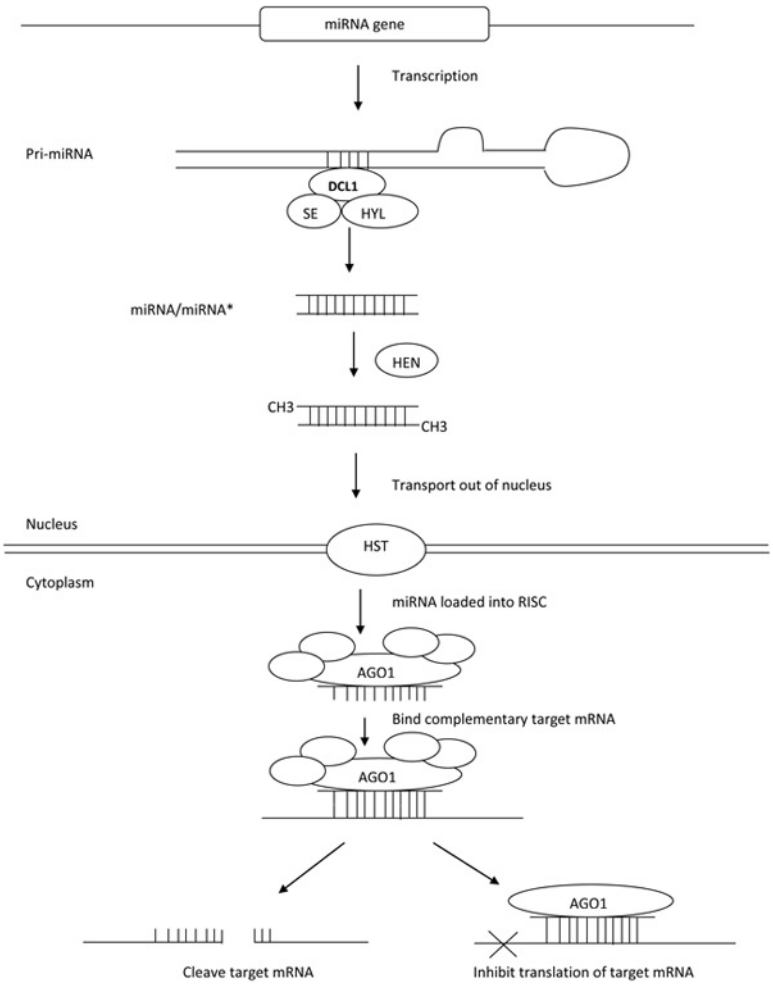


Figure 2.1: **An overview of miRNA biogenesis and function in Arabidopsis** miRNAs are transcribed from a gene and processed by DCL1, SE and HYL1 into an RNA duplex (miRNA/miRNA*). The duplex is methylated by HEN and transported out of the nucleus by HST. The miRNA portion of the duplex binds AGO1 to form RISC. The miRNA bound in RISC base pairs with a target mRNA that is complementary to the miRNA. The target mRNA is repressed by either cleavage or translational inhibition. This figure is reproduced from Phelps-Durr (2010) [97].

2.2.2 siRNAs

Short interfering RNAs (siRNAs) are another class of sRNA found in plants and animals. They are derived from long dsRNA, have a 2nt overhang at

the 3' end and are ~21nt in length. In contrast to miRNAs, exogenous (originating externally) siRNAs exhibit perfect complementarity with their target [31],[108]. An example of exogenous siRNA activated RNA silencing was found by the experimental introduction of dsRNA into a cell [37]. Other examples include transgenes and viruses [50].

There are several subtypes of endogenous siRNAs, one of which is currently thought to only be found in plants and is called a trans-acting short interfering RNA (ta-siRNA). In plants, microRNAs can trigger trans-acting siRNA biogenisis. Trans-acting siRNA (TAS) genes are transcribed in to RNA and are then targeted by miRNAs in either one position or two positions and cleaved. These cleavages and, in particular, the cleavage position on the transcript set the start of a phase window i.e. a small RNA is produced that starts 1nt after the end position of the previously excised sRNA. An RNA dependant RNA polymerase 6 (RDR6) processes a cleaved transcript turning it into double stranded RNA. The dsRNA is then recognised and processed by a Dicer-like 4 (DCL4) enzyme. The DCL4 enzyme cuts the transcript in 21nt increments (phase window) to produce mature ta-siRNAs [4],[9]. The 21nt ta-siRNAs may now be incorporated into RISC where they guide the complex to mRNA targets and repress their translation or cause cleavage [124].

2.2.3 Other types of sRNAs

There are many other classes and sub-classes of sRNA that can enter into the RNA silencing pathway and regulate mRNA expression [8],[59]. For example, natural antisense transcript siRNAs (nat-siRNAs) are derived from transcripts that contain complementary regions which can overlap to form dsRNA [19] and the overlapping regions trigger the production of 21-24nt sRNAs through dicer-like proteins [85]. Another class of sRNA is the PIWI-interacting RNAs (piRNAs) which are only found in animal systems [6, 43, 49, 126]. Their name derives from their interaction with PIWI proteins that are mainly observed in the germline [122].

2.3 Next generation sequencing

DNA sequencing is a powerful tool in determining the nucleotide sequence of DNA. The advent of next generation sequencing (NGS) technologies [78] in the mid to late 2000's has given rise to an explosion in the ability to produce sequencing data as never seen before. This can be attributed to the ongoing technological improvements which has driven down both the financial cost and time required to carry out high throughput sequencing experiments. Two of the most popular NGS platforms are Roche/454 and Illumina/Solexa which among others, respectively employs the FLX and HiSeq sequencing instruments.

2.3.1 Generation of clusters

To use the sequencing instruments a sample library is prepared using a propriety library preparation kit such as an Illumina branded product or a third party kit. The library of single stranded DNA fragments are washed across the surface of a flowcell for amplification. A flowcell is a transparent glass surface similar to a microscope slide [52]. The flowcell is decorated with adaptors complementary to those which were ligated to the DNA fragments during library preparation and are attached to the flowcell through covalent bonds. Adaptor-ligated DNA-fragments provide a template which bond to the complementary adaptors attached to the flowcell. Bridge amplification, also called bridge polymerase chain reaction (PCR) [52], is performed and the resulting double-stranded DNA is denatured to leave single stranded templates anchored to the flowcell. This process results in several million dense clusters of clonal DNA fragments (copies) ready for sequencing.

2.3.2 Sequencing by synthesis

Sequencing occurs in chemistry cycles to determine each nucleotide base and their order in the target DNA. During each cycle, nucleotides ‘A’, ‘G’, ‘C’ and ‘T’ each containing a unique florescent group are incorporated. The incorporated nucleotides bond to a complementary nucleotide on each clonal DNA fragment. DNA fragments are excited with a laser. The newly

bonded nucleotide emits florescence within its cluster which is captured by the sequencer camera. The colour-coded light is detected and captured by the camera and the nucleotide base is identified. Subsequent cycles continue to identify bases in order to determine the nucleotide sequence of the template DNA fragments. This process is multiplexed and happens in parallel so that each DNA template sequence within the sample library can be identified.

The final output is a file containing all sequences within the submitted library preparation. The file usually takes the form of FASTQ format [28] which contains base-call quality scores (also known as phred quality scores [34],[33]). Data repositories are available such as the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) to publicly index and store sequence data [65]. Due to the ever-increasing size of NGS data as well as the cost of storage, the SRA allows several data compression formats, such as sequence read archive (SRA) format and standard flowgram format.

2.4 sRNA target prediction

To understand the function of a sRNA, an important step is to identify its potential targets. Computational plant sRNA target prediction tools have proved useful in identifying targets [29, 17, 5, 35] and generally attempt to

model some biological characteristic discovered *in vitro* such as positional base-paring properties. However, computational target prediction methods tend to rely heavily on the near-perfect complementarity between the sRNA and the predicted target transcript. Predictions generated by such tools tend to produce varying levels of false positive results, therefore further experimental validation is required. The tools come in the form of local command line or web-based applications and tend to comprise of an established sequence alignment tool at the applications core. The core algorithms are usually wrapped in code for pre/post processing where biologically relevant supporting statistics are calculated e.g. minimum free energy (MFE), or core alignments are filtered based upon current biological understanding e.g. targets discarded based upon positional mismatches within a duplex that are not found experimentally. Below we present several examples of popular plant sRNA target prediction tools.

2.4.1 psRNATarget

psRNATarget is a web server that can be used for plant sRNA target prediction [29] and builds upon a the Samuel Roberts Nobel Foundation's previous tool for plant sRNA target prediction called miRU [132]. The tool uses reverse complementary matching to identify potential targets through the Smith-Waterman [116] algorithm implementation and the scoring system

used by miRU [132]. The web-tool can also evaluate target site accessibility by calculating un-paired energy (UPE) around the sRNA target site using the RNAup program within the Vienna RNA Package [79]. Similar to TAPIR (see 2.4.2), psRNATarget is able to identify potential transcript translational repression as in the case of target mimicry by allowing and reporting interactions that contain additional nucleotides within the central region of the sRNA/transcript duplex. The web-tool uses a backend pipeline operated upon a distributed computing platform. The software is not publicly available for download. The web-interface offers three methods of user input: user-submitted sRNA searched against preloaded transcripts; preloaded sRNAs searched against user-submitted transcripts or user submitted sRNAs searched against user submitted transcripts. All user inputs are required to be in FASTA format. The output is a browser-viewable or downloadable list of sRNA/target pairs. As this tool is web-based, the user suffers from the disadvantage of reliance on network connectivity, third-party web-server uptime, bandwidth usage and the tradeoff between data size and the time required to upload/download data. However, for researchers with limited access to computer facilities, the web-based nature of this tool could be considered a benefit.

2.4.2 TAPIR

TAPIR is a plant miRNA target prediction tool offered as a web-server and downloadable tool [17]. The underlying methods used by the tool are the FASTA algorithm (not to be confused with FASTA format) [95] and the RNAhybrid algorithm [102]. The web-tool can be used in two modes, "fast" mode or "precise" mode using the two different backend algorithms respectively where the precise mode using RNAhybrid is much slower in compute time than fast mode. The precise mode is a key feature of TAPIR and can be useful for predicting miRNA target mimicry. In brief, target mimicry [39] is a level of sRNA regulation where multiple transcripts share sequence similarity and can be targeted by the same sRNA. However, the target mimic has several additional nucleotides between bases 10 and 11 of the sRNA within the binding site, resulting in a bulge at the position where cleavage would normally occur. Because of the bulge, the target mimic is not cleaved, but instead sequesters the sRNA, therefore preventing the sRNA from regulating other transcripts sharing sequence similarity to the mimic. The web-tool takes as input user supplied sRNA(s) and transcript(s) in FASTA format and outputs sRNA/target interaction predictions along with supporting statistics such as minimum free energy (MFE) and alignment score. The alignment score system used by TAPIR is based on that suggested by Allen et. al. [5]. As the tool is downloadable and offered as a

web-interface, the developers have made it possible for researchers to scale their analysis in-line with the dataset size and compute resources available to them.

2.4.3 TargetFinder

TargetFinder is a downloadable, command line based, plant sRNA target prediction tool written in the Pearl scripting language [5, 35]. It makes use of the FASTA35 [96] program to make a alignments between sRNAs and transcripts. The tool takes as input a single sRNA sequence and a list of transcripts in FASTA format. To predict sRNA/target interactions, the tool first attempts to identify alignments between an input query sRNA sequence and the supplied reference transcripts. Each alignment identified is converted into an RNA duplex (sRNA/target) and given an interaction score. The interaction score calculated is based upon observations made on experimentally validated sRNA/target interactions where certain mismatched positions and types of base pair mismatches e.g. G:U pairs, effect the score. Predictions are output to the terminal and comprise the sequence identifiers, score and interaction duplex. The position-dependent scoring system implemented within this tool has successfully been used by several other tools and methodologies in plant sRNA target prediction [17, 38, 89].

2.5 sRNA target validation

Computational methods used to find sRNA targets tend to suffer from a high number of false positive predictions [88]. Therefore computational predictions usually require further experimental validation. Here we describe two methods that can be used to validate sRNA/mRNA interactions.

2.5.1 Low throughput small RNA target validation

In plants, a common feature of the sRNAs we have described is that they can silence mRNAs in a sequence specific manner through endonucleolytic cleavage. The examination of mRNA cleavage products is one of the steps necessary for sRNA/target interaction validation. A method known as RLM-5' RACE (RNA linker mediated 5' rapid amplification of cDNA ends) can be used to experimentally validate sRNA mediated cleavage by identifying mRNA cleavage fragments/products for a particular mRNA. The technique ligates a sequence adaptor i.e. attaches an adaptor through an enzymatic process using covalent bonds, to the 5' end of the target mRNA fragment which is characteristically uncapped, i.e. the nucleotide base at the 5' end of the mRNA fragment has an exposed phosphate making the fragment ligation competent. The target mRNA fragment is reverse transcribed into cDNA and amplified through polymerase chain reaction (PCR) [107]. The resulting PCR products can then be sequenced and the mRNA cleavage

fragment identified.

To evidence sRNA function, the cleavage fragment can be aligned to the reference mRNA and the first nucleotide at the 5' end of the fragments are expected to align to same position as the cleavage site of the complementary sRNA, i.e. between base positions 10 and 11 of the sRNA [73]. This method is low throughput as the 5' RACE protocol needs to be performed for every predicted cleavage site on each gene of interest. The methods also requires prior knowledge of the flanking region adjacent to each predicted cleavage site. Therefore, due to the time and resources required to carry out the 5' RACE protocol, it is practical for only a limited number of sRNA target validations. A new high-throughput technique called Parallel Analysis of RNA Ends (PARE) may be used to identify and validate sRNA/target interactions on a much larger scale, which we now describe.

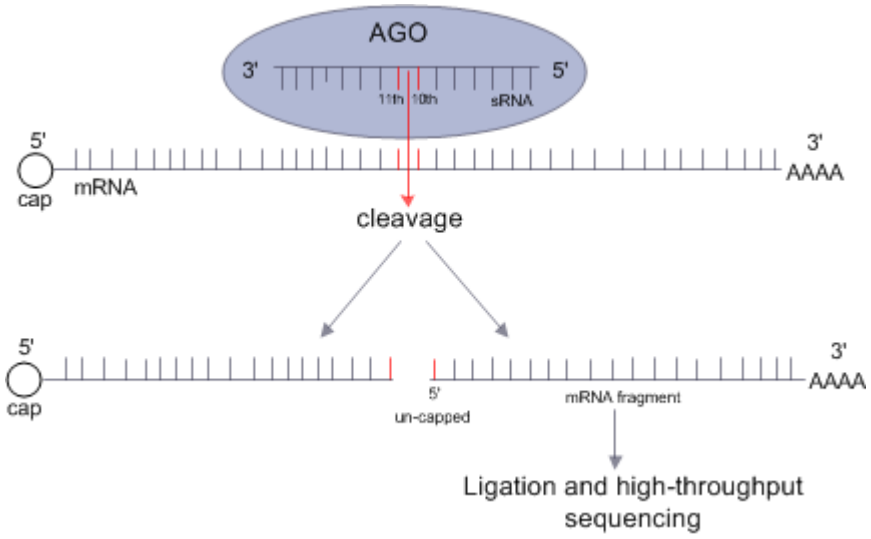
2.5.2 High-throughput sRNA target validation

In 2008, German et al. [41] described a new technique, called Parallel Analysis of RNA Ends (PARE) or degradome sequencing, which can be used to globally sample cleaved mRNA fragments using high-throughput sequencing technology. This new technique can be used to identify new sRNA/mRNA interactions [1],[42]. During the transcription process, mRNAs are given an altered nucleotide known as a 5' cap. This altered nucleotide is also known

as a 7-methylguanosine cap. Its purpose is to protect the mRNA from exonucleases. Exonucleases are enzymes that degrade RNA. When mRNAs are subjected to sRNA mediated endonucleolytic cleavage, the mRNA is sliced and the fragment upstream of the cleavage site degrades, yet the downstream (3' of the sRNA) remains stable. The remaining stable mRNA fragments do not have a 5' cap, but instead have a 5' monophosphate and are said to be uncapped at the 5' end (see figure 2.2a).

This new experimental technique selectively clones all uncapped RNA molecules which have a 3' poly-A tail, but unlike 5' RACE does not require any knowledge of which mRNA is being targeted. Therefore, this method can provide a snapshot of the mRNA degradation profile within the sample. The snapshot of degraded mRNA fragments obtained using this method has been termed the degradome. When the fragments within the degradome are realigned to a reference/template mRNA *in-silico*, there is evidence of clear peaks at the cleavage site of a mRNA corresponding to the position of cleavage by a sRNA (see figure 2.2b). The degradome data provides support for the interaction between sRNAs and their complementary mRNA targets and this method has been successfully used to identify miRNA targets in a variety of organisms [1, 3, 93].

A



B

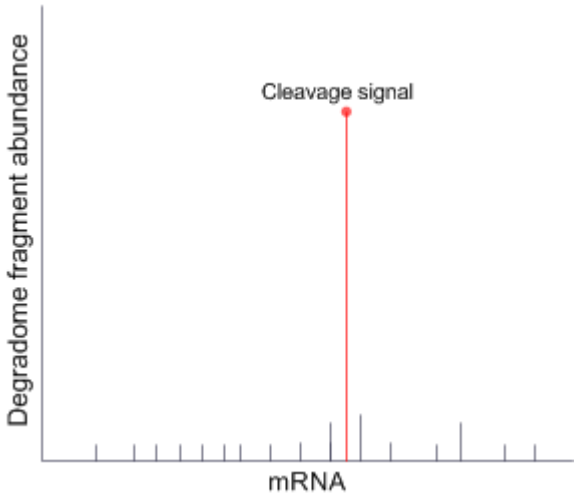


Figure 2.2: (A) An mRNA has a 5' cap (5' 7-methylguanosine) structure and a 3'-poly A tail. An sRNA is loaded into an Argonaute (AGO) protein and can target the mRNA which may lead to endonucleolytic cleavage. The mRNA fragments that are un-capped (5' monophosphate) after cleavage can be obtained using high-throughput sequencing methods. (B) Cleavage that has been mediated by an sRNA can be seen as a cleavage signal (peak) in the mRNA fragment abundance when they are realigned to the mRNA.

2.6 Sequence alignment

Sequence alignment is a fundamental operation within all NGS sequence analysis strategies to identify sRNAs and their targets, and so we briefly review several short read alignment tools. We carried out some benchmarking to obtain performance timings using two datasets. A reference sequence set containing 39,640 transcripts (cDNA TAIR9 cdna 20090619) was obtained from The Arabidopsis Information Resource (TAIR) [118]. A sRNA query dataset to be aligned to the reference transcripts comprise 900,000 unique, randomly generated, short sequences. The short query sequences were within the size range of 20-25nt in length. We chose this dataset as it represented the dimensions of the experimental data available to us. We will call the reference transcripts and short query sequences Dataset A.

We ran Dataset A through RMAP, MicroRazerS, SOAP 2, PASS and PatMaN on a machine with the following specification: Dell Power Edge 2950, Quad Core XEON L5420 processor (2.5GHZ), 32GB RAM (667MHZ), 4 x 500GB SERIAL ATAU 7.2K 3.5" hard disk drive, running Linux operating system CentOS version 5.3. Due to each tool employing different algorithmic properties, parameter options, number of allowable mismatches and gaps, we will discuss each tool separately. A summary is given in table 2.1.

2.6.1 RMAP

RMAP is a tool which has been specifically designed for aligning short reads produced from next-generation sequencing technology and has recently been updated with improvements to mapping accuracy and memory requirements. The algorithm is essentially an approximate pattern matching technique using seeds, where a seed is a variable length substring of a query read [114],[110] .

The tool uses two mapping criteria. The first is that all unknown characters within a reference sequence, for example, the character 'N', are always counted as a mismatch. The second criterion uses base-call quality scores such as those scores provided in the FASTQ formatted files from NGS output (see 2.3.2). Recognised characters within a query read are considered either high quality or low quality depending upon their position and quality score. Characters in low quality positions act as wild cards and will result in a match [115].

In comparison to the other tools (see table 2.1), RMAP performs fast alignments and we found that it could process 900,000 query reads, with up to 10 mismatches, against a reference data set of 39,640 sequences (Dataset A) in approximately 2 minutes. We found it to be the fastest alignment tool with mismatches as compared to the other alignment tools we discuss here. However, during this test we found that some alignments reported by other

tools were not reported by RMAP. Using settings suggested within the user documentation, the tool was unable to perform an alignment for several of the query reads, therefore producing false negative results. Further adjustment of the tool's parameter settings failed to produce the alignments expected.

2.6.2 MicroRaZerS

MicroRazerS is a new addition to the SeqAn library [30] and is a tool which has been specifcally designed for aligning small RNA reads and, in particular, microRNAs to a reference dataset. The application employs a q-gram counting strategy, where a q-gram of length q is a sub-string of a query read. An index of q-grams is built and is then used to scan the reference data and filter potential alignments based upon the number of q-grams shared between reads and reference sequences. Alignments are then carried out upon the filtered data using a seed approach [32]. We found that we could process Dataset A in 17 minutes 21 seconds.

Though this tool can map reads to sequences allowing for a minimal number of mismatches, there are no options to configure the number of mismatches allowed during the mapping process. This means that if we were to use this tool, we would be dependent upon the MicroRazerS implementation of rules used to identify sRNAs, and the rules used are not

clearly defined.

2.6.3 SOAP 2

SOAP 2 [70] is an updated version of SOAP [110],[69], which has been re-implemented using Burrows Wheeler Transformation (BWT) compression[22] to improve its speed and memory requirements. The tool builds a BWT index table of the reference sequences and scans over it using a seed technique whilst allowing for up to 2 mismatches within the seed and a user specified number of mismatches within the remainder of the string. We found that we could process Dataset A in 3 minutes 35 seconds.

Though this tool performs fast alignments with mismatches and gaps, its level of accuracy produces false negative results similar to RMAP. We would surmise that its level of accuracy is linked to the use of seeds, the seed length and number of allowable mismatches within the seed. Unfortunately seed configuration is not parameterised. A potential problem with this application is its dependence upon 64-bit architectures which could limit its use.

2.6.4 PASS

PASS is a short read aligner which uses a seed word technique to perform the mappings between query and reference sequences allowing for gapped and

un-gapped alignments [23]. The algorithm aligns short reads to reference sequences by building an index of seed-words which is aligned with a pre-computed score table (PST). The PST is supplied with the software and provides score values for mismatches and gaps which indicates the quality of the alignment. The algorithm takes three steps in the alignment process. Firstly, the indexed seed words are used to scan the query reads. Secondly, it checks to see if the seed/read match can be extended to a full alignment and finally it refines the alignment and its mapping score. We found that we could process Dataset A in 5 minutes 28 seconds.

The tool allows mismatches but does not provide runtime parameters for their alteration. Though the documentation associated with the tool does not explicitly say how many mismatches are allowed, through testing, it has been found to allow up to 7 mismatches when aligning a short read of 24 nucleotides in length. An advantage to this application is the multithreading support which makes good use of multi-cored processors. It is written in C++ and supported on both Windows and Linux platforms.

2.6.5 PatMaN

PatMaN is a short read alignment tool which uses a keyword search tree [98]. The algorithm builds a search tree such that each query read is placed into the tree as a path from root to leaf, where edges represent nucleotides

and leaf nodes contain identifiers for query reads. The algorithm traverses the tree evaluating each nucleotide within the reference sequence allowing for a user defined number of mismatches and gaps. We found that we could process Dataset A in 9 seconds with 0 mismatches allowed and 4 hours, 26 minutes, 21 seconds with 4 mismatches allowed.

Table 2.1: Timing comparison for short read sequence alignment tools

Tool	Timing (hours, mins, secs)	Total mismatches	Configurable mismatches
RMAP	1m 47s	10	yes
MicroRazerS	17m 21s	undefined	no
SOAP 2	3m 35s	2	partial
PASS	5m 28s	7	no
PatMaN	0m 9s	0	yes
PatMaN	4h 26m 21s	4	yes

2.7 Discussion

Since the discovery of RNA silencing in the early 1990's, the sRNA field has become a diverse and rapidly expanding field of research. The advent of next generation sequencing and subsequent improvements in this technology provides researchers with a rich source of information relating to sRNAs. With this in mind, it is likely that many sRNAs and their targets, and potentially even new classes of sRNA are yet to be discovered. As sRNAs have been found to regulate gene activity in response to drought [91] and a new high-throughput technique for sRNA analysis is available to us, we can

use this knowledge to computationally analyse samples of RNA subjected to water stress versus those that are grown under optimal conditions. However, we first need computational tools to carry out this large-scale analysis. In the next chapter we describe PARESnip, a software tool that we designed to perform this task. In the following chapter we describe the application of this new tool to a degradome experiment devised to understand water stress in plants.

Chapter 3

High throughput sRNA/target interaction identification and validation using the degradome.

3.1 Summary

This chapter describes the multithreaded software application PARESnip (Parallel Analysis of RNA Ends - Snip) that we designed to analyse and validate sRNA/target interactions through the RNA degradome. In the next chapter we will use it to analyse sRNA/mRNA interactions involved in plant water stress. We start with describing the background followed by a detailed look at the methods we used to create the tool. We then provide the results from several degradome analyses and identify over 4000 sRNAs

and their targets. This work was published in the journal Nucleic Acids Research [38].

3.2 Background

As mentioned in the last chapter, high throughput sequencing has become a de facto standard for the analysis of sRNA samples [36],[53],[83]. Typically, a single experiment will produce millions of sRNA reads capturing a snapshot of the expression profile of the sRNAome in a single sample [92],[110]. As described in chapter 2, recent technological advances have enabled researchers to conduct high throughput target identification experiments in plants by using an approach called Parallel Analysis of RNA Ends (PARE) [41]. However, computational tools to analyse such data are both scarce and limited in functionality.

CleaveLand [2] was the first tool developed specifically to analyse degradome data, and it has been successfully used to identify micro RNA (miRNA) targets in a variety of organisms [1],[3],[71],[93]. Due to the algorithms implemented in CleaveLand and the size of sRNA and degradome data sets (typically millions of sequences) it is impractical to analyse all possible sRNA/degradome interactions using this software in a reasonable timescale without a large degree of parallelization across multiple machines. As a consequence the tool is generally used to find cleaved targets of a small

number of sRNAs, such as known or candidate miRNAs. This means that users typically have to ignore the vast majority of sRNA reads in such analyses and have to assume some prior knowledge of which sRNAs are likely to have targets. As a result many legitimate sRNA mediated mRNA cleavages could potentially be missed. While this is acceptable for users interested in looking for targets of known miRNAs, it greatly restricts the possibility to get a sense of all of the sRNA regulatory interactions leading to mRNA cleavage. In addition, CleaveLand is a command line based application that can only be used in a Linux/UNIX environment. This excludes a large number of potential users who do not have access to, or expertise in, such environments.

To the best of our knowledge, only two other methods have been developed for identifying sRNA/target interactions evidenced through the degradome in addition to CleaveLand; SoMART [68] and SeqTar [133]. SoMART is a collection of web server tools for processing sRNAs. To process degradome data, the user first needs to predict sRNAs that could potentially target a user supplied transcript with the Slicer detector tool. The dRNA mapper tool can then be used to align degradome sequences to the transcript sequence. The user then has to manually compare the output from Slicer detector and dRNA mapper to identify cleaved targets. To automatically process more than one transcript the user would therefore have to develop additional methods and post processing software. In addition,

the SoMART website is restricted to a prescribed list of sRNA and degradome databases. SeqTar attempts to broaden the alignment rules used in CleaveLand between sRNAs and their potential targets so as to identify miRNA targets. As with CleaveLand, SeqTar suffers from the fact that its underlying algorithms make it impractical to analyse all possible sRNA/degradome interactions in a reasonable timescale without a large degree of parallelization across multiple machines. Moreover, SeqTar is not available in a publicly downloadable package, which greatly reduces its potential user base.

In this chapter we describe a new, user friendly, cross platform degradome analysis tool, PAREsnip, which enables flexible and comprehensive high throughput target analysis, allowing users to identify genome wide networks of sRNA/target interactions resulting in transcript cleavage. As well as being able to analyse data sets like CleaveLand PAREsnip is also able to process entire sRNAome and transcriptome data sets in a short timeframe on a typical desktop computer.

3.3 Methods

3.3.1 Input

For a specific organism the inputs for PAREsnip are:

- mRNA dataset (transcriptome),
- transcript degradation fragments obtained from a PARE experiment (degradome),
- small RNA dataset (sRNAome) and
- the genome sequence.

The first three inputs are required but the genome is optional. When included, the genome is used during the data-filtering process described later. All of the inputs must be in FASTA format and must only contain the characters A, C, G, T and U. Sequences containing unknown characters and ambiguity codes are discarded as they cannot be accurately aligned later. FASTQ to FASTA and adaptor removal tools are provided within the UEA sRNA Workbench [89],[117]. An overview of the steps involved in processing the input data is shown in Figure 3.1.

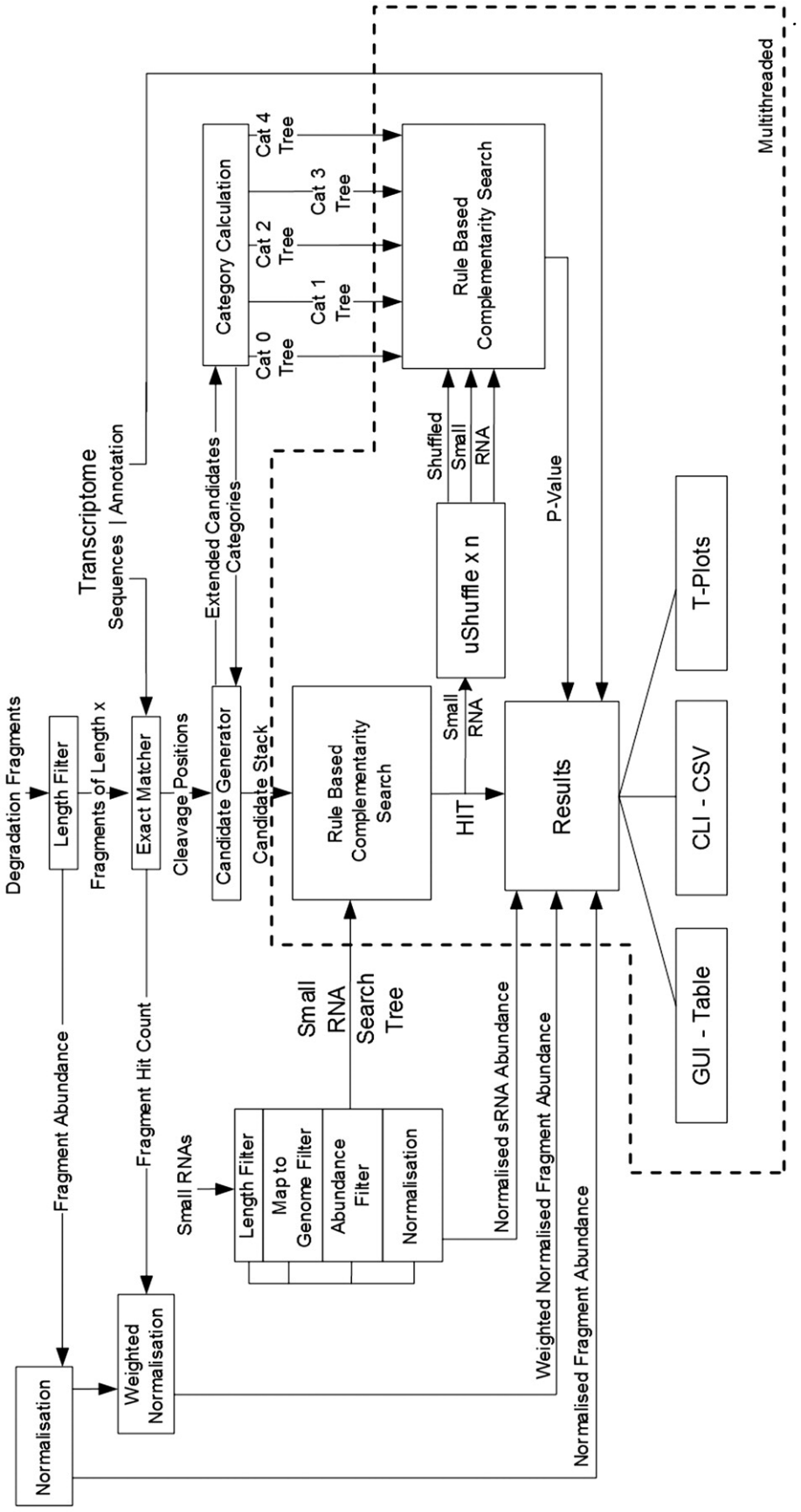


Figure 3.1: **Schematic of PAREsnip.** Boxes represent functions and solid arrowed lines represent data flow. The functions and dataflow operating concurrently using multithreading are enclosed with a dotted line.

3.3.2 Data filtering

Several user-configurable filters based on: sequence length, sequence abundance and sequence complexity may be applied to the sRNAome. If a sequence has an exact full-length match to known tRNA or rRNA, it will be omitted. T/rRNA sequences are obtained from Rfam [48] and EMBL/Genbank [58] sequence databases. If a genome is provided, sRNA sequences are mapped to it using PatMaN [98]. Any sequences without a match to the genome are removed from further analysis, as they are likely to be either sequencing errors or sample contamination.

3.3.3 Signals of cleavage

Degradome fragments are exactly matched to the transcriptome and 5-end alignment positions are recorded. The degradome fragment abundance at any given position could represent an sRNA cleavage event at that position [42],[1]. Potential cleavage sites on a single transcript can be categorized according to degradome read abundance. Higher abundance reads are more likely to be the result of endonucleolytic cleavage as opposed to random degradation products, which are more likely to accumulate at a lower background level. PAREsnip uses the 5-category system defined in CleaveLand (version 2) [2], which are:

- Category 0 is defined as a signal having greater than one raw read at

the position. The abundance at that position is equal to the maximum on the transcript, and there is only one maximum.

- Category 1 is the same as Category 0 in all aspects apart from there is more than one maximum on the transcript. This means that there are two or more signals on the transcript with the same strength (abundance).

- Category 2 is defined as a signal having greater than one raw read at the position. The abundance at that position is less than the maximum, but greater than the median abundance for that transcript.

- Category 3 is defined as a signal having greater than one raw read at the position and the abundance at that position is less or equal to the median value for that transcript.

- Category 4 is defined as only one raw read at the position.

The categorization of the signal strength is based on either the raw abundance or weighted abundance of degradation fragments; the latter is the default PAREsnip setting. Weighted abundance is calculated by dividing the abundance of a degradome fragment (tag) by the number of positions across all transcripts to which the tag has aligned. The strongest signals, described as Categories 0, 1 and 2, convey the strongest empirical evidence for true cleavage products [1]. The weaker Categories 3 and 4 signals could

be difficult to distinguish from background noise and random degradation. It is therefore possible for the user to exclude any of the five categories before commencing an analysis in PAREsnip.

Figure 3.2: (A) Applying the binding rules to the partitioned 4-way tree. Small RNAs are encoded into a 4-way tree. The tree is partitioned based on the nucleotides at positions 10 and 11 in the pattern sequence to be searched for. As the tree is searched, sRNA/target binding rules are applied. (B) Searching the partitioned 4-way tree. To search for a pattern within the tree we start at level 10 denoted as (1), which corresponds to the 10th nucleotide in a small RNA (counted from the 5 end). The tree is followed towards the root performing Watson and Crick base pairing denoted as (2). At each traversal, the binding rules are checked. If the root is reached successfully the algorithm jumps back to (1) and begins a pre-order walk down the tree, denoted as (3). While walking down the tree, if the rules are broken, then the traversals of that branch stop. If a terminator node is reached, then a successful alignment has been made and an sRNA/target interaction discovered.

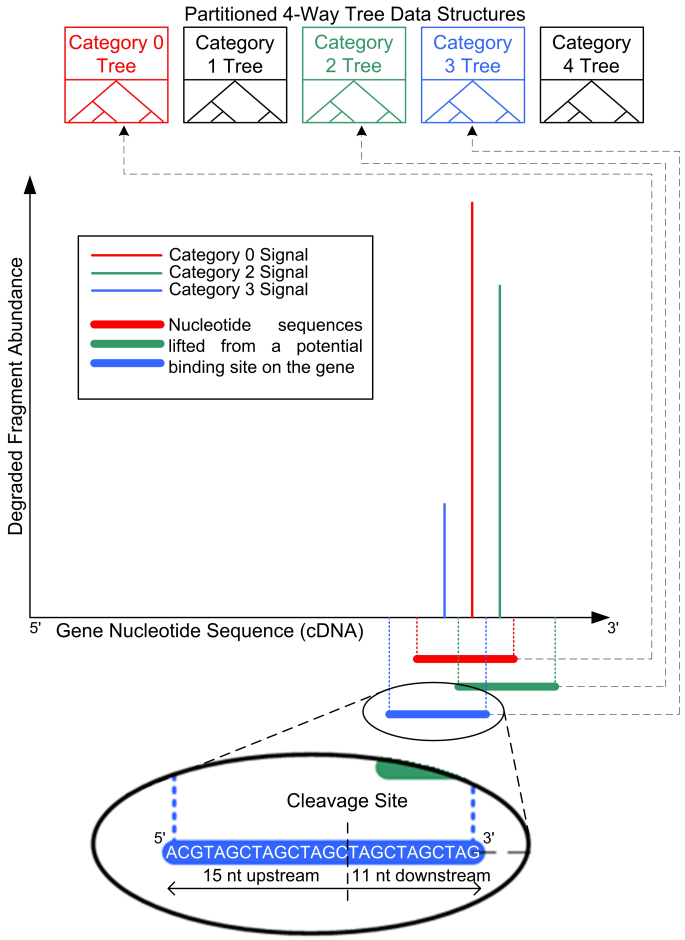


Figure 3.3: Data structure created from degradome fragments mapped to transcripts. Bars represent 5 ends of degradome fragments aligned to a transcript. Degradome signals are characterized by category. A sub-sequence of 26nt is extracted from the transcript based on the cleavage site. The sub-sequence is encoded into a partitioned 4-way tree according to the assigned category.

3.3.4 Data structures

Small RNA sequences are encoded into unique paths within a trie [44], which is an m-way search tree data structure. Since RNA and DNA sequences are described by the symbols (A,C,G,T or A,C,G,U) we use a 4-way tree (Figure

3.2 A). Edges represent nucleotide bases and nodes offer path choice through the tree. Many short sequences share a similar nucleotide composition. By encoding all sequences into a 4-way tree, those that share a similar composition will lie on the same path until the similarity ends and new branches are created. A terminator node marks the end of a path and therefore an sRNA sequence encoded within the tree. This structure allows us to remove sequence and subsequence redundancy, therefore reducing our search space and memory footprint. Also, the number of nucleotide/edge comparisons required when attempting to search for a sequence within the tree is reduced.

Once the sRNAs are encoded in the tree, target searches can be performed. The starting node for each search is the 10th node because we know that position 10 of the sRNA/target duplex must be complementary in order to cleave a target [73],[106]. Therefore pairs of nodes at levels 10 and 11 within the 4-way tree are collected and placed into labelled bins (table 3.1) according to the pairs nucleotide composition. There are a total of 16 bins that correspond to the 16 possible dinucleotide combinations. Searches for sRNAs that could cause cleavage at a given degradome peak position are initiated by identifying the bin corresponding to nucleotides 10 and 11 of the candidate sequence. The tree is then traversed from nucleotide 10 towards the root. We place a restriction that once a walk up the tree from an entry point has occurred, the parent node of the entry point obtained from

the bin may never be visited again during the current search and only descendent nodes of the entry point may be traversed. This restriction ensures that unnecessary nucleotide comparisons are not computed. We partition the tree by hiding all paths that have starting nodes in any of the other 15 labelled bins.

Table 3.1: **Organisation of partitioned 4-way tree entry points.** Nodes at levels 10 and 11 within a 4-way tree data structure are collected and placed into labelled bins. There are a total of 16 bins as there are a total of 16 possible dinucleotide combinations. The label for each bin is the nucleotide at level 10 followed by the nucleotide at level 11. The bins hold entry points into the tree data structure. Entry nodes within a bin are used to partition the 4-way tree.

Bin Label	Bin Number	Tree Level 10	Tree Level 11
AA	1	A	A
AC	2	A	C
AG	3	A	G
AT	4	A	T
CA	5	C	A
CC	6	C	C
CG	7	C	G
CT	8	C	T
GA	9	G	A
GC	10	G	C
GG	11	G	G
GT	12	G	T
TA	13	T	A
TC	14	T	C
TG	15	T	G
TT	16	T	T

The organization of the data in this way lends itself to the fast mapping of sequences in an all-against-all search because only a small fraction of the millions of sequences obtained from a high-throughput sequencing experiment, that are encoded into the 4-way tree, have the potential to be aligned with the candidate pattern. This is possible as we know that the 10th and 11th nucleotides of the sRNA, which sit at levels 10 and 11 in the tree, must match the 10th and 11th nucleotide of the search pattern exactly [106]. This contributes to the computational speed of PAREsnip.

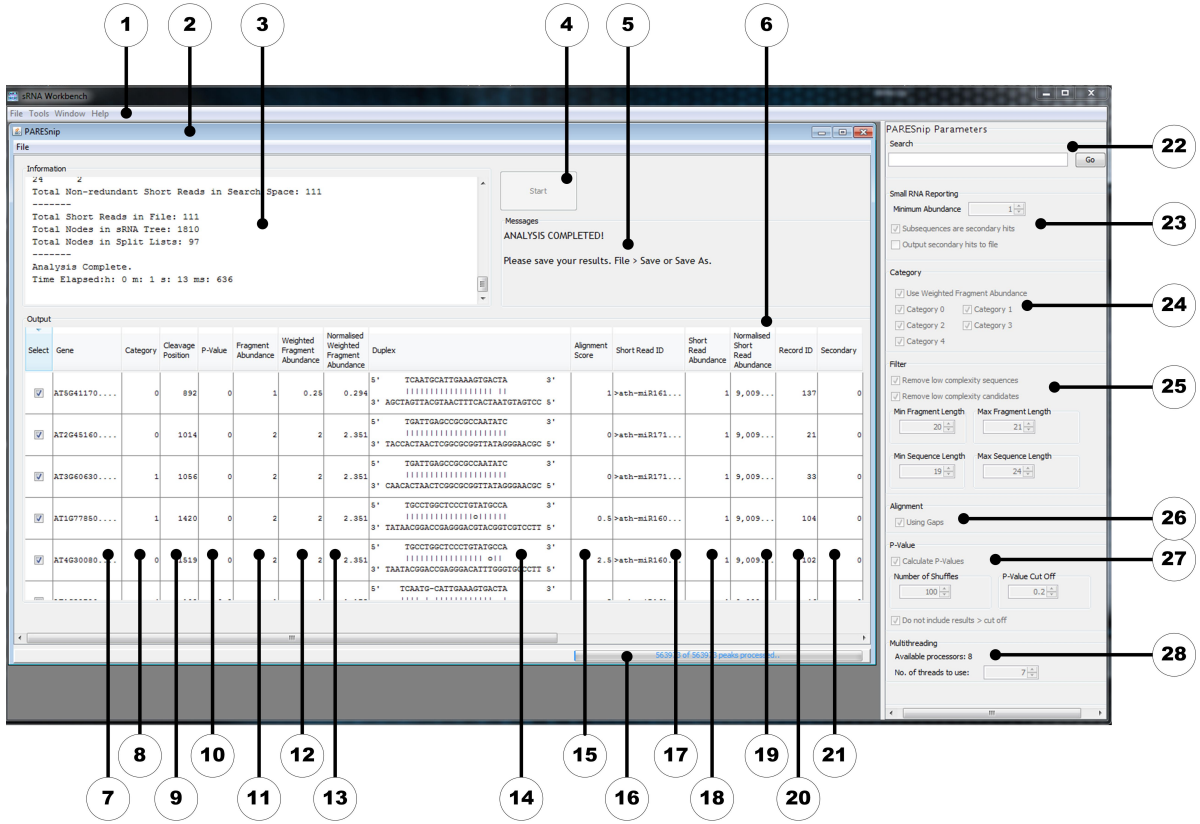


Figure 3.4: **PARESnip's Graphical User Interface** Elements of the interface are numbered 1 to 28. (1) UEA sRNA Workbench. (2) PARESnip. (3) Statistics related to the input data. (4) Starts an analysis. (5) Help messages to the user. (6) Main output table. (7) Gene annotation. (8) Cleavage category (signal strength). (9) Nucleotide position of cleavage. (10) p-value. (11) Raw abundance of degradation fragments aligned to position. (12) Weighted degradation fragment abundance aligned to position. (13) Normalised weighted abundance of fragments aligned to position. (14) Visual sequence alignment. (15) Total alignment score (G:U pairs + mismatches + indels). (16) Analysis progress bar. (17) Annotation of sRNA. (18) Abundance of sRNA. (19) Normalised abundance of sRNA. (20) Unique identifier for each record. (21) Total subsequences of sRNA which align to this position. (22) Search tabular output for text. (23) Abundance and subsequence filter for sRNAs. (24) Signal calculation option and signal strength reporting options. (25) Filters for low complexity and length of degraded fragments and sRNAs. (26) Allow a single gap in an alignment. (27) Number of shuffles to be used and cut-off value when calculating p-values. (28) Number of processors available and number of processors to be used.

3.3.5 Search algorithm

The core of PAREsnips operation is what we call the Rule-Based Complementarity Search algorithm. It is a method of traversing the partitioned 4-way tree, searching for sRNA sequences that could potentially cleave a transcript accounting for the degradome peak at a given position. The method is designed to make as few nucleotide comparisons as possible and will disregard the large sections of the 4-way tree that will never produce a valid alignment, based on a set of previously described targeting rules [106],[5]. The rules used by the search algorithm are user configurable and the default settings are:

- No more than four mismatches between sRNA and target (G-U bases count as 0.5 mismatches).
- No more than two adjacent mismatches in the sRNA/target duplex.
- No adjacent mismatches in positions 2-12 of the sRNA/target duplex (5' end of the sRNA).
- No mismatches in positions 10-11 of sRNA/target duplex.
- No more than 2.5 mismatches in positions 1-12 of the sRNA/target duplex (5' of sRNA).

The algorithm requires a candidate pattern on which to execute its rules. The pattern is the reverse complement of the first 11-nt downstream and

up to 15-nt upstream from the position of a categorized degradome cleavage signal on the transcript. The algorithm looks at the two nucleotides either side of the cleavage position in the pattern and identifies the appropriate bin (Table 3.1). The algorithm retrieves a starting node from the bin and traverses a single path up the tree to the root (Figure 3.2 B). As it does so, it makes a nucleotide comparison between the pattern and the edge in the path and tests the rule set (Figure 3.2 A). If at any point one of the rules is broken, the search is aborted, the starting node discarded and the next starting node is obtained from the bin. If, on the other hand, the algorithm successfully reaches the root of the tree without breaking any of the rules, then it returns to the entry point and begins a pre-order walk through the tree. A history of alignment records is kept while the tree is traversed. Each record is composed of nucleotide matches, mismatches and single gaps along with a running alignment score. A mismatch contributes 1.0 to the score, unless it is a G-U (wobble) pair in which case it contributes 0.5 to the score. A gap in the alignment contributes a value of 1.0 to the score. If a terminator node is found, then the algorithm must have reached it without breaking the rules in one or more of the alignment records kept in its history. In this case the algorithm examines its history of alignment records and selects the alignment with the lowest score and places it onto a communal stack of identified valid alignments. If at any point a rule is broken during a traversal and there is no valid alignment in its maintained history, the

algorithm no longer continues down its current path. When there are no more paths to traverse, the algorithm looks in the bin and if there are any remaining starting nodes, it will obtain the next starting node from the bin and repeat the procedure until the bin is empty. The stack of valid alignments represents possible sRNA/target interactions. Each interaction within the stack is passed on to the system to calculate the P-value before being reported to the user.

3.3.6 Calculating p-values

For each sRNA/target duplex reported by PAREsnip, a P-value is calculated. The P-value gives us a score that indicates how likely the reported duplex occurred by chance. The P-value calculation methods are based on those published in CleaveLand (version 2.0) [2] but use our Rule-Based Complementarity Search algorithm and partitioned 4-way trees during the calculation. For every position, on every mRNA containing a cleavage signal a 26-nt sequence representing the sRNA-binding site is extracted and placed into one of five possible category trees (Figure 3.3). The category trees are the same in structure and function to the partitioned 4-way tree used to encode sRNAs, but instead contain sections of mRNAs where cleavage has occurred.

The sRNA for each sRNA/target alignment on the stack of valid align-

ments is randomly shuffled and mapped to all target sites encoded into a 4-way tree (Figure 3.2 A). The chosen 4-way tree corresponds to the category given to the output sRNA/target record. The random shuffles of the sRNA preserve dinucleotide frequency and are generated by the third-party Java programme uShuffle [55]. The user may define the number of shuffles to be used (the default is 100) and the resulting P-value is the number of times the randomly shuffled sRNA aligns to a target site encoded within the category tree. The P-value is provided as a decimal. For example, if 100 shuffles were used and 5 of those aligned to a target site of the same category, then the resulting P-value would be 0.05. An alignment below the user-specified P-value cut-off is accepted as valid and output to file or to the user interface.

3.3.7 Output

PAREsnip displays results in a tabular format where each row in the table shows an sRNA/target interaction. The columns show alignment category, P-value, binding score and abundance information along with a visual sequence alignment of sRNA and target mRNA. Statistics relating to the input data set are provided such as sequence count and sequence length distribution. When the tool is operated in GUI mode (Figure 3.4), a results table is displayed and updated as interactions are found. Columns and rows

may be sorted and re-arranged and the data in the table may be saved as comma separated value (csv) format. If the user operates the tool from the command line, the table is saved straight to disk in csv format, which can be imported directly into most spreadsheet and statistical packages. PAREsnip lets the user generate and investigate publication quality t-plots through the UEA sRNA Workbench tool called VisSR [89],[117].

3.3.8 Availability

PAREsnip is a multi-platform, multi-threaded (Figure 3.1) application written in Java and is released as part of the UEA sRNA Workbench [89],[117] (http://srna-workbench.cmp.uea.ac.uk). It may be run from the command line or a graphical user interface (GUI).

3.4 Results

3.4.1 Benchmarking

To measure the runtime performance of PAREsnip we simulated 10 sRNA data sets of increasing size. The sRNAs were generated by extracting 1924nt sequences centred on cleavage positions within the Arabidopsis thaliana transcriptome (TAIR 10 representative gene model) [118]. Transcripts, cleavage positions and sRNA sequence lengths were selected at random.

The performance of PAREsnip was measured by using the simulated sRNAs with the A. thaliana transcriptome and the publicly available PARE degradome library GSM278370 A. thaliana Col-0 wild-type seedlings [1],[11]. We observe a linear time operation with a peak memory requirement of 5.5 gigabytes.

We also benchmarked the performance of CleaveLand (version 2) and compared the runtime with that of PAREsnip (Table 3.2). We found that PAREsnip significantly outperformed CleaveLand for the considered data sets. Note that, even though there is a version 3 of CleaveLand, we compared PAREsnip with version 2 since the target prediction step of version 3 only receives a single sRNA sequence for analysis, and therefore cannot be practically used on larger numbers of sRNAs without developing additional software. Even so, to get a rough idea of the performance of CleaveLand (version 3), we obtained an average runtime of 87s per sRNA sequence for 10 simulated sRNAs, which is roughly 3 times faster than version 2, but still significantly slower than PAREsnip.

3.4.2 Comparison with CleaveLand

As CleaveLand is currently the only publicly available tool for degradome analysis, we compared all miRNA targets reported by CleaveLand (version 2) [2] with those reported by PAREsnip using two data sets. We ob-

Table 3.2: **Run time for PAREsnip and CleaveLand**

Number of sRNAs	CleaveLand Timing	PAREsnip Timing
10	46 min 6s	29s
25	1h 55 min 25s	30s
50	3 h 51 min 35s	31s
1000	-	2 min 3s
10 000	-	10 min 14s
20 000	-	19 min 11s
40 000	-	39 min 8s
60 000	-	53 min 9s
80 000	-	73 min 24s
100 000	-	87 min 16s

tained all known mature A. thaliana miRNAs from miRBase (release 17) [60] and analysed them using both tools, seeking targets within the transcriptome (A. thaliana representative gene model TAIR release 10) [118] using two publicly available degradome libraries: GSM278335 and GSM278370 A. thaliana Col-0 wild-type inflorescence tissue taken from Gene Expression Omnibus (GEO) [1],[11]. A collection of previously validated miRNA targets obtained from the literature [42],[47],[54],[86],[35] and the MPSS database [90] (Supplementary Table S1, see Chapter 3.5) were used to identify previously validated miRNA targets reported by both tools.

The results are summarized in Figure 3.5 (full results in Supplementary Tables S2 and S3, see Chapter 3.5). As can be seen, PAREsnip reports either the same number or slightly more previously validated targets than

CleaveLand. The interactions reported by PAREsnip and not by CleaveLand or vice versa are due to the random factor within the P-value systems used by both tools. For example, in contrast to CleaveLand, PAREsnip uses dinucleotide random shuffles when calculating a P-value through the use of uShuffle [55]. Furthermore, differences between the interactions predicted by the two tools are probably also due to the reporting of hits that contain a mismatch at position 10 (from 5' of sRNA), multiple gaps within a duplex and more than 2.5 mismatches or adjacent mismatches within the seed region (positions 112 5' of sRNA) of the duplex. Again, in contrast to CleaveLand, these features within a duplex are not permitted by the Rule-Based Complementarity Search algorithm used by PAREsnip.

Figure 3.5: Venn diagram showing the comparison of results produced by CleaveLand and PAREsnip. The Venn diagram shows the intersection of predictions made by PAREsnip and CleaveLand and is a summary of the results within Supplementary Tables S2 and S3 (see Chapter 3.5).

3.4.3 Filtering by p-value

To examine the usefulness of the P-value computed by PAREsnip as a confidence score upon which predicted interactions can be excluded, we ran it on all known mature A. thaliana miRNAs, GSM278370 [1],[11] degradome and the A. thaliana transcriptome (representative gene model, TAIR release 10) (32) with increasing P-value thresholds. The predictions were compared with previously validated interactions (Supplementary Table S1, see Chapter 3.5) to provide an insight into the number of validated interactions retained along with the number of other interactions reported in relation to the increasing threshold (Figure 3.6). Note that a P-value cut-off of 1 captures all possible predictions. PAREsnip reported a total of 91 validated and 1026 non-validated interactions using a P-value cut-off of 1. We find that a threshold of 0.05 captures 94.5pc of possible validated interactions (a loss of 5.5pc validated interactions) while capturing 7.6pc of the total non-validated interactions. In light of this and other similar experiments we have chosen a default P-value setting for PAREsnip of 0.05.

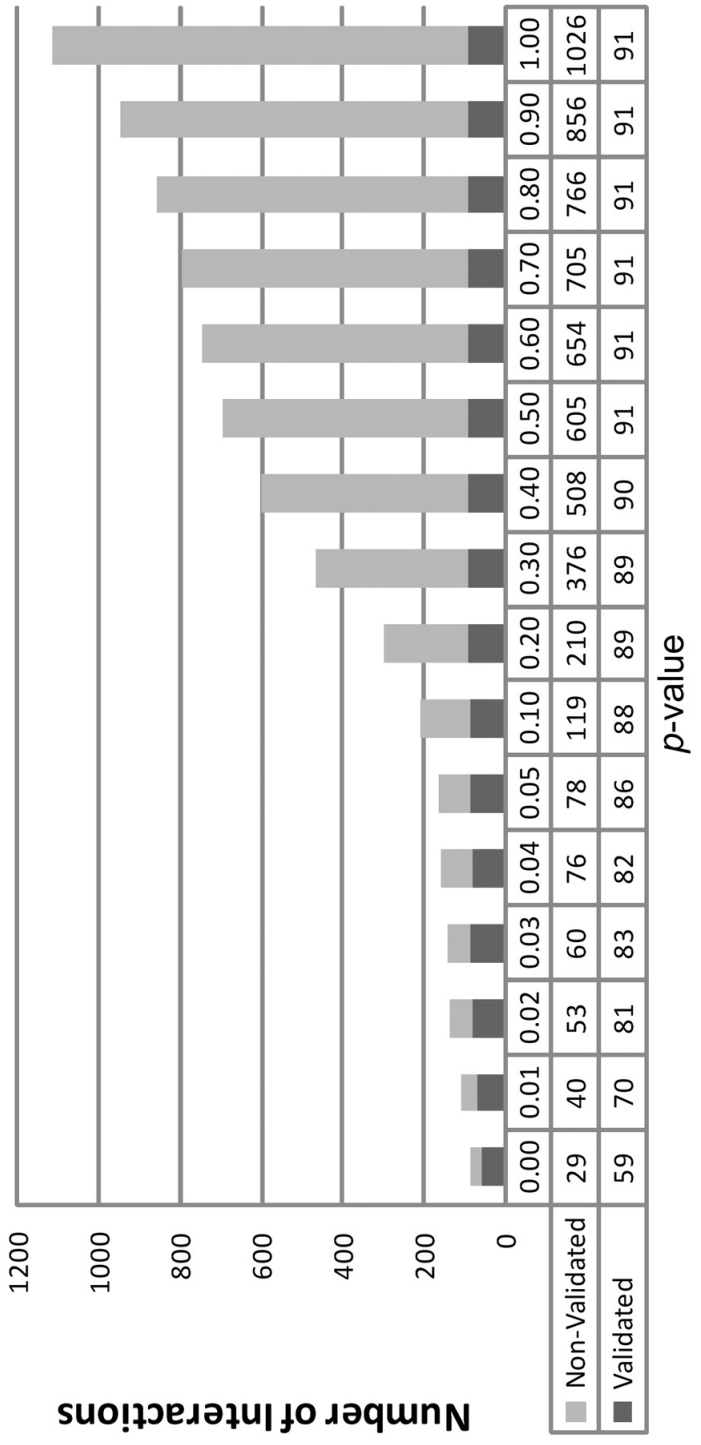


Figure 3.6: Interactions reported by PAREsnip with P-value increases. Starting from the smallest P-value of 0.00, we see a progressive increase in the number of small RNA/mRNA interactions reported. The P-value cut-off of 0.05 captures 94.5pc of total validated interactions reported by PAREsnip and is the default setting.

3.4.4 Genome-wide discovery of sRNA/target interactions

Small RNA sample libraries obtained from a high-throughput sequencing experiment typically contain millions of sequences. To look for interactions on a genome-wide scale, including all sRNAs obtained from a high-throughput sequencing experiment, we used PAREsnip to analyse the following data sets: sRNAome GSM342999 A. thaliana Col-0 biological replicate 1 inflorescence tissue [11],[87]; degradome GSM278335; transcripts: A. thaliana (representative gene model TAIR release 10) [118]. For this and every subsequent analysis the following settings were used: a maximum of 4.0 mismatches, 100 dinucleotide shuffles and a P-value threshold of 0.05. Within these data, PAREsnip reported 36,351 interactions. Despite the support found for these interactions, in particular the degradation signal, observed sRNA, sequence specificity within each duplex and low P-value, it is difficult to believe that so many interactions are genuine. Therefore the combined restrictions of mismatch positions, the number of permitted mismatches and P-value filter, on their own, do not appear to be sufficient measures to extract valid interactions above the noise when performing an analysis on such a large scale. It is likely that many degradome signals are not the product of sRNA-induced cleavage but are instead random degradation fragments that happen to also be complementary to one or more of the millions of sRNA inputs. To address this problem we employed cross-sample

conservation with the aim of reducing the number of reported targets. The rationale behind this approach is that both degradome fragments and sRNA sequences that are products of random degradation are unlikely to be conserved between biological replicates whereas bona fide cleavage signals and functional sRNAs are likely to be present across samples.

To explore this approach we used PAREsnip to independently analyse two sRNA biological replicates GSM342999 (set B1) and GSM343000 A. thaliana Col-0 biological replicate 2 inflorescence tissue (set B2) [11],[87] along with the degradome GSM278335. The results were compared and only the conserved interactions across the two samples were retained. For an interaction to be conserved the interaction must share the same target transcript, cleavage site and sRNA sequence. In set B1 36351 interactions were identified (Supplementary Table S4a and b, see Chapter 3.5) and in set B2 26098 interactions (Supplementary Table S5a-c, see Chapter 3.5). By comparing the interactions between the sets we found 7273 conserved interactions. To ascertain whether such a result could occur by chance, we carried out the same experiment again but using simulated sRNA sets containing randomly generated sequences. The simulated sets (set R1 and R2) maintained the same characteristics as the real sRNA libraries, including unique and redundant sequence count and sequence length distribution. The sequences themselves were randomly generated by sampling from the Arabidopsis genome sequence. Set R1 identified a total of 21,783 interac-

tions and R2 identified 21,862 interactions. Comparing the interactions of R1 and R2 using the same conservation criteria we found that no interactions were conserved. This indicates that sRNAs being observed in multiple samples (biological replicates) could provide a method for extracting reliable hits above noise with some measure of confidence.

We extended the conservation method to include signals of degradation so that a reliable interaction should contain degradation products that are conserved across multiple degradome library samples as well as the sRNA being conserved across multiple sRNAomes. We analysed two data sets: Set D1 comprised sRNAome-GSM342999 and degradome-GSM280226 A. thaliana Col-0 inflorescence tissue [11],[42] and set D2 comprised sRNAome-GSM343000 and degradome-GSM280227 A. thaliana xrn4 inflorescence tissue [11],[42]. Reference transcripts were the A. thaliana representative gene model (TAIR release 10) [118]. Within sets D1 and D2 we found a total of 65110 and 49938 interactions, respectively. The 65110 interactions are shown in Supplementary Table S6a-d (see Chapter 3.5), and the 49938 interactions are shown in Supplementary Table S7a-c (see Chapter 3.5). Based on the previously validated interactions (Supplementary Table S1, see Chapter 3.5), 163 and 179 interactions within the total number of interactions found in sets D1 and D2, respectively, had been previously experimentally validated. When comparing the results of sets D1 and D2 we found a total of 4466 conserved interactions. Of the validated interactions, 149 were

conserved giving an above 80pc retention rate. The 4466 conserved interactions meet the binding rules criteria for mismatch positioning within the sRNA/mRNA duplex and have a mismatch score of 4 or less. They have a P-value of 0.05 or less and the sRNA and positional cleavage signal are conserved across multiple samples.

3.5 Supplementary tables

The supplementary tables S1 through to S7 mentioned within this chapter contain large amounts of data and it is not practical to include them in print. However, for completeness, we provide a brief description of the data contained within each supplementary table below and the data tables are freely available for download from Nucleic Acids Research Online (http://nar.oxfordjournals.org/content/40/13/e103/suppl/DC1).

Supplementary table S1 shows the compiled results from a literature review that was carried out to obtain experimentally validated and predicted mRNA targets for all known miRNAs. All known miRNAs were obtained from miRBase [60] and a total of 707 validated or predicted miRNA targets were obtained from 11 independent studies [10],[18],[25],[35],[36],[42], [47],[54],[86],[90],[93].

Supplementary tables S2 and S3 show the results produced by PAREsnip and CleaveLand when analysing two degradome libraries. Re-

sults that had been previously experimentally validated are identified within the tables.

Supplementary tables S4 through to S7 show the raw unprocessed results produced by the PAREsnip tool in 4 degradome analyses.

3.6 Discussion

We have described a novel, freely-available application called PAREsnip, designed for the identification of cleaved targets from sRNA and degradome data sets generated using next-generation sequencing technologies. The tool can also be used on small-scale experiments. PAREsnip is a user-friendly GUI-based, cross-platform (Windows, Linux, MacOS) application that enables biologists to run the application and analyse their data without the need for dedicated bioinformatics support or specialized computer hardware. We have also made a command-line version of the tool available for users who wish to incorporate PAREsnip into computational pipelines.

We have shown that PAREsnip performs at least as well as current methods in detecting validated miRNAmRNA interactions in published data sets and that it runs significantly faster than the competition on a standard desktop computer. The speed of PAREsnip opens up new avenues in the sRNA field as it enables users to look for targets of all sequenced sRNAs rather than a subset of sequences that they suspect might have a target (such as

annotated miRNAs and trans-acting small interfering RNAs).

We have demonstrated that degradome and sRNA data are inherently noisy (probably due to background mRNA degradation) and that searching a random sRNA data set with the same properties as a real input data set against the degradome can lead to a comparable number of predicted target interactions. This makes it difficult to separate real targets from false positives when running on high-throughput data. However, by using biological replicates of sRNA and degradome data sets we appear to be able to remove spurious degradation products, as they are highly unlikely to be conserved between two or more samples. We show that by using this conservation method on a random sRNA set no targets are predicted (resulting in zero false positives), whereas when applying it to a real set we retrieve 4466 high-confidence interactions and recover 80pc of the previously validated targets present.

PAREsnip is extensively user-configurable; this allows users to customize search parameters and binding rules in order to make searches more liberal or stringent. It was recently reported that several new miRNA targets were discovered and validated using more relaxed binding rules implemented in the SeqTar algorithm [133]. By relaxing the stringency of the binding rules PAREsnip can also be used to search more deeply for individual miRNA targets. Conversely, tightening the rules will lead to a reduction in the number of candidates reported when run across entire sRNA sets. This

flexibility also allows users to customize searches and could allow them to optimize parameters for searching degradome data sets such as those published by Bracken [20] and Karginov [57].

While the use of published binding rules and P-value filtering provides a strong set of predicted sRNA/target interactions it is difficult to estimate an accurate false positive rate. One of the reasons is that currently there is no experimental method to directly test sRNA/target interactions. The only method is the 5' RACE to map the non-capped 5 end of individual mRNA fragments. However, this method is based on the same principle as the PARE/degradome library generation and so it is questionable whether it can be used to validate the high-throughput results. In fact, since 5RACE experiments focus on a small region of an mRNA, it is more likely to yield an artefact than the unbiased PARE/degradome library approach.

3.7 Conclusion

PAREsnip can be used to search for genome-wide interactions between all sRNAs and transcripts as well as predicting targets of small groups of miRNAs. This high-throughput approach to degradome analysis opens a new avenue for researchers interested in identification of sRNA targets. Due to its speed and efficiency PAREsnip removes the need for users to know in advance which sequences are likely to have a target and instead allows

users to generate complete networks of sRNA target interactions. By using replicates and applying a conservation rule we predict over 4000 putative sRNA/mRNA interactions in the Arabidopsis sets we analysed. This suggests that sRNA-mediated targeting and cleavage of transcripts may be even more widespread than previously anticipated and provides a useful new tool for experimentalists to study such interactions in more depth. In the following chapter we will use PAREsnip to analyse sRNA/mRNA interactions involved in plant water stress.

Chapter 4

Analysis of the RNA degradome during plant water stress.

4.1 Summary

In this chapter we describe the analysis of stress response RNA degradome datasets for barrel medic (*Medicago truncatula*), a model legume species. These datasets were obtained from the high-throughput sequencing of four experimental degradome libraries prepared using the PARE protocol. The libraries were prepared by biologist Dr. Gyorgy Szittya (Dalmay group, UEA). The data provides us with four snapshots of mRNA degradation at two distinct intervals of water stress. The intervals are control and dehydration. This data provides us with a rich source of information that we can analyse using the new PAREsnip tool. We used the tool to find

previously described and potentially novel sRNA/target interactions and compare levels of sRNA mediated activity across the water stress intervals.

4.2 Background

One of the consequences of a warmer climate is drier conditions for crops. It has been projected that by 2050 seasonal average temperatures will be higher than ever experienced in the past century [13]. Drier conditions and temporary extreme weather conditions causing drought can threaten crop yield and adversely affect food production, farm income and food security worldwide [45]. Plants can temporarily adapt to water shortage to survive suboptimal conditions. However, harvestable yields produced by stress tolerant species is the focus of agriculture rather than plant survival. In particular, agriculture prefers species that can tolerate abiotic stress such as drought during the relatively short, but important growing periods.

Plants can respond to changes within their environment by regulating gene expression at the post-transcriptional level. Recent studies have demonstrated changes in sRNA mediated gene regulation in response to environmental stress factors such as cold [135],[121], salinity [40],[74] and drought [134],[67] in a number of plant species such as Rice (*Oryza sativa*), Wheat (*Triticum aestivum*), Poplar (*Populus euphratica*) and the model plant *Arabidopsis thaliana*. A better understanding of how plants respond

to environmental stress factors through sRNA mediated gene regulation could allow experimentalists to potentially modify important crop plants and improve their resilience to environmental change. For example, Ni et. al. (2012) [91] demonstrated that within a transgenic line of *Arabidopsis thaliana*, the over expression of the miRNA miR394a and the subsequent down regulation of its mRNA target (Glyma08g11030) improved the plant's tolerance to temporary severe water stress conditions. They showed that the transgenic plants over expressing miR394a lost water more slowly than wild-type plants and upon rehydration, the majority of transgenic plants were able to recover and continue growing, unlike wild-type plants which had only a 25% survival rate.

Legumes are important crop plants, accounting for one third of the worlds primary crop production, covering 12% to 15% of the worlds arable surface [15],[46]. The model species barrel medic is a legume that has a small diploid genome and has been used to study sRNA activity using high-throughput sequencing techniques [119] and in particular, sRNA activity during plant water stress [125]. Wang et. al (2011) [125] identified a total of 40 (32 known and 8 novel) miRNAs that are responsive to water stress in barrel medic. However, the study focused heavily upon differential expression analysis of miRNAs rather than their mRNA targets. Currently, miRBase [60] holds over 600 miRNA sequences for *Medicago truncatula*, but little is known about their function. To better understand how im-

portant crop plants such as legumes respond to water stress through gene regulation, and identify the function of known miRNAs, several genome-wide RNA degradation profiles of barrel medic were sequenced at different stages of dehydration. The RNA degradation profile for roots and leaves at control and dehydrated states were obtained using the PARE [41] protocol. With the high-throughput tool PAREsnip, we analysed the sequenced barrel medic degradomes and identified potential stress responsive mRNAs that are differentially cleaved by miRNAs under water stress conditions.

4.3 Methods

4.3.1 Sequencing of the Medicago degradome

Tissue samples were extracted from Medicago truncatula leaves and roots under control (hydrated) and water stress (dehydrated) conditions. Four degradome libraries were prepared from the 4 tissue samples using the PARE [41] protocol. High-throughput sequencing of the libraries was carried out by BaseClear on the Illumina platform using the Genome Analyzer II instrument. The resulting RNA degradation profiles in FASTQ format were:

- Control leaf (hydrated): dataset named CTA,
- Stress leaf (dehydrated): dataset named SWA,
- Control root (hydrated): dataset named CTR,

- Stress root (dehydrated): dataset named SWR.

4.3.2 Medicago genome, transcriptome and miRNAs

The Medicago genome and transcriptome (reference mRNAs) version Mt3.5v4 [130] were downloaded from the J.Craig Venter Institute (JCVI) file transfer protocol (FTP) server. The FASTA formatted transcriptome comprise full length cDNA sequences and their associated gene annotations obtained from gene predictions. JCVI define a full length cDNA sequence as the expressed regions (exons) and the untranslated regions (UTRs) but does not include the intragenic regions (introns). All mature Medicago truncatula miRNAs (348 total unique mature miRNA sequences) were downloaded from miRBase (release 18) [60].

4.3.3 Data preparation

As the miRNAs downloaded from miRBase and the transcripts and genome dowloaded from JCVI were in FASTA format already, only minor preparation was required. Adaptor sequences were removed from degradation fragments in each of the 4 libraries (CTA, SWA, CTR, SWR) using the UEA sRNA Workbench Adaptor Remover tool (version 2.3.2). The FASTQ formatted degradome files were converted to FASTA format required by the PAREsnip degradome analysis tool. The redundant and non-redundant

(unique) degradome read counts as well as size distribution for each of the libraries are summarised in Figure 4.1:

	CTR		SWR		CTA		SWA	
Fragment Length (nt)	Total reads	Distinct reads	Total reads	Distinct reads	Total reads	Distinct reads	Total reads	Distinct reads
16	8,245	4,165	9,866	4,349	4,284	2,331	3,998	2,243
17	16,940	7,768	20,390	7,998	8,376	4,180	10,636	5,073
18	18,278	9,360	25,504	9,741	14,337	7,178	15,735	7,894
19	91,991	65,398	152,317	89,441	135,385	79,075	116,257	74,607
20	8,724,582	3,964,103	16,163,523	4,891,285	13,923,701	4,338,860	12,589,561	3,963,231
21	6,604,880	3,152,534	12,421,234	3,922,326	14,100,074	4,100,594	10,946,527	3,482,931
22	13,349	7,852	23,472	12,385	27,454	12,537	22,157	11,606
23	5,896	3,129	8,335	3,696	10,489	4,435	8,256	3,523
24	4,835	2,697	7,685	3,474	8,867	3,901	7,044	3,103
25	3,179	1,956	6,476	3,039	7,121	3,342	5,465	2,539
26	2,505	1,734	7,479	3,402	8,429	3,873	5,582	2,910
27	1,653	1,237	5,944	3,102	7,416	3,766	4,510	2,597
28	777	648	3,716	2,192	5,208	2,848	2,868	1,911
Total	15,497,110	7,222,581	28,855,941	8,956,430	28,261,141	8,566,920	23,738,596	7,564,168

Figure 4.1: Summary of degradome library contents for water stress samples. CTR = control root. SWR = stress root. CTA = control leaf. SWA = stress leaf.

4.3.4 Analysis pipeline

To carry out the PAREsnip analysis on the four PARE libraries, three inputs in the form of a degradome, a set of miRNAs and a transcriptome (mRNAs) were required. More specifically, we used the following inputs:

- Analysis 1: Degradome CTA, all miRBase Medicago miRNAs, transcriptome Mt3.5v4,
- Analysis 2: Degradome SWA, all miRBase Medicago miRNAs, transcriptome Mt3.5v4,
- Analysis 3: Degradome CTR, all miRBase Medicago miRNAs, transcriptome Mt3.5v4, and
- Analysis 4: Degradome SWR, all miRBase Medicago miRNAs, transcriptome Mt3.5v4.

The parameter settings used for each of the above degradome analyses using PAREsnip can be found in Appendix A.

The conservation method described in Chapter 3.4.4 was used to identify miRNA/ mRNA target interactions that are conserved between control and stress samples for root as well as miRNA/mRNA target interactions that are conserved between control and stress samples for leaf. The final output of the computational pipeline consisted of six datasets identifying degradome evidenced miRNA/mRNA interactions. The six datasets show

interactions found in: dehydrated root sample only; dehydrated leaf sample only; hydrated root sample only; hydrated leaf sample only; hydrated and dehydrated root samples; hydrated and dehydrated leaf samples (Figure 4.2).

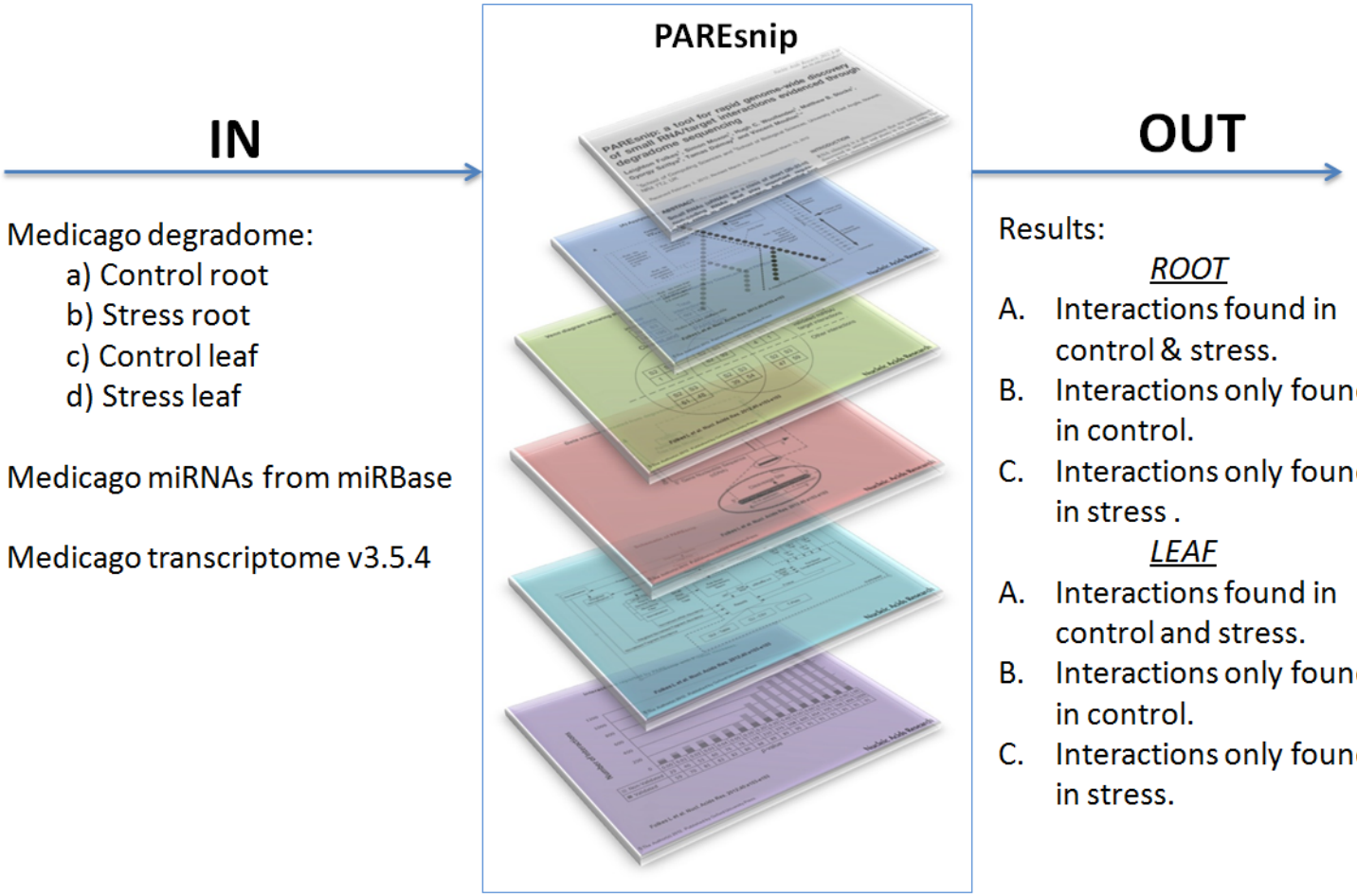


Figure 4.2: Overview of a degradome data analysis pipeline used using PAREsnip.

4.3.5 Candidate selection

From the interactions conserved between control and stress samples, we selected for deeper investigation those target transcripts that were cleaved by miRNAs that had a differentially cleaved fragment abundance between the hydrated and dehydrated states. To identify a change in abundance of cleavage products between conserved interactions found in both control and stress states we calculated a $\log_2$ ratio:

$$\log \text{ ratio} = \log_2(\frac{\text{Stress}_\text{i} + \text{offset}}{\text{Control}_\text{i} + \text{offset}}),$$

where Stress and Control are parallel lists of cleaved fragment abundance for conserved miRNA/mRNA interactions for the dehydrated and hydrated states, respectively. i is the element within each list and offset is an integer providing a background correction for low cleavage fragment abundance levels as suggested by Mohorianu et. al. (2011) [84]. A log ratio of 1 means a two-fold change in cleavage fragment abundance between states and a log ratio of 2 means a four-fold change etc. A positive value indicates enrichment of cleavage fragments in the dehydrated state, whereas a negative value indicates enrichment of cleavage fragments in the hydrated state.

4.4 Results

4.4.1 Signals of degradation

High-throughput sequencing of the control (CTR) and stress (SWA) degradomes prepared from roots resulted in 7,116,637 and 8,813,611 distinct 20-21nt signatures, respectively. For leaf samples, sequencing resulted in 8,439,454 and 7,446,162 distinct 20-21nt signatures in control (CTA) and stress (SWA) samples, respectively. The degradome sequences for both root libraries were matched to the genome and we found 4,429,787 (62%) and 5,603,150 (64%) of the distinct 20-21nt sequences matched. This roughly agrees with the results reported by another study in *Arabidopsis thaliana* where a similar count of the total number of PARE degradome signatures were found matching the genome [42]. However, in leaf samples, more distinct 20-21nt signatures matched the genome than in root, with 8,439,454 (81%) and 7,446,162 (84%) matches for CTA and SWA, respectively. For all four degradome libraries, the percentage of distinct signatures matching to the genome at a single location ranged from 85% to 86%. This also roughly agrees (~10% less) with results reported by another study in *Arabidopsis thaliana* where there was a similar number of signatures mapping to one location [42].

PAREsnip was used to analyse the degradomes. The degradome analyses predicted over 2,000 miRNA/ mRNA interactions in root samples and

over 3,000 miRNA/mRNA interactions in leaf samples, the difference between interaction hits in root and leaf samples reflect the difference in the total number of genome matches found within the samples. This is a high number of predictions and we surmise that several factors contributed towards this. Firstly, the vast majority of the degradome assisted predictions were category 4 interactions (only a single degradation fragment) and were not conserved between samples. Secondly, the confidence values such as cleavage signal strength, p-value and raw degradome fragment abundance was initially set to capture all possible cleavage events. We therefore filtered the results using more stringent confidence values. Only interactions with a signal strength of category 0, 1 or 2, a p-value of 0.03 and a raw degradome read abundance of 2 or more were retained.

A total of 366 interactions with strong signals of degradation that fulfilled the confidence value criteria were identified. Of those 366 interactions, 106 were found to be conserved between stress and control libraries in root, and 158 interactions were found to be conserved between stress and control libraries in leaf (Figure 4.3).

	A) Stress & Control	B) Only in Control	C) Only in Stress
Root	106	31	15
Leaf	158	39	17

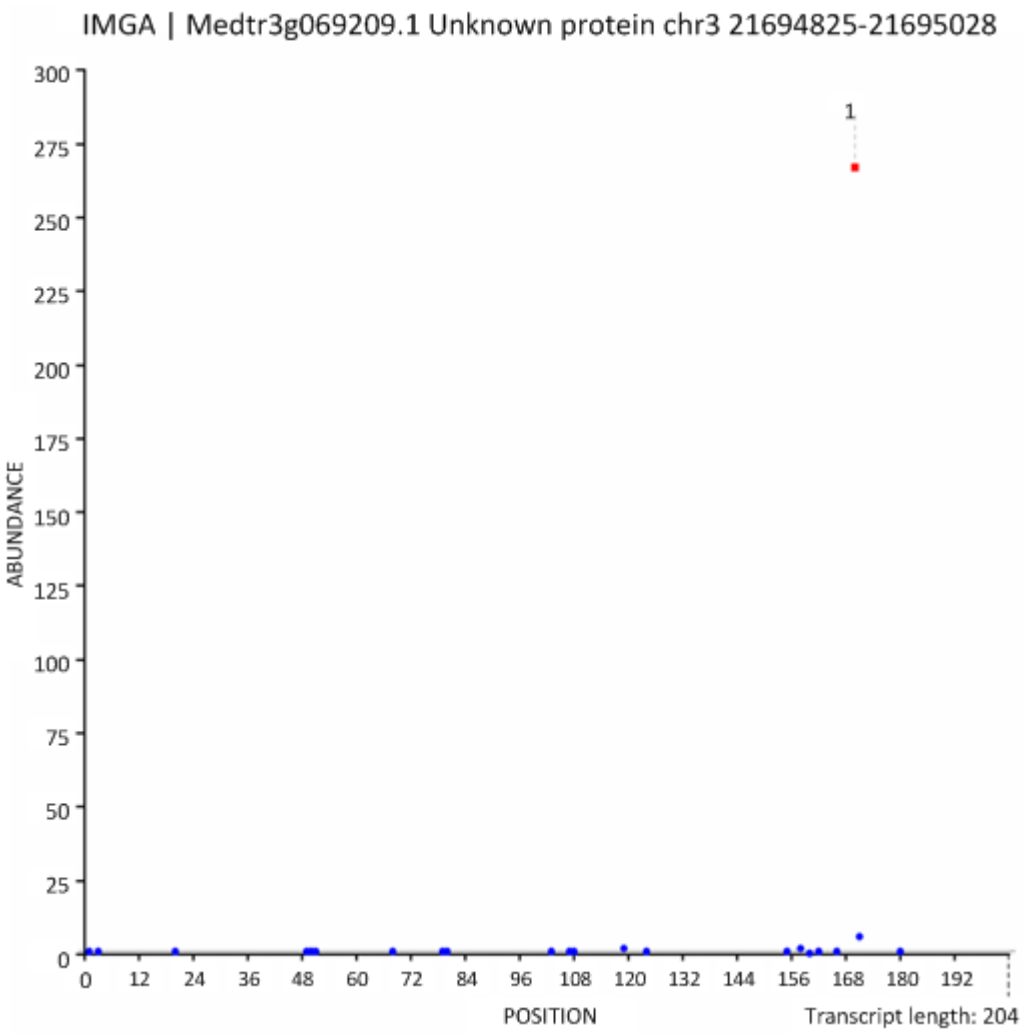
Figure 4.3: Summary of interactions identified with strong signals of degradation.

4.4.2 Differentially cleaved genes

Thirteen of the 366 interactions demonstrated convincing evidence that the mRNAs targeted could potentially be stress responsive (Table 4.1). The thirteen potential stress response interactions met the confidence value criteria in either control or stress analyses and they had an increase or decrease in cleavage products between control and stress states. From the thirteen interactions, two candidates were chosen for further investigation as they were found in both root and leaf degradome samples for both water stress and control.

The first candidate selected was the miRNA miR-172 which was found to target the gene Medtr2g093060.1 (Figure 4.5). The degradation product abundance was >2 fold less in stress than in control samples (Table 4.1). Therefore the gene Medtr2g093060.1 is possibly up regulated in response to water stress. The second candidate selected was miRNA miR1509b, which was found to target the gene Medtr3g069290.1 (Figure 4.4). It had an increase of degradation products in the dehydrated leaf and root samples when compared to hydrated state (Table 4.1). Therefore, the mRNA is likely to be down regulated in response to water stress.

A



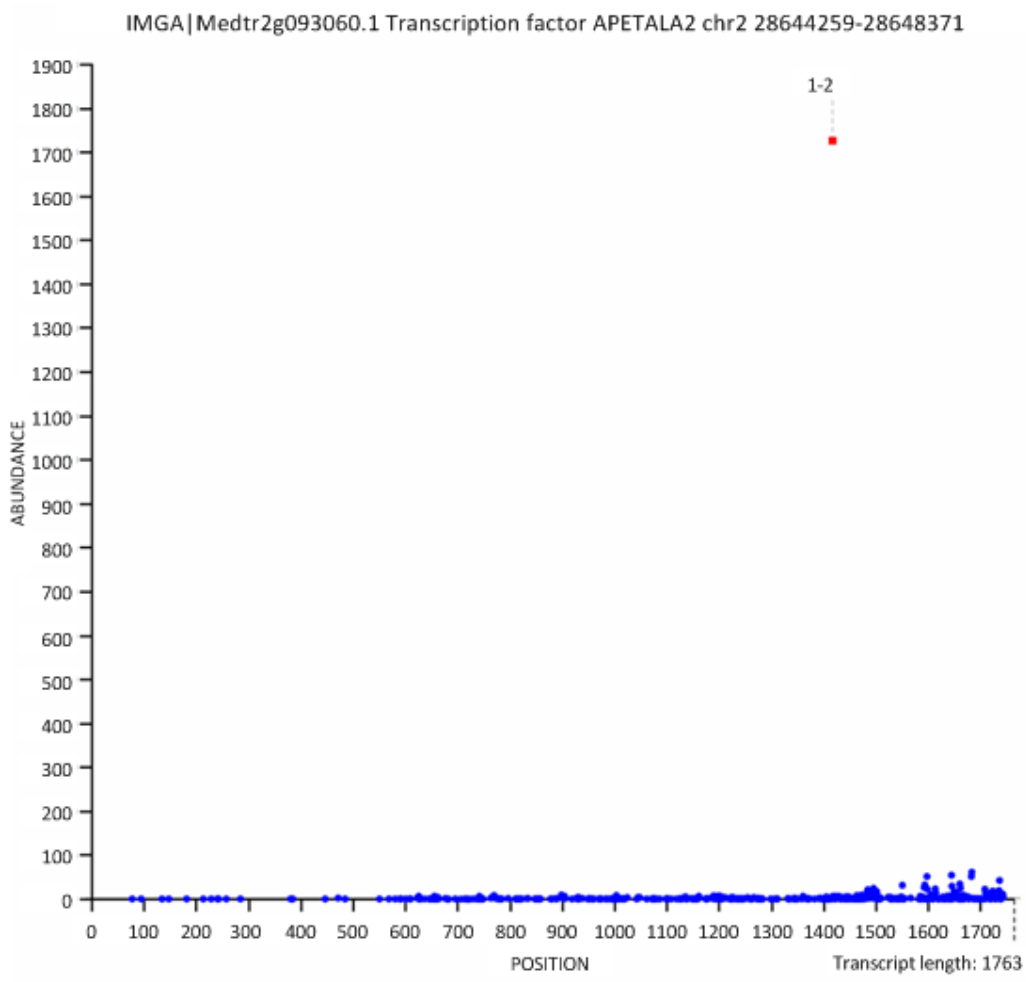
B

```
Category: 0 1 2 3 4
Degradome alignment: ●  Median: ——

0  #1   Position:170 Abundance: 267.00(deg) 1(sRNA)
     5'     TTAATCTAGGAAATTACACTCG       3'  ID:mtr-miR1509b
            |||||| |o||||||||||| |           Score: 2.5
     3' TATCAATTAGTTTCTTTAATGTGAACTATCTT 5'  p-value: 0.0
```

Figure 4.4: A: A t-plot showing the degradation activity for the transcript Medtr3g069290.1. It identifies the cleavage site of mtr-miR1509b (red point). The x axis gives nucleotide positions along the transcript. The y axis gives the abundance of cleavage fragments. B: The interaction data showing miR1509b/Medtr3g069290.1 alignment duplex, raw cleavage product abundance, alignment score and p-value.

A



B

```
Category: 0 1 2 3 4
Degradome alignment:   Median: ——

0  #1   Position:1416 Abundance: 1727.00(deg) 2(sRNA)
5'     AGAATCTTGATGATGCTGCAT        3'  ID:mtr-miR172b
       ||||||o||||||||||||||            Score: 1.5
3' TACCTCTTAGGACTACTACGACGTCTTCTCTT 5'  p-value: 0.0

0  #2   Position:1416 Abundance: 1727.00(deg) 1(sRNA)
5'     AGAATCCTGATGATGCTGCAG        3'  ID:mtr-miR172
       ||||||||||||||||||||||           Score: 0.0
3' TACCTCTTAGGACTACTACGACGTCTTCTCTT 5'  p-value: 0.0
```

Figure 4.5: A: A t-plot showing the degradation activity for the transcript Medtr2G093060.1. It identifies the cleavage site of mtr-miR172 and mtr-172b (red point). The x axis gives nucleotide positions along the transcript. The y axis gives the abundance of cleavage fragments. B: The interaction data showing miR172/Medtr2G093060.1 alignment duplex, raw cleavage product abundance, alignment score and p-value.

Table 4.1: Interaction data table shows conserved interaction cleavage fragment abundance for control and stress states. The acronyms are: CTR=control root, SWR=water stress root, CTA=control leaf, SWA=water stress leaf, FA=fragments abundance, WFA=weighted fragment abundance, NWFA=normalised weighted fragment abundance, FC=fold change, CC=cleavage category.

Gene	miRNA	CC	P-val	Score	CTR FA	CTR WFA	CTR NWFA	SWR FA	SWR WFA	SWR NWFA	FC $offset = 20$	FC $offset = 10$
Medtr2g093060.1	miR172	0	0.0	0	1163	1163	76.857	196	196	6.878	-1.849	-2.363
Medtr2g093060.1	miR172b	0	0.0	1.5	1163	1163	76.857	196	196	6.878	-1.849	-2.363
Medtr7g010860.1	mtr-miR2089	0	0.0	4	1592	1592	105.207	867	867	30.427	-1.312	-1.511
Medtr7g010860.1	mtr-miR2118	0	0.0	2.5	1592	1592	105.207	867	867	30.427	-1.312	-1.511
Medtr4g026490.1	mtr-miR171f	0	0.0	2.5	156	156	10.309	11	11	0.386	-0.572	-0.967
Medtr7g083610.1	mtr-miR393	0	0.0	2	154	154	10.177	47	47	1.649	-0.479	-0.792
Medtr3g069290.1	mtr-miR1509b	0	0.0	2.5	4	4	0.264	205	205	7.149	0.424	0.744
contig 74895 1.1	mtr-miR396a	0	0.0	3	69	69	4.560	5	5	0.175	-0.284	-0.517
contig 74895 1.1	mtr-miR396b	0	0.0	2	69	69	4.560	5	5	0.175	-0.284	-0.517
contig 237721 1.1	mtr-miR396a	0	0.0	3	13	13	0.859	9	9	0.316	-0.038	-0.074
contig 237721 1.1	mtr-miR396b	0	0.0	2	13	13	0.859	9	9	0.316	-0.038	-0.074
					CTA FA	CTA WFA	CTA NWFA	SWA FA	SWA WFA	SWA NWFA		
Medtr3g069290.1	mtr-miR1509b	0	0.0	2.5	15	15	0.536	267	267	11.383	0.612	1.021
contig 49827 2.1	mtr-miR5555	0	0.0	3	111	111	3.967	397	397	16.926	0.624	0.947
contig 74895 1.1	mtr-miR396b	0	0.0	2	3507	3507	125.325	309	309	13.174	-2.131	-2.546
contig 74895 1.1	mtr-miR396a	0	0.0	3	3507	3507	125.328	309	309	13.174	-2.131	-2.546
Medtr2g093060.1	mtr-miR172	0	0.0	0	1727	1727	61.717	502	502	21.402	-0.981	-1.191
Medtr2g093060.1	mtr-miR172b	0	0.0	1.5	1727	1727	61.717	502	502	21.402	-0.981	-1.191
contig 7892 1.1	mtr-miR5232	0	0.0	3	456	456	16.296	123	123	5.244	-0.524	-0.787
contig 237721 1.1	mtr-miR396b	0	0.0	2	263	263	9.399	38	38	1.620	-0.443	-0.739
contig 237721 1.1	mtr-miR396a	0	0.01	3	263	263	9.399	38	38	1.620	-0.443	-0.739
Medtr4g026490.1	mtr-miR171f	0	0.0	2.5	257	257	9.184	60	60	2.558	-0.372	-0.611
Medtr7g083610.1	mtr-miR393b	0	0.0	2	739	739	26.409	377	377	16.073	-0.363	-0.482
Medtr7g010860.1	mtr-miR2118	0	0.0	2.5	6374	6374	227.784	4718	4718	201.145	-0.164	-0.171
Medtr7g010860.1	mtr-miR2089	0	0.01	4	6374	6374	227.784	4718	4718	201.145	-0.164	-0.171

4.4.3 Genes containing AP2 domains subfamily members

Of our two candidate genes selected for deeper investigation, we first consider the candidate Medtr2g093060.1 targeted by miR172. The gene has been annotated (Mt3.5v4 [130]) as belonging to the APETELA 2 family. This is a large family of genes encoding transcription factors (TFs) and have been well described in both Arabidopsis and Rice [103],[104] [109].

The common feature of an APETELA 2 family member is the existence of a DNA binding domain called AP2/ERF, which is roughly 60-70 amino acids in length. In *Arabidopsis thaliana*, genes containing AP2/ERF domains have been classified into five subfamilies based on the number and sequence of AP2/ERF domains on the gene [103],[104]. Two of the subfamily members called Dehydration Responsive Ethylene Binding (DREB) and Ethylene Responsive Factors (ERF) contain a single AP2/ERF domain. TFs belonging to the DREB subfamily in other plant species have been identified as responsive to water stress conditions [62]. In rice, the AP2 family has been classified into four subfamilies [109]. The classifications are based on sequence similarity within the domain.

To consider if our gene of interest has the domain organisation of DREB/ERF i.e. a single AP2 domain, we searched our query gene against Pfam [100] and PROSITE [111],[112] databases. Pfam showed one significant hit and one insignificant hit on the gene for the AP2/ERF domain. PROSITE

showed only one significant hit for the domain. The database search showing our candidate gene as having only a single AP2/ERF domain indicates that the *Medicago truncatula* transcript Medtr2g093060.1 belongs to one of these subfamilies.

Little is known about our second candidate gene Medtr3g069290.1 and indeed it has been annotated by the transcriptome curators as "unknown protein" (Mt3.5v4 [130]). However, the miRNA miR1509b that we found to target Medtr3g069290.1 can provide some insight. The miRNA is conserved in other legume species such as Soybean (*Glycine max*) [66] and has been found to be responsive to abiotic stress. In particular, in Wild-soybean (*Glycine soja*), the miRNA miR1509b is up-regulated when the plant is subjected to aluminum stress conditions [131]. This identifies miR1509b as a stress responsive miRNA and from our data in barrel medic, the increased down regulation of its target in response to water stress could imply a similar up regulation of the miRNA.

4.5 Discussion

In this chapter, we have described the how we used our new software tool PAREsnip to analyse the RNA degradome of an important crop species subjected to water stress. We have identified a total of thirteen potentially stress responsive interactions and considered two of those candidate

interactions in detail.

The genes of interest within our two candidate interactions are likely to be involved in water stress response and we have presented some compelling evidence to support this. The candidate interactions are conserved between four degradome samples and there is a significant change in cleavage products from control to dehydrated states. Furthermore, these findings are supported through similarities with abiotic stress response in other plant species. However, more conclusive evidence for our two candidate stress responsive interactions could be obtained by the validation of the cleavage by 5' RACE and by testing the mRNA expression level through qRT-PCR in both control and dehydrated states. Also, if the sRNA dataset could be obtained, then we could test whether the expression level of miR-172 and miR-1509b go down and up respectively during dehydration and go some way towards explaining the decreased and increased cleavage products.

Chapter 5

Small RNA interaction networks evidenced through the degradome.

5.1 Summary

This chapter describes the software application PAREnets (Parallel Analysis of RNA Ends - networks) that we designed to discover, analyse and visualise sRNA regulatory interaction networks that are evidenced through PARE data. We start by providing a brief background followed by a detailed look at the methods we used to create the tool. We then present some results of an analysis where we used the tool and identified a number of regulatory interaction networks. This work is an adapted version of a manuscript that

is in preparation for publication.

5.2 Background

Recent studies on sRNA interactions have shown that many sRNAs do not operate independently, but instead can form part of larger, more complex, regulatory networks. However, most studies within the literature have been carried out on either a singular instance of a network, or a tiny subset of all sRNAs, such as miRNAs. Others have focused on large-scale analyses but based on computational predictions using sequence complementarity and not empirical evidence. For example, Chen et. al. (2007) [26] describes singular instances of regulatory networks such as a regulatory cascade that is initiated by the miRNA miR173 resulting in the production of ta-siRNAs from two TAS genes. Meng et. al. (2011) [81] considered all known miRNAs and identified several regulatory networks involving co-regulation of transcripts by miRNA and miRNA*s. MacLean et. al. (2010) [76] considered all sRNAs and hypothesized the existence of large-scale sRNA networks and predicted several interaction networks containing specialized hubs of activity. However, the networks they generated were based on computational prediction alone and they conceded that their networks may contain false-positive interactions.

At the time of MacLean et. al.'s study, it was difficult to validate

sRNA/target interactions in a high-throughput way, with such validations being restricted to low-throughput methods such as 5' RACE (Chapter 2.5.1). With the new genome-wide method of sRNA target validation (PARE) and the development of our PAREsnip tool (Chapter 4) that can be used to rapidly analyse the degradome and validate sRNA/target interactions, it is now possible to attempt to discover large scale regulatory interaction networks based on experimental evidence i.e. the degradome, rather than predictions based on sequence complimentarity as well as single instances of networks. We know that sRNAs play important roles in diverse processes such as pathogen response [101], development [88], [99], reproduction [123] and stress response (Chapter 4) and we reason that large scale regulatory networks of sRNA interactions are also involved in such diverse processes. For example, sRNA networks have been found in the vegetative and reproductive stages in the life cycle of rice [82]. Also, tissue specific regulatory networks have also been identified; for example, Ma et. al. (2013) [75] found that some sRNA/target sub-networks were highly accumulated in the roots of rice.

Considering the potential importance and growing number of regulatory networks being found, there is a clear lack of computational tools able to make use of the degradome as a resource and discover sRNA interaction networks. This is not surprising considering that degradome sequencing is a relatively new high-throughput validation method. However, with a grow-

ing interest from the sRNA community in carrying out analyses involving regulatory networks, there is a clear need for such tools. Indeed, to the best of our knowledge, no tool exists that has been designed to discover, analyse and visualize sRNA regulatory networks. A computational method has been suggested [80], and several methods have been described and used to discover sRNA networks,[82],[76],[81], but these methods have relied heavily upon manual filtering, ad-hoc in-house computational pipelines, and are not publicly available.

In this chapter we describe the design and use of a new user-friendly software tool that we have developed and provisionally called PAREnets (Parallel Analysis of RNA Ends and networks). It allows users to build and visualize sRNA interaction networks which are evidenced through genome-wide degradome analysis. By using the degradome to identify sRNAs and their targets on a large scale, we hope to facilitate the discovery of new regulatory interaction networks using a tool that requires very little computational expertise. In the next section we will describe the methods that we have developed within the software.

5.3 Methods

5.3.1 Input

The tool takes as input two data files. The first is a FASTA file containing reference transcripts. Annotation for each transcript need to be on the first line of each transcript record, followed by the transcript sequence itself on subsequent lines of a record. The second input file is the output from a PAREsnip based degradome analysis in comma separated value (csv) format. An overview of the steps involved in processing the input data is shown in Figure 5.1. The diagram in shown in Figure 5.1 provides a design schematic for the dataflow and operations performed upon the data.

5.3.2 Output

The tool has two forms of output. Firstly, a text file that contains the nodes and edges of the network and secondly, images in .png format of networks that have been selected by the user. Though the tool is able to output network data and images to file, the tool's intended primary mode of operation is through the use of a graphical user interface (GUI). The networks drawn by the tool are presented to the user through an interactive GUI (Figure 5.2) that can aid analyses by making the vast amount of data contained within the networks more humanly understandable.

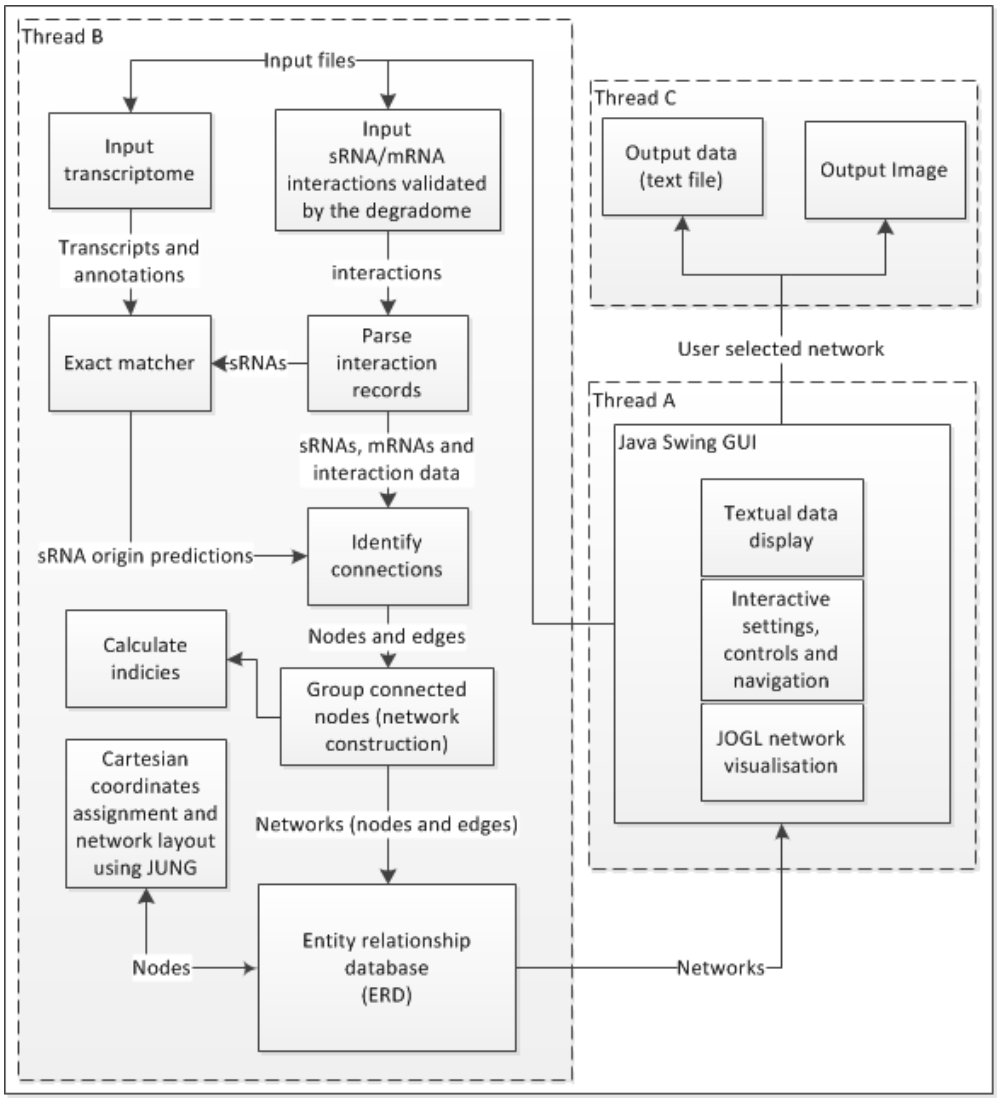


Figure 5.1: Schematic of PAREnets. Boxes represent functions and solid arrowed lines represent data flow. The functions and dataflow operating concurrently using multithreading are enclosed within dotted lines. There are three individual units of execution that are designed to operate concurrently by using the multithreading code that is built into the framework of the Java programming language. The three units of execution are for the graphical user interface (Thread A), data processing (Thread B) and data output (Thread C). Using multithreading helps a computer system to maintain a responsive graphical user interface by taking advantage of central processing units that have multiple cores.

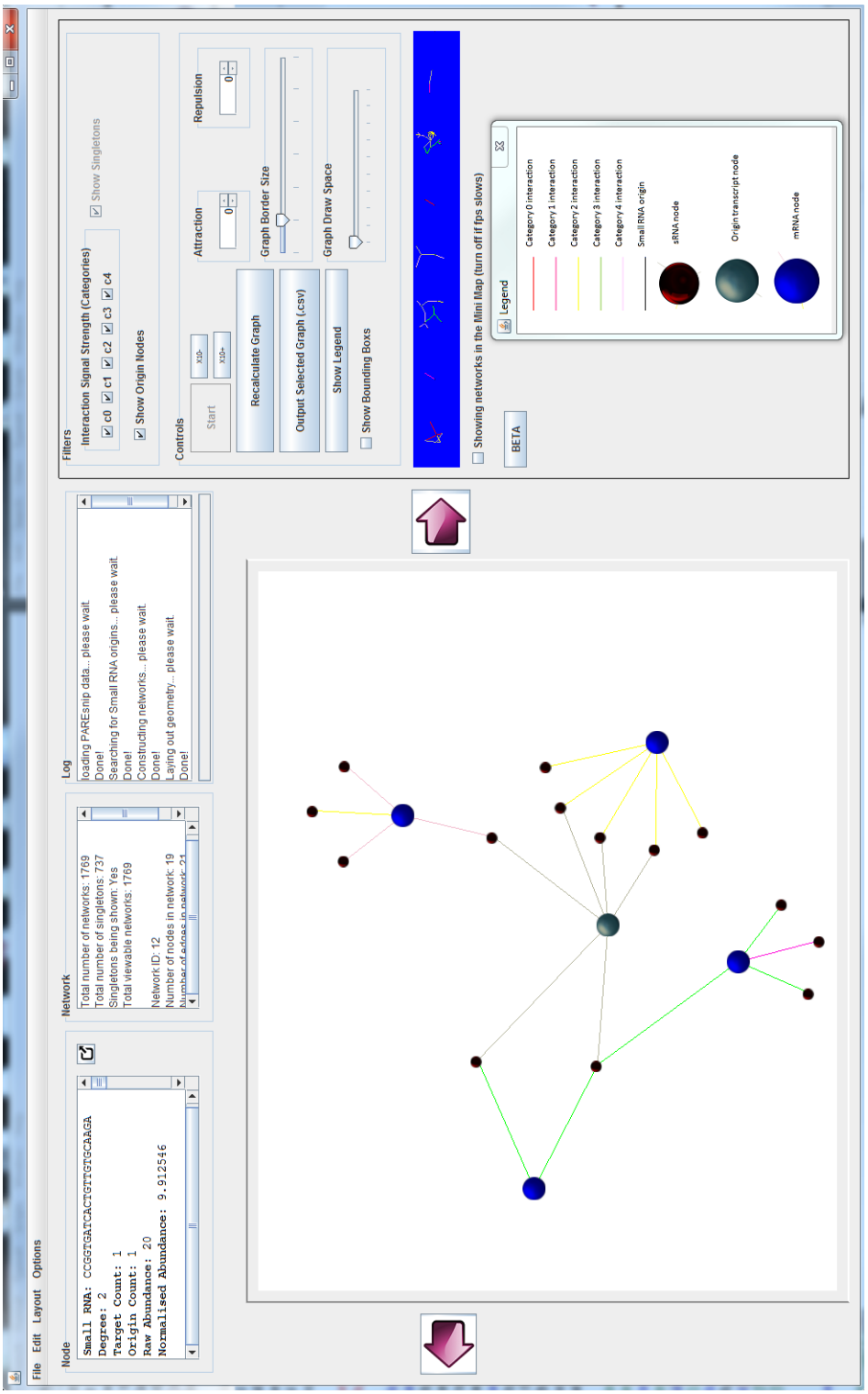


Figure 5.2: Screenshot of the graphical user interface for the PAREnets (beta) tool. The image shows an example of an sRNA network generated from Arabidopsis data that can be viewed and interacted with by the user in 3D space.

5.3.3 Nodes and edges within sRNA regulatory networks

A network is comprised of RNA molecules that are connected through either degradome validated cleavage events or predicted points of origin. Nodes within a network are sRNAs or mRNAs. sRNAs and mRNAs are connected through edges. The type of edge that connects them describes the type of interaction event between them. The six types of interaction event and therefore six edge types connecting sRNAs with mRNAs within a network are described as follows (see Chapter 3.3.3):

- degradome evidenced category 0 cleavage event (strong signal),
- degradome evidenced category 1 cleavage event (strong signal),
- degradome evidenced category 2 cleavage event,
- degradome evidenced category 3 cleavage event,
- degradome evidenced category 4 cleavage event (weak signal) and
- predicted transcript of sRNA origin.

Transcript nodes can be considered as either cleaved by an sRNA (target), giving origin to an sRNA (origin) or both sRNA target and origin.

5.3.4 Interaction filtering

Several filters based on cleavage signal strength and node type can be activated and deactivated by the user. Activation of a signal filter will remove nodes and edges that are connected to a network that correspond to the cleavage signal strength or category described in Chapter 3.3.3. Nodes of predicted sRNA origin and their corresponding edges can be removed by activating a control within the GUI.

5.3.5 Network construction

The massive number of interactions identified by PAREsnip and validated by the degradome are output in large data tables. These large data tables contain the interaction records that can be linked to form a larger network (Figure 5.3 B).

To begin the construction of a network, the interaction records are parsed and individual components of each record are extracted into memory. To predict the potential origin of a sRNA, the tool maps all of the parsed sRNAs against all of the input transcripts using exact matching criteria. If a sRNA exactly maps to a transcript, a new interaction record is created for the prediction.

Individual interactions are then grouped together and placed into bins. Each bin contains all of the interactions for a network. Each interaction

within a bin shares either an sRNA, mRNA or origin transcript with another interaction. To do this, the tool selects an interaction record and subsequently searches all other records for connections. An interaction record is added to a bin if it contains either the same sRNA, mRNA or origin. The search continues until all possible connections are exhausted. Each connection found is placed into a bin. With each additional connection placed into to a bin, a search is carried out for connections to that addition. Figure 5.3 (A) shows a network example where all interactions within one bin are used to generate a network visualisation. A database using an entity relationship model (ER model) [27] stores and dynamically manages the interactions within the bins (Figure 5.1). Depending upon the input data, the tool is able to rapidly generate and manage thousands of bins and therefore thousands of networks.

5.3.6 Graphical display using Open GL

To prepare the network for interactive display, primitive geometry is attached to each of the sRNAs, mRNAs and origins within the ERD. The geometry is the visual representation of the data within each interaction and is displayed as a polygon filled area [51]. The interactive display of the networks has been programmed using the OpenGL application programmers interface (API) [128] and the Java programming language binding

for the OpenGL API hosted by the JOGL (Java Open Graphics Library) open-source project.

To organise the 2D coordinates for the geometry display, we use the Java Universal Network/Graph (JUNG) Framework [77]. Within this framework there are four layout algorithms that the tool uses to position the nodes in each of the networks. The four layout algorithms are:

- Fruchterman-Reingold force-directed algorithm (FRLayout),
- Kamada-Kawai algorithm (KKLayout),
- a self-organizing map layout algorithm (ISOMLayout) and,
- a circle layout algorithm that positions vertices equally spaced on a regular circle.

Once the nodes are displayed using the initial layout coordinates obtained from the JUNG Framework, one can select any node and re-position it within the viewable screen. This is achieved by using colour-picking methods [51]. The updated node position is recorded within the database. If a node is selected, the node's sequence, interaction, abundance and annotation data is displayed. This allows quick and easy information gathering from the network.

Displaying thousands of filled polygons within thousands of networks and still maintain the interactivity of the software tool without lock-ups

presents a challenge. To overcome this we used an axis-aligned bounding box method to clip any network geometry that sits outside the viewable space. This method is traditionally known as clipping [105]. Briefly, nodes and sub-networks that are outside the viewable area of the tools 3D viewport are completely, though temporarily, removed from the tool's rendering processes. This means that only a small subset of the total number of nodes and edges within all of the networks are being rendered at any one time.

5.4 Results

5.5 Network analysis

In Chapter 3.4.4 we used PAREsnip to analyse two data sets (D1 and D2) that were sequenced from *A. thaliana* biological replicates. From sets D1 and D2 we recovered 65,110 and 49,938 sRNA/mRNA interactions. The results were compared and we found a total of 4,466 interactions that were conserved across both replicates. We will call the 4,466 conserved interactions reported by PAREsnip dataset P1.

To construct, visualise and discover regulatory networks within the inflorescence tissue of *A. thaliana* we used our network tool to analyse the dataset P1. The transcripts used for this network analysis were *A. thaliana* representative gene model (TAIR release 10) [118].

Within the data, we identified a total of 697 regulatory networks and 937 singletons. A singleton is a single interaction that was not placed within a network and was therefore discarded. To give a feel for the scale of the networks, we identified the number of nodes within each network and created a size distribution (Table 5.1). Of the 697 networks, 630 (90%) of the networks contained between 3 and 10 nodes and two networks contained more than 90 nodes.

Table 5.1: **Number of nodes within a network**

Node Count	# Networks
3-10	630
11-20	39
21-30	9
31-40	6
41-50	5
51-60	2
61-70	2
71-80	1
81-90	1
>90	2

Only 29 (4%) of the 697 networks contained one or more known miRNAs or miRNA subsequences (isomirs). One of the 29 known miRNA mediated networks is provided as an example in Figure 5.4. Ninety of the sRNA effectors within the networks that we have identified are yet to be described.

Ninety seven percent of the sRNA networks that we have identified using

our tool are to the best of our knowledge not yet described. An example of such a sRNA interaction network is shown in Figure 5.5.

5.5.1 Network validation

The interactions generated by the network tool are validated by the sequencing of degradomes and sRNAomes of two biological replicates. Therefore, the cleavage signal, degradation fragments and sRNAs are found within two plants that are grown under the same conditions. Furthermore, the interactions have a p-value of less than 0.05 and meet the previously described targeting rules. This provides compelling evidence for the interactions discovered by the networks tool.

5.5.2 Availability

The degradome assisted network analysis and discovery tool that we have described in this chapter is a multi-platform, multi-threaded, application written in Java and will be released as part of the UEA sRNA Workbench [89],[117] (http://srna-workbench.cmp.uea.ac.uk).

5.6 Discussion

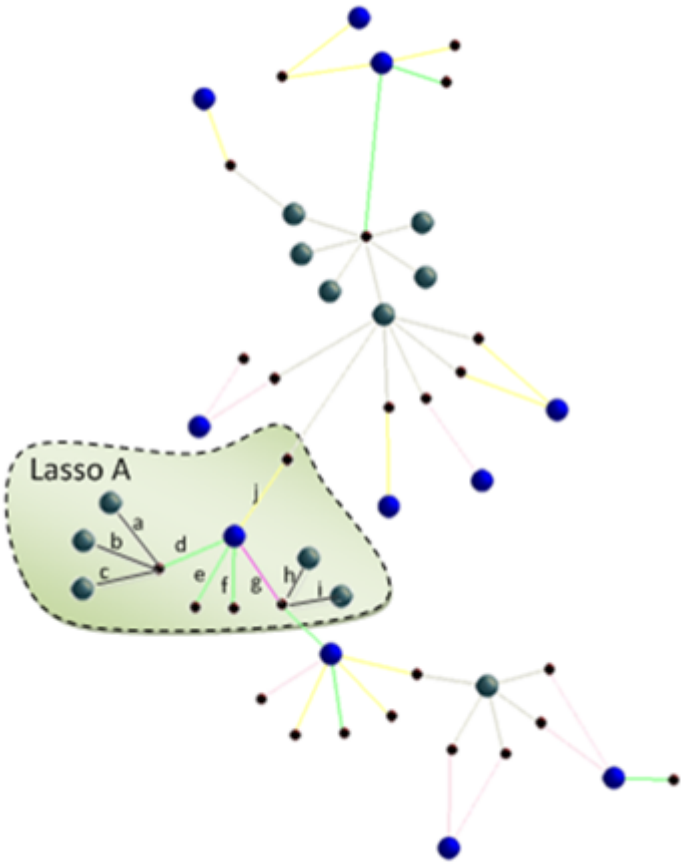
In this chapter we have described a novel, freely available software solution that can be used to discover and analyse sRNA networks. It makes use of

the rich source of evidence that we call the degradome and it can do this because it builds upon the power of PAREsnip. It is the first tool capable of performing a genome-wide network analysis using the degradome within a reasonable time-frame.

Several of the networks that we have identified contain interactions with a low level abundance of degradation products. However, as the degradation signals and sRNAs are conserved across plants, it could indicate an "inactive" network that is laying dormant. Possibly inactive until a response or need for the network to function arises. For example, a network may become active during processes such as growth and development or stress response. The networks with low-level degradation signals could give a promising clue to some interesting functions laying dormant. Also, cross referencing gene annotations for nodes within a network could lead to hypotheses about what functions a particular network may be involved in.

We have identified 697 networks within the previously published PAREsnip results, of which 668 are potentially un-described and could provide a starting point for further analyses. The degradome assisted network discovery and analysis tool is a user-friendly GUI-based, cross-platform (Windows, Linux, MacOS) application that enables biologists to run the application and analyse their data without the need for dedicated bioinformatics support or specialized computer hardware.

A



B

Validated Cleavage Events

Interaction	Target	Category	Score	P-val	Pos	FA	Duplex
d	AT4G01150.1	3	4	0.01	311	8	5' AATAGTTCAACTGTTTGTGATTCC 3' 3' GAAGTTGTTAAGTTGAAAAACGTTA-GGTAAG 5'
e	AT4G01150.1	3	2.5	0.03	258	5	5' GAAAGGGGATCCGGTTAAA 3' 3' GCCACTTTCTCCTAG-CCGGTTTCCTCGCTCC 5'
f	AT4G01150.1	3	4	0.00	276	5	5' AAGCCTCAACGGAACGGAGAAACT 3' 3' TTTGTTCGAAGTTGCCTTGCCACTTTCTCCTA 5'
g	AT4G01150.1	4	3	0.04	491	1	5' GAGGACACAGAGTTGACTGC 3' 3' TCGTCTCCT-TGTCTCAACTACCGTGGTTGTT 5'
j	AT4G01150.1	2	2.5	0.00	364	11	5' GATGTTTGTTGGTATCAATGGT 3' 3' ACTACT-CAAGCAACCATAGTTACCTGCTCCA 5'

Predicted Origin Events

Interaction	Origin	Srna	Start	End	Origin Transcript Annotation
a	AT4G09380.1	AATAGTTCAACTGTTTGTGATTCC	2949	2972	transposable element gene | chr4:5946583-5950874
b	AT4G28970.1	AATAGTTCAACTGTTTGTGATTCC	2808	2831	transposable element gene | chr4:14283564-14287590
c	AT2G23720.1	AATAGTTCAACTGTTTGTGATTCC	2929	2952	transposable element gene | chr2:10088774-10092878
h	AT2G14990.1	GAGGACACAGAGTTGACTGC	706	725	transposable element gene | chr2:6475058-6480183
i	AT4G10830.1	GAGGACACAGAGTTGACTGC	877	896	transposable element gene | chr4:6650724-6654712

Figure 5.3: A: Network example showing predicted and validated interactions from conserved degradation signals described in Chapter 3.4.4. The large nodes are transcripts and small nodes are sRNAs. The coloured edges are validated signals of degradation. A black edge connects a sRNA to its predicted transcript of origin. A large grey node is a transcript of predicted sRNA origin. A large blue node is a cleaved transcript supported by the degradome. B: A table showing an example of the data used to construct a section of the network captured in A:Lasso A.

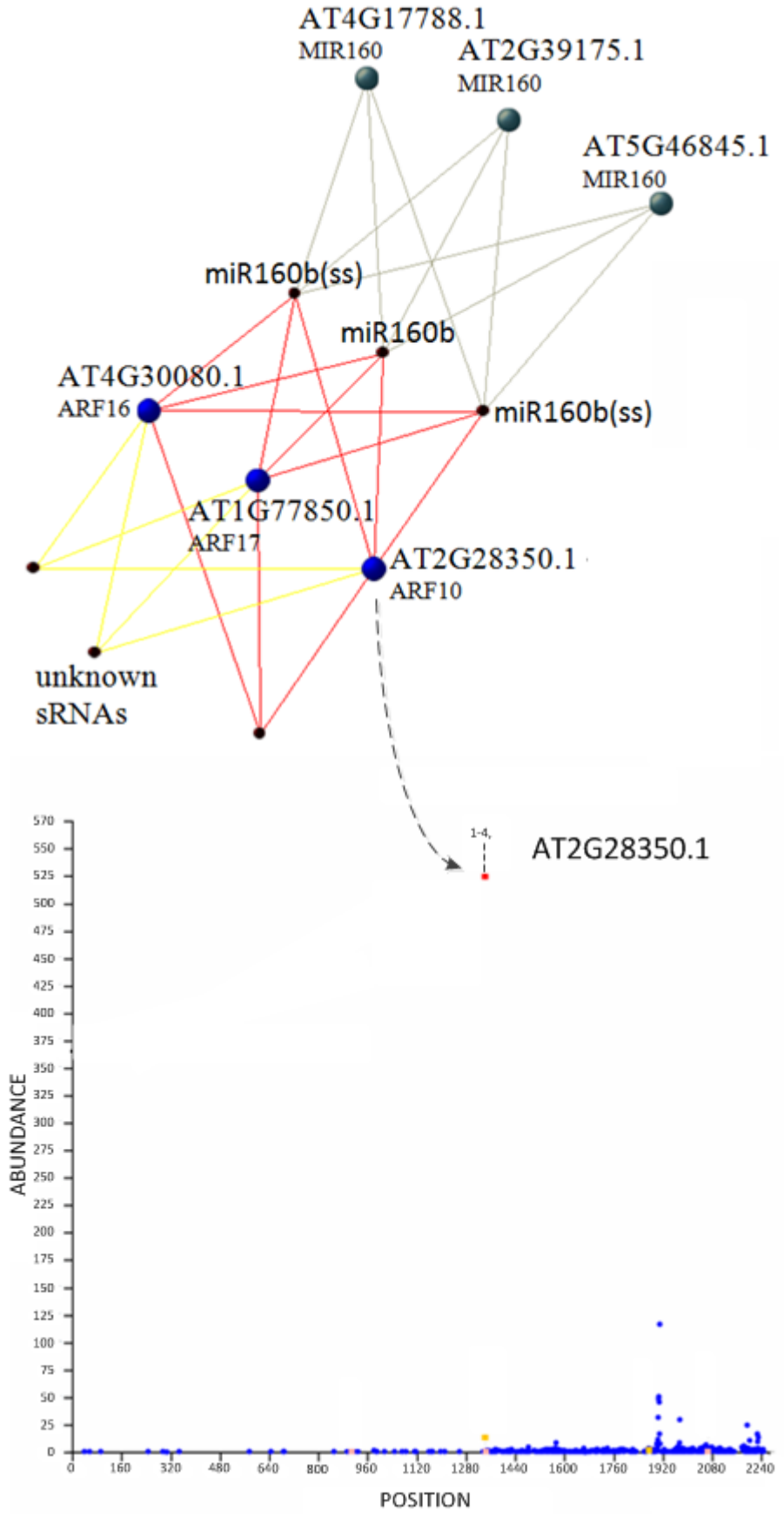


Figure 5.4: A network image and t-plot generated using dataset P1. It shows an example of co-regulation for a known miRNA network (miR160 family).

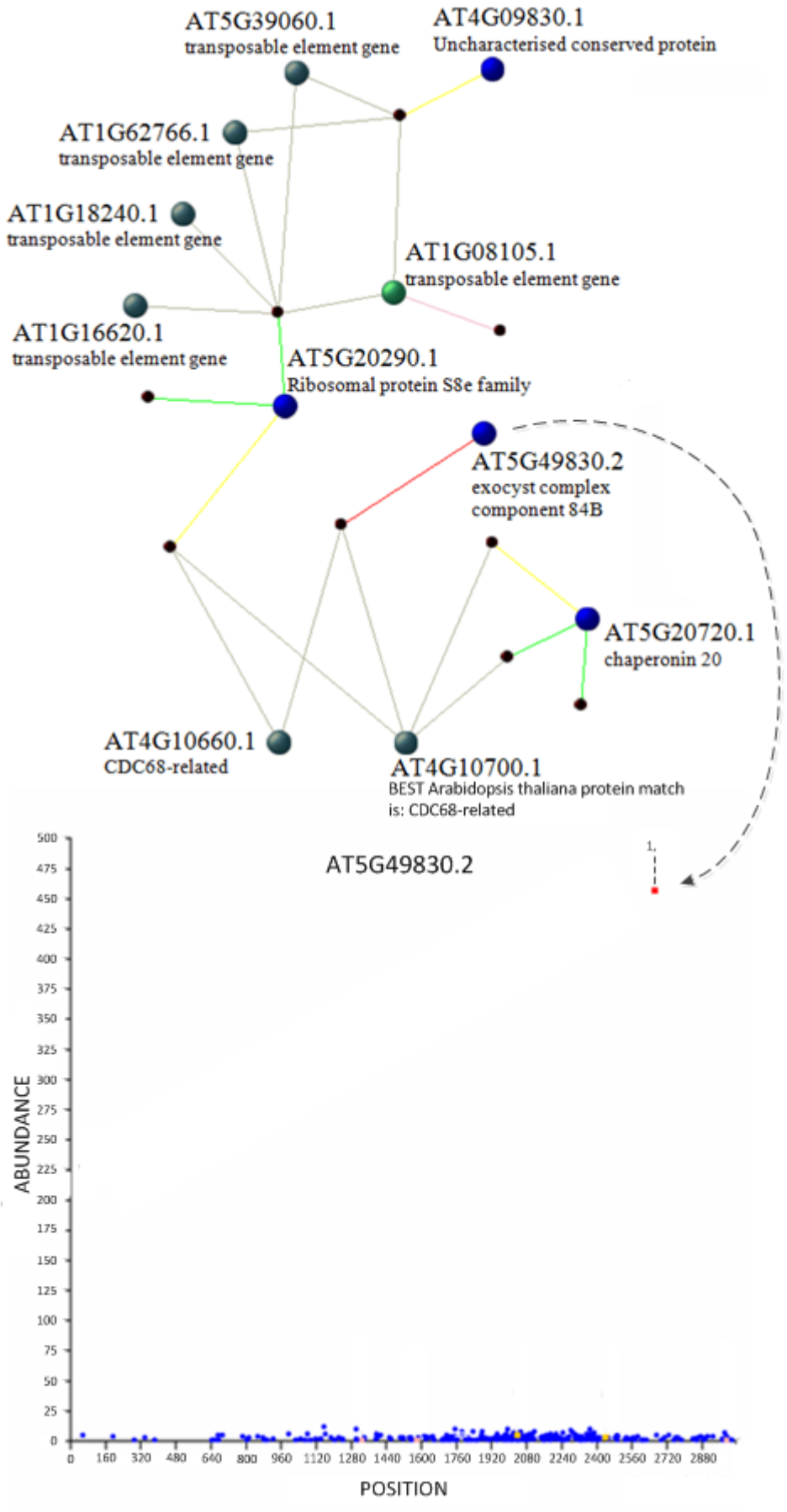


Figure 5.5: A network image and t-plot generated using dataset P1. It shows an example of co-regulation for a novel sRNA regulatory network.

Chapter 6

Conclusion and future work.

The future work that we present in this chapter is an adapted version of a grant proposal written in collaboration with Professor Tamas Dalmay (School of Biological Sciences - UEA) and Professor Vincent Moulton (School of Computing Sciences - UEA).

6.1 Summary

In this chapter we describe a road-map for a future project that builds upon the genome-wide analysis capability of PAREsnip. We propose a new software tool that could be used to detect the functional evidence of mRNA cleavage within the degradome to support the prediction of novel miRNAs. We finish this chapter with conclusions for this thesis and making some closing remarks.

6.2 Background

As we described earlier in Chapter 2.2.1, miRNAs are a class of sRNA that are produced from a stem-loop structure (Figure 2.1) which contains characteristic biogenesis features such as a two nucleotide 3' overhang on the miRNA and miRNA* as well as a limited number of gaps and mismatches within the duplex.

Computational tools are used to analyse sRNAs that are obtained from high-throughput sequencing experiments, and identify miRNAs based on their biogenesis features. However, the computational thresholds for specific biogenesis features such as length of stem, the size and number of bulges and the number of mismatches are based upon a small number of initially discovered miRNAs. More specifically, programs such as miRCat [117] and miR-Deep [129] only consider biogenesis features and therefore have to use strict thresholds to avoid making a large number of false positive predictions and consequently they can miss functional miRNAs.

As we described earlier in this thesis (Chapter 2.5.2), the degradome provides a snap-shot of a plant's mRNA degradation profile, giving quantitative evidence of sRNA mediated cleavage. This is currently the most widely accepted functional data for miRNAs [1],[42]. The PAREsnip tool that we developed (Chapter 4) allowed us for the first time to search the degradome data for cleavage products potentially caused by all the sRNAs

in a given tissue. In Chapter 3.4.4 we identified more than 4000 interactions caused by about 3500 unique sRNAs of which only 149 were known miRNAs. A preliminary analysis of the non-annotated sRNAs potentially causing cleavages revealed that hundreds of them "look like miRNAs (with stem-loop structures and miRNA/miRNA* duplexes with 2 nt overhang) but are just below the threshold for some of the criteria used to annotate miRNAs. This raises the possibility that the currently used thresholds to annotate miRNAs are too strict and that there could be many more plant miRNAs than we currently know. We hypothesise that there is a large population of miRNAs that are not annotated as such based on only the biogenesis features using the current strict criteria. We propose that these currently missed potential miRNAs could be annotated as miRNAs by including the functional readout from degradome analysis and using slightly relaxed thresholds for biogenesis criteria.

6.3 Future work objectives

Currently there are no computational tools available that can analyse both the biogenesis and functional data available to identify novel miRNAs. We suggest that by developing a new tool that combines degradome analysis with relaxed biogenesis thresholds for miRNA prediction, we could identify many new miRNAs and miRNA-like sRNAs. The new relaxed miRNA pre-

diction thresholds could be experimentally validated using mutant plants. Potential objectives for this project would be to

- develop a new program to predict miRNAs based on biogenesis and functional data,
- generate sRNA and degradome libraries from a model organism such as *Arabidopsis thaliana*, and
- predict new miRNAs using the tool and experimentally validate the new miRNA predictions.

6.4 Suggested development of a new program called miR-PARE

We suggest the development of a new program that could be used to predict miRNAs based on biogenesis and functional data. The new miR-PARE program could predict novel miRNAs by combining both biogenesis and functional data using the core algorithms of PAREsnip [38] and the miRNA prediction tool called miRCat [117]. A suggested pipeline is shown in Figure 6.1.

mir-PARE would take eight input files in FASTA format: a transcriptome, a genome and three biological replicates of sRNA and PARE libraries, respectively. The software would first identify sequences in the PARE and

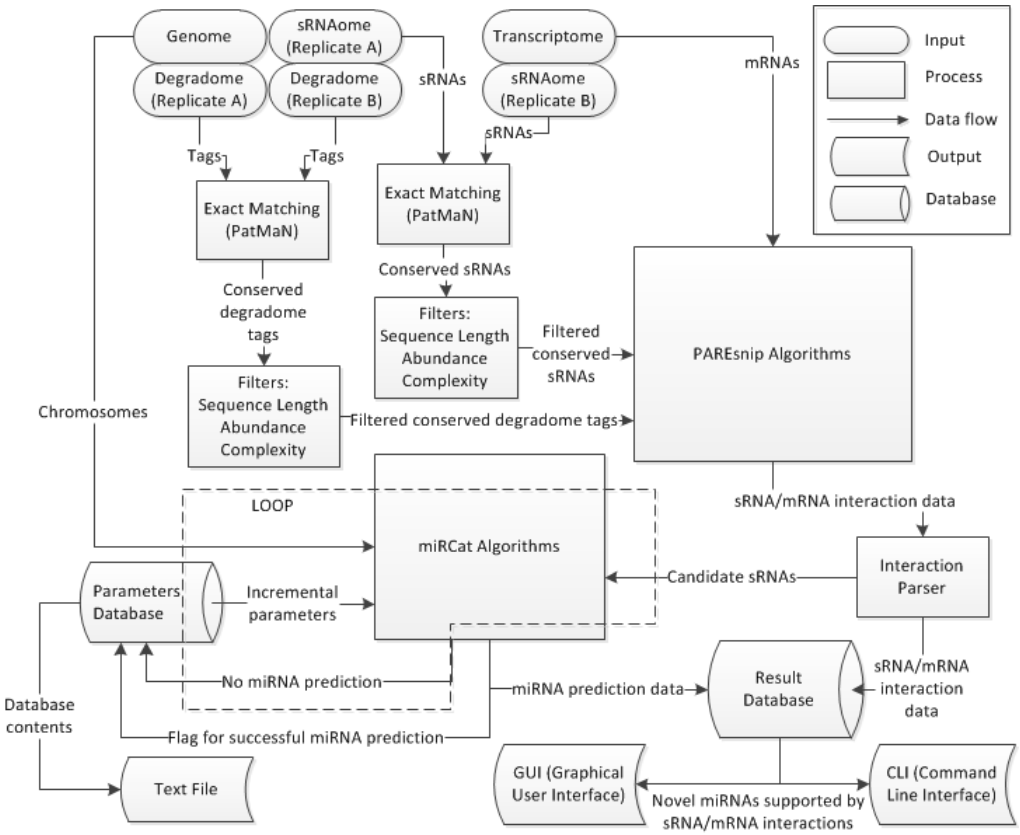


Figure 6.1: An overview of the suggested miR-PARE program

sRNA datasets that are conserved between the biological triplicates using freely available short-read alignment software such as PatMaN [98] (Chapter 2.6.5) and filter out all other sequences. This initial step has been shown as important in reducing noise within the data [38]. Further filters would then be applied to the sRNA sequences to remove low complexity candidates and sequences outside a user adjustable size range.

Next, the software would use the PAREsnip algorithms to identify abundant mRNA cleavage fragments and sRNAs that can anneal to those sites (i.e., predicted sRNAs/mRNA target pairs). The details of the sRNA/mRNA

interactions found by PAREsnip would be stored for output later and the identified sRNAs would be input to the miRCat algorithm. miRCat can be modified to apply a cyclical approach where each sRNA would be processed multiple times using increasingly less-strict parameter values on each pass until either a hairpin structure would be predicted, or the maximum parameter thresholds would be reached (there are 16 adjustable parameters). A database linked to miRCat would be used to store and monitor the parameter combinations that are applied to the miRCat algorithm. When a hairpin structure would be predicted, the parameter combination that was used would be stored in the database and the tool would link this with the miRNA/mRNA interactions as well as the miRNA biogenesis predictions and the results stored within the database.

The primary output of the tool could be the miRNA stem-loop secondary structure and mature-miRNA/mRNA duplex as well as supporting statistics such as p-values and sequence abundance data. These results could be output through a user-friendly Graphical User Interface (GUI). Secondary outputs could be the contents of the parameter database and sRNAs that are identified by PAREsnip but did not result in a stem-loop prediction.

6.5 sRNA and degradome libraries

AGO1 (Figure 2.1) is the key component of the complex that cleaves the mRNA targets so we could generate PARE data for the ago1 mutant plants and use this data to filter out false positive targets. We could do this because potential sRNA targets identified in the wild type PARE library should show a smaller peak in the ago1 library if the target cleavage is indeed sRNA dependent. We could generate libraries from the dcl1 mutant plants because mutation in the DCL1 gene causes a significant down-regulation of miRNAs. Therefore miRNA candidates identified using the sRNA and PARE datasets from the wild type plants would hopefully show a reduced accumulation in the dcl1 plants compared to the wild type. This control would enable us to determine the optimum thresholds for the different parameters miRCat relies upon.

6.6 Prediction of new miRNAs using miR-PARE

miR-PARE could be used to analyse new datasets and the ago1 and dcl1 datasets could be used to optimise the tool's thresholds. Before running miR-PARE, we would compare the PARE data obtained from wild type and ago1 plants and filter out all of the potential sRNA targets (all degradome peaks) that are not reduced in ago1 compared to wild type. The remaining potential targets would be processed by miR-PARE using a cycli-

cal approach i.e., sRNA output by the PAREsnip algorithm would be processed by the miRCat algorithm with increasingly less strict parameters until either the sRNA is predicted as a miRNA or the maximum threshold reached. However, altering the parameters poses a challenge as they are not generally independent. For example, increasing the allowable length of a hairpin structure may not yield a feasible prediction unless the number of mismatches and/or gaps is increased. Therefore this step would need to be carried out many times with different combinations of maximum thresholds for each parameter. The different reiterations of miR-PARE are expected to yield different lists of miRNA predictions. We could exploit the dcl1 sRNA data to decide which sets of thresholds were the best. The expression level of every predicted miRNA (obtained with different thresholds) in the wild-type library would then be compared to the dcl1 mutant library. If a predicted miRNA showed a significant reduction in expression level in the dcl1 library compared to wild type, then the parameters stored within the database which were used to predict the miRNA would contribute towards a threshold window.

6.7 Future work discussion

We have described a framework for a new computational tool that could be used to identify novel miRNAs based upon functional evidence, and

suggested the sequencing of two mutant datasets that could be used to support the configuration and use of the tool. Though this suggestion for future work is based upon the development of a computational tool, we envisage a collaboration with biologists would be required to enable the production of the libraries from the mutant plants. We would hope that this project is realised sometime in the future and our suggested program contributes towards the annotation of many new miRNAs.

6.8 Thesis conclusions

In recent years, the advancements in next generation sequencing technologies has allowed us to gain an ever increasing number of biologically interesting insights into the world of RNA silencing. However, the dramatic increases in the quality, depth and amount of sequencing data being generated by NGS technology challenges the community to develop robust computational tools, that are scalable, and can perform rapid execution of their NGS data analysis.

In this thesis, we have gone some way to meet this challenge by developing the new PAREsnip tool, that for the first time has made it possible to rapidly analyse sRNA/target interactions on a genome-wide scale within a practical time frame using only modest computing resources. It is difficult to predict what other challenges future advances in sequencing technology

may bring, but if advancements in sequencing technology continue on their upwards trend and the size and quality of high-throughput sequencing data continues to increase, then it is likely that there will be an even greater demand for robust analysis tools that can manage the sheer volume of data being produced.

We used the PAREsnip tool to analyze publicly available datasets and we were able to confidently identify over 4000 sRNA/target interactions. We then developed a tool to see if those interactions could form part of a larger regulatory interaction network and identified more than 600 networks within the data. Our findings imply that there are many sRNAs which operate within regulatory networks that are yet to be annotated and the diversity of sRNA biogenisis is not fully described by the existing annotation criteria. In this chapter, we have suggested a novel approach to finding potentially hundreds of novel miRNAs that could go some way to explaining some of the functional sRNAs identified within our data. However, a challenge that still lays ahead is to confidently annotate the remaining functional sRNAs and this holds the potential promise of discovering a new class of sRNA.

NGS technologies have proved to be a valuable resource for providing insights into how plants respond to stressful environmental conditions. A better understanding of how plants respond to conditions such as water stress can help towards mitigating and adapting to the environmental impacts of climate change.

We are getting closer to understanding the complexities of how plant sRNAs interact with mRNAs and how these interactions work within plants. We hope that this will lead to exciting improvements in crop plants that will make them more resilient to environmental stress factors in the future.

Bibliography

[1] Charles Addo-Quaye, Tifani W Eshoo, David P Bartel, and Michael J Axtell. Endogenous siRNA and miRNA targets identified by sequencing of the arabidopsis degradome. *Current Biology: CB*, 18(10):758–762, May 2008. [cited at p. 20, 21, 31, 36, 37, 50, 51, 54, 103]

[2] Charles Addo-Quaye, Webb Miller, and Michael J. Axtell. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics*, 25(1):130–131, January 2009. [cited at p. 31, 36, 47, 50]

[3] Charles Addo-Quaye, Jo Ann Snyder, Yong Bum Park, Yong-Fang Li, Ramanjulu Sunkar, and Michael J Axtell. Sliced microRNA targets and precise loop-first processing of MIR319 hairpins revealed by analysis of the physcomitrella patens degradome. *RNA (New York, N.Y.)*, 15(12):2112–2121, December 2009. [cited at p. 21, 31]

[4] Edwards Allen, Zhixin Xie, Adam M. Gustafson, and James C. Carrington. microRNA-Directed phasing during Trans-Acting siRNA biogenesis in plants. *Cell*, 121(2):207–221, April 2005. [cited at p. 11]

[5] Edwards Allen, Zhixin Xie, Adam M. Gustafson, and James C. Carrington. microRNA-Directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 121(2):207–221, April 2005. [cited at p. 14, 17, 18, 45]

[6] Alexei Aravin, Dimos Gaidatzis, Sebastien Pfeffer, Mariana Lagos-Quintana, Pablo Landgraf, Nicola Iovino, Patricia Morris, Michael J Brownstein, Satomi Kuramochi-Miyagawa, Toru Nakano, Minchen Chien, James J Russo, Jingyue Ju, Robert Sheridan, Chris Sander, Mihaela Zavolan, and Thomas Tuschl. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099):203–207, July 2006. [cited at p. 12]

[7] Milo J. Aukerman and Hajime Sakai. Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-Like target genes. *The Plant Cell Online*, 15(11):2730–2741, November 2003. [cited at p. 8]

[8] Michael J. Axtell. Classification and comparison of small RNAs from plants. *Annual review of plant biology*, 64:137–159, 2013. [cited at p. 12]

[9] Michael J. Axtell, Calvin Jan, Ramya Rajagopalan, and David P. Bartel. A Two-Hit trigger for siRNA biogenesis in plants. *Cell*, 127(3):565–577, November 2006. [cited at p. 11]

[10] Tyler W. H. Backman, Christopher M. Sullivan, Jason S. Cumbie, Zachary A. Miller, Elisabeth J. Chapman, Noah Fahlgren, Scott A. Givan, James C. Carrington, and Kristin D. Kasschau. Update of ASRP: the arabidopsis small RNA project database. *Nucleic acids research*, 36(Database issue):D982–985, January 2008. [cited at p. 59]

[11] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, 39(Database issue):D1005–1010, January 2011. [cited at p. 50, 51, 54, 56, 57, 58]

[12] David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004. [cited at p. 9]

[13] David S. Battisti and Rosamond L. Naylor. Historical warnings of future food insecurity with unprecedented seasonal heat. *Science*, 323(5911):240–244, January 2009. [cited at p. 65]

[14] David Baulcombe. RNA silencing in plants. *Nature*, 431(7006):356–363, 2004. [cited at p. 9]

[15] Vagner A. Benedito, Ivone Torres-Jerez, Jeremy D. Murray, Andry Andriankaja, Stacy Allen, Klementina Kakar, Maren Wandrey, Jerome Verdier, Helene Zuber, Thomas Ott, Sandra Moreau, Andreas Niebel, Tancred Frickey, Georg Weiller, Ji He, Xinbin Dai, Patrick X. Zhao, Yuhong Tang, and Michael K. Udvardi. A gene expression atlas of the model legume medicago truncatula. *The Plant Journal*, 55(3):504513, 2008. [cited at p. 66]

[16] Emily Bernstein, Amy A. Caudy, Scott M. Hammond, and Gregory J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366, January 2001. [cited at p. 7]

[17] Eric Bonnet, Ying He, Kenny Billiau, and Yves Van de Peer. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*, 26(12):1566–1568, June 2010. [cited at p. 14, 17, 18]

[18] Filipe Borges, Patrcia A Pereira, R Keith Slotkin, Robert A Martienssen, and Jrg D Becker. MicroRNA activity in the arabidopsis male germline. *Journal of experimental botany*, 62(5):1611–1620, March 2011. [cited at p. 59]

[19] Omar Borsani, Jianhua Zhu, Paul E. Verslues, Ramanjulu Sunkar, and Jian-Kang Zhu. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in arabidopsis. *Cell*, 123(7):1279–1291, December 2005. [cited at p. 12]

[20] Cameron P. Bracken, Jan M. Szubert, Tim R. Mercer, Marcel E. Dinger, Daniel W. Thomson, John S. Mattick, Michael Z. Michael, and Gregory J. Goodall. Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage. *Nucleic acids research*, 39(13):5658–5668, July 2011. [cited at p. 62]

[21] Peter Brodersen and Olivier Voinnet. The diversity of RNA silencing pathways in plants. *Trends in Genetics*, 22(5):268–280, May 2006. [cited at p. 8]

[22] Michael Burrows and David. J. Wheeler. A block-sorting lossless data compression algorithm. 1994. [cited at p. 26]

[23] Davide Campagna, Alessandro Albiero, Alessandra Bilardi, Elisa Caniato, Claudio Forcato, Svetlin Manavski, Nicola Vitulo, and Giorgio Valle. PASS: a program to align short sequences. *Bioinformatics (Oxford, England)*, 25(7):967–968, April 2009. [cited at p. 27]

[24] Elisabeth J. Chapman and James C. Carrington. Specialization and evolution of endogenous small RNA pathways. *Nature Reviews Genetics*, 8(11):884–896, November 2007. [cited at p. 8]

[25] Padmanabhan Chellappan, Jing Xia, Xuefeng Zhou, Shang Gao, Xiaoming Zhang, Gabriela Coutino, Franck Vazquez, Weixiong Zhang, and Hailing Jin. siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic acids research*, 38(20):6883–6894, November 2010. [cited at p. 59]

[26] Ho-Ming Chen, Yi-Hang Li, and Shu-Hsing Wu. 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, February 2007. [cited at p. 85]

[27] Peter Pin-shan Chen. The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems*, 1:936, 1976. [cited at p. 93]

[28] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger FASTQ file format for sequences with

quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, April 2010. [cited at p. 14]

[29] Xinbin Dai and Patrick Xuechun Zhao. psRNATarget: a plant small RNA target analysis server. *Nucleic acids research*, 39(Web Server issue):W155–159, July 2011. [cited at p. 14, 15]

[30] Andreas Dring, David Weese, Tobias Rausch, and Knut Reinert. SeqAn an efficient, generic c++ library for sequence analysis. *BMC Bioinformatics*, 9(1):11, January 2008. [cited at p. 25]

[31] Sayda M. Elbashir, Javier Martinez, Agnieszka Patkaniowska, Winfried Lendeckel, and Thomas Tuschl. Functional anatomy of siRNAs for mediating efficient RNAi in drosophila melanogaster embryo lysate. *EMBO J*, 20(23):6877–6888, December 2001. [cited at p. 11]

[32] Anne-Katrin Emde, Marcel Grunert, David Weese, Knut Reinert, and Silke R Sperling. MicroRazerS: rapid alignment of small RNA reads. *Bioinformatics (Oxford, England)*, 26(1):123–124, January 2010. [cited at p. 25]

[33] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome research*, 8(3):186–194, March 1998. [cited at p. 14]

[34] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome research*, 8(3):175–185, March 1998. [cited at p. 14]

[35] Noah Fahlgren, Miya D. Howell, Kristin D. Kasschau, Elisabeth J. Chapman, Christopher M. Sullivan, Jason S. Cumbie, Scott A. Givan, Theresa F. Law, Sarah R. Grant, Jeffery L. Dangl, and James C. Carrington. High-throughput sequencing of arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PloS one*, 2(2):e219, 2007. [cited at p. 14, 18, 51, 59]

[36] Noah Fahlgren, Christopher M Sullivan, Kristin D Kasschau, Elisabeth J Chapman, Jason S Cumbie, Taiowa A Montgomery, Sunny D Gilbert, Mark Dasenko, Tyler W H Backman, Scott A Givan, and James C Carrington. Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA (New York, N.Y.)*, 15(5):992–1002, May 2009. [cited at p. 31, 59]

[37] Andrew Fire, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature*, 391(6669):806–811, February 1998. [cited at p. 11]

[38] Leighton Folkes, Simon Moxon, Hugh C. Woolfenden, Matthew B. Stocks, Gyorgy Szittya, Tamas Dalmay, and Vincent Moulton. PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic acids research*, 40(13):e103, July 2012. [cited at p. 18, 31, 105, 106]

[39] Jose Manuel Franco-Zorrilla, Adrin Valli, Marco Todesco, Isabel Mateos, Mara Isabel Puga, Ignacio Rubio-Somoza, Antonio Leyva, Detlef Weigel, Juan Antonio Garca, and Javier Paz-Ares. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature genetics*, 39(8):1033–1037, August 2007. [cited at p. 17]

[40] Peng Gao, Xi Bai, Liang Yang, Dekang Lv, Xin Pan, Yong Li, Hua Cai, Wei Ji, Qin Chen, and Yanming Zhu. osa-MIR393: a salinity- and alkaline stress-related microRNA gene. *Molecular Biology Reports*, 38(1):237–242, January 2011. [cited at p. 65]

[41] Marcelo A. German, Shujun Luo, Gary Schroth, Blake C. Meyers, and Pamela J. Green. Construction of parallel analysis of RNA ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nature Protocols*, 4(3):356–362, 2009. [cited at p. 20, 31, 67]

[42] Marcelo A. German, Manoj Pillay, Dong-Hoon Jeong, Amit Hetawal, Shujun Luo, Prakash Janardhanan, Vimal Kannan, Linda A. Rymarquis, Kan Nobuta, Rana German, Emanuele De Paoli, Cheng Lu, Gary Schroth, Blake C Meyers, and Pamela J Green. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nature Biotechnology*, 26(8):941–946, August 2008. [cited at p. 20, 36, 51, 58, 59, 75, 103]

[43] Angelique Girard, Ravi Sachidanandam, Gregory J. Hannon, and Michelle A. Carmell. A germline-specific class of small RNAs binds mammalian piwi proteins. *Nature*, 442(7099):199–202, July 2006. [cited at p. 12]

[44] Michael T. Goodrich and Roberto Tamassia. *Data Structures and Algorithms in Java*. John Wiley and Sons, USA, 4th edition, 2005. [cited at p. 40]

[45] Jemma Gornall, Richard Betts, Eleanor Burke, Robin Clark, Joanne Camp, Kate Willett, and Andrew Wiltshire. Implications of climate change for agricultural productivity in the early twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554):2973–2989, September 2010. [cited at p. 65]

[46] Peter H. Graham and Carroll P. Vance. Legumes: Importance and constraints to greater use. *Plant Physiology*, 131(3):872–877, March 2003. [cited at p. 66]

[47] Robert Grant-Downton, Gael Le Trionnaire, Ralf Schmid, Josefina Rodriguez-Enriquez, Said Hafidh, Saher Mehdi, David Twell, and Hugh Dickinson. MicroRNA and tasiRNA diversity in mature pollen of arabidopsis thaliana. *BMC genomics*, 10:643, 2009. [cited at p. 51, 59]

[48] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R. Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*, 33(Database issue):D121–124, January 2005. [cited at p. 36]

[49] Shane T. Grivna, Brook Pyhtila, and Haifan Lin. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proceedings of the National Academy of Sciences*

of the United States of America, 103(36):13415–13420, September 2006. [cited at p. 12]

[50] Andrew J. Hamilton and David C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, N.Y.)*, 286(5441):950–952, October 1999. [cited at p. 11]

[51] Donald Hearn and Pauline M. Baker. *Computer graphics with OpenGL*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004. [cited at p. 93, 94]

[52] Robert A. Holt and Steven J. M. Jones. The new paradigm of flow cell sequencing. *Genome Research*, 18(6):839–846, June 2008. [cited at p. 13]

[53] David Stephen Horner, Giulio Pavesi, Tiziana Castrignan, Paolo D'Onorio De Meo, Sabino Liuni, Michael Sammeth, Ernesto Picardi, and Graziano Pesole. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in bioinformatics*, 11(2):181–197, March 2010. [cited at p. 31]

[54] Li-Ching Hsieh, Shu-I Lin, Arthur Chun-Chieh Shih, June-Wei Chen, Wei-Yi Lin, Ching-Ying Tseng, Wen-Hsiung Li, and Tzyy-Jen Chiou. Uncovering small RNA-mediated responses to phosphate deficiency in arabidopsis by deep sequencing. *Plant physiology*, 151(4):2120–2132, December 2009. [cited at p. 51, 59]

[55] Minghui Jiang, James Anderson, Joel Gillespie, and Martin Mayne. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, 9:192, 2008. [cited at p. 48, 52]

[56] Martin Jinek and Jennifer A. Doudna. A three-dimensional view of the molecular machinery of RNA interference. *Nature*, 457(7228):405–412, January 2009. [cited at p. 8]

[57] Fedor V. Karginov, Sihem Cheloufi, Mark M. W. Chong, Alexander Stark, Andrew D. Smith, and Gregory J. Hannon. Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, drosha, and additional nucleases. *Molecular cell*, 38(6):781–788, June 2010. [cited at p. 62]

[58] Ilene Karsch-Mizrachi, Yasukazu Nakamura, and Guy Cochrane. The international nucleotide sequence database collaboration. *Nucleic acids research*, 40(Database issue):D33–37, January 2012. [cited at p. 36]

[59] V. Narry Kim, Jinju Han, and Mikiko C. Siomi. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, 10(2):126–139, February 2009. [cited at p. 12]

[60] Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(Database issue):D152–157, January 2011. [cited at p. 51, 59, 66, 68]

[61] Yukio Kurihara and Yuichiro Watanabe. Arabidopsis micro-RNA biogenesis through dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12753 – 12758, 2004. [cited at p. 9]

[62] Charu Lata and Manoj Prasad. Role of DREBs in regulation of abiotic stress responses in plants. *Journal of Experimental Botany*, 62(14):4731–4748, October 2011. hello. [cited at p. 81]

[63] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993. [cited at p. 9]

[64] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–4060, October 2004. [cited at p. 9]

[65] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(suppl 1):D19–D21, January 2011. [cited at p. 14]

[66] Christine Lelandais-Brire, Loreto Naya, Erika Sallet, Fanny Calenge, Florian Frugier, Caroline Hartmann, Jerome Gouzy, and Martin Crespi. Genome-wide medicago truncatula small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. *The Plant cell*, 21(9):2780–2796, September 2009. [cited at p. 82]

[67] Bosheng Li, Yurong Qin, Hui Duan, Weilun Yin, and Xinli Xia. Genome-wide characterization of new and drought stress responsive microRNAs in populus euphratica. *Journal of Experimental Botany*, 62(11):3765–3779, July 2011. [cited at p. 65]

[68] Feng Li, Ryan Orban, and Barbara Baker. SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. *The Plant journal: for cell and molecular biology*, 70(5):891–901, June 2012. [cited at p. 32]

[69] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5):713–714, March 2008. [cited at p. 26]

[70] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, August 2009. [cited at p. 26]

[71] Yong-Fang Li, Yun Zheng, Charles Addo-Quaye, Li Zhang, Ajay Saini, Guru Jagadeeswaran, Michael J Axtell, Weixiong Zhang, and Ramanjulu Sunkar. Transcriptome-wide identification of microRNA targets in rice. *The Plant journal: for cell and molecular biology*, 62(5):742–759, June 2010. [cited at p. 31]

[72] Zachary Lippman and Rob Martienssen. The role of RNA interference in heterochromatic silencing. *Nature*, 431(7006):364–370, September 2004. [cited at p. 8]

[73] Cesar Llave, Zhixin Xie, Kristin D. Kasschau, and James C. Carrington. Cleavage of scarecrow-like mRNA targets directed by a class of arabidopsis miRNA. *Science*, 297(5589):2053–2056, September 2002. [cited at p. 8, 9, 20, 41]

[74] Wenjing Lu, Jincai Li, Fangpeng Liu, Juntao Gu, Chengjin Guo, Liu Xu, Huiyan Zhang, and Kai Xiao. Expression pattern of wheat miRNAs under salinity stress and prediction of salt-inducible miRNAs targets. *Frontiers of Agriculture in China*, 5(4):413–422, November 2011. [cited at p. 65]

[75] Xiaoxia Ma, Chaogang Shao, Huizhong Wang, Yongfeng Jin, and Yijun Meng. Construction of small RNA-mediated gene regulatory networks in the roots of rice (oryza sativa). *BMC genomics*, 14(1):510, 2013. [cited at p. 86]

[76] Daniel MacLean, Nataliya Elina, Ericka R. Havecker, Susanne B. Heimstaedt, David J. Studholme, and David C. Baulcombe. Evidence for large complex networks of plant short silencing RNAs. *PLoS ONE*, 5(3), March 2010. [cited at p. 85, 87]

[77] Joshua Madadhain, Danyel Fisher, Padhraic Smyth, Scott White, and Yan-Biao Boey. Analysis and visualization of network data using jung. *Journal of Statistical Software*, 10:1–35, 2005. [cited at p. 94]

[78] Elaine R. Mardis. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9:387–402, 2008. [cited at p. 12]

[79] Ulrike Mckstein, Hakim Tafer, Jrg Hackermller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNARNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006. [cited at p. 16]

[80] Yijun Meng, Chaogang Shao, and Ming Chen. Toward microRNA-mediated gene regulatory networks in plants. *Briefings in bioinformatics*, 12(6):645–659, November 2011. [cited at p. 87]

[81] Yijun Meng, Chaogang Shao, Lingfeng Gou, Yongfeng Jin, and Ming Chen. Construction of MicroRNA- and MicroRNA*-mediated regulatory networks in plants. *RNA Biology*, 8(6):1124–1148, November 2011. [cited at p. 85, 87]

[82] Yijun Meng, Chaogang Shao, Huizhong Wang, Xiaoxia Ma, and Ming Chen. Construction of gene regulatory networks mediated by vegetative and reproductive stage-specific small RNAs in rice (*Oryza sativa*). *New Phytologist*, 197(2):441–453, January 2013. [cited at p. 86, 87]

[83] Michael L. Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January 2010. [cited at p. 31]

[84] Irina Mohorianu, Frank Schwach, Runchun Jing, Sara Lopez-Gomollon, Simon Moxon, Gyorgy Szittya, Karim Sorefan, Vincent Moulton, and Tamas Dalmay. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant Journal*, 67(2):232246, 2011. [cited at p. 74]

[85] Dov Moldovan, Andrew Spriggs, Elizabeth S. Dennis, and Iain W. Wilson. The hunt for hypoxia responsive natural antisense short interfering RNAs. *Plant Signaling & Behavior*, 5(3):247–251, March 2010. [cited at p. 12]

[86] Dov Moldovan, Andrew Spriggs, Jun Yang, Barry J. Pogson, Elizabeth S. Dennis, and Iain W. Wilson. Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in arabidopsis. *Journal of experimental botany*, 61(1):165–177, 2010. [cited at p. 51, 59]

[87] Taiowa A. Montgomery, Seong Jeon Yoo, Noah Fahlgren, Sunny D. Gilbert, Miya D. Howell, Christopher M. Sullivan, Amanda Alexander, Goretti Nguyen, Edwards Allen, Ji Hoon Ahn, and James C. Carrington. AGO1-miR173 complex initiates phased siRNA formation in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20055–20062, December 2008. [cited at p. 56, 57]

[88] Simon Moxon, Runchun Jing, Gyorgy Szittya, Frank Schwach, Rachel L. Rusholme Pilcher, Vincent Moulton, and Tamas Dalmay. Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Research*, 18(10):1602–1609, October 2008. [cited at p. 19, 86]

[89] Simon Moxon, Frank Schwach, Tamas Dalmay, Dan Maclean, David J. Studholme, and Vincent Moulton. A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics (Oxford, England)*, 24(19):2252–2253, October 2008. [cited at p. 18, 34, 49, 97]

[90] Mayumi Nakano, Kan Nobuta, Kalyan Vemaraju, Shivakundan Singh Tej, Jeremy W. Skogen, and Blake C. Meyers. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic acids research*, 34(Database issue):D731–735, January 2006. [cited at p. 51, 59]

[91] Zhiyong Ni, Zheng Hu, Qiyan Jiang, and Hui Zhang. Overexpression of gma-MIR394a confers tolerance to drought in transgenic arabidopsis thaliana. *Biochemical and biophysical research communications*, 427(2):330–335, October 2012. [cited at p. 28, 66]

[92] Helio Pais, Simon Moxon, Tamas Dalmay, and Vincent Moulton. Small RNA discovery and characterisation in eukaryotes using high-throughput approaches. *Advances in experimental medicine and biology*, 722:239–254, 2011. [cited at p. 31]

[93] Vitantonio Pantaleo, Gyorgy Szittya, Simon Moxon, Laura Miozzi, Vincent Moulton, Tamas Dalmay, and Jozsef Burgyan. 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, June 2010. [cited at p. 21, 31, 59]

[94] Mee Yeon Park, Gang Wu, Alfredo Gonzalez-Sulser, Herve Vaucheret, and R. Scott Poethig. 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, March 2005. [cited at p. 9]

[95] William Pearson. Finding protein and nucleotide similarities with FASTA. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002. [cited at p. 17]

[96] William R. Pearson and David J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448, April 1988. [cited at p. 18]

[97] Tara L. Phelps-Durr. MicroRNAs in Arabidopsis. *Nature Education*, 3(9):51, 2010. [cited at p. 10, 144]

[98] Kay Prfer, Udo Stenzel, Michael Dannemann, Richard E Green, Michael Lachmann, and Janet Kelso. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics (Oxford, England)*, 24(13):1530–1531, July 2008. [cited at p. 27, 36, 106]

[99] Amada Pulido and Patrick Laufs. Co-ordination of developmental processes by small RNAs during leaf development. *Journal of Experimental Botany*, 61(5):1277 –1291, March 2010. [cited at p. 86]

[100] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, November 2011. [cited at p. 81]

[101] Andres Quintero, Alvaro L. Perez-Quintero, and Camilo Lpez. Identification of ta-siRNAs and cis-nat-siRNAs in cassava and their roles in response to cassava bacterial blight. *Genomics, Proteomics & Bioinformatics*, 11(3):172–181, June 2013. [cited at p. 86]

[102] Marc Rehmsmeier, Peter Steffen, Matthias Hchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, October 2004. [cited at p. 17]

[103] Yoh Sakuma, Qiang Liu, Joseph G Dubouzet, Hiroshi Abe, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. DNA-binding specificity of the ERF/AP2 domain of arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochemical and biophysical research communications*, 290(3):998–1009, January 2002. [cited at p. 81]

[104] Abdelaty Saleh and Montserrat Pages. Plant AP2/ERF transcription factors. *Genetika*, 35(1):37–50, 2003. [cited at p. 81]

[105] Daniel Sanchez-crespo and ProQuest. *Core Techniques and Algorithms in Game Programming*. New Riders Publishing ; Pearson Education [Distributor], Berkeley; Old Tappan, 2003. [cited at p. 95]

[106] Rebecca Schwab, Javier F. Palatnik, Markus Riester, Carla Schommer, Markus Schmid, and Detlef Weigel. Specific effects of microRNAs on the plant transcriptome. *Developmental cell*, 8(4):517–527, April 2005. [cited at p. 41, 43, 45]

[107] Elizabeth Scotto-Lavino, Guangwei Du, and Michael A Frohman. Amplification of 5' end cDNA with 'new RACE'. *Nature protocols*, 1(6):3056–3061, 2006. [cited at p. 19]

[108] Dimitri Semizarov, Leigh Frost, Aparna Sarthy, Paul Kroeger, Donald N. Halbert, and Stephen W. Fesik. Specificity of short interfering RNA determined through gene expression signatures. *Proceedings of the National Academy of Sciences*, 100(11):6347 –6352, May 2003. [cited at p. 11]

[109] Akhter Most Sharoni, Mohammed Nuruzzaman, Kouji Satoh, Takumi Shimizu, Hiroaki Kondoh, Takahide Sasaya, Il-Ryong Choi, Toshihiro Omura, and Shoshi Kikuchi. Gene structures, classification and expression models of the AP2/EREBP transcription factor family in rice. *Plant and Cell Physiology*, 52(2):344–360, February 2011. [cited at p. 81]

[110] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008. [cited at p. 24, 26, 31]

[111] Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, 3(3):265–274, September 2002. [cited at p. 81]

[112] Christian J. A. Sigrist, Edouard de Castro, Lorenzo Cerutti, Beatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenar-

ios. New and continuing developments at PROSITE. *Nucleic acids research*, 41(Database issue):D344–347, January 2013. [cited at p. 81]

[113] Haruhiko Siomi and Mikiko C. Siomi. On the road to reading the RNA-interference code. *Nature*, 457(7228):396–404, January 2009. [cited at p. 8]

[114] Andrew D. Smith, Wen-Yu Chung, Emily Hodges, Jude Kendall, Greg Hannon, James Hicks, Zhenyu Xuan, and Michael Q. Zhang. Updates to the RMAP short-read mapping software. *Bioinformatics*, 25(21):2841–2842, November 2009. [cited at p. 24]

[115] Andrew D. Smith, Zhenyu Xuan, and Michael Q. Zhang. Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC bioinformatics*, 9:128, 2008. [cited at p. 24]

[116] Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981. [cited at p. 15]

[117] Matthew B. Stocks, Simon Moxon, Daniel Mapleson, Hugh C. Woolfenden, Irina Mohorianu, Leighton Folkes, Frank Schwach, Tamas Dalmay, and Vincent Moulton. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15):2059–2061, August 2012. [cited at p. 34, 49, 97, 103, 105]

[118] David Swarbreck, Christopher Wilks, Philippe Lamesch, Tanya Z. Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, Tom Meyer, Robert Muller, Larry Ploetz, Amie Radenbaugh, Shanker Singh, Vanessa Swing, Christophe Tissier, Peifen Zhang, and Eva Huala. The arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, 36(suppl 1):D1009–D1014, January 2008. [cited at p. 23, 49, 51, 56, 58, 95]

[119] Gyorgy Szittya, Simon Moxon, Dulce M. Santos, Runchun Jing, Manuel P. S. Fevereiro, Vincent Moulton, and Tamas Dalmay. High-throughput sequencing of medicago truncatula short RNAs identifies eight new miRNA families. *BMC Genomics*, 9(1):593, December 2008. [cited at p. 66]

[120] Guiliang Tang, Brenda J. Reinhart, David P. Bartel, and Phillip D. Zamore. A biochemical framework for RNA silencing in plants. *Genes & Development*, 17(1):49 –63, January 2003. [cited at p. 9]

[121] Zhonghui Tang, Liping Zhang, Chenguang Xu, Shaohua Yuan, Fengting Zhang, Yonglian Zheng, and Changping Zhao. Uncovering small RNA-Mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing. *Plant Physiology*, 159(2):721–738, June 2012. [cited at p. 65]

[122] Travis Thomson and Haifan Lin. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annual review of cell and developmental biology*, 25:355–376, 2009. [cited at p. 12]

[123] Frederic Van Ex, Yannick Jacob, and Robert A. Martienssen. Multiple roles for sRNA during plant reproduction. *Current Opinion in Plant Biology*, 14(5):588–593, October 2011. [cited at p. 86]

[124] Herve Vaucheret, Franck Vazquez, Patrice Crete, and David P. Bartel. 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, May 2004. [cited at p. 9, 11]

[125] Tianzuo Wang, Lei Chen, Mingui Zhao, Qiuying Tian, and Wen-Hao Zhang. Identification of drought-responsive microRNAs in medicago truncatula by genome-wide high-throughput sequencing. *BMC Genomics*, 12(1):367, July 2011. [cited at p. 66]

[126] Toshiaki Watanabe, Atsushi Takeda, Tomoyuki Tsukiyama, Kazuyuki Mise, Tetsuro Okuno, Hiroyuki Sasaki, Naojiro Minami, and Hiroshi Imai. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes & development*, 20(13):1732–1743, July 2006. [cited at p. 12]

[127] James. D. Watson and Francis H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953. [cited at p. 8]

[128] Mason Woo, Jackie Neider, Tom Davis, and Dave Shreiner. *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Version 1.2.*

Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 1999. [cited at p. 93]

[129] Xiaozeng Yang and Lei Li. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics (Oxford, England)*, 27(18):2614–2615, September 2011. [cited at p. 103]

[130] Nevin D. Young, Frederic Debelle, Giles E. D. Oldroyd, Rene Geurts, Steven B. Cannon, Michael K. Udvardi, Vagner A. Benedito, Klaus F. X. Mayer, Jerome Gouzy, Heiko Schoof, Yves Van de Peer, Sebastian Proost, Douglas R. Cook, Blake C. Meyers, Manuel Spannagl, Foo Cheung, Stephane De Mita, Vivek Krishnakumar, Heidrun Gundlach, Shiguo Zhou, Joann Mudge, Arvind K. Bharti, Jeremy D. Murray, Marina A. Naoumkina, Benjamin Rosen, Kevin A. T. Silverstein, Haibao Tang, Stephane Rombauts, Patrick X. Zhao, Peng Zhou, Valerie Barbe, Philippe Bardou, Michael Bechner, Arnaud Bellec, Anne Berger, Helne Bergs, Shelby Bidwell, Ton Bisseling, Nathalie Choisne, Arnaud Couloux, Roxanne Denny, Shweta Deshpande, Xinbin Dai, Jeff J. Doyle, Anne-Marie Dudez, Andrew D. Farmer, Stephanie Fouteau, Carolien Franken, Chrystel Gibelin, John Gish, Steven Goldstein, Alvaro J. Gonzlez, Pamela J. Green, Asis Hallab, Marijke Hartog, Axin Hua, Sean J. Humphray, Dong-Hoon Jeong, Yi Jing, Anika Jcker, Steve M. Kenton, Dong-Jin Kim, Kathrin Klee, Hongshing Lai, Chunting Lang, Shaoping Lin, Simone L. Macmil, Ghislaine Magdelenat, Lucy Matthews, Jamison McCorrison, Erin L. Monaghan,

Jeong-Hwan Mun, Fares Z. Najar, Christine Nicholson, Celine Noirot, Majesta OBleness, Charles R. Paule, Julie Poulain, Florent Prion, Baifang Qin, Chunmei Qu, Ernest F. Retzel, Claire Riddle, Erika Sallet, Sylvie Samain, Nicolas Samson, Iryna Sanders, Olivier Saurat, Claude Scarpelli, Thomas Schiex, Beatrice Segurens, Andrew J. Severin, D. Janine Sherrier, Ruihua Shi, Sarah Sims, Susan R. Singer, Senjuti Sinharoy, Lieven Sterck, Agns Viollet, Bing-Bing Wang, Keqin Wang, Mingyi Wang, Xiaohong Wang, Jens Warfsmann, Jean Weissenbach, Doug D. White, Jim D. White, Graham B. Wiley, Patrick Wincker, Yanbo Xing, Limei Yang, Ziyun Yao, Fu Ying, Jixian Zhai, Liping Zhou, Antoine Zuber, Jean Denarie, Richard A. Dixon, Gregory D. May, David C. Schwartz, Jane Rogers, Francis Quetier, Christopher D. Town, and Bruce A. Roe. The medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524, December 2011. [cited at p. 68, 81, 82]

[131] Qiao-Ying Zeng, Cun-Yi Yang, Qi-Bin Ma, Xiu-Ping Li, Wen-Wen Dong, and Hai Nian. Identification of wild soybean miRNAs and their target genes responsive to aluminum stress. *BMC Plant Biology*, 12(1):182, October 2012. [cited at p. 82]

[132] Yuanji Zhang. miRU: an automated plant miRNA target prediction server. *Nucleic acids research*, 33(Web Server issue):W701–704, July 2005. [cited at p. 15, 16]

[133] Yun Zheng, Yong-Fang Li, Ramanjulu Sunkar, and Weixiong Zhang. Se-

qTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic acids research*, 40(4):e28, February 2012. [cited at p. 32, 61]

[134] Liguo Zhou, Yunhua Liu, Zaochang Liu, Deyan Kong, Mei Duan, and Lijun Luo. Genome-wide identification and analysis of drought-responsive microRNAs in oryza sativa. *Journal of Experimental Botany*, 61(15):4157–4168, October 2010. [cited at p. 65]

[135] Xuefeng Zhou, Guandong Wang, Keita Sutoh, Jian-Kang Zhu, and Weixiong Zhang. Identification of cold-inducible microRNAs in plants by transcriptome analysis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1779(11):780–788, November 2008. [cited at p. 65]

Appendices

Appendix A

Parameter definitions

Below shows a list of PAREsnip parameter options followed by the parameter used (bold) and a description of the parameter.

- min sRNA abundance: **1**. Targets will not be sought for any sRNA with a raw abundance less than the specified minimum,
- subsequences are secondary hits: **no**. One sRNA sequence which hits a transcript cleavage site may also have several sRNA subsequences;
- output secondary hits to file: **no**. Secondary hits are recorded and may be output to file ;
- use weighted fragments abundance: **yes**. Categories can be calculated using either raw degradation fragment abundance or weighted degradation fragment abundance. Weighted fragment abundance is the raw fragment abundance divided by the number of times the fragment

aligned to the transcriptome. Ticking this check-box tells PARESnip to calculate categories using weighted fragment abundance;

- category 0: **yes**. Peaks identified as category 0 will be included as potential sRNA cleavage sites;
- category 1: **yes**. Peaks identified as category 1 will be included as potential sRNA cleavage sites;
- category 2: **yes**. Peaks identified as category 2 will be included as potential sRNA cleavage sites;
- category 3: **yes**. Peaks identified as category 3 will be included as potential sRNA cleavage sites;
- category 4: **yes**. Peaks identified as category 4 will be included as potential sRNA cleavage sites;
- discard tr rna: **yes**. Any sRNA which has a full length match to t/rRNA will be discarded from the search;
- discard low complexity srnas: **yes**. Candidate targets are removed if of low complexity. A low complexity sequence contains 2 or fewer unique nucleotides;
- discard low complexity candidates: **yes**. Sequences within the sRNAome are discarded if the composition of the sequence is of low

complexity. A low complexity sequence contains 2 or fewer unique nucleotides;

- min fragment length: **20**. Any degradation fragment having fewer nucleotides that this threshold will be discarded;
- max fragment length: **21**. Any degradation fragment having more nucleotides that this threshold will be discarded;
- min sRNA length: **19**. Any sRNA having fewer nucleotides than this threshold will be discarded;
- max sRNA length: **24**. Any sRNA having more nucleotides than this threshold will be discarded;
- allow single nt gap: **yes**. A sRNA/target duplex may contain a single nucleotide gap;
- allow mismatch position 11: **no**. A sRNA/target duplex may contain a mismatch at position 11 (5' sRNA);
- allow adjacent mismatches: **no**. A sRNA/target duplex may contain more than 2 adjacent mismatches after position 12 (3' sRNA);
- max mismatches: **4.0**. The maximum number of mismatches permitted within an sRNA/target duplex (G:U pairs = 0.5 mismatches);

- calculate pvalues: **yes**. PARESnip will calculate p-values for each interaction reported;
- number of shuffles: **100**. The number of dinucleotide shuffles to be used for p-value calculation;
- pvalue cutoff: **1.0**. The p-value threshold. A p-value calculation will not continue past this threshold;
- do not include if greater than cutoff: **yes**. Interactions with a p-value exceeding the threshold will not be reported;
- number of threads: **7**. The number of threads PARESnip should use to perform the analysis. More threads reduce the time taken to complete an analysis.

List of Figures

List of Tables