

**Computational methods for
functional analysis of plant small
RNAs using the RNA degradome**



Joshua Thody

School of Computing Sciences

University of East Anglia

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2020

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

Acknowledgements

First and foremost, I would like to thank my supervisory team, Professor Vincent Moulton, Professor Tamas Dalmay and Dr. Leighton Folkes for the invaluable advice, guidance and support that they have given me over the course of my studies. I would also like to thank the Norwich Research Park and the BBSRC for the financial support they have provided over the last four years.

Much of this work would not be possible without invaluable contribution from collaborators, in particular Dr. Irina Mohorianu and Dr. Ping Xu. I would also like to thank Dr. Ping Xu and her group for hosting me in Shanghai as part of our research collaboration. Additionally, I would like to thank Dr. Matthew Stocks and Dr. Irina Mohorianu for their support and guidance, particularly at the beginning of my studies. I would like to thank members of the Dalmay lab, specifically Rocky and Thomas, for their patience and understanding when explaining biological concepts.

Special thanks go out to my partner, Hannah, my parents and also Hannah's parents for their continued support and also the encouragement to pursue a PhD in the first place.

I am grateful to be surrounded by a wonderful group of friends both at UEA and outside of academia, without which I would not have succeeded. Finally, I would like to thank all members of CMP support for their technical assistance and advice over the last few years.

Declaration

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

Joshua Thody

September 2020

Publications

Stocks, Matthew B., Irina Mohorianu, Matthew Beckers, Claudia Paicu, Simon Moxon, **Joshua Thody**, Tamas Dalmay, and Vincent Moulton. "The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs." *Bioinformatics*, Volume 34, Issue 19, Pages 3382–3384

Joshua Thody*, Leighton Folkes*, Zahara Medina-Calzada, Ping Xu, Tamas Dalmay, and Vincent Moulton. "PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules." *Nucleic Acids Research*, Volume 46, Issue 17, Pages 8730-8739

Joshua Thody, Vincent Moulton, and Irina Mohorianu. "PAREameters: a tool for computational inference of plant miRNA–mRNA targeting rules using small RNA and degradome sequencing data." *Nucleic Acids Research*, Volume 48, Issue 5, Pages 2258-2270

Joshua Thody, Leighton Folkes and Vincent Moulton. "NATpare: a pipeline for high-throughput prediction and functional analysis of nat-siRNAs." *Nucleic Acids Research*, Volume 48, Issue 12, Pages 6481–6490

* Joint first authors.

Abstract

Small RNAs (sRNAs) are a broad class of short regulatory non-coding RNAs that play critical roles in many important biological pathways. They suppress the translation of messenger RNAs (mRNAs) by directing the RNA-induced silencing complex to their sequence-specific mRNA target(s). In plants, this typically results in mRNA cleavage and subsequent degradation of the mRNA. Cleaved mRNA fragments can be captured on a genome-wide scale using a high-throughput sequencing technique called degradome sequencing, which can then be used to identify causal sRNAs.

Recent improvements to sequencing technologies have resulted in typical sequencing experiments now producing millions of unique reads. This has led to new challenges in bioinformatics regarding the computation time and resources required to perform sRNA and degradome data analyses. In this thesis, we present three new sRNA and degradome analysis tools that we have developed called PAREsnip2, PAREameters and NATpare.

PAREsnip2 is a tool we developed to predict sRNA targets, on a genome-wide scale, using degradome data and configurable targeting rules. Employing novel sequencing encoding and data structures, PAREsnip2 outperforms existing tools in computation time, at times by more than two orders of magnitude, with minimal computational resource requirements.

PAREameters is a computational method for inference of plant microRNA targeting rules, using the degradome, that can then be employed by PAREsnip2.

Benchmarking on multiple *A. thaliana* datasets show that the computationally inferred criteria outperform currently used criteria in terms of sensitivity on all datasets while maintaining precision on most.

NATpare is a tool for high-throughput prediction and functional analysis of nat-siRNAs using the degradome. NATpare is the first tool of its kind to combine nat-siRNA prediction with functional analysis using the degradome. Compared to current methods, our new algorithm speeds up computation time by over two orders of magnitude when analysing an *A. thaliana* dataset. We also demonstrate that it is the only computational method able to complete analyses of non-model organisms within a reasonable time frame.

We exemplify the use of these computational methods by performing functional analysis of CMV D-satRNA derived sRNA in *S. lycopersicum* to better understand their role in virus induced plant death.

Access Condition and Agreement

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

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

Table of contents

List of figures	xiv
List of tables	xx
1 Introduction	1
2 Background	6
2.1 Summary	6
2.2 DNA and RNA	6
2.3 RNA interference	8
2.3.1 MicroRNAs	9
2.3.2 Small interfering RNAs	11
2.3.2.1 Secondary small interfering RNAs	13
2.3.2.2 Natural anti-sense transcript small interfering RNAs	14
2.4 Sequencing of biological samples	14
2.4.1 Roche/454 sequencing	15
2.4.2 Illumina sequencing	16

2.4.3	Ion Torrent sequencing	17
2.4.4	Pacific Biosciences	17
2.4.5	Sequencing data repositories	17
2.5	From biology to bioinformatics	18
2.5.1	Quality control and processing of sequencing data	18
2.5.2	Computational classification of plant sRNAs	22
2.5.2.1	miRNA prediction	22
2.5.2.2	Secondary siRNA and ta-siRNA prediction	23
2.5.2.3	NATs and nat-siRNA prediction	23
2.5.3	Computational prediction of small RNA targets in plants	24
2.6	Validation of sRNA targets	27
2.7	Discussion	29
3	High-throughput sRNA-mRNA target prediction using the degradome	30
3.1	Summary	30
3.2	Background	31
3.2.1	CleaveLand	31
3.2.2	SeqTar	35
3.2.3	PAREsnip	36
3.2.4	sPARTA	38
3.2.5	Web-based tools	40
3.2.6	Issues with current methods	40
3.3	Methods	41

3.3.1	Data input	42
3.3.2	Sequence filtering	44
3.3.3	Binary encoding of sequence input	44
3.3.4	Target candidate generation	45
3.3.5	Region extraction and candidate grouping	46
3.3.6	Predefined and user configurable targeting rules	47
3.3.7	Computing valid region alignment matrices	48
3.3.8	Three-stage candidate filtering	49
3.3.9	Target search and results filtering	49
3.3.10	Implementation and output	51
3.3.11	Degradome library construction	51
3.3.12	Sequence datasets	52
3.4	Results	52
3.4.1	Sequencing data	52
3.4.2	Computational performance benchmarking	53
3.4.3	Prediction performance benchmarking	54
3.4.4	Evaluation of the optional filtering methods	56
3.4.5	Genome-wide analysis of degradome datasets	58
3.5	Discussion	60
3.6	Conclusion	63
4	Computational inference of plant microRNA targeting rules using the degradome	64

4.1	Summary	64
4.2	Background	65
4.3	Methods	67
4.3.1	The PAREameters pipeline	67
4.3.2	miRNA prediction	70
4.3.3	Target prediction using permissive criteria	70
4.3.4	miRNA–mRNA duplex analysis and inference of targeting criteria	71
4.3.5	Implementation of PAREameters	73
4.3.6	Datasets	74
4.4	Results	75
4.4.1	Evaluation of inferred targeting rules in <i>A. thaliana</i>	75
4.4.2	Evaluation of data input size and retain rate on sensitivity and precision	77
4.4.3	Consistency of attribute distributions and inferred criteria across miRNA subsets in <i>A. thaliana</i>	82
4.4.4	Evaluation of miRNA targeting criteria in non-model or- ganisms	86
4.4.5	Employing data-driven targeting criteria on non-model or- ganisms	90
4.5	Discussion	91
4.6	Conclusion	94

5	High-throughput prediction and functional analysis of nat-siRNAs using the degradome	96
5.1	Summary	96
5.2	Background	97
5.3	Methods	99
5.3.1	Data input and configuration	101
5.3.2	Sequence filtering	103
5.3.3	Search space reduction	103
5.3.4	NAT pair search	104
5.3.5	Categorization of candidate nat-siRNAs	105
5.3.6	NAT alignment distribution and sRNA alignment densities	107
5.3.7	Functional analysis of candidate nat-siRNAs	107
5.3.8	Implementation and output	108
5.3.9	Sequencing datasets	108
5.4	Results	109
5.4.1	Benchmarking and comparison with NATpipe	109
5.4.2	Comparing the expression of nat-siRNAs in <i>A. thaliana</i> control and salt stress treated samples	112
5.4.3	Investigation into the function of <i>cis</i> - and <i>trans</i> -nat-siRNAs in different <i>A. thaliana</i> tissues	116
5.5	Discussion	118
5.6	Conclusion	120

6	Functional analysis of necrogenic CMV D-satRNA derived sRNA in <i>Solanum lycopersicum</i>	121
6.1	Summary	121
6.2	Background	122
6.3	Methods	125
6.3.1	Plant materials and growth	125
6.3.2	RNA extraction, library construction and sequencing	125
6.3.3	Pre-processing of sequencing libraries	126
6.3.4	Identification of necrogenic sRNA	126
6.3.5	Target prediction	127
6.3.6	Target validation	127
6.4	Results	128
6.4.1	Sequencing data	128
6.4.2	Necrogenic D-satRNA derived sRNA	131
6.4.3	Identification of host mRNA targets	131
6.4.3.1	SCC1P2	133
6.4.3.2	ERF4	134
6.4.3.3	CSFP	138
6.4.4	Target site conservation in lethal and non-lethal infection	139
6.4.4.1	SCC1P2	141
6.4.4.2	ERF4	142
6.4.4.3	CSFP	143
6.4.5	Validation of targets	144

6.5	Conclusion	145
7	Future work and thesis conclusion	148
7.1	Summary	148
7.2	Future work	148
7.2.1	Combining existing workflows	148
7.2.2	Further work into sRNA targeting criteria	149
7.2.3	Further work into necrogenic sRNA	150
7.2.4	Impact of CMV infection on host gene expression	151
7.3	Thesis conclusion	152
	References	200

List of figures

2.1	The central dogma of molecular biology. Demonstrating the process of DNA replication, transcription of DNA to RNA and the translation of RNA into functional proteins. Figure obtained from Wikipedia under CC-BY-SA 3 license.	8
2.2	Part of the RNAi pathway focusing on the involvement of RISC. This includes RISC binding double-stranded RNA, degrading one of the strands and using the other to target complementary messenger RNA resulting in cleavage and subsequent mRNA degradation.	10
2.3	Hierarchical classification system for endogenous plant sRNAs. Thick black lines indicate hierarchical relationships. Figure from Axtell <i>et al.</i> [9] with permission from Annual Reviews, Inc.	12
2.4	Percent of mismatched and G:U base-pairs at each target position in the rule development set. (B) Minimum Free Energy (MFE) ratio of target-miRNA duplexes from the rule development set. Every miRNA-mRNA duplex in the rule development set had a MFE ratio of at least 0.73. Figure from Allen <i>et al.</i> [3] with permission from Elsevier.	26

2.5	(a) The distribution of mismatches, gaps and G:U base pairs from 155 genuine miRNA-mRNA target duplexes in <i>A. thaliana</i> . (b) <i>A. thaliana</i> miR172a and its target , At4g36920, illustrating the alignment scoring system used to predict targets. The coloured box highlights positions 2 through 13, relative to the miRNA 5' end, indicating the region where penalty scores are doubled. Figure from Fahlgren and Carrington [61] with permission from Springer Nature.	27
3.1	A sRNA is loaded into an Argonaute (AGO) protein and can target the mRNA leading to endonucleolytic cleavage. The resulting mRNA fragments that are un-capped at the 5' end 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 when they are realigned to the mRNA reference sequence.	32
3.2	An overview of different stages of the PAREsnp2 algorithm. (A) Shows the inputs and processing steps performed to predict sRNA targets evidenced through degradome sequencing. (B) Shows the process of encoding sequence data into a number system. (C) Visual representation of the three-stage candidate filtering process. Regions are labelled R and target regions are labelled TR.	43
3.3	The number of interactions reported when using MFE as a filter. As the MFE filter ratio increases, there is a reduction in the number of captured sRNA-mRNA interactions. A cut-off score of 0.70 captures 98% of the possible validated interactions.	57

3.4	The number of interactions reported when using p -value as a filter. As the cut-off decreases, there is a reduction in the number of captured sRNA–mRNA interactions. The default cut-off score of 0.05 captures 85.6% of the possible validated interactions.	58
4.1	PAREameters pipeline. The input and output data are represented by continuous rounded rectangles, processes are represented by straight rectangles and the different steps of the analysis are represented by dashed rounded rectangles. PAREameters takes as input two types of sequencing samples, paired sRNA and degradome, a genome with corresponding annotations and current miRBase miRNA annotations. The output is a set of data-inferred thresholds for a rule-based prediction of miRNA–mRNA interactions using e.g. PAREsnip2. The sRNAome and degradome inputs are the experiment-specific datasets whereas the genome, transcriptome and annotated miRNA inputs are part of the species annotation. . .	69
4.2	Side-by-side comparison of property distributions for conserved and species-specific miRNAs in <i>A. thaliana</i> . Using experimentally validated miRNA–mRNA interactions as input, we calculated the position-specific properties (A) and the MFE ratio distribution (B) for the conserved and species-specific miRNA–mRNA interactions. The significance of the differences in the localization of gaps, G:U pairs and mismatches were assessed using offset χ^2 tests and the contribution of individual categories was evaluated using Fisher exact tests. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov–Smirnov test, which reported a p -value of 8.57×10^{-10}	83

5.4	Venn diagram showing the overlap of nat-siRNA predictions between all tissues within the D3 dataset.	118
6.1	Symptoms of CMV and D-satRNA infection in <i>S. lycopersicum</i> . Panel A is the control (Mock), B is CMV infected, C is CMV D-satRNA infected and D is CMV D-satRNA infected <i>S. lycopersicum</i> . CMV D-satRNA infection induces necrosis while Dm-satRNA attenuates some symptoms of CMV infection.	124
6.2	Venn diagram showing the overlap of unique sRNA containing at least one of the necrogenic nucleotides in each of the CMV D-satRNA infected sRNA libraries.	132
6.3	T-plot showing the degradation activity for SCC1P2 in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.	134
6.4	T-plot showing the degradation activity for SCC1P2 in the D2 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.001.	135
6.5	T-plot showing the degradation activity for SCC1P2 in the D3 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.001	135
6.6	T-plot showing the degradation activity for ERF4 in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.002	137
6.7	T-plot showing the degradation activity for ERF4 in the D2 degradome dataset. The cleavage signal is Category-2 and the interaction has a <i>p</i> -value of 0.082.	137

6.8	T-plot showing the degradation activity for ERF4 in the D3 degradome dataset. The cleavage signal is Category-2 and the interaction has a <i>p</i> -value of 0.121	138
6.9	T-plot showing the degradation activity for CSFP in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.001	139
6.10	T-plot showing the degradation activity for CSFP in the D2 degradome dataset. The cleavage signal is Category-0 and the interaction has a <i>p</i> -value of 0.001.	140
6.11	T-plot showing the degradation activity for CSFP in the D3 degradome dataset. The cleavage signal is Category-2 and the interaction has a <i>p</i> -value of 0.041	140
6.12	Target validation results for CSFP in CMV D-satRNA and Dm-satRNA infected <i>N. benthamiana</i> . Panel A is the predicted target site, panel B is the cleavage signal, panel C is fluorescent intensity in D-satRNA and panel D is fluorescent intensity in Dm-satRNA. .	146
6.13	Target validation results for SCC1P2 in CMV D-satRNA and Dm-satRNA infected <i>N. benthamiana</i> . Panel A is the predicted target site, panel B is the cleavage signal, panel C is fluorescent intensity in D-satRNA and panel D is fluorescent intensity in Dm-satRNA. The results from this experiment show a clear reduction in the fluorescent intensity in D-satRNA when compared to Dm-satRNA.	147

List of tables

3.1	The 2-bit encoding of nucleotides	45
3.2	Features within a sRNA–mRNA alignment which are used during the duplex alignment process and their default values but can also be configured by the user.	48
3.3	Summary statistics (number of reads) from the sequencing of three <i>A. thaliana</i> degradome replicates (NR = non-redundant).	53
3.4	Benchmarking results for both time and memory usage in Gigabytes (GB) from running each tool using the generated small RNA datasets. If the entry is DNF it means that the tool did not complete the analysis within the 10 day cut-off. A ‘-’ means that we did not attempt to run the tool. Benchmarking results show that PAREsnip2 was able to complete analysis considerably faster than all other tools with low resource requirements.	54

3.5	The results from the accuracy performance benchmarking of each tool over the three biological replicates. V = validated targets, NV = non-validated and %PV = percentage of possible validated targets that could be found. Results show PAREsnip2 captures a larger number of the experimentally validated <i>A. thaliana</i> targets compared to other publicly available tools using both sets of default targeting criteria.	56
4.1	The PAREsnip2 parameter values for the Allen <i>et al.</i> , manually inferred and PAREameters permissive criteria. The Allen <i>et al.</i> criteria were previously inferred in 2005 [3]. The manually inferred criteria was inferred on a set of 387 experimentally validated <i>A. thaliana</i> interactions. The permissive parameters are used initially by PAREameters to find high-confidence (HC) interactions. The inferred criteria are then extracted from HC interactions using the retain rate parameter. MM = mismatch, CR = core regions (positions 2-13 of miRNA), MFE = minimum free energy.	71
4.2	The PAREameters inferred criteria for each of the <i>A. thaliana</i> datasets. MM = mismatch, CR = core region (2-13nt of miRNA) and MFE = minimum free energy. datasets. MM = mismatch, CR = core region (2-13nt of miRNA), adj = adjacent and MFE = minimum free energy.	76
4.3	Comparison of sensitivity and specificity between the Allen <i>et al.</i> criteria and the PAREameters inferred criteria on the <i>A. thaliana</i> datasets. Allen = Allen <i>et al.</i> rules, Inf. = PAREameters inferred criteria, V = validated, NV = non-validated, Se = sensitivity and PPV = precision. An increase in the achieved Se is observed for the inferred criteria over all <i>A. thaliana</i> datasets.	78

4.4	The median sensitivity (Se) and precision (PPV) values for the cross-validation experiments on the <i>A. thaliana</i> datasets. The cross validation was done on a 75/25% split for training and testing, respectively. Each analysis was repeated 50 times and the median value was recorded.	79
4.5	The median sensitivity (Se) and precision (PPV) values for the training-size experiment on the <i>A. thaliana</i> D1 datasets. For each dataset, an increase in training-size resulted in an overall increase in sensitivity.	80
4.6	Offset χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in <i>A. thaliana</i> . The contribution of specific properties, such as mismatches (MM), G:U pairs and gaps are assessed using Fisher exact tests. Values at or below the significance threshold (0.05) and highlighted in bold.	85
4.7	Sensitivity on cross pairwise comparisons for criteria inferred on conserved or species-specific miRNAs for the validated <i>A. thaliana</i> interactions. The targeting criteria were inferred using a retain rate of 0.85 and a considerable decrease in sensitivity was observed for the mismatched pairs i.e. training on conserved interactions and testing on specific interactions.	86
5.1	The configurable parameters for NATpare. The values used during analysis can be changed by modifying the input configuration file or by using the command line when running the tool.	102

5.2	Computation performance comparison between NATpipe and the newly developed NATpare pipeline when evaluated on the simulated datasets. If the tool did not finish within 10 days it was recorded as did not finish (DNF).	110
5.3	Top 10 reported <i>G. max cis</i> -NATs with the highest number of unique reported nat-siRNAs by Zheng <i>et al.</i> [235] and the prediction results from NATpare and NATpipe	112
5.4	The differentially expressed nat-siRNAs, as reported by iDEP, in the <i>A. thaliana</i> seedling salt-stress dataset. 10 of the 29 sequences originated from NAT pairs where one of the transcripts is annotated as a potential natural antisense gene. The transcript that gives rise to the largest number of nat-siRNAs is currently annotated as unknown RNA and the corresponding NAT has an unknown function. Adjusted <i>p</i> -values were obtained using a false discovery rate of 0.1 and were expressed to 3 significant digits. Any extreme <i>p</i> -values (i.e. $p < 0.001$) were reported as $p < 0.001$	115
6.1	The number of redundant, non-redundant and <i>S. lycopersicum</i> genome aligned reads in each of the sRNA libraries.	129
6.2	The number of redundant, non-redundant and transcriptome aligned reads (positive direction only) in each of the PARE libraries. . . .	129
6.3	The number of redundant, non-redundant and virus aligned reads in each of the sRNA libraries.	130
6.4	The number of redundant and non-redundant reads that align to the D-satRNA sequence and contain at least one of the necrogenic nucleotides in each of the CMV D-satRNA infected sRNA libraries.	131

6.5	The number of unique target sites and their categories reported by PAREsnip2 when analysing the potential necrogenic sRNAs and corresponding degradome in each of the CMV D-satRNA infected libraries.	132
6.6	The number of unique target sites and their categories reported by PAREsnip2 that are also conserved between at least two replicates and have a cleavage signal with abundance ≥ 5	132
6.7	The mutations at the target site of the conserved SCC1P2 homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the <i>S. lycopersicum</i> target site are highlighted. S = survives infection and D = dies from infection. . .	142
6.8	The mutations at the target site of the conserved ERF4 homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the <i>S. lycopersicum</i> target site are highlighted. S = survives infection and D = dies from infection. . .	143
6.9	The mutations at the target site of the conserved CSFP homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the <i>S. lycopersicum</i> target site are highlighted. S = survives infection and D = dies from infection. . .	144

Chapter 1

Introduction

Small RNAs (sRNAs) are short (19-24 nucleotide) RNA molecules that are involved in many important and diverse biological pathways such as growth and development, disease resistance, and stress response [194, 54]. The mechanisms in which they function, a process known as RNA interference, where sRNAs regulate the expression of their target messenger RNAs (mRNAs), was discovered by [67], who were awarded a Nobel Prize for their work. In plants, sRNA mediated gene regulation typically happens through messenger RNA cleavage and these cleavage products can be captured on a genome-wide scale using a high-throughput sequencing technique called degradome sequencing. Recent advances in next-generation sequencing technologies have resulted in increased availability, higher throughput and reduced cost. Consequently, this has enabled generation of sRNA and related sequencing data from a wide range of species, tissues and conditions [58] in addition to sequencing datasets in general growing larger in both size and read count.

High-throughput sequencing has become one of the de facto experimental techniques for identifying sRNAs. However, it can still be quite challenging to identify and confirm their targets, a fundamental step in understanding their

function. Degradome data has proven to be a valuable resource that can be used to quantify mRNA cleavage products and to help identify possible causal sRNAs. Many computational methods exist for the identification of sRNA targets, some of which also incorporate degradome data into their prediction pipeline. However, these methods present various weaknesses, in particular with their computation time and resource requirements, but also their prediction accuracy. In this thesis, we develop user-friendly software for sRNA and degradome analysis that is scalable with recent sequencing datasets.

For each tool presented, we provide detailed descriptions of the methods implemented and perform in-depth computational analyses and benchmarking that demonstrates their usefulness to the field of sRNA research. Through collaboration with experimental biologists, we also utilize our software to perform computational analyses in *Solanum lycopersicum* that helps answer their biological questions. Moreover, further experimental validation of the predictions made using our tools provides verification of the computational methods developed and presented in this thesis.

We hope that our software contributions will enable the use of specialist bioinformatics tools without the need for any computational expertise and in doing so, will contribute towards new discoveries within RNA silencing pathways in all manner of experimental contexts.

We now give an overview of the contents of each chapter in this thesis.

Chapter 2. In this chapter, we provide an introduction into the relevant biological background information necessary to the work presented in the rest of this thesis. We focus on sRNA biogenesis and function, more specifically, their role in post-transcriptional gene silencing in plants. We then provide an overview of methods to obtain sequencing data from biological samples, which leads onto a discussion about quality control of sequencing data. This is followed by a descrip-

tion of some of the software tools available for classification of sRNAs. We then introduce tools available for sequence-based sRNA target prediction and discuss the search parameters they use. Finally, we introduce a high-throughput technique for validating sRNA targets in plants called the degradome.

Chapter 3. In this chapter, we introduce how degradome data can be used to support computational prediction of sRNA targets in plants. This leads onto a review of current methods and tools available for degradome analysis. We then introduce a new algorithm and software tool, called PAREsnip2, that can be used to quickly and efficiently identify sRNA targets on a genome-wide scale using configurable targeting rules. We evaluate the computational and prediction performance of PAREsnip2 and compare the results to those of currently available methods. The algorithms and software implementation, experimental testing and the generation of results were my contribution to this work. The idea to develop a new tool for this type of analysis was conceived jointly between myself, Dr. Leighton Folkes and Prof. Vincent Moulton. The growing of the plants and generation of the sequencing data was performed by members of the Dalmy lab at UEA.

Chapter 4. In this chapter, we introduce PAREameters, a tool for computational inference of plant microRNA (miRNA) targeting criteria using degradome sequencing data. This tool was developed to assist the user when selecting targeting parameters for predicting sRNA targets using PAREsnip2. We then evaluate current, manually inferred and computationally inferred criteria on a set of high-confidence experimentally validated *Arabidopsis thaliana* miRNA targets in multiple datasets. We investigate the differences in inferred targeting criteria of conserved and species-specific miRNAs, which then leads onto the analysis of targeting criteria in non-model organisms. Finally, we compare the differences in inferred criteria between high and low confidence miRNA targets. The algorithms and software implementation, experimental testing and the generation of results were my contribution to this work. The idea to develop a new tool for this type

of analysis was conceived jointly between myself, Dr. Irina Mohorianu and Prof. Vincent Moulton. The R code used for the statistical analyses presented in this chapter was written by Dr. Irina Mohorianu.

Chapter 5. In this chapter, we introduce a new tool, called NATpare, for classification and functional analysis of natural anti-sense short interfering RNAs (nat-siRNAs) using degradome data. We begin by discussing the use of degradome data besides from sRNA target prediction. We then introduce the new software pipeline and accompanying algorithm. We then discuss the evaluation process and compare the results to that of other publicly available tools. Finally, we exemplify the use of NATpare by performing an investigation into nat-siRNAs identified in different tissues and stress conditions. The idea to develop a new tool for this type of analysis was conceived jointly between myself, Dr. Leighton Folkes and Prof. Vincent Moulton.

Chapter 6. In this chapter, we perform analyses on sRNA and degradome libraries obtained from *S. lycopersicum* infected with *Cucumber mosaic virus* (CMV) and D-satellite RNA (D-satRNA) using the UEA sRNA Workbench. We begin by introducing the virus and satellite RNAs and the effect they have on various plant species. We then explain the data we are using, how it was obtained, how it was processed and the results of the quality checking. This is followed by an investigation into the possible function of necrogenic D-satRNA derived sRNAs using PAREsnip2. Through the degradome, we identify a number of putative targets for the sRNAs containing necrogenic nucleotides that are found exclusively within the D-satRNA libraries. We then investigate these target genes in more detail by comparing the target site to homologous genes found in surviving plants. Finally, we present the results from some experimental validation of these candidates. All computational analyses presented in chapter are my contribution to this work.

Chapter 7. In this final chapter, we discuss some future directions and extensions to this work

Chapter 2

Background

2.1 Summary

This chapter provides an introduction to the biology and a review of computational methods relevant to the work presented within this thesis. This starts with a description of RNA silencing, as it is fundamental to the rest of this work, which includes a brief account of the biogenesis and function of sRNAs. This leads onto an introduction to the computational side of this work, starting with the generation, quality control and processing of sequencing data obtained from biological samples. We then discuss the currently available tools for the classification of sRNAs from sequencing data. Finally, we discuss current computational methods used for the prediction of sRNA targets, the parameters they employ, and the high-throughput methods used to validate these predictions.

2.2 DNA and RNA

Deoxyribonucleic acid (DNA) is a nucleic acid sequence that stores the genetic instructions used in the development, functioning, and reproduction of all known

organisms, usually referred to as the organisms genome. The information in DNA is stored as a sequence made up of four nucleotide (nt) bases: adenine (A), guanine (G), cytosine (C), and thymine (T). It is composed of two separate strands where the nucleotides are bound together through Watson-Crick base pairs, i.e. pairs in the form of guanine:cytosine (G:C) and adenine:thymine (A:T) hydrogen bonds [210], and this results in double-stranded DNA that forms a helix structure. To represent direction on a sequence of nucleotides, the terms five prime (5') and three prime (3') are used, with 5' referring to the start and 3' referring to the end of the sequence.

Ribonucleic acid (RNA) is another type of nucleic acid sequence that is produced from a DNA template through a process called transcription. This is the first step of gene expression, in which a section of DNA is copied into RNA by the enzyme RNA polymerase, with the resulting RNA sequence being either coding or non-coding RNA (ncRNA). Coding RNAs, also known as messenger RNAs (mRNAs), serve as a template for protein synthesis through translation, which is part of the central dogma of biology [45], shown in Figure 2.1. Alternatively, the transcribed RNA may encode for a non-coding RNA, such as a microRNA (miRNA), transfer RNA (tRNA) or ribosomal RNA (rRNA), each with their own specific functions within a cell [55].

Although DNA and RNA are both nucleic acids sequences, they differ in a number of important ways. Firstly, RNA contains the base uracil (U) rather than thymine (T). Second, DNA is a blueprint for genetic information contained within the organism, whereas RNA employs this information to produce proteins or functional ncRNAs. Finally, DNA consists of two strands, arranged in a double helix [209], whereas RNA is usually only single stranded and folds upon itself to form different structures depending on its required function [170]. The way the RNA sequence folds and the structure that is formed is dependent on the intra-molecular

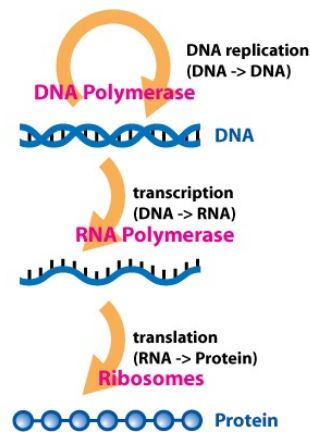


Fig. 2.1 The central dogma of molecular biology. Demonstrating the process of DNA replication, transcription of DNA to RNA and the translation of RNA into functional proteins. Figure obtained from Wikipedia under CC-BY-SA 3 license.

interactions between the nucleotides through complementary Watson–Crick base pairs [210].

2.3 RNA interference

RNA interference (RNAi), also known as RNA silencing, was discovered in both animals and plants in the 1990s [156, 66, 67]. The work in this thesis is focused exclusively on plants in which there exists at least three different RNAi pathways [16]. A common feature of each pathway is the excision of double stranded RNA (dsRNA) by RNase-III-type enzymes called Dicers [16]. In plants, there exists at least 4 different Dicer-like (DCL) proteins, each with a specific role within the RNAi pathway. Cleavage of dsRNA by one of the Dicer-like proteins results in the production of double stranded ncRNAs called small RNAs (sRNAs) which are in the range of 19-25 nucleotides (nt) [20]. After the cleavage process, the new double stranded fragments are separated into two single stranded RNAs. One of these single stranded RNAs, known as the guide RNA strand, is loaded into a member of the Argonaute (AGO) protein family, whilst the other strand is degraded. AGO forms part of a larger and complex system known as the RNA Induced Silencing

Complex (RISC), where mRNA targets are identified based on complementary Watson-Crick base pairing between the mRNA sequence and the guide sRNA sequence [226]. Once a target has been found, RISC can silence it through one of three mechanisms [34].

In plants, the sRNA-directed mRNA targets are generally silenced through cleavage and degradation due to the high degree of complementarity between the guide sRNA and the mRNA sequences [8]. This cleavage is highly specific and usually occurs between nucleotide positions 10 and 11 of the sRNA sequence [139, 59]. A simplified representation of this process is presented in Figure 2.2. The other two mechanisms are cytoplasmic siRNA silencing and suppression of transcription by DNA methylation [16], however these methods are not relevant to the work presented in this thesis, which is primarily focused on silencing through mRNA cleavage. Whilst the RNAi pathway is common to all sRNAs, the specific details differ based on the class of sRNA being processed. The following section summarises the differences in biogenesis and function of different classes of sRNA.

2.3.1 MicroRNAs

First discovered within the nematode model organism *Caenorhabditis elegans* in 1993 [119], miRNAs are a class of endogenous sRNA typically around 21nt in length. They are unique in that they are derived from a longer, single stranded precursor sequence that folds into an imperfect hairpin type structure known as a hairpin RNA (hpRNA). They can be found in plants, animals and even some viruses, and they play important roles in post-transcriptional regulation of gene expression [32, 108]. The biogenesis and mode of action of miRNAs differ in both plants and animals. In plants, the production of a mature miRNA is a multi-step process and starts with a single stranded primary miRNA (pri-miRNA) that is transcribed from a miRNA gene by an RNA polymerase II enzyme [120]. The pri-miRNA

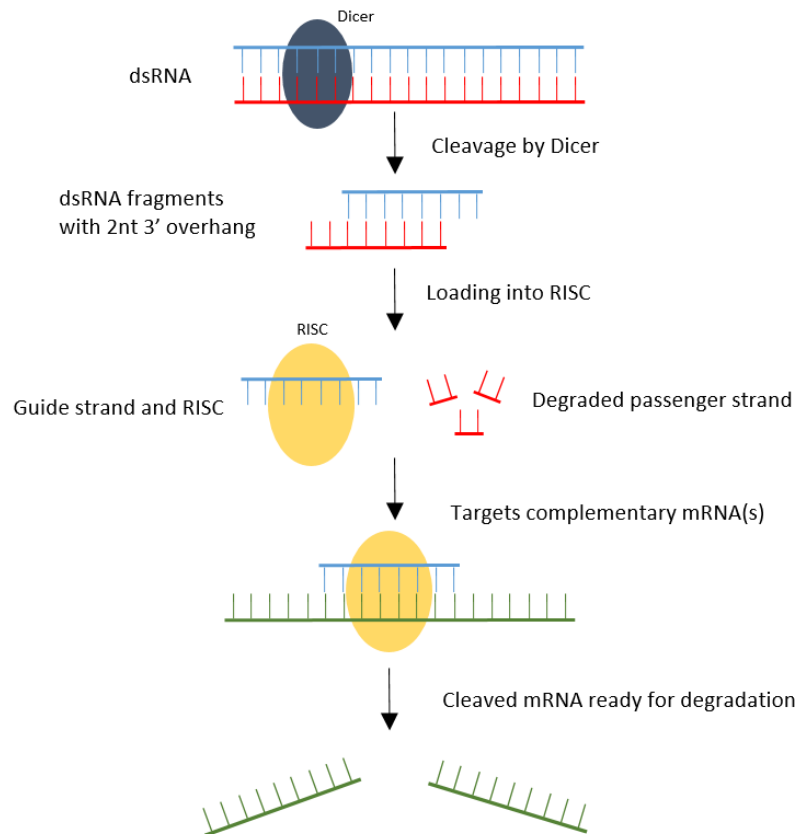


Fig. 2.2 Part of the RNAi pathway focusing on the involvement of RISC. This includes RISC binding double-stranded RNA, degrading one of the strands and using the other to target complementary messenger RNA resulting in cleavage and subsequent mRNA degradation.

then folds to form a hairpin structure and is processed by a DCL enzyme resulting in a precursor miRNA (pre-miRNA) with 2nt 3' overhang [91]. A DCL enzyme, typically DCL1, then further processes the pre-miRNA into a double stranded RNA (dsRNA) duplex consisting of the mature miRNA strand and its complementary sequence called the miRNA star (miRNA*) strand. Finally, the dsRNA duplex is separated by a helicase enzyme and the mature strand is loaded into a member of the AGO protein family and forms the RISC [149].

The primary mechanism of miRNA-mediated gene silencing in plants is through mRNA cleavage, however translational repression has also been observed [207]. Over the last 15 years, much emphasis has been placed on identifying plant miRNAs

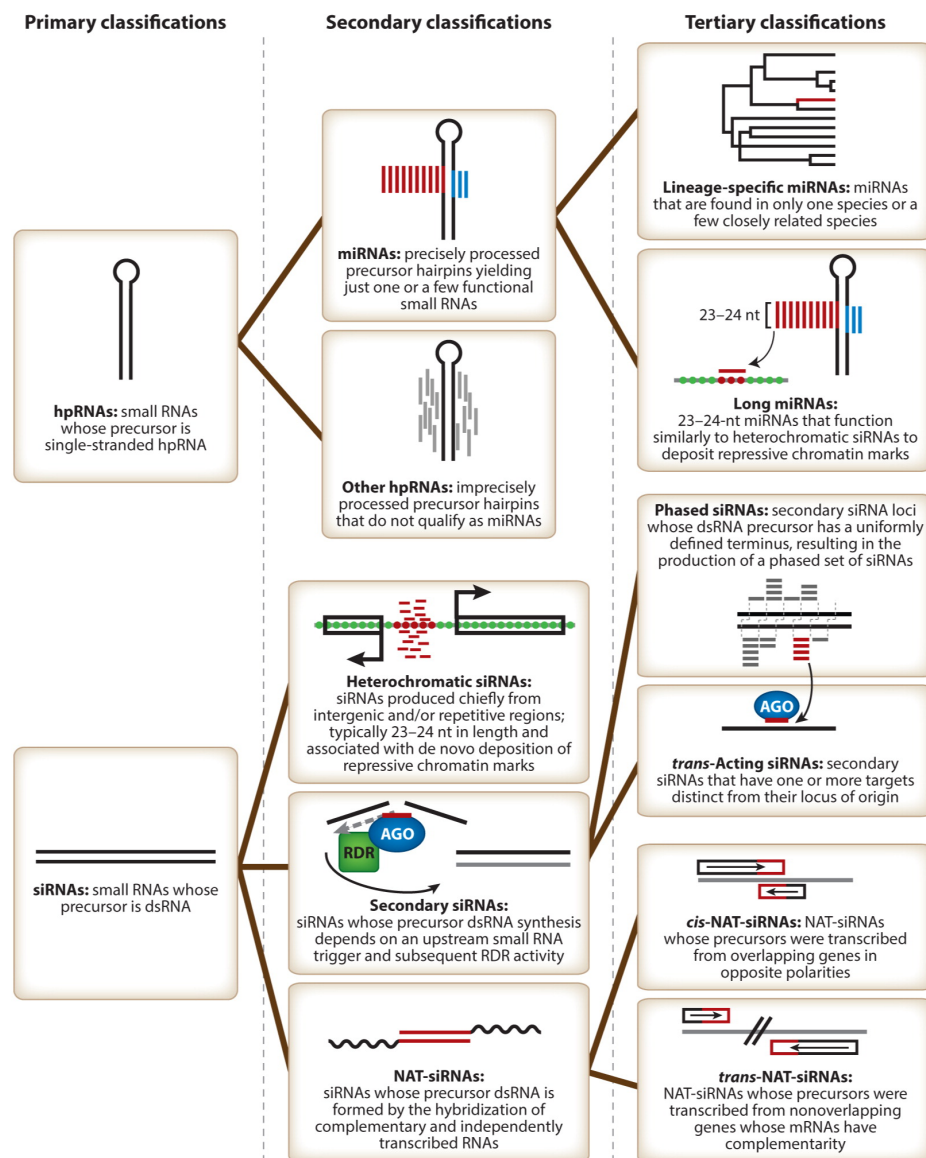
and investigating their involvement in plant development stages and stress response. The rapid growth of the miRNA field led to the miRBase database being created which provides comprehensive miRNA sequence data, annotation and predicted gene targets to the scientific community [77]. As of 2020, miRBase, which is currently on version 22, contains miRNA sequences from 271 organisms consisting of 38 589 hairpin precursors and 48 860 mature miRNA sequences [110].

Previous studies in both in model and non-model plants show that miRNAs play crucial roles in many biological processes including development, growth and response to different environmental abiotic and biotic stresses. In *Arabidopsis thaliana*, for example, miR167 targets auxin response factors (ARFs) in order to regulate the emergence of shoot-borne roots [81] and both miR156 and miR172 work together to regulate developmental timing and juvenile to adult transition [213]. It has also been shown that many miRNAs, including miR156, miR164, miR168, miR171, miR393, miR396 and miR398, are associated with a broad range of plant defense responses to stresses including drought, salt, and cold (see review paper [174]).

2.3.2 Small interfering RNAs

Found in both plants and animals, small interfering RNAs (siRNAs) are similar to miRNAs in that they are produced from the processing of long dsRNA precursors with 2nt 3' overhang and are involved in the RNAi pathway [33]. The primary difference between miRNAs and siRNAs is that miRNAs are derived from a single stranded hpRNA, whereas siRNAs are formed by the intermolecular hybridization of two complementary RNA sequences [9]. The dsRNA precursors of siRNAs can arise from the hybridization of sense and anti-sense transcripts, from the folding of an inverted-repeat sequence, from the hybridization of two unrelated RNA molecules with highly complementary sequences or, most commonly, following

synthesis by RNA-dependent RNA polymerases (RDR) [25], as shown in Figure 2.3. We now briefly introduce the biogenesis of different classes of siRNAs, with the exception of heterochromatic small interfering RNAs (het-siRNAs) as they do not induce mRNA cleavage.




 Axtell MJ. 2013.
Annu. Rev. Plant Biol. 64:137–59

Fig. 2.3 Hierarchical classification system for endogenous plant sRNAs. Thick black lines indicate hierarchical relationships. Figure from Axtell *et al.* [9] with permission from Annual Reviews, Inc.

2.3.2.1 Secondary small interfering RNAs

Secondary siRNAs are a type of sRNA derived from dsRNA precursors that are generated from prior sRNA processing [9]. Cleavage of mRNA transcripts can lead to synthesis of dsRNA by a RDR, which can then be processed by a DCL protein into secondary siRNAs. Current understanding of their biogenesis requires initiation by miRNAs or other secondary siRNAs, RDR6 and DCL4 [9]. Secondary siRNAs can be separated into two classifications: phased siRNAs (phasRNA) and *trans*-acting siRNAs (ta-siRNA), based on their origin and function. In both cases, they are able to function similarly to miRNAs by directing the RISC to induce cleavage of target mRNAs. The difference between these two classes is that ta-siRNA are able to induce the cleavage of mRNAs in *trans*, i.e. genes other than that of their originating mRNA [63].

The first miRNA-triggered ta-siRNA producing loci were initially identified and characterized in *A. thaliana* and these secondary ta-siRNA were found to suppress the expression of genes that were unrelated to their originating transcript [164, 206]. Currently, four families of ta-siRNA producing loci have been identified in *A. thaliana*: TAS1 and TAS2, cleaved by miR173, and TAS3 and TAS4, which are cleaved by both miR390 and miR828 [164, 206, 3, 211, 168]. Additional TAS genes, TAS5-10, have been identified or predicted in other plant species, suggesting that many secondary siRNA-producing loci may not yet be known [6, 124, 231, 241]. Since the discovery of TAS derived siRNA, the importance of these secondary siRNA has been the focus of much attention, for example, in 2006 Fahlgren *et al.* reported that juvenile-to-adult phase transition is controlled by TAS3 derived ta-siRNAs through negative regulation of ARF3 mRNA [62]. It has also been shown that failure of the ta-siRNA pathway to regulate ARF3 and ARF4 underlies tomato shoestring leaves, a symptom often associated with plant virus infection [223].

2.3.2.2 Natural anti-sense transcript small interfering RNAs

Natural anti-sense transcript small interfering RNAs (nat-siRNAs) are a unique class of siRNA that originate from the overlapping region of two complementary transcripts. These nat-siRNAs are induced by abiotic and biotic stresses [104, 26, 97] or accumulate in specific developmental stages [172, 238]. The founding example was identified in *A. thaliana*, where a pair of *cis*-NATs, SRO5 and P5CDH, were shown to be involved in the response to salt tolerance through the RNAi pathway [26]. During salt stress, SRO5 is expressed and can form a complementary overlapping region with the constitutively expressed P5CDH, which is then further processed by a specific biogenesis pathway to produce a 24nt nat-siRNA. This nat-siRNA then directs the cleavage of P5CDH, which is subsequently used as a template by RDR6 to produce dsRNA that is then processed by DCL1 to produce 21nt secondary nat-siRNAs, triggering a reinforcement phase [26]. Further information relating to this class of sRNA can be found in Chapter 5.

2.4 Sequencing of biological samples

Sequencing is the process of determining the order of the four nucleotide bases, A, C, T and G, that comprise a DNA sequence and is crucial to biological research amongst other fields [85]. The original methods for DNA sequencing were developed in 1977 and are now considered as first generation sequencing techniques. The first was called chemical cleavage sequencing and was developed by Maxam and Gilbert [147]. The second was called Sanger sequencing and was developed by Sanger and collaborators [176] building on a previously developed approach called plus and minus sequencing [175]. Chemical cleavage sequencing was not widely used due to the use of hazardous chemicals and the large amount of DNA that was required. Sanger sequencing however, due to its simplicity and reliability, became

the dominant method for sequencing at the time [82]. Further development into sequencing technologies resulted in the development of the first automatic sequencing machine in 1987 by Applied Biosystems, called the AB370, based on Sanger sequencing. The Sanger sequencing technique has been used in several sequencing projects of different plant species including *A. thaliana* [92], *Oryza sativa* (rice) [75] and *Glycine max* (soybean) [178]. Further improvements to these automatic sequencing instruments and their software aided the completion of the human genome in 2001 [42], which led to the development of Next Generation Sequencing (NGS) technologies [181], also referred to as High-throughput Sequencing (HTS) or Second Generation Sequencing (SGS).

NGS technologies differ from the original sequencing technologies in that they are massively parallel, high-throughput and lower in cost, and this development has enabled the generation of sequencing data on a massive scale [145]. Below, we briefly describe some of these technologies.

2.4.1 Roche/454 sequencing

Roche/454 sequencing was the first to achieve commercial introduction and is an approach that uses pyrosequencing, a technique that detects light emitted when additional nucleotides are added to a complementary strand of DNA being synthesized from a template sequence. This approach to sequencing is known as sequencing-by-synthesis [71]. In pyrosequencing, when an additional nucleotide is ligated it results in the release of a pyrophosphate, which initiates a number of subsequent downstream reactions that ends with the production of light by the enzyme luciferase. Through the detection of light after each subsequent additional nucleotide, the sequence of the DNA fragment is determined [144]. The use of a microtiter plate allows for a large number of reactions to occur in parallel, considerably increasing the sequencing throughput.

The Roche/454 sequencing technology is able to generate relatively long reads which are subsequently easier to map to a reference genome. However, it does have shortfalls when there exists homopolymer regions, i.e. large regions of a single nucleotide, resulting in insertion and deletion errors. This is because the identification and length of homopolymer regions are determined by the intensity of the light emitted and signals with very high or very low intensity levels may lead to miscalculating the number of nucleotides [90].

2.4.2 Illumina sequencing

Developed initially by Solexa and then later purchased by Illumina, the Illumina sequencing technology is a sequencing-by-synthesis approach and is currently the most used technology in the NGS market. The first sequencing machine released by Illumina/Solexa was the Genome Analyzer and was able to produce very short reads, roughly 35nt in length, and gave researchers the power to sequence 1Gbp of data in a single run. More recently, the output of the Illumina sequencing machines is much higher, around 600 Gbp, and the read lengths are longer, roughly 100bp, in length [113]. In brief, the Illumina sequencing method starts with the libraries being randomly fragmented and adaptors ligated to both ends of each fragment. Next, clusters are generated by loading the fragments onto a flow cell containing short sequences that are complementary to the library adaptors. Each fragment is then amplified into clonal clusters through a process called bridge amplification. During the sequencing process, the addition of a single nucleotide through synthesis emits a light signal which is detected by a camera and then translated into a nucleotide sequence through computer algorithms.

2.4.3 Ion Torrent sequencing

Life Technologies released the Ion Torrent's semiconductor sequencing technology in 2010. The preparation and sequence process are similar to that of the Roche/454 pyrosequencing platform. However, during the sequence synthesis process, instead of identifying light signals, Ion Torrent's semiconductor sequencing measures the pH changes induced by the release of hydrogen ions during DNA extension [173], which are then converted into a voltage signal and used to generate the nucleotide sequence [169]. One of the major advantages to the Ion Torrent sequencer is that they are able to produce reads with lengths up to 600bp.

2.4.4 Pacific Biosciences

Pacific Biosciences developed the first sequencer that uses single-molecule real-time sequencing (SMRT) and is an example of a third-generation sequencing technology. It uses the same light labeling process as other SBS technologies but does it in real time when the nucleotide additions occur rather than in cycle. Similar to other methods, the detection of the light emitting nucleotides makes it possible to determine the sequence composition. Compared to SGS, this approach has the advantage of being very fast to prepare [136] and allowing for sequencing of very long reads, currently averaging roughly 10 kbp but up to 60 kbp [39]. However, this approach has a high error rate, around 13% [113], consisting of predominantly insertion and deletion errors.

2.4.5 Sequencing data repositories

Given the increased use of sequencing in all manner of experimental and research contexts, a number of public repositories have been made available to freely store sequencing data and make it accessible to the wider community. Owing to the

current increase in throughput of modern sequencing machines, these repositories are vital for researchers to store and share their data. In the context of nucleic acid sequences, the predominant data repositories are the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) [122], the NCBI Gene Expression Omnibus (GEO) [56], the European Nucleotide Archive (ENA) [121] and GenBank [19].

2.5 From biology to bioinformatics

Over the last 20 years, computational techniques have been fundamental to biological research [146]. With the ever increasing amount of biological data that we have available, it is impractical for manual analyses and instead, a large number of computational methods and tools have been developed to both aid our understanding and to extract meaningful information from our data. In this section, we discuss the steps taken to process sequencing data obtained from a typical sequencing experiments and introduce some of the software tools that are relevant to this work.

2.5.1 Quality control and processing of sequencing data

In this thesis, we focus exclusively on data obtained from Illumina sequencing machines. Here, we discuss the steps taken for performing quality control and processing of sequencing data.

Typically, the data obtained from an Illumina sequencing experiment will be in FASTQ [41] format with the 5' adaptor removed but the 3' adaptor still present. Removal of the adaptor sequence is a crucial step in the quality control process and ensures that a valid sequence will map back to the reference genome. This is commonly done by matching the first ~ 8 nt of the adaptor sequence to the generated read, trimming the adaptor away and returning the remaining sequence for further

processing. Any sequences where the adapter can not be found are discarded. In the case where high-definition (HD) adaptors [219] are used, after trimming the adapter sequence an additional 4 nucleotides from both the 5' and 3' end are removed. After adapter trimming it is necessary to filter out some additional sequences, such as those that contain unknown bases, usually denoted as 'N', as matching them back to the reference genome may not provide accurate alignment. In addition, reads should be filtered based on their expected lengths, for example with sRNA and degradome sequencing, where the expected length ranges are 21-24nt [25] and 20-21nt [74], respectively.

After performing adapter trimming and length filtering, the next step is to perform further quality control by aligning the reads to the reference genome, with any sequences failing to align being discarded. Typically, due to the read length of sRNA and degradome data, mismatches and gaps are not permitted during the alignment process. Many short read sequencing alignment tools make use of hash table or tree based data structures as a core part of their algorithm to improve computation time [127]. Common software tools for short read alignment include PatMaN [165], Bowtie [117], [116], BWA [125], SOAP2 [132] and Bowtie2 [116].

Pattern Matching in Nucleotide databases (PatMaN) [165] is a short read alignment tool that allows for both gaps and mismatches within the alignment search. The algorithm starts by building a tree of all the query sequences such that every short read is placed into the tree as a path from the root node that ends at a leaf node containing the identifier for the query sequence. It then performs a non-deterministic search on the tree to find matches within the reference sequence. An advantage of PatMaN is that it does not require any pre-processing of the input data.

Bowtie is a fast and memory efficient alignment tool designed for the alignment of short read sequences to large genomes [117]. It indexes the reference sequence

using a system based on the Burrows-Wheeler Transform (BWT) [30] and the full-text minute-space (FM) index [64, 65], which allows it to keep a small memory footprint. The Bowtie algorithm introduced two novel extensions to an existing exact matching algorithm for searching in an FM index, developed by Ferragina and Manzini [64], that allow for the technique to be applied to short read alignment. The first of these extensions was the development of a quality aware backtracking algorithm that allows for mismatches and also favours high quality alignments. The second extension was a double indexing strategy to avoid backtracking unnecessarily. At the time of publication, Bowtie showed a large performance advantage over other tools available, which included MAQ [128] and SOAP [131], when aligning short reads to the human genome with comparable accuracy.

Short oligonucleotide alignment program (SOAP) 2 [132] is a revised version of the original SOAP algorithm [131]. The new algorithm uses BWT indexing of the reference sequence as a way to reduce the memory footprint. Exact matching is performed by constructing a hash table to search for the location of a read within the BWT reference index. In order to find non-exact alignments, a split-read strategy was developed. This works by splitting the read into a number of fragments, based on the number of mismatches allowed, and then counting the number of mismatches contained within the fragments. For example, to allow for one mismatch the read is split into two fragments. The mismatch can then exist in, at most, one of the two fragments. This method was able to give considerable performance increases compared to the original SOAP algorithm with a decreased memory requirement [132]. It was then compared to the other BWT alignment tool at the time, Bowtie, and it was found that they had very similar results when it came to the accuracy of the alignments and the memory required during the alignment process.

Similarly to Bowtie, the Burrows-Wheeler Aligner (BWA) [125] is based on the BWT FM index that enables fast exact matching. BWA supports gapped alignments and the default output alignment format is SAM (Sequence Alignment/Map format)

[126]. BWA was compared against Bowtie, MAQ and SOAP2. At the time of publication, SOAP2 and Bowtie were the other BWT based short read aligners, whilst MAQ indexes reads using a hash table. The results from [125] showed that on simulated data BWA was more accurate than Bowtie and SOAP2, with similar accuracy to MAQ. In terms of memory footprint, BWA and Bowtie were very similar, both outperforming SOAP2. MAQ achieved the lowest memory footprint on the simulated dataset and was identified to be linear with respect to the number of reads to be aligned.

Bowtie 2 [116] extends the FM index based approach of the previous version of Bowtie to allow for gapped alignments. Alignment gaps can occur from sequencing errors or from true insertions and deletions, and the original Bowtie algorithm will fail to align reads that contain gaps resulting in the alignment being missed. Furthermore, the inclusion of gapped alignments within the search greatly increases the size of the search space, substantially slowing the aligners dependent on index based approaches. Bowtie 2 attempts to resolve this issue by dividing the algorithm into two steps. The first step is to extract seeds from the query reads and to perform a gap free seed alignment, that uses the speed and memory advantages of the FM index found in the original Bowtie algorithm, in order to align the seed to the reference. The second step is to extend the seed alignment into a full alignment by performing dynamic programming that benefits from the efficiency of single instruction multiple data parallel processing that is available on modern processors. The benchmarking results from [116] show that Bowtie 2 was able to perform sensitive gapped alignments without any significant computational penalties and it was able to improve on the previous Bowtie algorithm in terms of speed and percentage of reads aligned.

2.5.2 Computational classification of plant sRNAs

In this section, we introduce software tools that can be used to classify sRNA sequences.

2.5.2.1 miRNA prediction

As discussed previously, miRNAs are known regulators of essential biological processes in plants and their biogenesis is key to their discovery. The surge of research interest in miRNAs led to the development of a number of miRNA prediction tools, each attempting to closely model the miRNA biogenesis pathway during prediction. Early tools developed for prediction of plant miRNAs include miRCat [155], miRDeep-P [222], miRDeepFinder [215], miRPlant [5] and more recently, miRCat2 [158]. These tools have been used to successfully predict plant miRNA in several organisms, including grapevine [160], wheat [79] and tomato [159].

Typically, these tools will align the sRNA sequences to the genome, extract longer sequences from the alignment site and attempt to fold them into a hairpin like structure using an RNA folding algorithm such as RNAfold [88], RNALfold [89] or RANDfold [24]. The candidate sRNAs that successfully fold into hairpins are then further processed using rule based models to minimize the reporting of false positives [150, 12]. Standard filtering criteria for miRNA candidates include: discarding those that match to the genome multiple times as genuine miRNAs are unlikely to be derived from highly repetitive regions of the genome [150, 111]; those that do not fit the typical length of a mature miRNA sequence (21-23nt); and where candidates do not have miRNA-like read alignment e.g. candidates filtered on the presence of a miRNA star (miRNA*) sequence resulting from precise processing of the pre-miRNA sequence [38]. Indeed, each tool may also have their own additional criteria that candidates must adhere to, for example the entropy-

based detection of miRNA loci designed to cope with the high sequencing depth of current NGS datasets implemented within miRCat2 [158].

2.5.2.2 Secondary siRNA and ta-siRNA prediction

Unlike miRNAs with their well defined secondary structure characteristic, secondary siRNA precursors, such as TAS genes, require alternative computational prediction strategies [153]. Typically, this type of computational analysis uses sRNA sequencing data and a genomic reference to look for phased alignment patterns. The UEA sRNA Workbench [192], ShortStack [10], pssRNAMiner [48] and shorttran [80] all implement slight variations of the method described by Chen *et al.* [37] for predicting phased sRNA. This approach identifies sRNA alignment clusters and the occurrence of phased patterns within them. In an attempt to better model the DCL processing of dsRNA, the UEA sRNA Workbench and ShortStack both introduced a 2nt shift that adjusts the start position of the sRNA located on the opposite strand during the prediction process.

2.5.2.3 NATs and nat-siRNA prediction

Annotated genomes/transcriptomes have been used in conjunction with sRNA sequencing datasets to predict NATs and NAT-siRNAs in a number of organisms [167, 141, 235, 233]. Recently, NATs were identified from public sequencing data in 69 plant species and a database called PlantNATsDB was constructed [35]. This database includes information regarding sRNAs originating from overlapping and non-overlapping regions of NAT transcript pairs. However, computational tools for the prediction of NATs and NAT-siRNAs are limited in number. Currently, two methods exist for this type of analysis: the NASTI-seq R package [133] and NATpipe [224].

NASTI-seq focuses exclusively on *cis*-NAT identification and uses strand-specific RNA sequencing data as input. NATpipe, which is currently the only tool developed for identifying both *cis*- and *trans*-NATs, uses transcript sequences as input and performs a BLAST [105] search to identify candidate NATs with annealing potential. These are then subject to an RNAplex [195] analysis to examine their secondary structure against a set of criteria [224]. If sRNA data is also provided as input, NATpipe will align these to the NAT sequences looking for phasing patterns, similar to that of ta-siRNA or phasiRNA.

2.5.3 Computational prediction of small RNA targets in plants

In this section, we introduce some of the widely used criteria for the prediction of plant sRNA targets. We then describe briefly some of the software tools and web servers that are available for predicting plant sRNA targets.

The majority of plant sRNA target prediction tools use fixed rule-based targeting criteria inferred from experimental observations. The first set of criteria derived for plant miRNA target prediction was published by Jones *et al.* [100] and then further refined by Allen *et al.* in 2005 [3]. The criteria were inferred from an analysis of 94 experimentally validated miRNA targets in *A. thaliana* and two defining features were identified. First, the position and frequency of miRNA-target mismatches were recorded and second, the predicted stability of the miRNA-target duplex was determined. All miRNA-target duplexes within the set of 94 validated targets contained four or less unpaired bases, four or less G:U pairs, up to one gap, and a total of seven or fewer total unpaired and G:U bases. Figure 2.4A shows the distribution of mismatches and G:U pairs across the miRNA sequence for all miRNA-mRNA duplexes. It was observed that positions 2-13 formed a core segment with relatively few mismatches compared to positions 1 and 14-21. A scoring system was defined where mismatches and gaps were scored as 1 and G:U

pairs were scored as 0.5. The difference compared to the previously published criteria [100] was the inclusion of a x2 score multiplier for mis-paired bases within the core segment. Using this criteria, a score of ≤ 4 captured 91 out of 94 validated targets.

To calculate the stability of the miRNA-mRNA duplex, the minimum free energy (MFE) ratio was calculated. This was determined using a hypothetical duplex consisting of the miRNA sequence and a perfectly complementary target sequence for each miRNA within the set of validated targets. Next, the minimum free energy of each actual miRNA-target duplex was determined using RNAFold [88]. The MFE ratio was then calculated by dividing the MFE of the actual duplex by the MFE of the perfectly complementary duplex for each of the 94 miRNA-mRNA interactions. It was shown that 89 out of 94 validated miRNA-mRNA duplexes in the rule set had an MFE ratio of at least 0.73, as shown in Figure 2.4B.

A later study, published in 2010, by Noah Fahlgren and James Carrington [61], performed a similar analysis but on a larger set of 155 experimentally validated miRNA-mRNA interactions. The postulated scoring system did not differ from that of the previous study [3], however by analysing a larger set of validated targets, it was observed that there exists mis-paired bases at position 10 of the miRNA in some experimentally validated targets, as shown in Figure 2.5.

Below, we briefly introduce and discuss popular computational methods and tools used to predict sRNA targets using sequence-based complementarity criteria.

Plant Small RNA Target Analysis Server (psRNATarget) [49] is a web server that can be used for plant sRNA target prediction and builds upon a previous tool for target prediction called miRU [234]. The motivation behind its development was that the majority of target prediction tools developed at the time of publication were specifically developed for animal sRNAs [130], which are significantly different to plant sRNAs in the target recognition process [15]. The alignment algorithm

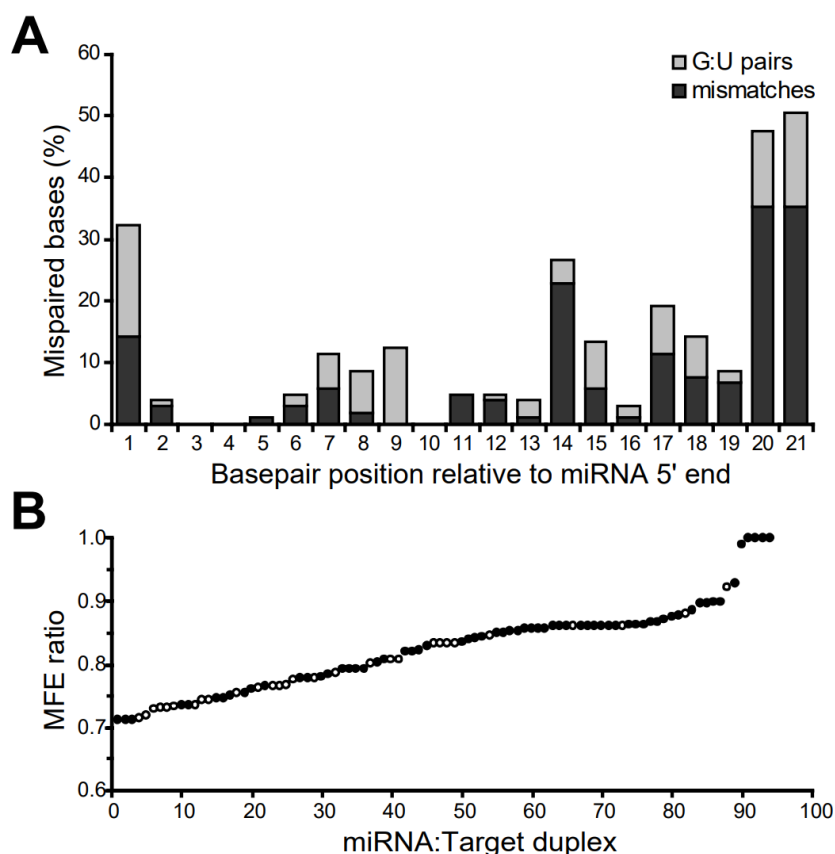


Fig. 2.4 Percent of mismatched and G:U base-pairs at each target position in the rule development set. (B) Minimum Free Energy (MFE) ratio of target-miRNA duplexes from the rule development set. Every miRNA-mRNA duplex in the rule development set had a MFE ratio of at least 0.73. Figure from Allen *et al.* [3] with permission from Elsevier.

used by psRNATarget is an implementation of the Smith-Waterman algorithm [186] called ssearch [162]. psRNATarget gives the option of two sets of default criteria for prediction, V1 [49] and V2 [50], the former uses the same scoring system as miRU [234] complemented with an analysis of the target site accessibility using the RNAup program [78]; the latter is based on the V1 criteria with an increased size of the seed region, from 2–8nt to 2–13nt based on a previous study [9].

TAPIR [23] is another popular tool that follows similar prediction methods used in psRNATarget. It allows a fast search using the FASTA algorithm [163], for which the ubiquitous FASTA format was first designed [135], and for filtering results uses RNAhybrid [112]. Targetfinder [61] and Target-align [216] are other

similar methods that fall into the same category, i.e. using implementations of the Smith-Waterman algorithm to identify possible targets based on sequence complementarity.

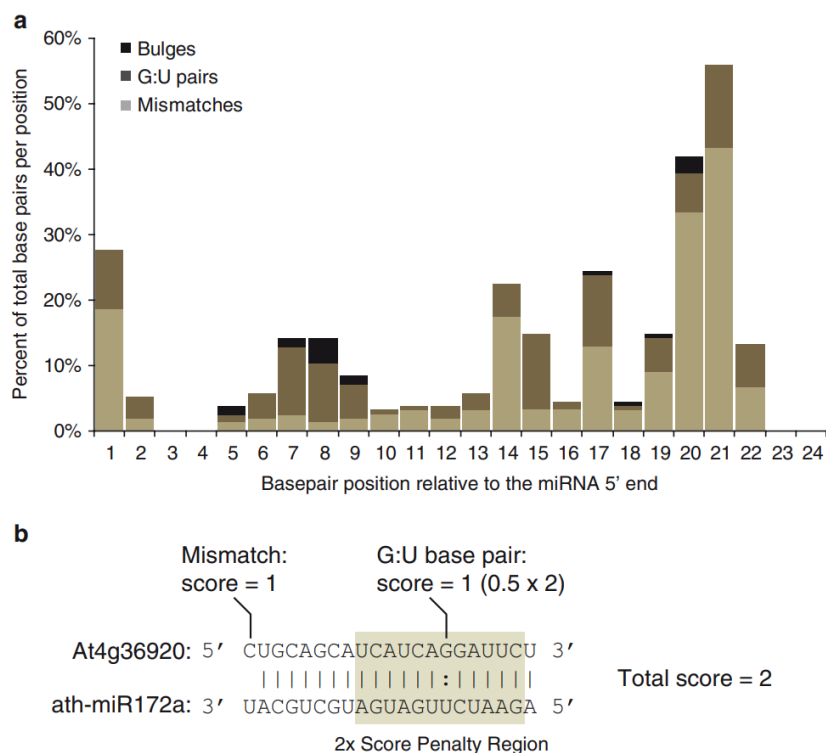


Fig. 2.5 (a) The distribution of mismatches, gaps and G:U base pairs from 155 genuine miRNA-mRNA target duplexes in *A. thaliana*. (b) *A. thaliana* miR172a and its target, At4g36920, illustrating the alignment scoring system used to predict targets. The coloured box highlights positions 2 through 13, relative to the miRNA 5' end, indicating the region where penalty scores are doubled. Figure from Fahlgren and Carrington [61] with permission from Springer Nature.

2.6 Validation of sRNA targets

An important step in understanding the biological function of a sRNA is to identify its targets. As we have seen, most computational tools for plant sRNA target prediction use techniques that search for complementarity between a sRNA sequence and a potential target sequence [229]. These types of prediction use stringent position-based targeting rules that tend to report a high number of predictions and offer

little flexibility. Whilst these results will almost certainly contain genuine targets, many of the predictions may be false positives [229]. Therefore, the predicted targets must undergo further experimental validation before we can be confident that the interaction is genuine.

As discussed above, the typical process of post-transcriptional gene silencing by sRNAs in plants is through mRNA cleavage and subsequent degradation. Experimental analysis of sRNA-directed AGO-mediated mRNA cleavage showed that it generally happens between the 10th and 11th position of the sRNA [139, 59]. The resulting upstream fragment of a cleaved mRNA degrades very quickly, however the downstream fragment is stable *in vivo* owing to the presence of the poly-A tail [51]. One such technique for validating sRNA targets is 5' rapid amplification of cDNA ends (RACE) [70] that works by identifying cleavage fragments for a specific mRNA. However, this method is time consuming as it must be performed for every cleavage site on each mRNA of interest. In addition, it also requires prior knowledge of the flanking regions adjacent to the expected cleavage site. Thus, this technique is best suited for validation of a small number of sRNA targets.

To identify and sequence the degraded mRNA cleavage products on a genome-wide scale, a number of different techniques have been developed, including Parallel analysis of RNA ends (PARE) [74], genome-wide mapping of uncapped and cleaved transcripts (GMUCT) [76] and nanoPARE [179]. However, for simplicity, we shall refer to the output of these techniques as the degradome for the rest of this thesis.

PARE, developed in 2008 [74], is a high-throughput technique for identifying sRNA mediated mRNA cleavage products on a genome-wide scale. The protocol combines a modified 5'RACE [70] with high-throughput deep sequencing to create libraries that contain 3' cleaved mRNA fragments. The cleavage products of sRNA-directed AGO activity differ from other isolated mRNAs as they lack a 5' cap and are therefore ligation competent. This technique selectively clones all uncapped

RNA molecules which have a 3' Poly-A tail, resulting in a snapshot of the mRNA degradation profile. GMUCT [76], also published in 2008, is, in essence, the same technique as PARE in that it uses a modified 5' RACE to identify sRNA-mediated mRNA cleavage sites. One issue with these methods is that they often require large amounts of input RNA, typically only obtainable from bulk samples. Consequently, nanoPARE [179] was developed as an alternative to conventional degradome techniques that can accurately profile mRNA 5' ends on a genome-wide scale using low amounts of total input RNA.

2.7 Discussion

In this chapter, we have provided an introduction to RNA interference, an overview of plant sRNA biogenesis and a brief description of the different classes of plant sRNA that exist for regulating gene expression in plants. We discussed various sequencing techniques and tools used to identify different classes of sRNA from HTS data. We then introduced methods for the prediction of plant sRNA targets and the parameters that they use. Finally, we introduced the degradome, a high-throughput strategy for the validation of sRNA targets in plants. In the next chapter, we expand on this validation method and introduce a new algorithm and software tool for degradome analysis.

Chapter 3

High-throughput sRNA-mRNA target prediction using the degradome

3.1 Summary

In the previous chapter, we introduced high-throughput experimental methods for validating sRNA targets in plants called degradome sequencing. In this chapter, we describe a software tool, called PAREsnip2, that we developed to predict sRNA-mRNA target interactions from degradome sequencing data using configurable targeting rules. Although PAREsnip2 uses a different approach, we give it this name since it is freely available in the UEA sRNA Workbench [192] where its predecessor, PAREsnip [68], is also implemented. We start by introducing how degradome data can be used to support sRNA target prediction, followed by the current software tools available for this type of analysis alongside their their shortcomings. Next, we describe the input, output and the methods we developed for PAREsnip2. We then

evaluate the performance of the tool and discuss results from several degradome analyses.

This chapter is an adapted version of the work published in *Nucleic Acids Research* [198].

3.2 Background

As discussed in Section 2.5.3, an important step in understanding biological function of a sRNA is to identify and validate its targets. Most computational tools for plant sRNA target prediction use techniques that search for complementarity between a sRNA sequence and a potential target-sequence using stringent, position based targeting rules, such as those derived by Allen *et al.* [3]. Whilst these results will almost certainly contain genuine targets, many of the predictions may be false positives [157]. Therefore, the predicted targets must undergo further experimental validation. Degradome sequencing captures the uncapped 5' ends of cleaved mRNA sequences, giving a snapshot of the mRNA degradation profile, and can be used as evidence to identify causal sRNAs, see Figure 3.1.

We now discuss tools developed specifically for this type of analysis.

3.2.1 CleaveLand

CleaveLand [1] was the first tool developed specifically for the analysis of degradome data and it has been used to successfully identify miRNA targets in a number of plant organisms [160, 134, 187, 123, 2]. The tool is implemented using the Perl programming language and the most recent version of the tool, CleaveLand4, was published in 2014 [29]. The first stage of the CleaveLand algorithm is to align the degradome data to the reference transcriptome. This is done using bowtie [117] and if needed, the bowtie indices for the transcript are built with

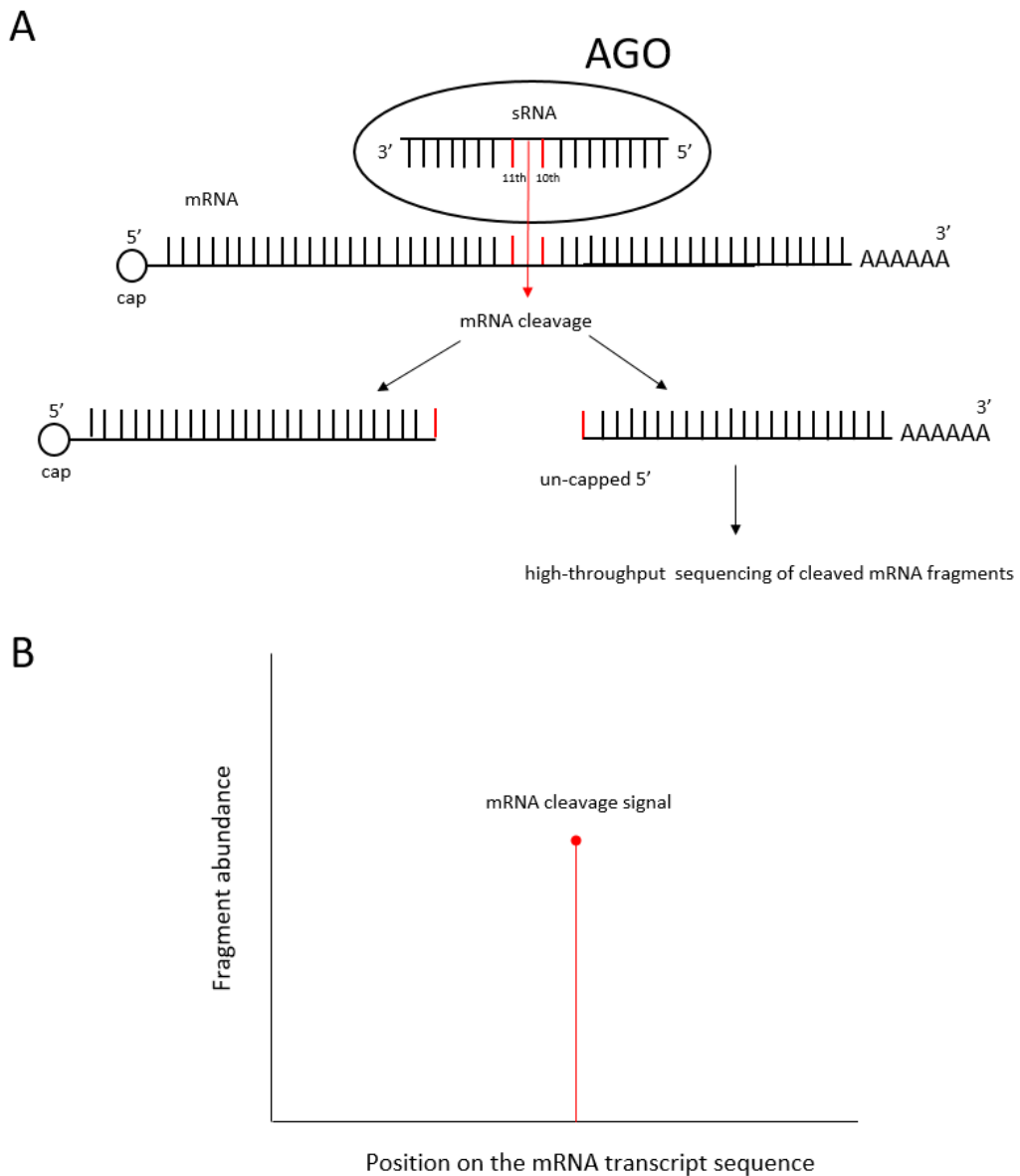


Fig. 3.1 A sRNA is loaded into an Argonaute (AGO) protein and can target the mRNA leading to endonucleolytic cleavage. The resulting mRNA fragments that are un-capped at the 5' end after cleavage can be obtained using high-throughput sequencing methods. (B) Cleavage that has been mediated by an sRNA can be seen as a cleavage signal when they are realigned to the mRNA reference sequence.

bowtie-build using default parameters. The bowtie alignment parameters allow up to 1 mismatch and align to only the forward strand of the transcriptome. In the case of multiple valid transcript alignments for a given degradome read, only one is randomly selected and reported. Alignments are then processed to quantify

the strength of the cleavage signal at each alignment site using a category system. Specifically, the category system is defined as follows:

Category 4: just one read at that position

Category 3: greater than one read, but below or equal to the mean depth of coverage on the transcript

Category 2: greater than one read, above the mean depth, but not the maximum on the transcript

Category 1: greater than one read, equal to the maximum on the transcript, when there is more than one position at maximum value

Category 0: greater than one read and is the maximum value on the transcript, when there is only one maximum value

The next stage of the CleaveLand algorithm is to find potential target sites of a given sRNA. This is done using a Perl script, which is provided with CleaveLand, called GStar that is a wrapper and parser for RNAplex [195], a tool created to search for possible interactions between two RNA sequences using an energy model [201, 239, 240]. GStar employs the search functionality of RNAplex to align sRNA sequences against the reverse complement of a set of transcript sequences. For each input sRNA sequence, the MFE, described in Section 2.5.3, of a perfectly complementary sequence is calculated under default parameters. Next, the same sRNA is analyzed against the entire transcriptome input and potential target sites where the MFE ratio, defined as the target site MFE divided by the perfect MFE, is greater than a given cutoff are kept for further processing. Reported sites are then processed to identify the putative target site, which is the position on the transcript opposite position 10 of the sRNA, and also the alignment score at the target site. This score is based on the position-specific properties, as defined by Allen *et al.* [3] and described in the previous chapter. Specifically, mismatched bases or gaps,

are penalized with a score of 1, G-U wobbles are penalized with a score of 0.5 and penalties are double within positions 2-13 of the sRNA.

After the reporting of potential target sites with GSTAr, the results are combined with those from the degradome read alignment stage to identify any matches opposite the predicted cleavage site (position 10 of the sRNA). If so, analysis progresses for that sRNA-mRNA target interaction and a p -value is calculated. Prior to the development of CleaveLand3, the p -value was calculated using a random shuffle system to indicate how likely the reported duplex occurred by chance. This was done by randomly shuffling the sRNA sequence and counting the number of times it produced a valid alignment duplex with all other peaks of the same category. This was repeated over a number of reshuffles, e.g. 100, and the p -value was reported as the proportion of the shuffles that successfully aligned. Since the development of CleaveLand3, the p -value is calculated using a cumulative binomial distribution function to determine how likely a given degradome hit is to occur by chance, given in equation 3.1, where n represents the number of predicted targets between 0 and a threshold, i.e. the number of predicted targets with score ≥ 0 and ≤ 5 , for a given sRNA. The probability, p , of a given position on a transcript having a cleavage peak of a certain category is $\frac{c}{L-(t*l)}$, where c is the number of peaks with a given category, L is the sum of all transcript lengths, t is the number of transcripts and l is the average length of the degradome fragments. Any sRNA-mRNA interactions that pass the p -value filtered are reported to the user.

$$P(X > 0) = 1 - \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{x-n} \quad \text{where} \quad p = \frac{c}{L-(t*l)}. \quad (3.1)$$

3.2.2 SeqTar

Sequencing-based sRNA target prediction (SeqTar) [236] is a method developed to identify miRNA targets that combines degradome data with a modified Smith-Waterman [186] alignment algorithm. The motivation for the development of SeqTar was that the only publicly available method for degradome analysis, CleaveLand, employed the stringent Allen *et al.* [3] targeting criteria that was known to miss genuine miRNA-mRNA targets. As described previously, this scoring scheme does not allow mismatches or G:U pairs at positions 10 or 11 and suggests discarding targets with an alignment score greater than 4. However, it has been shown that alignments that do not fit this criteria can also guide the cleavage of their target transcripts. For example, ath-miR159a can induce cleavage of AT5G18100 despite having a score of 6.5, corresponding to 4.5 mismatches [74]; ath-miR390 successfully guides the cleavage of TAS3b transcript although the complementary site has a score of 7, corresponding to 6.5 mismatches [11]; miR167 can lead to cleavage of Os06g03830 despite having a mismatch at position 11 [134]; and ath-miR173 successfully induces cleavage of AT1G50055 even though position 10 of the miRNA at the target site is a mismatch [3]. Instead of using a rule-based model and to reduce false positive predictions, SeqTar implemented two statistical methods to control the number of alignment predictions reported by the Smith-Waterman alignment. SeqTar was evaluated on a set of experimentally validated miRNA targets and the results compared to the version of CleaveLand available at the time. The results showed that SeqTar outperformed CleaveLand by identifying 43% and 42% more interactions in *Arabidopsis thaliana* and *Oryza sativa*, respectively, suggesting that SeqTar was a more effective method for degradome-supported miRNA target prediction.

3.2.3 PAREsnip

PAREsnip [68] is the first cross-platform bioinformatics tool for the analysis of degradome data. It was developed to overcome the speed limitations with CleaveLand and to enable high-throughput analysis of recent sequencing datasets within a reasonable time frame. At the time, due to the algorithms implemented within CleaveLand, it was impractical to analyse all possible sRNA targets and instead was used to find cleaved targets of a small number of sRNAs, such as known or candidate miRNAs. For performing degradome analysis on a given organism, the input for PAREsnip are a reference transcriptome, a degradome dataset, a sRNA dataset, and, optionally, a reference genome. The input sRNAs are then filtered to remove low abundance or low complexity reads and if a genome is provided, any reads that do not align are discarded. Degradome reads are then aligned to the reference transcriptome and sorted into a category system as defined in CleaveLand version 2, which are:

Category 4: just one read at that position

Category 3: greater than one read at the position and the abundance at that position is less or equal to the median value for that transcript.

Category 2: greater than one read at the position and the abundance at that position is greater or equal to the median value for that transcript.

Category 1: greater than one read, equal to the maximum on the transcript, when there is more than one position at maximum value

Category 0: greater than one read and is the maximum value on the transcript, when there is only one maximum value

In addition to the raw abundance, PAREsnip included the option to determine the category for a given transcript signal using weighted fragment abundance. Weighted abundance is calculated by dividing the abundance of a degradome read

by the number of positions across all transcripts to which the sequence aligns and this is the default configuration for PAREsnip.

Given that a DNA/RNA sequence can be made up of four nucleotides (A, C, G and T/U), a core part of the PAREsnip algorithm is the construction of a 4-way search tree using the four letter alphabet. The tree is then used to encode each sRNA into a unique path within the search tree. This means that similar sequences will lie on the same path until the similarity ends, reducing the possible search space. Once the tree has been constructed and the degradome reads aligned, the search for potential targets can be performed. The pairs of nodes at level 10 and 11 of the tree are collected and put into one of 16 bins, representing the possible 2nt sequences. Searches for sRNAs that could cause cleavage at a given degradome peak start by identifying the bin corresponding to nucleotides 10 and 11 of the candidate sequence and the tree is then traversed from nucleotide 10 towards the root. As it does this, it performs a nucleotide comparison between the sRNA and the target sequence checking to see if any of the Allen *et al.* targeting rules [3] have been broken, and updating an alignment score if necessary. If no rules are broken, it then returns to position 10 and traverses towards the 3' end of the sRNA, again checking at each positions if any of the rules have been broken, and updating the score. Once it reaches a terminator node within the tree and if no rules have been broken, it records it as a potential target and is subject to further processing before being reported to the user. PAREsnip uses a sRNA shuffle system, similar to the one implemented in earlier versions of CleaveLand, to calculate the p -value of the potential target duplex. If the p -value for the predicted sRNA target is within a given threshold, ≤ 0.05 by default, it is reported to the user.

3.2.4 sPARTA

Small RNA-PARE Target Analyzer (sPARTA) [101] is a command line degradome analysis tool that is capable of predicting sRNA targets on a whole genome scale. The motivation behind the development of sPARTA was that the tools available at the time (CleaveLand, SeqTar and PAREsnip) assumed that there exists a positive correlation between complementarity in the canonical seed region and probability of actual cleavage and by using SeqTar it is not feasible to perform the analysis on a whole genome scale. Furthermore, these tools require a set of reference sequences as input, which typically comprise the annotated portion of the genome and new genomes can be poorly annotated in their initial release. Therefore, using the annotated portion of the genome alone could lead to potential targets within intergenic regions (IGRs) being missed [101]. Alongside this, recent studies have found that even in well annotated genomes, there are still targets being found in the IGRs [7, 98, 188]. Currently, other available tools will miss these targets without an alternative reference sequence input being compiled to include IGRs, however the creation of such a reference would require time and a level of bioinformatics expertise that some users may not have. In an attempt to solve this, sPARTA allows the user to input a genome and generic feature format 3 (GFF3) file to automatically extract reference sequences, allowing the user to search for targets within the intergenic regions without creating their own reference sequence.

The sPARTA algorithm has four stages, the first is the fragmentation of features from the input files, this is done using either the transcriptome or a GFF file and corresponding genome. The next step is to map the degradome reads to the feature set, this is done using Bowtie 2, where an FM index [64] is created for each component of the feature set. The third step of the sPARTA algorithm is the prediction of sRNA targets within the feature set using a built in target prediction module called miRferno. This module has two prediction modes allowing the user

to optimise for time versus sensitivity. The heuristic mode is designed to be fast but less sensitive and works by extracting multiple seeds of length 6nt in 4nt intervals from the sRNA sequence. These seeds are then aligned to the FM indexes from the feature set with a maximum allowed mismatch of 1. If an alignment is found, it is extended to complete the alignment of the small RNA. The exhaustive mode is developed for improved sensitivity and extracts a smaller seed of 4nt with a 3nt interval, improving the efficiency of finding targets. During target prediction, miRferno offers two scoring systems: standard and seed free. The standard system is based on previously experimentally validated targets and the complementarity rules based on the seed region [61]. The seed free scoring system was added as studies [236, 29] have shown that there are miRNA target interactions that differ from the canonical targeting rules. In this system, mismatches in positions 10 and 11 are allowed.

The final stage of the sPARTA algorithm is to combine the predicted targets and the degradome reads with the aim of validating potential targets through cleavage evidence in the degradome. When aligning the degradome reads to the transcript sequences, a similar category system to PAREsnip and CleaveLand is also implemented within sPARTA. A *p*-value is then calculated using a modified version of the method implemented within CleaveLand version 3 and 4. The difference is how the number of trials is chosen, in CleaveLand the number of trials is the number of predicted targets with a score between 0 and a some threshold. In sPARTA, the number of trials is the number of predicted targets within a score bracket, i.e. the number of predicted targets with score ≥ 5 and ≤ 6 . This change is useful for cases where miRNA-target interactions have weak complementarity or for when a single miRNA cleaves a large number of targets. Performance benchmarking showed a considerable improvement compared to CleaveLand in terms of computation time and it was also able to capture a larger number of miRNA targets, however no benchmarking comparing sPARTA to PAREsnip was performed.

3.2.5 Web-based tools

Alongside the previously described stand-alone downloadable tools, two web-based services have been developed for performing degradome analysis.

StarScan (sRNA target Scan) [138], was the first publicly available web-server for identifying sRNA targets from degradome sequencing data. On release, StarScan contained one hundred degradome libraries from 20 species with reference genome sequences and gene annotations obtained from the Ensembl Plants Database [106]. StarScan takes as input a set of sRNA sequences in FASTA format, the user then selects the species and degradome library to be used during analysis. StarScan implements a category and *p*-value system similar to that of CleaveLand4 and sPARTA. Predicted targets that pass the results filtering stages are reported to the user and include the cleavage position, sRNA and target gene names, transcript ID, target gene types (e.g. protein coding or ncRNA), cleavage site (position 9, 10 or 11 of the sRNA), penalty scores and the category of the degradation signal. In addition to the data obtained from plants, StarScan also provided the ability to perform degradome analysis on animal data, for example using the data obtained from a previous human study [183].

Web-based pipeline of RNA degradome (webPORD) is the most recent web-server for the analysis of degradome data [212]. It works in a similar way to StarScan but is currently only populated with data obtained from *Homo sapiens*, *Mus musculus* (mouse), *O. sativa* and *A. thaliana*.

3.2.6 Issues with current methods

Recent advances in high throughput sequencing technologies has resulted in larger, more complex genomes being sequenced such as *Pinus taeda* [237] or *Triticum aestivum* [43], both being many times larger than that of popular model organisms.

Moreover, not only are larger genomes being sequenced, but degradome and sequencing datasets in general are growing ever larger in size and read count, with a typical sequencing experiment now containing millions of distinct reads in a single sample. In addition, the need for multiple samples and replicates is becoming the de-facto standard for biological experiments, further adding to this sequence-data deluge.

All of the tools for degradome analysis mentioned above are unable to process the volume of data currently being produced without imposing considerable time and resource constraints. In addition, the accuracy of these tools is primarily determined by the targeting rules that they apply and each tool uses a different set of fixed rules, which reduces their flexibility. Indeed, the rules currently implemented by the tools are inferred from the analysis of experimentally validated miRNA targets in *A. thaliana*. This was first performed on 94 validated miRNA-target duplexes by Allen *et al.* [3], influenced by an earlier study [100], and then, through a similar approach, on a larger set of 155 validated target duplexes by Fahlgren and Carrington [61]. As our understanding of miRNA targeting improves, these rules may change, and so current tools risk becoming obsolete.

3.3 Methods

We now introduce a novel degradome analysis method and software tool, called PAREsnip2, that is scalable with current sequencing datasets. The PAREsnip2 algorithm is split into three main stages. The first stage is the input of the sequencing data and targeting rules, the second is the pre-processing steps (developed to improve the speed and efficiency of an analysis), and the third is the prediction of sRNA targets. A visual overview of the steps involved in performing an analysis on the input data is shown in Figure 3.2A. We now explain each stage of the algorithm in more detail.

3.3.1 Data input

To perform an analysis using PAREsnip2 for a specific organism, the user must input the following data:

- a reference file (transcriptome) in either FASTA format or Generic Feature Format version 3 (GFF3) with corresponding genome;
- a genome file (optional unless using GFF3 as reference);
- one or more sRNA library replicates;
- one or more degradome library replicates

A reference file and at least one sRNA and degradome library are required to perform an analysis. If the user chooses to use a GFF3 file as a reference then a corresponding genome must also be provided. When extracting the gene sequences from the genome using a GFF3, the user has the option to include or exclude untranslated regions (UTRs).

The sRNA and degradome libraries must be in redundant FASTA format with the adapters trimmed.

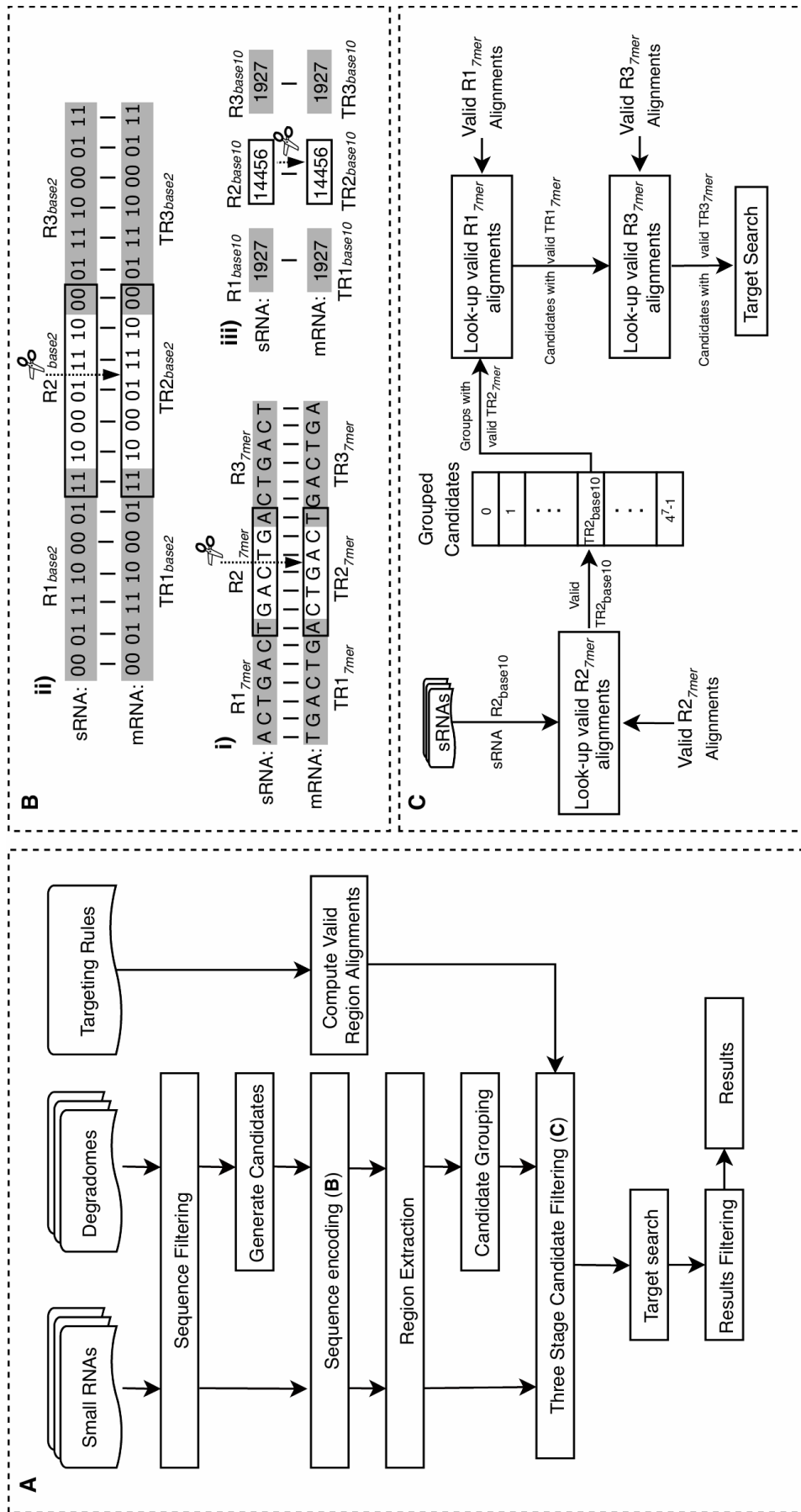


Fig. 3.2 An overview of different stages of the PAREsnip2 algorithm. (A) Shows the inputs and processing steps performed to predict sRNA targets evidenced through degradome sequencing. (B) Shows the process of encoding sequence data into a number system. (C) Visual representation of the three-stage candidate filtering process. Regions are labelled R and target regions are labelled TR.

3.3.2 Sequence filtering

Several optional filtering techniques can be applied to the input data to remove low quality reads, sequencing errors or sample contamination. First, any sequence containing ambiguous bases are discarded, as they cannot be accurately aligned. Second, a low complexity sequence filter is applied based on the sequence single, di- or tri-nucleotide composition. This works by discarding any sequences that contain more than 75%, 37.5% and 25% of a single, di- or tri-nucleotide composition, respectively. Third, we provide the functionality to filter sequences using conservation over multiple samples where sequences will only be considered if they are present within each sample. Finally, when a genome is provided, sRNA sequences can be aligned to the genome using PatMan [165], with any sequences that do not align being discarded.

3.3.3 Binary encoding of sequence input

A core component of the PAREsnp2 algorithm is the encoding of sequence data into a number system, shown in Figure 3.2B. Given that a sequence is composed of four nucleotide bases (A, C, G, T/U), it is possible to represent each nucleotide using two bits of computer memory, see Table 3.1, known as the base 2, or binary representation, of a nucleotide. We represent a whole sequence as a single decimal number by concatenating the binary representations of each nucleotide and converting the resulting, longer binary representation into decimal. We use this encoding technique to reduce the memory footprint of storing sequence data in memory and to speed up analysis. Furthermore, sRNA and mRNA sequences have an inverse encoding (Table 3.1), such that if a sRNA and mRNA sequence are represented by the same number then they will be perfectly complementary.

2-bit Representation	Small RNA	Messenger RNA
0 0	A	U or T
0 1	C	G
1 0	G	C
1 1	U or T	A

Table 3.1 The 2-bit encoding of nucleotides

3.3.4 Target candidate generation

To search for potential sRNA targets, we first generate a set of potential target-sequence candidates from the input data. The alignment of the degradome fragments to the reference gene sequences can inform us of potential sRNA cleavage events, with higher abundance fragments at a specific position more likely to be true cleavage signals. We developed a novel technique for exact match sequence alignment that uses the sequence encoding described above. First, degradome sequences are read from file, encoded as a number, and stored into a list. Once all the reads have been encoded and stored, the list is sorted into ascending order. Next, we split the reference sequences into subsequences using a sliding window and encode each of these into a decimal number. The size of the sliding window and the number of extracted subsequences are dependent on the accepted size range of the degradome reads. We then search the sorted list of encoded degradome fragments for the encoded reference subsequence using a binary search. If the number representing an encoded subsequence is found, an exact match has been identified at that position and is recorded. Once each reference sequence has been searched, the aligned degradome fragments are further processed to generate the set of target-sequence candidates. From the alignment position, we take 16nt towards both the 5' and 3' ends, resulting in a 32nt mRNA target-sequence candidate.

The newly generated target-sequence candidates are then sorted into one of five categories based on those previously defined in CleaveLand V4 [1] with a minor

modification. In our modification, we do not consider those fragments with an abundance of 1 during the average coverage calculation. This helps us to distinguish true lower abundance peaks from background degradation upon the transcript. An overview of the category system is provided below:

- Category-0 peaks are those that have greater than one read and are the maximum on the transcript when there is only one maximum;
- Category-1 peaks are those that have greater than one read and are the maximum on the transcript, but there is more than one maximum;
- Category-2 peaks are those that have greater than 1 read and are above the average fragment abundance on the transcript;
- Category-3 peaks are those that have greater than 1 read and are below or equal to the average fragment abundance on the transcript;
- Category-4 peaks are those that have just one read at that position on the transcript

3.3.5 Region extraction and candidate grouping

Three regions of length 7nt (*7mer*) are extracted from both the input sRNA sequences and the generated target-sequence candidates. These are named region R1, R2 and R3 for the sRNA and target region TR1, TR2 and TR3 for the target-sequence. The position of the extracted target-sequence regions are based on a potential cleavage position i.e. where the sRNA would align if there were no gaps or bulges within the duplex (Figure 3.2Bi). The extracted region sequences are then encoded into their decimal number format and stored for later use. Finally, the generated target-sequence candidates are grouped together using the decimal representation of their TR2 sequence such that any candidates sharing the same *7mer* at their TR2 will be grouped together.

3.3.6 Predefined and user configurable targeting rules

Since the discovery of miRNAs and their regulatory role in plants, there has been much discussion on the rules that should be used when predicting plant miRNA targets [100, 3, 101, 61, 29, 28, 161, 137, 107]. To the best of our knowledge, there are two generally accepted and widely used targeting rules for plant miRNAs. These rules are implemented within a position dependent scoring system based on the number of mismatches, G:U wobbles and target-bulged bases within the duplex. The first of these were inferred by Allen *et al.* in 2005 [3] and the second, through a similar approach with a larger set of validated targets, by Fahlgren and Carrington in 2010 [61], as described in the previous chapter. During a PAREsnip2 analysis, the user can choose between two sets of default targeting rules, either the Allen rules or the Fahlgren and Carrington rules. The difference between them is that the Fahlgren and Carrington rules permit a mismatch or G:U wobble at position 10 and 11 of the sRNA, based on our interpretation of their results [61]. However, these rules are based on a small set of experimentally validated miRNA targets and as more miRNA targets are experimentally validated, our understanding of these targeting rules may change. To address this, we offer the ability to search for potential targets based on a user configurable rule set. The rules that can be configured by the user and used during the search for potential targets are shown in Table 3.2.

Criteria	Allen <i>et al.</i>	Fahlgren & Carrington
Maximum score	4	4
Maximum adjacent mismatches	2	2
Maximum G/U Wobble Pairs	4	4
Maximum Mismatches	4	4
Mismatch Score	4	4
G/U Wobble Score	0.5	0.5
Gap Score	1	1
Permissible Mismatch Positions	all	all
Non-permissible Mismatch Positions	10, 11	none
Core Region Start Position	2	2
Core Region End Position	12	12
Maximum Mismatches Core Region	2	2
Maximum Adjacent Mismatches Core Region	1	1
Allow Mismatch Position 10	false	true
Position 10 Mismatch Score	1	1
Allow Mismatch Position 11	false	true
Position 11 Mismatch Score	1	1
Core Region Multiplier	2	2
Max Gaps Allowed	1	1
G/U Wobble Counts as Mismatch	false	false

Table 3.2 Features within a sRNA–mRNA alignment which are used during the duplex alignment process and their default values but can also be configured by the user.

3.3.7 Computing valid region alignment matrices

As discussed previously, we can represent biological sequences using a decimal number system. *7mers* that are comprised of a four-letter alphabet (A, C, G and T/U), where each nucleotide is encoded using 2 bits of computer memory, are represented by a decimal number between 0 and 16383. For each of the three regions, we create a 16384×16384 matrix that represents all possible combinations of alignments between *7mers*. Within these matrices, row numbers represent encoded sRNA *7mers* and column numbers represent encoded mRNA *7mers*. The

matrices are then populated by attempting to align the decoded sRNA and mRNA *7mers* using the user's chosen set of targeting rules. If a valid alignment is found within the matrix, we set that position to true otherwise it is set to false. This is repeated for every possible combination of alignments between *7mers* for each of the three regions.

3.3.8 Three-stage candidate filtering

We developed a three-stage candidate filtering technique to reduce the search space and therefore the computation time required to perform an analysis. When searching for degradome peaks potentially resultant of sRNA mediated endonucleolytic cleavage, we use the valid region alignment matrices to discard candidates that do not fit the chosen targeting rules (shown in Figure 3.2C). In the first stage of this technique, we consider only those target-sequence candidates where their TR2 *7mer* can successfully align to the R2 *7mer* of the sRNA. This is done by looking at the encoded sRNA R2 *7mer* row in the R2 valid region alignment table and taking all target-sequence candidates grouped on the columns set to true on that row.

In the second and third stages, we discard any target-sequence candidates where their TR1 or TR3 regions do not successfully align to the R1 or R3 regions of the sRNA. This is performed by first looking at the cell ($R1_{base10}$, $TR1_{base10}$) in the R1 valid region alignment matrix to see if it is set to true and if so, we do the same for the R3 and TR3 region, discarding any candidates if the cell values are set to false.

3.3.9 Target search and results filtering

Any target-sequence candidate that passes all stages of the three-stage candidate filtering process is aligned to the sRNA sequence using our duplex alignment algorithm employing the chosen targeting rules. When attempting to align a sRNA

to a potential target-sequence candidate, the search process starts at the cleavage site and then traverses towards the 5' end of the sRNA and at each position performs a nucleotide comparison between the two sequences. If the alignment towards the 5' end is successful, it then performs the same process towards the 3' end. If there is a mismatch, it will attempt to insert a gap and continue the alignment. If at any point one of the user's selected rules are broken then the alignment is discarded. This process will find all valid alignments based on the chosen targeting rules and the best possible alignment based on the users chosen criteria is selected. We first attempt to select the alignment that has the lowest alignment score and if there are multiple valid alignments with this score, the alignment with the fewest gaps is reported. If there are multiple alignments with the same number of gaps, the alignment with the fewest number of mismatches and G:U wobble pairs is reported.

Once a potential target has been identified, two optional filtering processes can be performed to improve the confidence of each prediction. The first is the application of a MFE ratio filter and the second is a p -value filter. The MFE is calculated using RNAPlex [195], which was shown to score favourably for sensitivity and precision when compared to other similar methods in a recent benchmarking of performance [202]. The MFE ratio is calculated by dividing the predicted target duplex MFE by the MFE of a perfectly complementary target site. Any predicted target site that has a MFE ratio less than a given cut-off is discarded. The default cut-off ratio is 0.7, as suggested by Allen *et al.* [3], but can be configured by the user. The second optional filtering process uses the binomial distribution p -value system implemented within CleaveLand V4 [1] but with the modification that the probability is calculated on a transcript by transcript basis.

3.3.10 Implementation and output

The algorithm has been implemented using the Java programming language and a user-friendly, cross-platform software package has been incorporated into the UEA sRNA Workbench (26). Analysis can be performed through the graphical user interface (GUI) or through the command-line interface (CLI) allowing PAREsnip2 to be used in other bioinformatics pipelines or workflows.

The results of PAREsnip2 are provided in comma-separated value (CSV) format, allowing them to be viewed in any CSV file viewer. They include information about the transcript peak such as cleavage position, abundance and weighted-abundance at the cleavage site, and the category of the peak on the transcript. A visual representation of the sRNA–mRNA duplex is displayed along with its alignment score. The sequence read abundance for small RNA and degradome data are provided in both raw and normalized values so that sequencing libraries can be compared. It is also possible to produce target plots from PAREsnip2 results using the T-plot tool contained within the UEA Small RNA Workbench [192].

3.3.11 Degradome library construction

Three *A. thaliana* degradome replicates were constructed using wild type Columbia (Col-0) plants grown at 22° with 16 hours light and tissue was harvested when plants were at growth stage 5, as defined by Boyes *et al.* [27]. For each replica, RNA was isolated from a pool of all leaves taken from nine plants with TRI reagent following manufacturer's instructions. This RNA was then used to construct degradome libraries following Zhai *et al.* protocol [230], with the only difference being that SuperScript II reverse transcriptase was used instead of Superscript III.

3.3.12 Sequence datasets

The sequencing datasets analysed in this chapter are described in Appendix A Table 1. Briefly, the transcriptome used in all of our analyses on *A. thaliana* was obtained from TAIR10 [115]. The computational performance benchmarking was carried out using a publicly available *A. thaliana* mature leaf degradome dataset [197]. Additionally, we simulated 9 sRNA datasets of increasing size to use as input data. These sRNAs were generated by first aligning the D1 reads to the reference and then extracting 19–24nt sequences centred on cleavage positions. Transcripts, cleavage positions and sRNA sequence lengths were selected at random.

The prediction performance benchmarking was performed using the three *A. thaliana* degradome replicates, which we described above, and *A. thaliana* mature miRNA sequences obtained from miRBase (v21) [77].

To perform genome-wide degradome analyses on *A. thaliana*, we obtained the corresponding sRNA libraries, which were previously published by our lab [158], for each of the *A. thaliana* degradome replicates. Additionally, we performed a genome-wide analysis on *T. aestivum* using publicly available sRNA and degradome datasets [196] and the *T. aestivum* transcriptome (cDNA) obtained from Ensembl Genomes [106].

3.4 Results

3.4.1 Sequencing data

We processed the raw data using tools provided within the UEA sRNA Workbench [192]. The adapter trimming tool was used to trim the adaptor sequences in each of the three degradome replicates. Next, using the Filter tool, we discarded sequences that contained any ambiguous bases and aligned the remaining sequences to the

genome (TAIR10) with no mismatches allowed. When mapping to the genome, 81%, 82% and 82% of trimmed reads successfully aligned in replicates D2A, D2B and D2C, respectively. Table 3.3 gives a summary of the statistics for the three replicates and Appendix A Figure 1 shows the read length distributions and as you would expect, the reads are primarily 20 and 21nt in length.

Replicate	Raw	Raw (NR)	Trimmed (NR)	Genome Aligned (NR)
D2A	45 581 525	15 267 190	11 114 679	9 009 977
D2B	34 915 085	13 385 729	10 103 828	8 316 470
D2C	26 067 832	10 199 905	7 715 372	6 337 667

Table 3.3 Summary statistics (number of reads) from the sequencing of three *A. thaliana* degradome replicates (NR = non-redundant).

3.4.2 Computational performance benchmarking

To measure the computational performance of the PAREsnip2 algorithm, i.e. the time and memory required to perform an analysis, we carried out computational benchmarking and compared our results to those of other publicly available methods. This benchmarking was performed on a desktop computer running Ubuntu 16.04 equipped with a 3.40GHz Intel Core i7-6800K six core CPU and 128GB RAM. Each tool was run using the authors default suggested parameters and for the fairest comparison, we included all filtering and pre-processing options available in PAREsnip2. Additionally, we set the number of threads to be used by the tools during the analyses to 12, except for CleaveLand as it is not an option.

For this benchmarking, we used the D1 dataset, the simulated sets of sRNA sequences and the TAIR10 cDNA transcriptome. Whilst the tools were performing the analysis on the simulated data, we monitored their peak memory usage using an in-house script and recorded the time they took to complete the analysis. The results of these analyses for both time and peak memory usage is shown in Table

3.4. Additionally, if the tool did not complete the analysis within 10 days, we recorded it as did not finish (DNF).

The results show that the newly developed PAREsnip2 algorithm substantially outperforms all the currently available tools on the simulated datasets. The largest dataset for which any of the existing tools could process in under 10 days contained 250 000 sequences. When performing analysis on this dataset, PAREsnip2 showed 319× improvement in computation time. Additionally, the results suggest that the computation time of PAREsnip2 grows linearly with the number of input sequences, taking just 1 h and 44 min to process the largest of the simulated datasets (1 000 000 sRNAs).

# Seqs	CleaveLand4	GB	PAREsnip	GB	sPARTA	GB	PAREsnip2	GB
1	19m 23s	1	9m 30s	58	12m 48s	25	5m 38s	5
10	27m 32s	1	9m 50s	58	12m 53s	25	5m 36s	5
100	1h 52m	1	12m 35s	58	13m 55s	25	5m 44s	5
1,000	15h 8m	1	44m 51s	58	1h 11m	26	6m 15s	6
10,000	6d 6h 48m	8	6h 25m	64	4d 6h 59m	37	6m 32s	6
100,000	DNF	-	2d 15h 16m	66	DNF	-	15m 1s	6
250,000	-	-	6d 10h 49m	68	-	-	29m 6s	7
500,000	-	-	DNF	-	-	-	53m 11s	8
1,000,000	-	-	-	-	-	-	1h 44m	8

Table 3.4 Benchmarking results for both time and memory usage in Gigabytes (GB) from running each tool using the generated small RNA datasets. If the entry is DNF it means that the tool did not complete the analysis within the 10 day cut-off. A ‘-’ means that we did not attempt to run the tool. Benchmarking results show that PAREsnip2 was able to complete analysis considerably faster than all other tools with low resource requirements.

3.4.3 Prediction performance benchmarking

To evaluate the prediction performance of each tool we collected a set of experimentally validated *A. thaliana* interactions by combining those previously published in

the literature [68, 191, 53] and those contained within miRTarBase [40] with any duplicates being removed. In total, we collected 616 validated interactions comprising 135 miRNAs. Out of these 135 miRNAs, 90 of them had unique sequences and were involved in 387 distinct miRNA–mRNA interactions. See Appendix A Table 2 for the complete list of curated validated targets.

Any of the validated interactions with a category-4 signal at the cleavage position on the transcript within the D2 degradome datasets were excluded from the benchmarking. These signals were excluded because it is difficult to distinguish between true miRNA cleavage products and random degradation with such low abundance. To identify the cleavage positions, we obtained the miRNA sequence from miRBase and the transcript sequence for each of the validated miRNA targets and performed the alignment between them using loose targeting rules, allowing a maximum of seven mismatches. In the case that multiple alignments were found between the miRNA and its target, we retained the alignment(s) with the best alignment score and MFE ratio. The position on the transcript opposite position 10 of the miRNA was recorded as the miRNA cleavage site. The category of the signal on the transcript was determined by aligning the D2 degradome datasets to the transcript and recording the abundance at the cleavage position. Out of a possible 387, we included 243, 239 and 224 validated interactions comprising 61, 60 and 58 miRNA sequences for datasets D2A, D2B and D2C, respectively.

We performed an analysis with each tool using the miRNA sequences contained within the validated set of miRNA–mRNA interactions, the *A. thaliana* transcriptome and the three D2 degradome datasets described previously. Each tool was run using the default parameters recommended by the authors but with category-4 interactions discarded as they were not considered previously. When benchmarking PAREsnip2, we performed the analysis using both sets of default targeting rules and the MFE filter with cut-off score of 0.7. The results produced by each tool when analyzing the three datasets were then compared against the set of validated targets

and are shown in Table 3.5. The results show that both sets of default targeting rules implemented within PAREsnip2 captured more of the experimentally validated interactions than the currently available tools. The differences between the results produced by the tools are likely due to variations in the implemented targeting rules and the filtering techniques applied. Additionally, the lower number of interactions reported by CleaveLand may be due to the way it handles degradome reads that map to multiple transcripts. If a degradome read aligns to more than one transcript, only one is randomly selected and reported by CleaveLand.

Tool Name	Replicate D2A			Replicate D2B			Replicate D2C		
	V	NV	%PV	V	NV	%PV	V	NV	%PV
sPARTA	171	120	70%	169	121	70%	162	127	72%
PAREsnip	177	48	73%	179	50	75%	167	57	75%
CleaveLand4	88	20	36%	95	26	40%	87	25	39%
PAREsnip2 Allen <i>et al.</i>	193	41	79%	191	39	80%	181	33	80%
PAREsnip2 Fahlgren & Carrington	219	48	90%	219	43	91%	205	37	91%

Table 3.5 The results from the accuracy performance benchmarking of each tool over the three biological replicates. V = validated targets, NV = non-validated and %PV = percentage of possible validated targets that could be found. Results show PAREsnip2 captures a larger number of the experimentally validated *A. thaliana* targets compared to other publicly available tools using both sets of default targeting criteria.

3.4.4 Evaluation of the optional filtering methods

To evaluate the success of the filtering techniques implemented within PAREsnip2, we repeated the prediction performance benchmarking on the D2B degradome dataset using the 60 miRNA sequences known to have existing targets, the default

Fahlgren and Carrington targeting rules, and increasing filtering cut-off values. The results of the MFE analysis are shown in Figure 3.3 and the results of the *p*-value analysis are shown in Figure 3.4.

When evaluating the MFE filter, we start with a cut-off score of 0.45, as this captures all possible interactions, and with increments of 0.05 thereafter, we record the number of validated and non-validated targets being captured. Using the initial value, we captured a total of 342 miRNA–mRNA interactions from 60 miRNAs with 223 being part of the validated set and 119 were non-validated. At the other end of the scale, by using a filter cut-off value of 1 we captured just 5 interactions, all of which are part of the validated set. The default value of the MFE ratio filter (0.70) for PAREsnip2 captures a total of 262 interactions and of these the filtering process kept 219 (98%) from the possible 223 validated interactions.

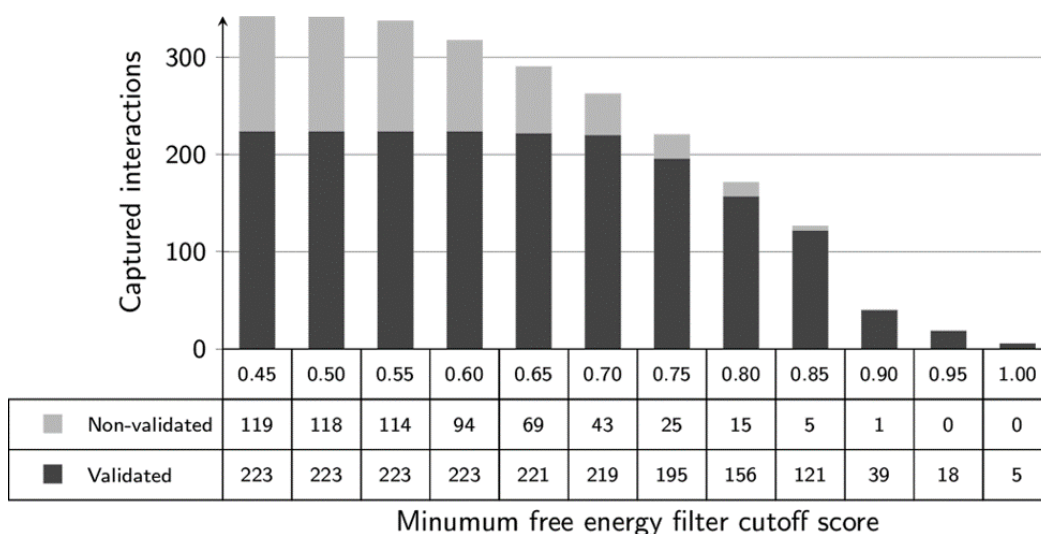


Fig. 3.3 The number of interactions reported when using MFE as a filter. As the MFE filter ratio increases, there is a reduction in the number of captured sRNA–mRNA interactions. A cut-off score of 0.70 captures 98% of the possible validated interactions.

Similarly, when evaluating the success of the *p*-value filter, we started with a cut-off score of 1, as this captures all possible interactions, and then repeated the analysis each time lowering the cut-off score and recorded the number of validated and non-validated targets being captured. A total of 342 interactions, with 223

validated and 119 non-validated, were captured using a cut-off score of 1 and a total of 174 interactions, with 165 validated and 9 non-validated, were captured using a score of 0.01. The default value for the p -value filter implemented within PAREsnip2 (0.05) captures a total of 209 interactions. Of these, the filtering process kept 191 from the possible 223 (85.6%) validated interactions.

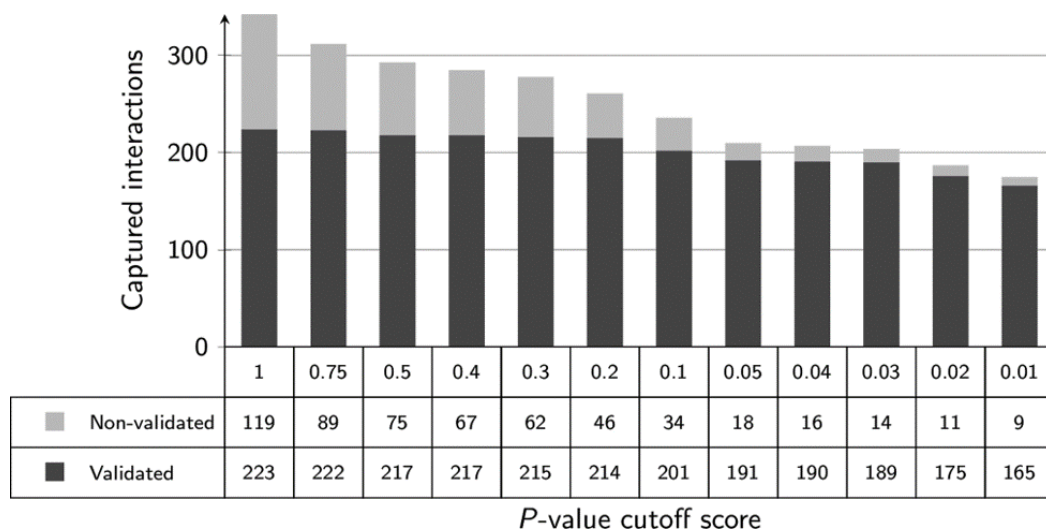


Fig. 3.4 The number of interactions reported when using p -value as a filter. As the cut-off decreases, there is a reduction in the number of captured sRNA–mRNA interactions. The default cut-off score of 0.05 captures 85.6% of the possible validated interactions.

3.4.5 Genome-wide analysis of degradome datasets

To illustrate the use of PAREsnip2, we carried out a genome-wide scale degradome analysis of dataset D2 using the sRNA–mRNA target interaction rules as described by Allen *et al.* [3]. For this analysis, we used the default stringent parameters, which discards category-4 signals and permits a minimum sRNA abundance of 5 reads. Additionally, the built-in conservation filter was used to increase confidence in the reported interactions. In total, PAREsnip2 captured 2008 sRNA–mRNA interactions (Appendix A Table 3), which comprised 960 category-0, 79 category-1, 511 category-2 and 458 category-3 interactions. To consider how the Allen *et al.* rules fared in capturing known interactions that have previously been validated, we

compared the results with the set of curated validated targets. We found that 178 of the validated targets were conserved within the three replicates of the dataset (degradome signal and miRNA sequence), and of these the Allen *et al.* targeting rules captured 132 (74%), which were predominantly category 0 interactions. Interestingly, 46 of the validated interactions within the sequencing data were missed. This could have been due to the stringency of the parameters that were used, or that fact that the Allen *et al.* rules were based on a small set of experimentally validated interactions and are somewhat outdated in their representation of the requirements of miRNA mediated cleavage activity. Therefore, to test this we repeated the analysis on the same dataset but using the Fahlgren and Carrington targeting rules where mismatches or G:U wobble pairs at positions 10 and 11 are allowed. This analysis identified 1072 category-0, 91 category-1, 611 category-2 and 529 category-3, making a total of 2303 interactions of which 151 (85%) of the possible validated interactions were captured (Appendix A Table 4). This shows a 11% improvement in identifying the known validated interactions compared to the Allen *et al.* targeting rules, which otherwise would have been missed. Performing this analysis using the Allen *et al.* rules took just 11 minutes and 32 seconds and the Fahlgren and Carrington targeting rules completed the analysis in 26 minutes and 48 seconds.

The timings for degradome analysis in *A. thaliana* led us to investigate the performance of PAREsnip2 on more complex species and larger genomes. The *T. aestivum* genome is much larger than *A. thaliana*, containing more than 155 000 transcript sequences within the genome annotation. We carried out a genome-wide analysis of the *T. aestivum* dataset (GEO accession GSE36867), which comprised a degradome of 4 306 082 non-redundant sequences and a corresponding sRNAome of 14 133 641 NR sequences. The default stringent parameters identified 25 063 interactions (Appendix A Table 5), which comprised 12 120 category-0, 1026 category-1, 5576 category-2 and 6341 category-3 interactions and completed in

just 31 minutes and 29 s. To investigate how using less stringent parameters would impact on the run-time performance of the tool, we repeated the analysis using the default flexible parameters. The tool identified 389,238 interactions (Appendix A Table 6), which comprised 83 409 category-0, 13 943 category-1, 79 935 category-2, 95 783 category-3 and 116 168 category-4 interactions with a run-time of 19 h and 39 min.

3.5 Discussion

In the age of genomics, the cost of sequencing has become cheaper and more accessible than ever before [203]. This had led to many more genomes being sequenced, some of which are much larger and significantly more complex than popular model organisms. Many genomes are used in large scale studies from human health [84] to food production [180]. Additionally, with the increasing number of reads being produced from sequencing experiments, the development of scalable and efficient algorithms for computational analysis of sequence data are becoming more and more important. In this chapter, we have developed a novel tool which is scalable with the increasing size and complexity of new genome releases and can perform a large scale degradome analysis using minimal computation resources. As an illustration, we ran our tool on wheat (*T. aestivum*), which in terms of base pairs is two orders of magnitude larger than *A. thaliana*. Using the default flexible parameters on the publicly available dataset described previously, the analysis took just 19 h and 39 min with a peak memory usage of 16GB and identified 389 238 targets by 169 636 sRNA sequences. In comparison, we terminated the execution of PAREsnip after 25 days on the same dataset, after which time it only reported 18% completion with a peak memory requirement of 175GB, far exceeding the resources you would expect to find in a typical desktop machine. Moreover, these results suggest that PAREsnip2 is the only tool capable

of performing degradome analysis over multiple biological replicates within a reasonable time scale.

Despite the improvements in computation time made possible with the newly developed algorithm, advancements to NGS technologies will continue to be developed, such as the illumina NextSeq 2000, which will be able to produce up to 1 billion reads per run. Therefore, changes to the current implementation of the algorithm may be required to avoid the tool becoming obsolete, such as harnessing the power of the GPU for the parallelizable target search.

As part of our performance comparison, we demonstrated that PAREsnp2 was able to outperform existing tools in terms of sensitivity when evaluated on a set of experimentally validated miRNA targets in *A. thaliana*. However, owing to limitations with the available data (i.e. the lack of an extensive set of true negatives), a full description of the performance of each tool using a confusion matrix was not possible. In the context of miRNA targets, a true negative is a miRNA-mRNA interaction with experimental evidence that the interaction does not occur and so this is often not reported in the literature. Furthermore, as the set of true positives used for this evaluation is almost certainly incomplete, further experimental validation of the perceived false positive predictions would provide a more accurate evaluation of the tools.

The miRNA targeting rules implemented within the currently available tools for degradome assisted target prediction are based on the analysis of experimentally validated miRNA targets in *A. thaliana*. These rules have been successfully applied to multiple other species during degradome analyses and sRNA target prediction with some predicted targets being further experimentally validated. However, probably in part due to the current lack of experimental evidence and to the best of our knowledge, no studies on miRNA targeting rules comparable to those performed on *A. thaliana* have been applied to other plant species. This may have resulted

in over-fitting our current understanding and implementation of these rules on *A. thaliana*. By providing the functionality to search for sRNA targets using configurable rules, users will be able to search for non-canonical targets that the existing rules would otherwise miss [236, 101, 29] and enable the potential to use a species specific set of rules if proven to be the case.

In its current form, PAREsnip2 is most suitable for the analysis of plant degradome datasets, as the primary mechanism for RNA silencing in plants is mRNA cleavage, whereas in animals the primary mechanism is translational repression. However, if the degradome data is available, PAREsnip2 could, in principle, be used for analysing sRNA mediated cleavage products in animals.

As is the case with many rule based systems, there exist a number of experimentally validated miRNA targets that do not fit the canonical set of targeting rules [236, 101, 29]. By adjusting the parameters so that these targets are found, PAREsnip2 may run the risk of increasing the rate at which false positives are reported. One potential solution to this would be to perform an analysis using a less stringent set of targeting rules alongside the built-in conservation filter. For example, if a high confidence, i.e. high abundance and low category peak, miRNA-target is reported across multiple biological replicates then further investigation, such as other experimental validation techniques, could be used to confidently determine if the reported interaction is real.

The PAREsnip2 algorithm has been implemented into a user-friendly and cross-platform (Windows, Linux and MacOS) application that enables users to analyse their data without the need for dedicated bioinformatics support or specialized computer hardware. Additionally, the tool can be run using the command line for users who wish to incorporate PAREsnip2 into more complex computational pipelines. Enabling the use of specialist bioinformatics software without the need

for any computational expertise will hopefully lead to new discoveries within RNA silencing pathways in all manner of experimental contexts.

3.6 Conclusion

In this chapter we introduced PAREsnip2, which is a fast and configurable software tool for analysing plant sRNA and degradome datasets. We discussed that the miRNA targeting rules implemented within the currently available tools are based on the analysis of experimentally validated miRNA targets in *A. thaliana*. Indeed, many predicted targets using these rules have been experimentally validated in other species. However, no studies investigating miRNA targeting rules of tissue specific or species specific miRNAs have been performed. In the following chapter, we employ the configurability of PAREsnip2 to investigate the differences between targeting criteria of multiple subsets of miRNAs.

Chapter 4

Computational inference of plant microRNA targeting rules using the degradome

4.1 Summary

In the previous chapter, we introduced PAREsnip2, a software tool for the analysis of plant sRNA and degradome datasets. PAREsnip2 has two sets of default targeting criteria, the Allen *et al.* rules, which were inferred in 2005 on 94 experimentally validated miRNA targets in *Arabidopsis thaliana* and the Fahlgren and Carrington rules, which are based on a larger set of 155 interactions [61]. However, these criteria may not be optimal across all datasets e.g. for specific organisms, tissues or treatments. In this chapter, we present a new tool, PAREameters, for data-driven inference of plant miRNA targeting criteria. Using publicly available sequencing datasets, we illustrate how PAREameters extracts information from paired sRNA and degradome sequencing data, in conjunction with miRNA annotations (e.g. from miRBase [110]), to infer criteria that results in increased sensitivity when evaluated

in *A. thaliana*. We show that different subsets of miRNA–mRNA interactions, such as those containing conserved or species-specific miRNAs, those found in monocots and dicots, and those identified in model and non-model organism, display variation in their target interaction properties. The tool is freely available, open source and provided as part of the UEA sRNA Workbench [192].

This chapter is an adapted and extended version of the work published in *Nucleic Acids Research* [200].

4.2 Background

Improvements to Next Generation Sequencing technologies have resulted in larger and more diverse experiments, including those that make use of multiple data types, for example, to increase prediction accuracy of regulatory interactions by combining sRNA sequencing and mRNA quantification [148]. These improvements have also led to the sequencing and annotation of different organisms' genomes and facilitated functional studies outside of the context of model organisms [69]. However, a vast proportion of our understanding of specific biological mechanisms is based on the study of model organisms, mostly due to their lower regulatory complexity and availability of extensive, varied, public sequencing datasets. Many computational methods designed for extracting information and features from sequencing data (e.g. sRNA classification and target prediction) often summarize the data-mining results into rule-based models, derived from experimental observations. However, this approach carries the risk of over-fitting a model (e.g. set of thresholds or accepted ranges) on specific sets of observations.

As discussed in Section 2.3, sRNAs play important roles in transcriptional and post-transcriptional gene regulation in eukaryotes [143]. In plants, the latter mode of action is achieved predominantly through miRNAs, which reduce the amount of

mRNA available for translation by directing the RISC to their sequence-specific mRNA target(s) and inducing cleavage and subsequent degradation of the mRNA [28]. The miRNA classification criteria were first proposed by Ambros *et al.* [4] and Meyers *et al.* [150]; however, more recently these criteria have been updated based on a substantial increase in publicly available sequencing datasets and known miRNA annotations by Axtell *et al.* [12]. For example, the new miRNA annotation criteria [12] increased the number of allowed mismatches and asymmetric bulges compared to the previous annotation model [4, 150]. In this chapter, we investigate the applicability and portability of the current miRNA target interaction model.

Most miRNA target prediction tools use fixed rule-based targeting criteria, the majority of which are variations of the rules inferred by Allen *et al.* [3] on experiment specific, low-throughput validated *A. thaliana* miRNA–mRNA interactions (discussed in Section 2.5.3). One particularly prominent problem with fixed, sequence-based targeting criteria is how they address miRNA–mRNA target sites that contain central mismatches [12], e.g. psRNATarget classifies all interactions containing central mismatches as translational repression ones [49, 50]. However, this contradicts the more refined set of potential outcomes illustrated in the literature, namely that central mismatches can induce mRNA cleavage [3], act as target-mimics [94, 137], cause translational repression [95] or simply be non-functional [129]. Thus, without additional data it is difficult to predict miRNA function based solely on complementarity patterns.

One such type of additional data is degradome sequencing [74, 179], which captures the 5' ends of downstream cleaved mRNAs, described in Chapter 3. Tools for predicting miRNA targets that combine the Allen *et al.* criteria, with minor variations, and degradome sequencing data are described in Chapter 3. The performance evaluation, over three biological replicates, that we performed (see Table 3.5) demonstrated that even the most sensitive tool, PAREsnip2, was only able to capture ~80% of the expressed and experimentally validated interactions when

using the Allen *et al.* criteria. Further analyses revealed that the remaining $\sim 20\%$ were missed mostly due to discrepancies in the number or position of mismatches, gaps, G:U pairs and the MFE ratio.

These results suggest that the current targeting criteria may be too stringent or over-fitted on a small set of organism, tissue or treatment specific experimentally validated miRNA–mRNA interactions. Analyses of miRNA–mRNA interactions in various organisms have shown that currently implemented criteria do not capture all known and expressed miRNA–mRNA interactions (e.g. in *A. thaliana* [29] and *Oryza sativa* [236]). This is further borne out by a preliminary analysis, where we show that, by following a similar approach for manually inferring targeting criteria as Allen *et al.*, parameters shown in Table 4.1, we achieve a sensitivity increase of $\sim 15\%$ when evaluating on experimentally validated interactions in *A. thaliana*, presented in Appendix B Table 1 and discussed in Section 4.4.1. In addition, the portability of current criteria across organisms and tissues has not yet been quantitatively evaluated. Furthermore, the sensitivity and precision of a set of predictions may differ based on the size or characteristics of the input data. For example, functional analysis of a specific miRNA may benefit from reduced precision, yet good sensitivity, to increase the number of candidates for further investigations; whereas an analysis on the entire set of sRNAs requires concerted high sensitivity and precision. We now present a method that aims to overcome some of these drawbacks.

4.3 Methods

4.3.1 The PAREameters pipeline

As mentioned above, our new method is called PAREameters. In Figure 4.1, we present an overview of the PAREameters pipeline. The input consists of synony-

mous sRNA and PARE samples; technical or biological replicates can be used for assessing technical variation and noise between samples or for the exclusion of spurious results. An annotated reference genome and transcriptome, and a set of known plant miRNAs (e.g. from miRBase [110]) are also required. The tool's output consists of miRNA predictions and their mRNA targets, based on a set of highly permissive parameters. PAREameters also provides a set of suggested targeting criteria, based on these predictions, but also provides the properties of these interactions as individual outputs. In doing so, the user can interpret the results manually to infer criteria that satisfy their sensitivity and precision requirements.

The first stage of the pipeline is to remove low quality reads, sequencing errors or to identify sample outliers. PAREameters includes several optional filtering methods: (i) sequences containing ambiguous bases (e.g. Ns) are discarded; (ii) a low sequence complexity filter is applied based on the single, di- or tri-nucleotide frequencies (described in Section 3.3.2), with set thresholds of 75%, 37.5% and 25%, respectively; (iii) all reads that do not align to the provided reference genome are discarded. We now explain each of the other stages of the pipeline in more detail.

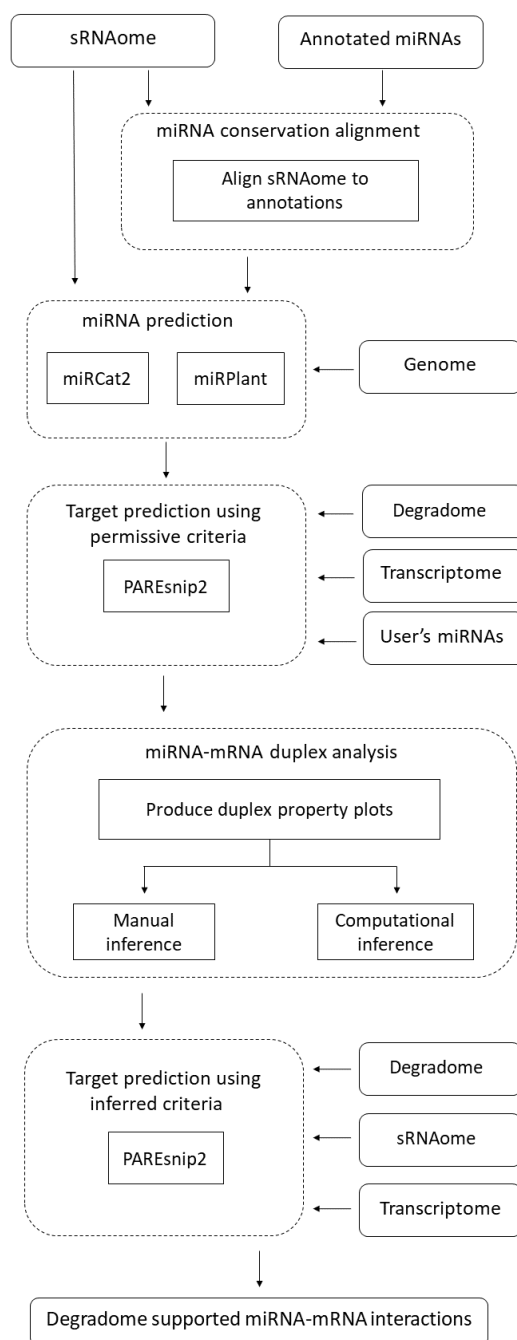


Fig. 4.1 PAREameters pipeline. The input and output data are represented by continuous rounded rectangles, processes are represented by straight rectangles and the different steps of the analysis are represented by dashed rounded rectangles. PAREameters takes as input two types of sequencing samples, paired sRNA and degradome, a genome with corresponding annotations and current miRBase miRNA annotations. The output is a set of data-inferred thresholds for a rule-based prediction of miRNA–mRNA interactions using e.g. PAREsnp2. The sRNAome and degradome inputs are the experiment-specific datasets whereas the genome, transcriptome and annotated miRNA inputs are part of the species annotation.

4.3.2 miRNA prediction

The miRNA candidates used as input for PAREameters are obtained via two approaches: (i) with focus on conserved miRNAs, the input sRNA samples are aligned (positive strand only) to all known plant miRNA sequences, obtained from miRBase [110], allowing up to two mismatches and no gaps. The selected sequences are then used as input to miRPlant [5]. Candidates that fulfill the criteria for miRNA prediction (default parameters) are then retained for the subsequent steps; (ii) with focus on all miRNAs (conserved and new) as predicted using miRCat2 [158] (default plant parameters) with the whole sRNA sample as input. All data pre-processing required steps to run the miRNA prediction tools, such as building the bowtie index [116] and organizing the sequencing data into non-redundant format, are handled by PAREameters.

4.3.3 Target prediction using permissive criteria

The sRNAs that satisfy miRNA biogenesis criteria (as described above) are provided as input to PAREsnip2 [198]. In addition and to compensate for the stringent criteria of miRNA prediction tools, the user can provide their own annotated miRNA entries if they have an abundance ≥ 5 (user-defined parameter) but did not fulfill the criteria of the prediction tools. The target prediction is then performed on the input data using a set of highly-permissive, user-configurable, parameters, shown in Table 4.1.

The miRNA–mRNA interactions predicted by PAREsnip2 are kept if the abundance of the peak of interest is ≥ 5 and are further classified into high-confidence (HC) or low-confidence (LC). For the former, the peak is the highest across the whole transcript (i.e. Category-0 or 1); for the latter, the peak is not the highest on the transcript (i.e. Category-2 or 3). The categorization of miRNA–mRNA interactions is based on the distribution of abundances of the degradome reads

Parameter	Allen <i>et al.</i>	Manually Inferred	PAREameters
Allow MM at pos 10	No	Yes	Yes
Allow MM at pos 11	No	Yes	Yes
Max # adj mm in CR	1	0	2
Max # MM in CR	2	2	3
Max score	4	5	6
Max # MM	4	4	6
Max # G:U	4	3	5
Max # adj MM	2	1	4
MFE ratio cut-off	0.7	0.65	0.6

Table 4.1 The PAREsnip2 parameter values for the Allen *et al.*, manually inferred and PAREameters permissive criteria. The Allen *et al.* criteria were previously inferred in 2005 [3]. The manually inferred criteria was inferred on a set of 387 experimentally validated *A. thaliana* interactions. The permissive parameters are used initially by PAREameters to find high-confidence (HC) interactions. The inferred criteria are then extracted from HC interactions using the retain rate parameter. MM = mismatch, CR = core regions (positions 2-13 of miRNA), MFE = minimum free energy.

aligned to each transcript, as described in the Section 3.3.4. Peaks with abundance less than 5 are excluded as it is difficult to distinguish between true miRNA cleavage products and random degradation at such low abundance.

When comparing the results of PAREameters, where similar results were observed for all replicates, only one was randomly selected to illustrate the conclusions for all the subsequent comparative analyses.

4.3.4 miRNA–mRNA duplex analysis and inference of targeting criteria

Valid miRNA–mRNA duplexes, based on the analysis of the degradome data coupled with the miRNA prediction, are characterized using specific properties, such as the number and location of mismatches, G:U wobble pairs and adjacent mismatches, the alignment score and the MFE ratio. The algorithm then infers a

set of targeting criteria that attempts to retain at least 85% (user-defined parameter) of the valid miRNA–mRNA duplexes. We chose the default value of the retain rate parameter based on the analysis of sensitivity gain against precision loss of inferred criteria across an incremental range of retain rate values on a benchmark leaf *A. thaliana* dataset comprising three replicates [198], presented in the results. The biological interpretation of the retain rate threshold is that a higher degree of complementarity between a miRNA and its target results in higher confidence that the interaction is genuine, whereas interactions with weaker complementarity may require further experimental validation before you can be confident that they are genuine.

Using a set of experimentally validated interactions as validation (Appendix A Table 1, described in Section 3.4.3), we focused on HC interaction pairs at known target sites with corresponding miRNAs. The validation classes: true positives (TP), false positives (FP) and positives (P) are used in a loose sense, i.e. TP consists of the predicted interactions with experimental validation, FP is the set of predicted interaction for which, currently, there is no experimental validation, and P is the set of experimentally validated interactions with corresponding HC peaks. For each set of targeting rules, we present the sensitivity as $Se = TP / P$ (proportion of predicted validated interactions) and the precision as $PPV = TP / (TP+FP)$ (proportion of validated interactions, out of the total number of reported interactions). In our evaluation, we did not include specificity as a measure of performance because the class of true negatives (TN) cannot be accurately determined. The set of TN comprises the interactions for which there is experimental evidence that interactions do not occur; since the current available information is based on positive events, i.e. experimental validation confirming the interaction happens within an experimental context, it is not possible to obtain a comprehensive set of TN data. Moreover, degradome based miRNA target prediction tools are validation-driven, i.e. they only report interactions that are predicted to be TP based on the defined criteria,

which makes it impossible to perform the specificity calculation as perceived TN results are not reported.

In addition, PAREameters provides a summary of the interaction properties, enabling the manual interpretation of the results and allowing the user to choose a set of targeting criteria that satisfies their choice of sensitivity and precision.

The significance of the distribution of properties with respect to a reference set of miRNA–mRNA interactions (Appendix A Table 1, described in Section 3.4.3) was calculated using offset χ^2 tests and the contribution of each feature was assessed using individual Fisher exact tests [152], e.g. when comparing conserved versus species-specific interactions, the former is considered the reference. The χ^2 tests were used to assess the overall differences in distributions, across all 21 positions, whereas the Fisher exact tests compared the values for each individual position, against the sum of values for all remaining 20 positions. To control false discoveries from multiple testing we corrected the reported p -values using the Benjamini-Hochberg correction [18] for all the χ^2 and Fisher’s exact tests. Finally, the relative distributions of miRNA–mRNA duplex MFE ratios [3, 198] were analyzed using Kolmogorov–Smirnov tests; briefly, the distributions were first sampled, without replacement, to the same number of entries (given the high number of measurements present in each of compared subsets, this did not distort the original MFE distributions); next, the cumulative distributions were directly compared using the Kolmogorov–Smirnov test and the p -value was reported. The significance threshold for all statistical tests was set at 0.05.

4.3.5 Implementation of PAREameters

We implemented the PAREameters tool in Java (version 8); the code used to create the plots and perform the significance tests is implemented in R (version 3.5.1, Apple Darwin) and is invoked from the PAREameters pipeline using system calls,

assuming a valid version of R is installed and correctly configured on the users PATH. All computational analyses and benchmarking were performed on a desktop machine running Ubuntu 18.04 equipped with a 3.40GHz Intel Core i7-6800K six core CPU and 128GB RAM. PAREameters is optimized both in run-time and computational resource usage, as shown in Appendix B Table 2; the analysis of a typical *A. thaliana* and *Triticum aestivum* sample completes in ~30 min and 1 day 10 h, with 6 and 10 GB memory (RAM) requirements, respectively. PAREameters is a user-friendly, cross-platform (Windows, Linux and MacOS) application that enables users to analyze sequencing datasets without the need of specialized support or dedicated hardware.

4.3.6 Datasets

The sequencing datasets analysed in this chapter are described in Appendix B Table 3. Briefly, the *A. thaliana* datasets comprise paired sRNA and PARE samples: wild-type leaf triplicates [158, 198], wild-type leaves in multiple growth stages [197] and wild-type flower, leaf, root and seedling of plants grown at 15°C [83]. The genome and transcriptome used for all *A. thaliana* were obtained from TAIR10 [115].

In addition to the *A. thaliana* datasets, we exemplify the usage of PAREameters on sRNA and corresponding PARE datasets from *Amborella trichopoda* leaf and opened female flower, *Glycine max* leaf [36], *Oryza sativa* inflorescence [214] and *T. aestivum* 2.2mm spikes [196]. The transcriptome and genome sequences for organisms other than *A. thaliana* were obtained from EnsemblPlants Release 43 [22].

Summaries of each sRNA dataset, such as the number of raw and unique reads, genome matching reads and the number of known miRNAs present (based on current miRBase (Release 22) [110] annotation) are presented in Appendix B Table

4. Summaries for each of the PARE data are presented in Appendix B Table 5 and include the number of transcriptome matching reads (positive strand only).

4.4 Results

4.4.1 Evaluation of inferred targeting rules in *A. thaliana*

We first illustrate the differences in sensitivity and precision between two sets of manually inferred criteria in *A. thaliana*. These criteria are those previously defined by Allen *et al.* [3] and those we manually inferred on a larger set of experimentally validated interactions (Appendix A Table 1). We then highlight the advantages of the data-driven approach implemented in PAREameters by presenting the increase in sensitivity of the computationally inferred targeting rules compared with the Allen *et al.* criteria when benchmarked on multiple *A. thaliana* datasets.

Using the *A. thaliana* leaf dataset D1, we employed two sets of targeting criteria, the Allen *et al.* criteria and criteria we manually inferred from a larger set of validated *A. thaliana* miRNA–mRNA interactions (Table 4.1). These criteria were provided as input parameters for PAREsnip2 for target prediction. The evaluation of these manually inferred criteria, presented in Appendix B Table 1, showed an increase in sensitivity between 11.43% and 19.82% when benchmarked on multiple *A. thaliana* datasets. Upon further inspection, the majority of validated interactions that were missed using the criteria we manually inferred were due to having an MFE ratio less than the selected cut-off value of 0.65. The MFE ratio quantifies the hybridization strength between the miRNA and its target and thus a higher cut-off value may result in interactions more likely to cause cleavage being reported.

The increase in performance of the manually inferred criteria may be due to overfitting on the larger set of interactions. In addition, due to the scarcity of validated

interactions, either as number of valid interactions or localization of specific modes of action in different cell types [140], these criteria may not be portable between various organisms or tissues. Therefore, we used the PAREameters tool to infer targeting criteria from the *A. thaliana* D1, D2 and D3 datasets. The resulting criteria, presented in Table 4.2, were then utilized by PAREsnip2 for target prediction and the results evaluated and compared to the predictions obtained using the Allen *et al.* criteria. The evaluation method used is identical to that of the manually inferred criteria. Specifically, for each dataset, the class of positive (P) data included experimentally validated miRNA–mRNA interactions with HC transcript peaks and corresponding miRNA sequence with abundance ≥ 5 .

Parameter	D1A	D1B	D1C	D2A	D2B	D2C	D3A	D3B	D3C	D3D
MM at pos 10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MM at pos 11	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Max adj MM CR	0	0	0	0	0	0	0	0	0	0
Max MM CR	1	1	1	1	1	1	1	1	1	1
Max score	4.50	4.50	5.00	5.00	5.00	4.50	4.00	4.50	4.50	4.50
Max MM	3	3	3	3	3	3	3	3	3	3
Max G:U	2	2	2	2	2	2	2	2	2	2
Max adj MM	1	1	1	1	1	1	1	1	1	1
MFE ratio cut-off	0.69	0.70	0.69	0.66	0.65	0.67	0.72	0.71	0.70	0.71

Table 4.2 The PAREameters inferred criteria for each of the *A. thaliana* datasets. MM = mismatch, CR = core region (2-13nt of miRNA) and MFE = minimum free energy. datasets. MM = mismatch, CR = core region (2-13nt of miRNA), adj = adjacent and MFE = minimum free energy.

The results, presented in Table 4.3, show that the computationally inferred criteria provides increased sensitivity compared to the Allen *et al.* criteria, whilst also maintaining precision on most datasets. Over all datasets, PAREameters inferred criteria with a median sensitivity of 88.5% (range: 82.8–89.4%) versus 81.4% (range: 75.6–84.6%) for the Allen *et al.* criteria. The median precision for the PAREameters inferred criteria was 91.3% (range: 80.1–96.8%) versus 91.4%

(range: 83.8–97.5%) for the Allen *et al.* criteria. We also evaluated the time and memory performance of PAREameters on each dataset. The run-time of the pipeline depends on the size of the input data (sequencing depth of the sRNA and PARE samples and the size of the reference genome). On *A. thaliana* D1, D2 and D3 datasets, the run-time range was 16 min and 52 s to 1 h 4 min (this excludes the time taken to build the bowtie index as this is only done once per species) and the memory usage varied between 5GB and 8GB (see Appendix B Table 2). The inference component of PAREameters is linear on the size of the sRNA and PARE input data.

4.4.2 Evaluation of data input size and retain rate on sensitivity and precision

We now demonstrate that the increase in sensitivity of the PAREameters inferred criteria when compared to the Allen *et al.* criteria is not a result of over-fitting on the input data by evaluating performance using a cross-validation approach. We then show how increasing the amount of training data may lead to a more accurate representation of inferred targeting criteria. Finally, we assess how the retain rate parameter impacts sensitivity and precision of the PAREameters inferred criteria.

Based on the properties of HC miRNA–mRNA duplexes with cleavage signal confirmation in the PARE data, PAREameters inferred targeting criteria that increased the sensitivity and retained precision versus existing fixed criteria when tested against the set of experimentally validated interactions in *A. thaliana*. To avoid the over-fitting of targeting criteria based on characteristics of the input data, we tested the stability of the inferred properties using a cross-validation technique and the set of experimentally validated *A. thaliana* miRNA–mRNA interactions on the D1, D2 and D3 datasets. Specifically, we used the HC interactions with corresponding miRNA sequences in each dataset as a starting point. We then randomly

Dataset	# miRNAs	# V	Allen V	Inf. V	Allen NV	Inf. NV	Allen Se	Inf. Se	Allen PPV	Inf. PPV	Se gain	PPV difference
D1A	37	129	105	112	9	8	81.4%	86.8%	92.1%	93.3%	5.4%	1.2%
D1B	38	131	109	116	11	10	83.2%	88.5%	90.8%	92.0%	5.3%	1.2%
D1C	35	121	95	107	12	14	78.5%	88.4%	88.7%	88.4%	9.9%	-0.3%
D2A	40	140	117	125	14	29	83.5%	89.3%	89.3%	81.1%	5.8%	-8.2%
D2B	38	137	113	121	13	30	82.4%	88.3%	89.6%	80.1%	5.9%	-9.5%
D2C	40	144	117	120	3	4	81.2%	83.3%	97.5%	96.7%	2.1%	-0.8%
D3A	32	79	64	68	4	7	81.0%	86.1%	94.1%	90.6%	5.1%	-3.5%
D3B	29	70	57	58	11	13	81.4%	82.8%	83.8%	81.6%	1.4%	-2.2%
D3C	36	111	84	98	6	7	75.6%	88.2%	93.3%	93.3%	12.6%	0.0%
D3D	35	104	88	93	3	3	84.6%	89.4%	96.7%	96.8%	4.8%	0.1%

Table 4.3 Comparison of sensitivity and specificity between the Allen *et al.* criteria and the PAREameters inferred criteria on the *A. thaliana* datasets. Allen = Allen *et al.* rules, Inf. = PAREameters inferred criteria, V = validated, NV = non-validated, Se = sensitivity and PPV = precision. An increase in the achieved Se is observed for the inferred criteria over all *A. thaliana* datasets.

split the HC validated interactions in each dataset to form two groups: the training group, containing 75% of the data, and the testing group, which contained the remaining 25%. PAREameters was used to infer parameters on the training set and these were employed by PAREsnip2 for target prediction on the test set. We then calculated the sensitivity and precision of the inferred parameters on the training set and on the test set. The random cross-validation was repeated 50 times and the results, presented in Table 4.4, show that PAREameters is able to infer targeting parameters with a median sensitivity of 77.5% (range: 67.0–81.3%) and precision 83.2% (range: 75.0–100.0%) when evaluated on the unobserved testing data.

Dataset	Test Inferred Se	Test Inferred PPV
D1A	81.3%	83.0%
D1B	78.1%	83.3%
D1C	80.0%	82.1%
D2A	79.0%	78.0%
D2B	77.0%	87.3%
D2C	67.0%	89.0%
D3A	68.4%	78.0%
D3B	77.0%	75.0%
D3C	78.0%	96.0%
D3D	76.0%	100%

Table 4.4 The median sensitivity (Se) and precision (PPV) values for the cross-validation experiments on the *A. thaliana* datasets. The cross validation was done on a 75/25% split for training and testing, respectively. Each analysis was repeated 50 times and the median value was recorded.

The decrease in sensitivity from our previous analysis likely originates from the fact we are inferring criteria from one set of miRNA–mRNA interactions and testing on a different set of miRNA–mRNA interactions. Whereas previously, we were inferring criteria from the whole set of PAREameters predicted HC miRNA–mRNA interactions. This further supports our hypothesis that miRNAs may have different modes of action or target complementarity requirements and demonstrates that

using just one set of fixed criteria may not be sufficient when performing miRNA target prediction.

To investigate further how increasing the amount of training data may lead to a more accurate representation of inferred targeting criteria, we evaluated the computationally inferred criteria produced by PAREameters on different sized subsets of the experimentally validated interactions contained within the D1 datasets. Starting with 10% of the validated data, followed by increments of 10% until the final value of 90%, we used PAREameters to infer criteria on the training subset and then evaluated those criteria on the remaining unseen data. Analysis on each subset was performed 50 times and the results shown in Table 4.5. On each dataset, increasing the amount of training data resulted in an overall increase in sensitivity. Intriguingly, the increase in training data resulted in a decrease in precision. However, this should not be seen as a negative result, as we've previously stated, the class FP is the set of predicted interactions for which, currently, there is no experimental validation. Indeed, the current class of positive data is almost certainly incomplete, therefore further experimental validation can only increase the sensitivity and precision values for the inferred criteria.

Training size	D1A Se	D1A PPV	D1B Se	D1B PPV	D1C Se	D1C PPV
10%	69.0%	95.2%	48.7%	97.2%	71.8%	90.9%
20%	70.9%	93.8%	59.6%	94.8%	66.7%	90.4%
30%	77.2%	92.4%	64.3%	94.5%	73.8%	89.4%
40%	79.2%	91.5%	75.6%	91.9%	70.8%	88.2%
50%	78.1%	89.9%	76.9%	89.7%	77.5%	86.4%
60%	75.5%	88.0%	77.9%	87.5%	77.1%	84.1%
70%	79.0%	85.3%	76.9%	85.3%	77.8%	83.8%
80%	82.0%	80.8%	80.8%	83.3%	79.2%	82.4%
90%	83.3%	78.2%	84.6%	79.3%	79.2%	80.0%

Table 4.5 The median sensitivity (Se) and precision (PPV) values for the training-size experiment on the *A. thaliana* D1 datasets. For each dataset, an increase in training-size resulted in an overall increase in sensitivity.

To assess how changes to the PAREameters retain rate parameter impact sensitivity and precision, we evaluated the computational inferred targeting criteria produced by PAREameters on the D1 dataset with increasing retain rate values. The results of this analysis are shown in Appendix B Table 6 and Appendix B Figure 1. Starting with an initial value of 0.5 and with increments of 0.05 thereafter, we recorded the number of validated and non-validated interactions being captured and determined the differences between Se and PPV for each incremental range. Next, we calculated the absolute value of the ratio between the increases in Se with respect to loss in PPV. For example, the Se and PPV values obtained using a retain rate value of 0.75 on the D1A dataset was 75.2% and 95.1%, respectively, and the Se and PPV values obtained using a retain rate value of 0.80 were 83.7% and 93.9%, respectively. This resulted in a Se increase of 8.5% and a loss in PPV of -1.2% for the 0.75–0.80 range and a Se/PPV ratio of 7.1, specifically, there was a 7.1x increase in Se with respect to the loss in PPV for this range increment. The optimal value for the retain rate parameter is obtained at the first increment range that results in a Se/PPV ratio < 1 (i.e. the loss in precision is greater than the increase in sensitivity), presented in Appendix B Table 7. In the *A. thaliana* D1 data used to exemplify the selection of the retain rate parameter, the first increment range with a Se/PPV ratio < 1 was the 0.85–0.90 range, which resulted in the value of 0.85 being selected as the default for the retain rate parameter.

Using the initial value on the D1A dataset, we capture a total of 30 miRNA–mRNA interactions, all of which are experimentally validated interactions. At the other end of the scale, using a retain rate of 1.0 captured 156 interactions, which comprised 128 validated and 28 non-validated. The default parameter value (0.85) captures a total of 120 interactions and provides a sensitivity value of 86.8% and precision value of 93.3%. A visual representation of these results of all three replicates in D1, which show similar results, can be found in Appendix B Figure 1. The increment range of 0.85–0.90 was the first to have a Se/PPV ratio less than 1 and

was consistent across three biological replicates. In experiments for which the values vary between samples, we recommend the usage of a consistent threshold across all samples of the experiment.

4.4.3 Consistency of attribute distributions and inferred criteria across miRNA subsets in *A. thaliana*

To evaluate the portability of targeting criteria (and distribution of properties) across miRNA subsets, we inferred criteria on a set of conserved and species-specific *A. thaliana* miRNAs [110] and their experimentally validated targets (Appendix A Table 1, described in Section 3.4.3). The group built on the conserved miRNAs comprised 201 miRNA–mRNA interactions from 42 unique miRNA sequences (Appendix B Table 8). The group built on miRNAs specific to the Brassicaceae family comprised 184 interactions from 47 unique miRNA sequences (Appendix B Table 9). The summaries of the position-specific property distributions, which include the localizations of gaps, mismatch, and G:U wobbles and the MFE ratio distributions for the conserved and specific miRNA interactions are presented in Figure 4.2 panel A and panel B, respectively. In Figure 4.2A, the Brassicaceae specific miRNAs show highly similar results to that of Allen *et al.* [3] (Figure 2.4), i.e. a large proportion of mismatches or G:U wobble pairs at position 1, no mismatches at the canonical positions 9 and 10 and relatively few mismatches in the 5' core region (positions 2–13) of the miRNA when compared to the 3' end. In contrast, the requirements for complementary of species-specific miRNAs appear to differ when compared to conserved miRNAs, especially at the miRNA 5' end, with mismatches being tolerated at positions 5, 8 and 9, in addition to the canonical position 10 of the miRNA.

To evaluate whether the differences in properties between specific-specific and conserved miRNA interactions in *A. thaliana* are significant, we performed χ^2 tests

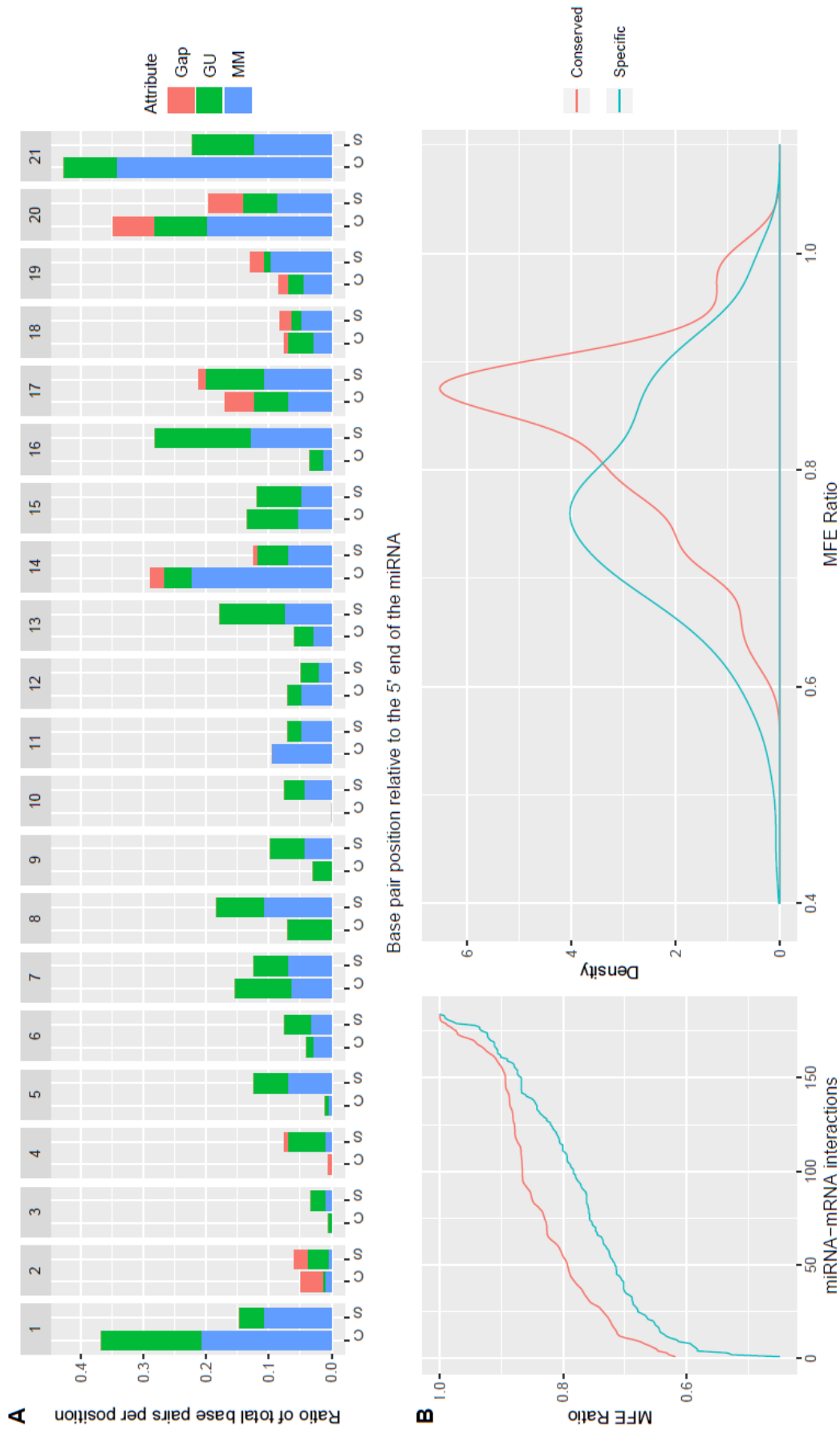


Fig. 4.2 Side-by-side comparison of property distributions for conserved and species-specific miRNAs in *A. thaliana*. Using experimentally validated miRNA-mRNA interactions as input, we calculated the position-specific properties (A) and the MFE ratio distribution (B) for the conserved and species-specific miRNA-mRNA interactions. The significance of the differences in the localization of gaps, G:U pairs and mismatches were assessed using offset χ^2 tests and the contribution of individual categories was evaluated using Fisher exact tests. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of 8.57×10^{-10} .

of significance using the conserved properties as the expected distribution and the species-specific properties as the observed distribution. Additionally, we use the Fisher's exact test to determine the specific property at each position responsible for the significance of the differences. The results of the significance analysis for the position-specific property distributions are presented in Table 4.6. Based on the χ^2 tests, significant differences between properties can be found at positions 1, 16 and 21. Based on the Fisher's exact test, positions 8, 14, 16 and 21 have significant differences in their proportion of mismatches. We also analyzed the differences in MFE ratio distributions between conserved and species-specific miRNAs, shown in Figure 4.2B, and the significance of the differences were evaluated using a Kolmogorov–Smirnov test, which reported a p -value of 8.57×10^{-10} . These results may suggest a higher complementarity requirement between conserved miRNAs and their targets than that of species-specific miRNAs.

To investigate the portability between criteria inferred exclusively on conserved or species-specific miRNA interactions, we evaluated the inferred rules of each set of interactions (all four pairwise combinations: conserved rules on conserved interactions, conserved rules on species-specific interactions and the similar pairs on the species-specific rules), using PAREsnp2. The results, presented in Table 4.7, show a consistent decrease in sensitivity for both the conserved and species-specific miRNAs when inferring criteria on the other subset of miRNA–mRNA interactions. Specifically, a decrease from 82.1% to 65.7% and 76.1% to 56.0% for the conserved and species-specific miRNA–mRNA interactions, respectively. Further investigation into these differences support our previous observation regarding the differences in MFE ratio of conserved and species-specific miRNA interactions, with the inferred values being 0.75 and 0.68, respectively, further supporting our previous observation regarding an increased complementarity requirement for conserved miRNAs. Another intriguing difference between the inferred criteria is

miRNA position	χ^2	MM	G:U	Gap
1	0.039	0.268	0.120	1.000
2	0.779	1.000	0.870	1.000
3	0.811	1.000	0.870	1.000
4	0.322	1.000	0.454	1.000
5	0.085	0.147	0.497	1.000
6	0.811	1.000	0.837	1.000
7	0.811	1.000	0.870	1.000
8	0.085	0.017	1.000	1.000
9	0.637	0.276	0.870	1.000
10	0.392	0.276	0.741	1.000
11	0.779	0.747	0.870	1.000
12	0.811	0.747	1.000	1.000
13	0.288	0.479	0.497	1.000
14	0.085	0.017	1.000	1.000
15	0.996	1.000	1.000	1.000
16	0.002	0.017	0.06	1.000
17	0.637	0.747	0.870	1.000
18	0.779	1.000	0.870	1.000
19	0.687	0.402	0.896	1.000
20	0.288	0.172	0.879	1.000
21	0.039	0.011	1.000	1.000

Table 4.6 Offset χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA–mRNA interactions in *A. thaliana*. The contribution of specific properties, such as mismatches (MM), G:U pairs and gaps are assessed using Fisher exact tests. Values at or below the significance threshold (0.05) and highlighted in bold.

an allowed mismatch or G:U pair at position 10 of the species-specific miRNAs.

Complete list of inferred parameters can be found in Appendix B Table 10.

Inferred on	Evaluated on	Possible	Captured	Sensitivity
Conserved	Conserved	201	165	82.1%
Specific	Conserved	201	132	65.7%
Specific	Specific	184	140	76.1%
Conserved	Specific	184	103	56.0%

Table 4.7 Sensitivity on cross pairwise comparisons for criteria inferred on conserved or species-specific miRNAs for the validated *A. thaliana* interactions. The targeting criteria were inferred using a retain rate of 0.85 and a considerable decrease in sensitivity was observed for the mismatched pairs i.e. training on conserved interactions and testing on specific interactions.

The differences between the properties of conserved and species-specific interactions highlight the need for customization in the set of criteria used for describing and capturing miRNA–mRNA interactions when conserved or species-specific miRNAs are involved.

4.4.4 Evaluation of miRNA targeting criteria in non-model organisms

Current miRNA targeting rules, inferred on interactions mostly consisting of conserved miRNAs from *A. thaliana* [3], have been applied to other species for target prediction [160, 134, 123, 103]. However, to the best of our knowledge, no comprehensive investigation into the suitability of these fixed targeting criteria has been performed in non-model organisms. The characterization of miRNA–mRNA interactions has been facilitated by both the increased complexity of experiments involving non-model plant species and through the analysis of RNA degradation profiles (PARE [74] sequencing and more recently NanoPARE [179]), which despite tech-

nical limitations, e.g. sequencing bias [190], can provide reliable high-throughput validation of miRNA-mediated cleavage sites.

To investigate the suitability and portability of the fixed Allen *et al.* criteria on non-model organisms and evaluate the scope for customized, organism-specific rules, we conducted an exploratory analysis using as input the HC degradome-supported miRNA–mRNA interactions reported by PAREameters. We compared the inferred rules for flower and leaf tissues in several organisms to produce a quantitative summary of the variation ranges of thresholds for the selected rules. Appendix B Table 11 shows these summaries of inferred criteria per organism; Figure 4.3A illustrates the position-specific distributions of G:U pairs, mismatches and gaps, and Figure 4.3B shows the MFE ratio distributions for the miRNA–mRNA duplexes from flower tissue across organisms in *A. thaliana*, *A. trichopoda*, *O. sativa* and *T. aestivum*. Similar plots for leaf tissue in *A. thaliana*, *A. trichopoda* and *G. max* are presented in Appendix B Figure 2.

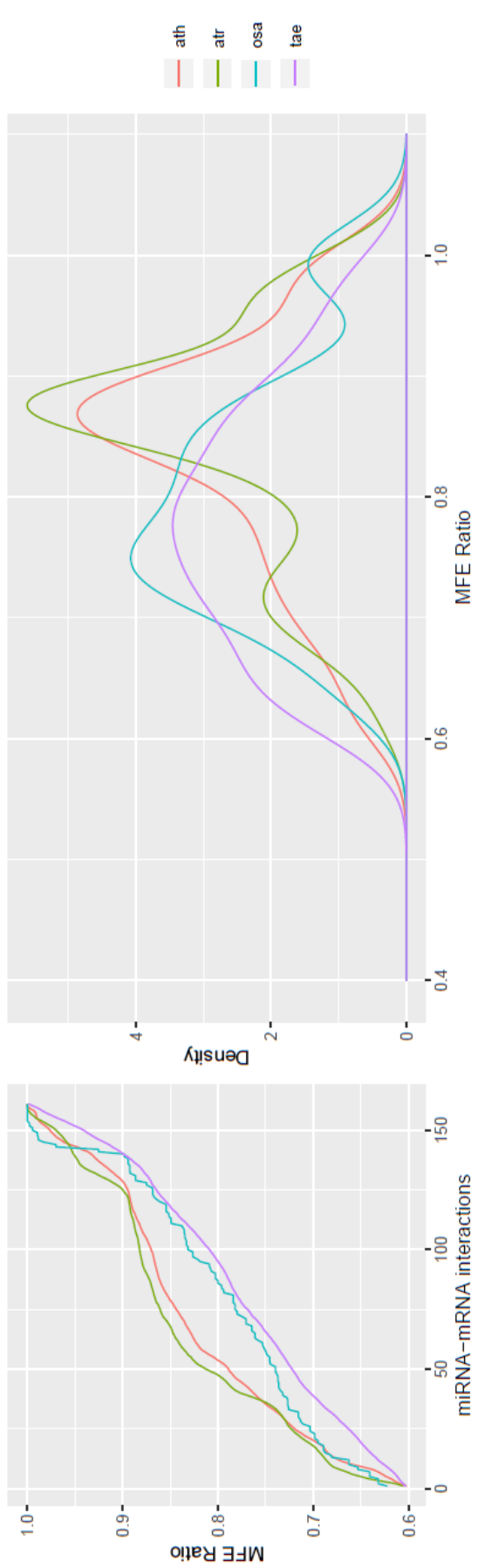
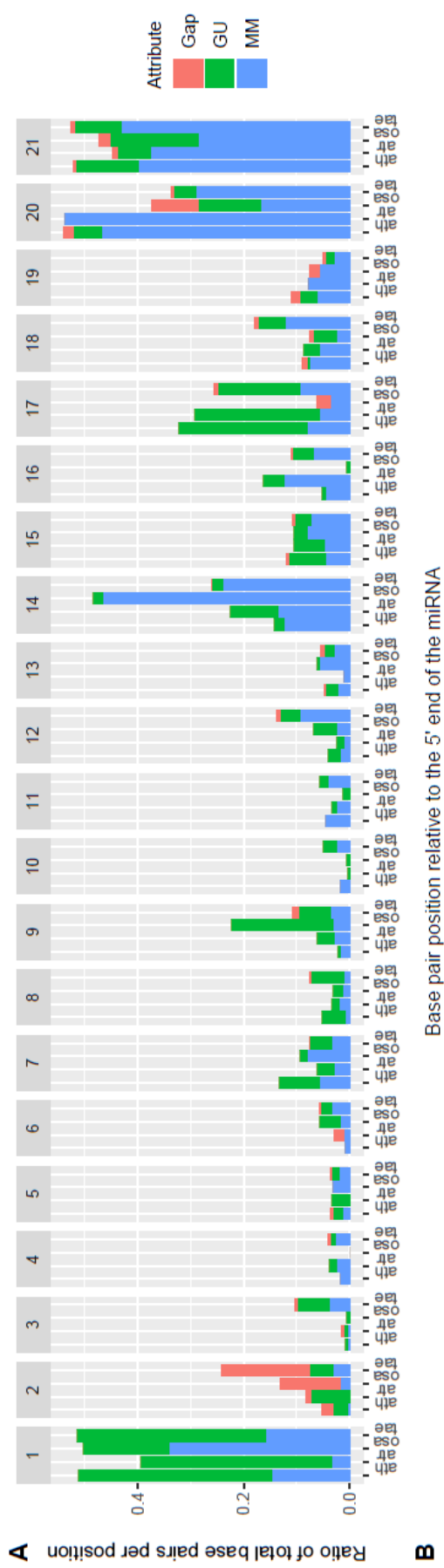


Fig. 4.3 Side-by-side comparison of flower miRNA-mRNA interaction property distributions in monocots and dicots. The position-specific properties (A) and MFE ratio distribution (B) of miRNA-mRNA interactions from flower tissues in *A. thaliana*, *A. trichopoda*, *O. sativa* and *T. aestivum*. The position-specific properties showed significant differences at certain positions and there is a clear separation in the MFE distributions between monocots and dicots.

The distributions of position-specific properties in flower tissue show interesting variations between species. To evaluate whether the non-model organism distributions differ from the *A. thaliana* distributions, we used the offset χ^2 test and a localized Fisher's exact test, presented in Appendix B Table 12. The former show significant differences at positions 1, 9, 14, 17 and 20 in *O. sativa* and position 2 in *T. aestivum*. The results of the localized Fisher's exact test show significant differences at positions 1, 14 and 20, and positions 1, 9 and 17 in *O. sativa* for mismatches and G:U wobble pairs, respectively. Moreover, the Fisher's exact test reported a significant difference in the proportion of gaps at position 2 in *T. aestivum*. Alongside the position-specific properties, the MFE ratio was also investigated as a discriminative feature (appendix B Figure 2B) and the Kolmogorov–Smirnov test was used to evaluate differences between distributions of different species. The distribution of MFE ratios and results of the statistical test, presented in Appendix B Table 13, illustrates the differences between monocots and dicots, with significant differences only reported when comparing different groups. The identification of these subtle differences when compared to *A. thaliana* support the conclusion that species-specific and data driven criteria could facilitate a better description of the miRNA–mRNA interactions.

The differences observed between conserved and species-specific miRNAs in *A. thaliana* prompted a similar investigation in other, non-model organisms. Similarly, as for *A. thaliana* miRNA interactions, we classified the miRNAs that had HC predicted interactions, as reported by PAREameters, into conserved or species-specific for each of the non-model organisms. Specifically, miRNAs present only in an individual clade, based on current miRBase annotations (Release 22) [110], were considered species-specific; otherwise they were classified as conserved. The conservation analysis was done against the current miRNA variants from miRBase, allowing up to two mismatches, at any positions, and no gaps. If a miRNA predicted on a non-model organism dataset did not match any miRNA variant in miRBase or

aligned only to a known species-specific miRNA, then it would be classified as a species-specific, otherwise the miRNA was classified as conserved. The summaries of the position-specific properties distributions and MFE ratio distributions for each of the non-model organisms are presented in Appendix B Figures 3–7. The results of the significance tests comparing the conserved and species-specific properties are presented in Appendix B Tables 14–18.

To illustrate the impact of the differences between targeting properties and subsequently inferred targeting criteria in non-model organisms, we focus on the results in *T. aestivum*, presented in Appendix B Figure 7 and Appendix B Table 18. Out of the 21 positions analysed, 7 had significant differences based on the χ^2 tests (the conserved properties were considered the expected distribution and the species-specific properties were the observed distribution), with three of these differences in the miRNA core region (positions 2, 3 and 12). Also showing a significant difference were the MFE ratio distributions, evaluated using the Kolmogorov–Smirnov test, which reported a p -value of $p < 0.001$. Also, other non-model organisms showed significant differences within the miRNA core region, for example in *O. sativa* inflorescence (Appendix B Figure 6 and Appendix B Table 17). Moreover, significant differences between the MFE ratio distributions are also observed in *A. trichopoda* flower (Appendix B Figure 4) and *G. max* leaves (Appendix B Figure 5).

4.4.5 Employing data-driven targeting criteria on non-model organisms

To evaluate the differences in number and identity of predicted miRNA targets when using the Allen *et al.* and PAREameters inferred criteria on the non-model organisms, we performed target prediction using PAREsnip2. The inferred criteria were able to capture a larger number of interactions; the only exception was

observed for the D6 (*O. sativa*) dataset for which 149 interactions from 42 miRNAs were found using the Allen *et al.* criteria and 115 interactions from 33 miRNAs using the inferred rules with an overlap of 100%. The larger number of interactions reported for the D5 (*G. max*) and D7 (*T. aestivum*) datasets when compared to D4 (*A. trichopoda*) and D6 (*O. sativa*) may have arisen from number of repeat regions or duplicated transcripts present within the current genome annotation.

We then investigated the overlap between the miRNAs and their interactions for each set of criteria, presented in Appendix B Table 19, and concluded that, except for D6 (*O. sativa*), a higher number of miRNAs and their interactions were specific to the inferred criteria, highlighting yet again the distance from the Allen *et al.* criteria. For the above analysis, we used the default retain rate of 0.85 so to explore its effect on the overlap between the Allen *et al.* criteria and the inferred criteria, we repeated the analysis using a retain rate value of 1, to capture all PAREameters reported HC interactions. All of the captured interactions using the Allen *et al.* criteria were a subset of the interactions captured by the PAREameters inferred criteria when using a retain rate of 1 (Appendix B Table 20); the increase in miRNAs with targets varies between 4 (D6) and 102 (D7) and the increase in reported interactions varies between 12 (D6) and 783 (D7), depending on the organism or dataset. These results further suggest that the Allen *et al.* criteria may have been too stringent, or inadequately calibrated for the specific organism or miRNAs in question.

4.5 Discussion

The comparison of validated miRNA–mRNA interaction properties between conserved and species-specific miRNAs in *A. thaliana* highlighted interesting and perhaps previously unknown differences. When investigating the features of conserved miRNA interactions, we observed similar patterns to that of Allen *et al.*

[3] regarding complementarity in the core region of the miRNA (2–13) and at the canonical position 10. This observation is further supported by a recent study of highly conserved miRNAs in *N. benthamiana* [137], where it was shown that a single mismatch at the 5' end of miR160 significantly diminished target site efficacy, and two or more consecutive mismatches at the 5' end fully abolished it. Furthermore, the authors highlighted that a single-nucleotide mismatch at positions 9 and 10, in addition to combinations of mismatches at positions 9, 10 and 11 led to the complete elimination of the responsiveness of miR164. However, the species-specific miRNAs tended to tolerate more flexibility at these positions. These results motivated a similar analysis in non-model organisms and the results of which did mirror the trends observed in *A. thaliana*. However, it is important to emphasize that these result from a series of predictions, and are subject to changes from additional validations. Nonetheless, this output highlights, yet again, the potential differences in the range of suitable thresholds used for predicting targets for subsets of miRNAs and reiterate the remark that one set of fixed criteria for inferring miRNA–mRNA targets may not be sufficient.

Throughout this chapter, we used exclusively the HC interactions, reported by PAREameters, for all comparative analysis. This is in part because the strongest degradation signal on a transcript is likely a result of miRNA cleavage and focusing on this subset of interactions increases the confidence in the prediction results. However, it has been shown that weaker/lower abundance degradation signals may also be caused by miRNAs; these can be captured during target prediction, albeit with lower prediction confidence. These lower abundance signals may be a result of lower miRNA expression, reduced cleavage efficiency or even sequencing bias [190]. Indeed, it is also possible that the degradation fragments may not be caused by miRNA cleavage but instead are a result of noise or random degradation of the transcript. It has been shown that real miRNA cleavage sites tend to be conserved across biological replicates and therefore, we further tested the hypothesis that

the properties of genuine miRNA–mRNA interactions will be consistent between biological replicates. To investigate this, we re-ran the analysis of the *A. thaliana* D1 dataset, allowing both HC and LC interactions to be reported, and compared the results, across replicates, using the same statistical evaluations, as described in the methods.

The outcome of these analyses, presented in Appendix B Figures 8 and 9, show a consistent decrease in the number of LC interactions reported compared to the number of HC interactions and a higher variability in distributions of properties, across replicates, for the LC interactions. This remark supports our previous observation that genuine miRNA cleavage signals are likely to have the strongest signal (Category-0 or 1) on transcripts. The consistency of the MFE ratio distributions and the position-specific properties of HC interactions between replicates is remarkable, with no significant differences in properties reported (Appendix B Table 21), supporting our previous hypothesis that genuine miRNA cleavage sites are conserved between biological replicates. Conversely, when comparing the property distributions of LC interactions between replicates, we observe a higher variation in the proportions of interactions with specific properties, however no significant differences were reported by the statistical tests (Appendix B Table 22). We speculate that the cause of these variations of properties between replicates is due to the higher proportion of putative false positive predictions, i.e. the Category-2 and 3 interactions comprise a combination of genuine target sites and random degradation illustrated by the lower abundance of the transcript degradation signals.

When performing an investigation into the differences between properties of conserved and species-specific miRNA interactions reported by PAREameters in the non-model organisms, we identified statistically significant differences in the *O. sativa* and *T. aestivum* datasets. However, as these results are based solely on predictions made using the degradome, it is difficult to determine if these observations are caused by genuine biological differences or if they are a result of

prediction artefacts. Nonetheless, we hope that these observations will prompt the creation of an extensive set of experimentally validated miRNA-mRNA interactions in a wide range of tissues and treatments from various species that could give a more conclusive answer to whether there exists differences between conserved and species-specific miRNA interactions.

In this chapter, we also highlighted that targeting criteria inferred on non-model organisms or subsets of interactions are less compatible with current fixed criteria and often lead to a decrease in sensitivity. Given the current, limited understanding of the miRNA-mRNA interactions in various species, it is difficult to propose a biological interpretation of these variations, however, based on the side-by-side analysis of various datasets, we can conclude that a customized selection of parameters may result in a higher precision output that could facilitate a more detailed overview of regulatory interactions and an in-depth assessment of the underlying regulatory networks. Furthermore, the differences observed in the flower tissue between monocots and dicots emphasize the usefulness of data-inferred, species and tissue specific thresholds. We have demonstrated that PAREameters is applicable for a wide variety of experimental designs in both model and non-model organisms and could enable further understanding of the subtle variations in miRNA-mRNA interactions in different species, tissues and treatments. In addition, this novel data-driven approach may enable new discoveries, i.e. regulatory sequences or modes of action, within the RNA silencing pathways.

4.6 Conclusion

In this chapter, we describe PAREameters, a novel approach and tool that enables data-driven inference of plant miRNA targeting criteria that can be used by PAREsnip2. Through refining the targeting criteria, the discovery and characterization of new miRNA-mRNA interactions per tissue or organism (both model and non-

model) becomes possible. When evaluating the performance of the PAREameters inferred criteria, we observed an increase in sensitivity compared to the Allen *et al.* criteria over all the *A. thaliana* datasets, whilst also maintaining precision on most datasets, when benchmarked against a set of experimentally validated miRNA–mRNA interactions. In the next chapter, we describe a new software tool, called NATpare, that we developed to predict nat-siRNAs from sRNA and degradome sequencing data.

Chapter 5

High-throughput prediction and functional analysis of nat-siRNAs using the degradome

5.1 Summary

Throughout this thesis, we have used degradome data as a resource for improving confidence when predicting sRNA targets. However, this data can also be used to capture cleavage products generated through Dicer-mediated processing of sRNA precursors, as demonstrated with miRNA biogenesis [134, 2, 225]. In this chapter, we describe a new software tool, called NATpare, that we developed to predict nat-siRNAs from sRNA and degradome sequencing data. NATpare takes sRNA, transcriptome and, optionally, degradome data as input and enables the identification of both *cis*- and *trans*-nat-siRNAs. It is scalable with the increasing size of modern sequencing datasets and enables comprehensive analysis of nat-siRNAs in more complex transcriptomes for the first time within a reasonable time frame. In addition, if corresponding degradome data is available, NATpare provides the reported nat-

siRNAs to PAREsnip2 for prediction of potential mRNA targets based on evidence within the degradome.

We start by introducing the background followed by a description of the methods that we used to create the tool. After this, we perform computational and prediction performance benchmarking of NATpare and compare the results with that of another publicly available tool for this type of prediction. We then perform prediction and differential expression analyses on control and stress treated samples in *Arabidopsis thaliana*. Finally, we perform functional analysis, using PAREsnip2, of *cis*- and *trans*-nat-siRNAs in multiple *A. thaliana* tissues before concluding with a discussion.

This chapter is an adapted version of "NATpare: a pipeline for high-throughput prediction and functional analysis of nat-siRNAs.", which is published in Nucleic Acids Research [199].

5.2 Background

Natural antisense transcripts (NATs) are endogenous RNA transcripts that share sequence complementary with other RNA transcript sequences [60]. They have been identified in multiple eukaryotes, including *Homo sapien*, *Mus musculus*, *Saccharomyces cerevisiae*, *Oryza sativa* and *A. thaliana* [204]. NATs include both protein coding (PC) and non-protein coding (NPC) transcripts [118] and can be classified into either *cis*-NATs or *trans*-NATs based on their genomic origin. *cis*-NATs are transcribed from the same genomic location but on opposite strands, resulting in sections of perfectly complementary dsRNA forming from the two transcript sequences. Conversely, *trans*-NATs originate from different genomic locations and can form imperfect dsRNA [204]. There are three types of NAT orientation that can form dsRNA: 5' overlap (head-to-head), 3' overlap (tail-to-tail)

and the complete enclosure of one transcript by the other (full overlap) [118], shown in Figure 5.1. Although current understanding is limited, research has suggested a variety of regulatory roles for NATs, such as RNAi, alternative splicing, genomic imprinting, and X-chromosome inactivation [204, 31, 26].

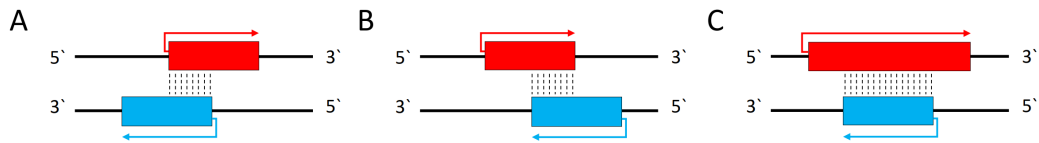


Fig. 5.1 The three types of NAT orientation that can form dsRNA: 5' overlap (head-to-head) (A), 3' overlap (tail-to-tail) (B) and the complete enclosure of one transcript by the other (full overlap) (C). Transcript sequences are always transcribed in the 5' direction and are represented by coloured arrows. Regions of complementarity between the two sequences are represented by dashed lines.

Over the last few years, much research attention has been focused on the biogenesis and function of nat-siRNAs [26, 87, 172, 227, 233]. The founding example was identified in *A. thaliana*, where a pair of *cis*-NATs, SRO5 and P5CDH, were shown to be involved in the response to salt tolerance through the RNAi pathway [26]. During salt stress, SRO5 is expressed and can form a complementary overlapping region with the constitutively expressed P5CDH, which is then processed by a biogenesis pathway dependent on Dicer-like 2 (DCL2), RNA-dependent RNA polymerase 6 (RDR6), Suppressor of Gene Silencing 3 (SGS3) and DNA-directed RNA polymerase IV subunit 1 (NRPD1) to produce a 24nt nat-siRNA. This nat-siRNA then directs the cleavage of P5CDH, which is subsequently used as a template by RDR6 to produce dsRNA that is then processed by DCL1 to produce 21nt secondary nat-siRNAs [26].

In 2012, Zhang *et al.* [231] performed a genome-wide analysis of plant nat-siRNAs in both *O. sativa* and *A. thaliana*, which revealed insights into their distribution, biogenesis and function. In this study, more than 17 000 unique siRNAs corresponding to *cis*-NATs from biotic and abiotic stress challenged *A. thaliana* and 56 000 from abiotic stress treated *O. sativa*. These siRNAs were enriched in

the overlapping region of NAT pairs and displayed either site-specific or distributed patterns.

Current tools available for the prediction of NATs and nat-siRNAs are limited in both number and functionality. NATpipe [224] suffers from limitations in its run-time and also requires a large number of third party dependencies that must be installed and configured by the user. This requires computational expertise that some users may not have. Additionally, NATpipe is developed to exclusively discover phased-distributed nat-siRNAs, however based on a previous study [231], nat-siRNAs production can also follow site-specific patterns and thus would be missed by NATpipe. Moreover, the results reported by NATpipe do not give any indications into the possible function of any predicted nat-siRNAs. Finally, based on our prediction performance benchmarking, limitations with the implementation of the NATpipe algorithm causes some known *cis*-NAT pairs and their corresponding *cis*-nat-siRNAs to be discarded.

5.3 Methods

The NATpare algorithm is split into four main stages with the final stage being optional and dependent on the input data. The first is the pre-processing of input sequencing data and the approaches taken to reduce the possible search space. The second stage is the identification of potential NAT pairs. In the third stage, potential nat-siRNAs are identified and additional quantitative information is extracted and reported. Finally, and if degradome data is provided, the candidate nat-siRNAs are subject to functional analysis using PAREsnip2 to search for potential mRNA targets. A visual overview of the steps involved for performing analysis on the input data is shown in Figure 5.2. We now explain each stage of the algorithm in more detail.

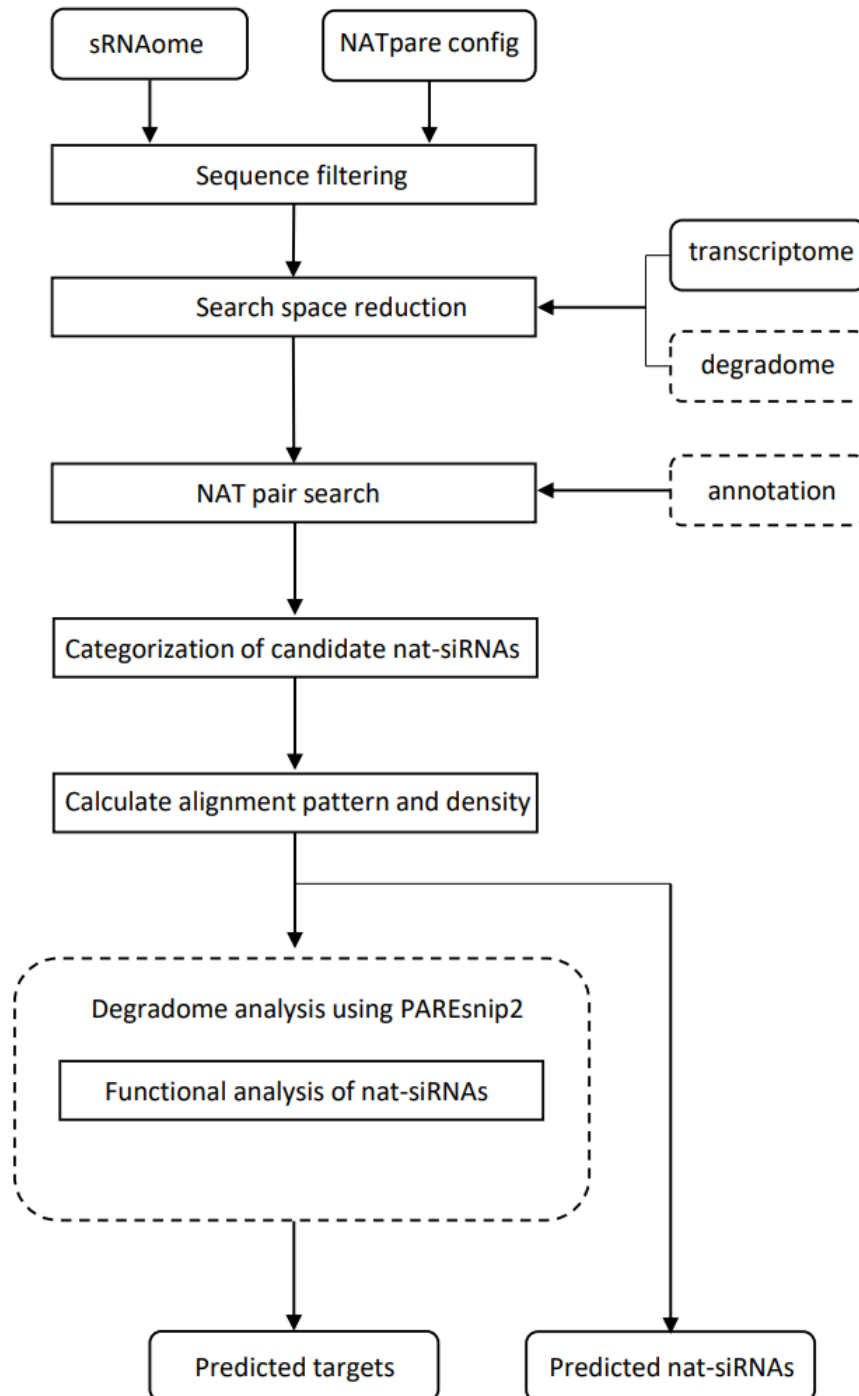


Fig. 5.2 A visual overview of the NATpare pipeline. Input and output data are represented by rounded rectangles and processes are represented by straight rectangles. Data input or processing steps surrounded by dashed lines are optional and dependent on the provided input data. NATpare takes as input HTS data (sRNA and degradome) along with a reference transcriptome and outputs a list of predicted nat-siRNA. Additional annotation information, in the form of a GFF3 file, can be used to annotate the predicted NATs (*cis* or *trans*).

5.3.1 Data input and configuration

To perform an analysis using NATpare for a specific organism, the user must input the following data:

- A reference file (transcriptome) in either FASTA or Generic Feature Format version 3 (GFF3) with the genome sequence in the GFF file;
- A genome file (optional unless using GFF3 as reference);
- A set of sRNAs in redundant FASTA format
- A degradome library in redundant FASTA format (optional)

A reference file and at least one sRNA library are required to perform analysis. If the user chooses to use a GFF3 file as a reference then a corresponding genome must also be provided. When extracting the gene sequences from the genome using information from the provided annotation (GFF3), the tool will include all splice variants of a given transcript that are detailed within the annotation. The input sRNA library must be in redundant FASTA format with the adaptors trimmed. Tools available to processing FASTQ files, such as adaptor trimming and other quality checking, can be found in the UEA sRNA Workbench [192], where NATpare is also implemented. When performing analysis, the user has the option to configure a number of parameters to meet their requirements, which are shown in Table 5.1. The most notable parameters are the number of expected sRNA phases, which is defined as the number of expected adjacent sRNAs, with or without overlap, that align to a given transcript for it to be reported, as shown in Figure 5.3, and the minimum overlap length between two NATs (i.e. the minimum overlap length considered possible to produce sRNAs).

Parameter	Default Value	Description
Minimum overlap length	100	Minimum length of the annealed region between NATs
Minimum sRNA phases	1	Minimum number of sRNA alignment phases (shown in Figure 5.3)
Minimum sRNA length	19	Minimum input sRNA length
Maximum sRNA length	24	Maximum input sRNA length
Minimum sRNA abundance	1	Minimum input sRNA abundance
Minimum tag length	19	Minimum length of degradome reads
Maximum tag length	21	Maximum length of degradome reads
<i>cis</i> only	true	Only search for NATs with perfect complementarity or from the same genomic location
Coverage ratio	80%	The percentage of overlap required between the BLAST and RNAPlex alignments
Largest bubble region	10%	Largest non-complementary region in a <i>trans</i> -NAT alignment cannot be longer than 10% of the total alignment
Low complexity filter	true	Discard input sequences based on their complexity
Genome alignment	true	If a genome is provided, discard any sRNAs that do not align

Table 5.1 The configurable parameters for NATpare. The values used during analysis can be changed by modifying the input configuration file or by using the command line when running the tool.

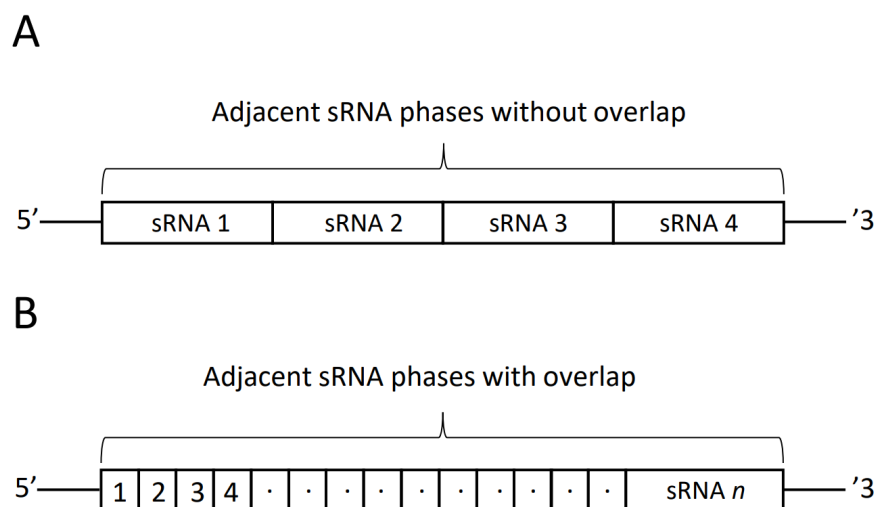


Fig. 5.3 The two types of adjacent sRNA alignment phases considered by NATpare. Adjacent sRNA phases without overlap are when the first position at the 5' end of an aligned sRNA is adjacent to the last position at the 3' end of another aligned sRNA. Adjacent sRNA phases with overlap are where sRNA sequences align contiguously to a given transcript.

5.3.2 Sequence filtering

Several optional filtering techniques can be applied to the input data to remove low quality reads, sequencing errors or sample contamination. First, any sequence containing ambiguous bases are discarded as they cannot be accurately aligned. Second, a low complexity sequence filter is applied based on the sequence composition, described in Section 3.3.2. Specifically, this works by discarding any sequences that contain more than 75%, 37.5% and 25% of a single, di- or tri-nucleotide composition, respectively. Finally, if a genome is provided, sRNA sequences can be aligned using PatMan [165], with any sequences that do not align being discarded.

5.3.3 Search space reduction

A core component of the NATpare algorithm is the pre-processing of the input data to reduce the possible search space and thus reduce the required run-time of a given

analysis. In the first step, the sRNA and optional degradome libraries are aligned to the provided transcript sequences in the positive direction with no mismatches allowed. For this, we use the Binary Search Alignment algorithm described in 3.3.4. Next, we extract sub-sequences based on the following criteria:

- Adjacent aligned sRNA sequences, either at the 5' end or 3' end, that meet the minimum number of expected phases (configurable parameter)
- If provided, degradome tags where the first position aligns adjacent to the 3' position of an aligned sRNA, which results in a ~ 40 nt sequence

The use of degradome data is to find DCL-mediated cleavage evidence and to determine those sRNA that may be site-specific, i.e. there is a preferential DCL cleavage site, based on the types of distribution patterns found in a previous study [233].

Once the longer sub-sequences that meet either of the above criteria have been extracted, we take their reverse complement and perform exact match sequence alignment to all other transcripts using PatMan [165]. This process gives us potential overlapping regions, that may give rise to sRNAs, between two transcripts and are then subject to more a comprehensive analysis.

5.3.4 NAT pair search

After the generation of the candidate NATs from the search space reduction technique, they are subject to an alignment search using BLAST [99]. If the alignment length is greater than or equal to the expected minimum, the NAT pair is then identified as either *cis* or *trans*. If a GFF3 file is provided as input this will be determined by the genomic origin of the two transcripts, otherwise it will be determined based on previously described criteria [224]. Specifically, if the overlapping region is perfectly complementary, it will be considered as a *cis*-NAT, otherwise it will be

considered as a *trans*-NAT, albeit without genomic location information. In the case of *trans*-NATs, the reported alignment is further analysed using RNAplex [195] to verify the annealing potential of the BLAST-predicted alignment at the secondary structure level. The results from RNAplex must meet the following criteria for the NAT pair to be considered for further analysis:

1. The reported annealing region should overlap with the BLAST reported complementary region by at least 80% (configurable parameter)
2. Any unpaired region within the annealing region should be no longer than 10% (configurable parameter) of the total length of the overlapping region

Unlike NATpipe, we only do the hybridization analysis if the reported BLAST alignment or genomic location information suggests that the NAT pairs work in *trans*. In addition and to compensate for the long processing time of RNAplex, if the length of either transcript of the NAT pair is greater than 5000nt, we omit the hybridisation step and instead just proceed with the reported BLAST alignment.

Once all of the candidate NATs have been processed, those passing all the required criteria are categorized into the following groups:

- High-coverage (HC): the complementary region is longer than 50% of the length of either transcript
- 100nt: the complementary region is 100nt or longer in length
- Low-coverage (LC): the complementary region is less than 100nt in length

5.3.5 Categorization of candidate nat-siRNAs

Once the overlapping regions between NATs have been determined, the pipeline extracts the sRNA sequences that aligned to these positions. Rather than just

providing the user with a set of aligned sRNAs, we developed a system to categorise each sRNA based on the current understanding of the sRNA biogenesis model. For this system, we also include degradome data (if provided) as this provides a snapshot of the mRNA degradation profile, which can include Dicer-mediated cleavage products [134, 2, 225]. In addition, by looking at the degradation profile, it can also give us an indication as to what mRNAs are currently being expressed, as the mRNA must be expressed in order to be degraded, and thus improve our nat-siRNA prediction model.

For each biogenesis group, we define the mature sRNA as the one originating from the transcript currently being investigated. For example, given the NAT pair consisting of transcripts A and B, when investigating sRNA alignments to transcript A, those sRNAs aligning to B will be considered the star sequences, and vice-versa when investigating transcript B, those aligning to A will be considered the star sequence.

- Group 1: sRNA and sRNA* sequence present with 2nt 5' overhang and both sequences supported by the degradome data
- Group 2: sRNA and sRNA* sequence present with 2nt 5' overhang and only mature sequence supported by the degradome data
- Group 3: sRNA and sRNA* sequence present with 2nt 5' overhang
- Group 4: sRNA present with degradome evidence but no sRNA*
- Group 5: Only the sRNA aligning to the overlapping region

5.3.6 NAT alignment distribution and sRNA alignment densities

To determine the distribution pattern of aligned sRNAs for a given NAT pair, we implemented a method described previously [231]. Specifically, starting from the first aligned sRNA closest to the 5' end of a transcript, sRNAs are clustered if their first nucleotide is within a 10nt long segment of the starting sRNA, with any cluster containing more than 5 reads being retained for further analysis. For each NAT, we record the number of clusters and the percentage of the unique reads in these clusters relative to the whole transcript. Alignments are considered to be site-specific if a transcript contains 10 or less clusters and the percentage of unique reads within these clusters is 50% or greater than that over the whole transcript, otherwise it is categorized to have a distributed pattern.

For each NAT pair, we also report the sRNA alignment density for the overlapping region and for the whole transcript. To do this, we implement the same methods as described previously [231]. Briefly, for each NAT pair, we counted the number of unique sRNAs, denoted as N_o , mapping to the overlapping region and the total number, denoted as N_g , mapping to both transcripts. We then measured the length of the overlapping region, denoted as L_o , and the sum of the length of both transcripts, denoted as L_g . Finally, the ratios N_o/L_o and N_g/L_g were reported as the sRNA alignment densities for the overlapping region and for the overall transcript sequences within the NAT pair, respectively.

5.3.7 Functional analysis of candidate nat-siRNAs

It has been shown that *cis*-nat-siRNAs can direct the cleavage of their mRNA targets [26]. Therefore, to provide further indication of the function of the reported nat-siRNAs and if degradome data is provided as input, we incorporate the predicted

nat-siRNAs into PAREsnip2. For target prediction, we allow the user to configure their own parameters or alternatively use the default configurations provided in PAREsnip2. Additionally, if the user has a version of R installed and is correctly configured on their PATH, the pipeline can automatically produce t-plots to provide a visual representation of the reported interactions.

5.3.8 Implementation and output

The algorithm has been implemented using the Java programming language and a user-friendly, cross-platform software package has been incorporated into the UEA sRNA Workbench [192]. Analysis using NATpare can be performed through the command-line interface as a standalone application or alternatively be incorporated into larger and more complex bioinformatics pipelines or workflows. The results of NATpare are provided in comma-separated value (CSV) format, allowing them to be viewed in any CSV file viewer.

5.3.9 Sequencing datasets

The sequencing datasets analysed in this chapter are described in Appendix C Table 1.

To enable a comprehensive evaluation of the NATpare tool, we performed computational benchmarking on multiple plant species with varying transcriptome sizes (Appendix C Table 2), including *A. thaliana*, *Solanum lycopersicum*, *O. sativa*, *Glycine max* and *Triticum aestivum*. The transcriptome used for all species in the computational performance benchmarking were extracted from genome and GFF files obtained from Ensembl Plants [22]. 100 000 sRNA sequences were used in the computational benchmarking for each species and were simulated from the overlapping region of two randomly selected *cis*-NAT pair, based on the genomic

information provided within the genome annotation. All generated sequences were 21nt in length and were randomly selected to be extracted from either transcript within the NAT pair.

For the prediction performance comparison between NATpare, NATpipe and those reported by a previous study, we used the *G. max* datasets (sRNA and transcriptome) described in Appendix C Table 1. The control and stress treated *A. thaliana* sRNA sequences that were used for the seedling salt stress analysis were obtained from [13] and the flower, root, seedling and leaf libraries, with corresponding degradome data, were obtained from [83]. For all *A. thaliana* analysis, besides from the computational benchmarking, we used the TAIR10 reference transcriptome [115].

5.4 Results

5.4.1 Benchmarking and comparison with NATpipe

To measure the computational performance of the newly developed NATpare algorithm, i.e. the time and memory required to perform an analysis, we carried out computational benchmarking and compared our results to those of the other publicly available method. This benchmarking was performed on a desktop computer running Ubuntu 18.04 equipped with a 3.40GHz Intel Core i7-6800K six core CPU and 128GB RAM.

For this benchmarking, we used the simulated set of sRNA sequences and the reference transcriptome, produced using the GFF file obtained from Ensembl [22], as described in the methods, for each species. The reason that we used simulated data is that it allows us to generate nat-siRNAs that we know should be captured by the tools and thus allows for the fairest possible comparison. As NATpipe

can only predict nat-siRNA originating from *cis*-NATs, we adjusted the NATpare parameters to also have this restriction. We recorded the time taken for each tool to perform analysis on the simulated data and the results of these analyses are shown in Table 5.2. If a tool did not complete the analysis within 10 days, we recorded it as did not finish (DNF). The results show that the newly developed algorithm substantially outperforms NATpipe on the simulated datasets in terms of computation time. For the *A. thaliana* dataset, the only dataset that NATpipe was able to complete within the 10 day cut-off limit, the newly developed method was able to complete the analysis 227 times faster. For all tested datasets, the memory requirement varied between 4GB and 8GB depending on the number of transcript sequence within the reference annotation. The timing results suggest that the time taken is dependent on the number of transcripts and transcript pairs that contain overlapping and complementary regions, for which the exact number is difficult to determine, particularly when you consider *trans*-NATs, as this information is not possible to obtain, even with a complete genome annotation, without thorough computational analysis. However, the results of the computational performance benchmarking demonstrate NATpipe’s speed limitations and the need for additional pipelines or software tools for the prediction of nat-siRNAs.

Species	Annotation	# Transcripts	NATpipe	NATpare
<i>S. lycopersicum</i>	SL3.0	33925	DNF	4m 52s
<i>O. sativa</i>	IRGSP-1.0	42378	DNF	5m 38s
<i>A. thaliana</i>	TAIR10	48359	1d 18h 34m	11m 15s
<i>G. max</i>	<i>G. max</i> v2.1	88412	DNF	1h 5m
<i>T. aestivum</i>	IWGSC	133744	DNF	13h 2m

Table 5.2 Computation performance comparison between NATpipe and the newly developed NATpare pipeline when evaluated on the simulated datasets. If the tool did not finish within 10 days it was recorded as did not finish (DNF).

Next, we wanted to evaluate the predictions reported by the tools on real sequencing data. However, unlike other classes of sRNA, such as miRNAs, there

is no extensive set of true positives to evaluate against. Nevertheless, a number of previous studies have manually predicted NATs and nat-siRNAs in both model and non-model plants, for example, *A. thaliana* [231], *G. max* [235] and *Z. mays* [218]. As NATpipe is currently the only publicly available tool for the prediction of nat-siRNAs, we performed an analysis on a publicly available *G. max* dataset and investigated the overlap in the number of nat-siRNAs reported by computational methods, NATpipe and NATpare, and those found previously during manual analysis [235]. For this analysis, we used the *G. max* cDNA reference transcriptome, obtained from Phytozome and the D1 sRNA dataset, as described in the methods. In addition, to compensate for the long processing time required by NATpipe and the fact that it is only able to predict *cis*-nat-siRNAs, we restricted the input transcript sequences only to those with perfectly complementary overlapping regions, as reported by a BLAST search using those transcripts previously found to produce nat-siRNAs [235] as input.

The results from the top 10 NAT pairs, based on number of generated nat-siRNAs, are presented in Table 5.3 and the rest in Appendix C Table 3, show that NATpare is able to capture a larger number of the previously reported nat-siRNAs in *G. max* compared to NATpipe. To investigate the overlap in results between the two tools, we compared the results and found that all of the NATpipe reported nat-siRNA were a subset of those reported by NATpare. In addition, further investigation into the NAT pairs missed by NATpipe showed that the RNAplex hybridization step of the algorithm did not always correspond to the alignment reported by BLAST, thus no results were reported, which supports our decision to perform RNA hybridization exclusively on *trans*-NATs. Interestingly, we observed differences between the numbers of reported nat-siRNAs from the previous study [235] and the prediction tools and consider this likely to be a result of minor discrepancies between the different filtering and prediction methods applied to the input sRNAs.

Gene A	Gene B	Overlap length	Zheng <i>et al.</i>	NATpare	NATpipe
Glyma13g11940	Glyma13g11970	542	1864	1802	0
Glyma13g11820	Glyma13g11830	428	1285	1406	0
Glyma13g11940	Glyma13g11950	147	724	576	0
Glyma13g11940	Glyma13g11960	118	509	487	0
Glyma11g30060	Glyma11g30070	392	244	237	209
Glyma13g21780	Glyma13g21790	355	28	28	0
Glyma15g06490	Glyma15g06500	156	26	26	0
Glyma17g23860	Glyma17g23870	174	18	11	11
Glyma03g22390	Glyma03g22400	276	17	17	16
Glyma15g37470	Glyma15g37480	764	15	15	0

Table 5.3 Top 10 reported *G. max cis*-NATs with the highest number of unique reported nat-siRNAs by Zheng *et al.* [235] and the prediction results from NATpare and NATpipe

5.4.2 Comparing the expression of nat-siRNAs in *A. thaliana* control and salt stress treated samples

The current understanding of NATs and nat-siRNAs is that they are expressed during certain stress conditions, development stages or disease response [26, 231, 232]. To illustrate the use of NATpare and to validate the results reported by the tool, we performed analysis on a publicly available dataset, D2, obtained from *A. thaliana* seedling under salt stress, a type of abiotic stress in which the plants response has been previously shown to involve nat-siRNAs [26]. Before performing analysis using NATpare and to increase confidence within the predictions, we discarded any sRNAs that were not conserved between at least 2 out of 3 biological replicates. Next, we further filtered the data to remove any known miRNAs or isomiRs by aligning the sRNAs to all known plant miRNAs, obtained from miRBase (release 22) [110], allowing up to 2 mismatches. In addition, we removed any sRNA that may have originated from tRNA or rRNA sequences using the filtering

methods implemented within the UEA sRNA Workbench [192]. The results for this analysis, including the breakdown of the NAT and nat-siRNA prediction categories, can be found in Appendix C Table 4. After performing analysis on the filtered data using NATpare, we then investigated the overlap between the control and treatment samples and the results show that there exists a clear separation in the reported nat-siRNAs between treatment and control, with just 281 overlapping sequences within the intersection, yet 877 and 581 being specific to control and treatment, respectively. As the biogenesis of nat-siRNAs require both transcripts to be expressed simultaneously within the same cell, the separation and differences in the number of nat-siRNAs that are reported between control and treatment may be due to transcriptional changes in response to the stress.

To investigate these results further, we performed differential expression analysis with iDEP [73], using the default parameters, which reported 31 differentially expressed (DE) nat-siRNAs using a false discovery rate of 0.1. These comprised of 29 up-regulated nat-siRNAs in the treatment datasets, presented in Table 5.4, and two up-regulated nat-siRNA in the control datasets. For each of the up-regulated nat-siRNAs identified in the treatment datasets, we examined the current annotation model (TAIR10) and found that 10 of the 29 sequences originated from a NAT pair where one of the transcripts is currently annotated as a potential natural antisense gene. Majority of the other up-regulated nat-siRNAs in the treatment datasets originated from transcripts annotated as either unknown protein or other RNA. Further analysis of all NAT pairs giving rise to DE nat-siRNAs, besides for AT5G01600.1 and AT5G01595.1, showed that the sRNA alignment density within the overlapping region was greater than that of the whole transcript, suggesting that sRNAs are more likely to originate from overlapping regions of these NATs.

Sequence	Originating gene	Annotation	Corresponding NAT	Annotation	log2fc	Adjusted p-value
CAAAAAGTCTGAATCGTCGAGG	AT3G41761.1	other RNA	AT3G41762.1	unknown protein	7.76	$p < 0.001$
CCGGCGACTTTCCGGCGATCGG	AT3G41761.1		AT3G41762.1		7.73	$p < 0.001$
CAAAAAGTCTGAATCGTCGAGGA	AT3G41761.1		AT3G41762.1		6.43	$p < 0.001$
AAAAAGTCTGAATCGTCGAGG	AT3G41761.1		AT3G41762.1		6.21	$p < 0.001$
AAAAAGTCTGAATCGTCGAGGA	AT3G41761.1		AT3G41762.1		6.13	$p < 0.001$
CCGGCGACTTTCCGGCGATCGGT	AT3G41761.1		AT3G41762.1		5.96	$p < 0.001$
CGGCGACTTTCCGGCGATCGG	AT3G41761.1		AT3G41762.1		5.54	$p = 0.002$
CCGGCCCGCGGATTTTCGCCCG	AT3G41761.1		AT3G41762.1		5.28	$p = 0.007$
AAAAAGTCTGAATCGTCGA	AT3G41761.1		AT3G41762.1		4.99	$p = 0.035$
GGCGACTTTCCGGCGATCGG	AT3G41761.1		AT3G41762.1		4.91	$p = 0.062$
CCGGCCCGCGGATTTTCGCCCG	AT3G41761.1		AT3G41762.1		4.22	$p = 0.061$
GGCGACTTTCCGGCGATCG	AT3G41761.1		AT3G41762.1		4.12	$p = 0.081$
AACTGCTGAATCGTCGAGG	AT3G41761.1		AT3G41762.1		3.69	$p = 0.035$
TCCGGCGACTTTCCGGCGATCGG	AT3G41761.1		AT3G41762.1		3.58	$p = 0.001$
AAAAAGTCTGAATCGTCGAGG	AT3G41761.1		AT3G41762.1		3.08	$p = 0.044$
CCGGCCCGCGGATTTTCGCC	AT3G41761.1		AT3G41762.1		2.70	$p = 0.027$
AAACTGCTGAATCGTCGAGGA	AT3G41761.1		AT3G41762.1		2.52	$p = 0.054$
CAAAAAGTCTGAATCGTCGAG	AT3G41761.1		AT3G41762.1		2.44	$p = 0.002$

Table continues on the next page.

Sequence	Originating gene	Annotation	Corresponding NAT	Annotation	log2fc	Adjusted <i>p</i> -value
TAAAGAGAGAACAAAGGATGGTT	AT1G05560.1	UDP- glucosyltransferase 75B1	AT1G05562.1	Potential natural antisense gene	4.46	<i>p</i> = 0.035
GACAAAGTAGAAAAAATGGCG	AT1G05560.1		AT1G05562.1		3.79	<i>p</i> = 0.026
AGTAGAAAAAATGGCGCCA	AT1G05560.1		AT1G05562.1		3.26	<i>p</i> = 0.007
CAAGTAGAAAAAATGGCGCC	AT1G05560.1		AT1G05562.1		3.16	<i>p</i> < 0.001
AAGTAGAAAAAATGGCGCC	AT1G05560.1		AT1G05562.1		2.07	<i>p</i> = 0.024
CAAGTAGAAAAAATGGCGC	AT1G05560.1		AT1G05562.1	UDP-	1.98	<i>p</i> = 0.027
TGAGAAATTTCCGGTTTGGTTT	AT1G05562.1	Potential natural antisense gene	AT1G05560.1	glucosyltransferase 75B1	5.18	<i>p</i> = 0.015
TTGTTTGTGTGGAAAGGTGTG	AT1G05562.1		AT1G05560.1		4.80	<i>p</i> = 0.098
AGACAGATTAGGTAACCTCGAA	AT1G05562.1		AT1G05560.1	Cytochrome b561/ferric	2.20	<i>p</i> = 0.035
GCGGCGGAGAAAGTATGTGGATA	AT3G59068.1	Potential natural antisense gene	AT3G59070.1	reductase transmembrane with DOMON related domain	4.91	<i>p</i> = 0.062
GCCACTACTCCCTCACGGCTCTGC	AT5G01600.1	ferretin 1	AT5G01595.1	other RNA	6.22	<i>p</i> < 0.001

Table 5.4 The differentially expressed nat-siRNAs, as reported by iDEP, in the *A. thaliana* seedling salt-stress dataset. 10 of the 29 sequences originated from NAT pairs where one of the transcripts is annotated as a potential natural antisense gene. The transcript that gives rise to the largest number of nat-siRNAs is currently annotated as unknown RNA and the corresponding NAT has an unknown function. Adjusted *p*-values were obtained using a false discovery rate of 0.1 and were expressed to 3 significant digits. Any extreme *p*-values (i.e. *p* < 0.001) were reported as *p* < 0.001.

5.4.3 Investigation into the function of *cis*- and *trans*-nat-siRNAs in different *A. thaliana* tissues

In a previous study by Yuan *et al.* [228], manual analyses of 40 publicly available *A. thaliana* sRNA datasets obtained from flower, leaf and seedling tissues identified 5385 nat-siRNAs that could be mapped to the overlapping region of a single *cis* or *trans*-NAT pair and were conserved between at least three of the 40 datasets. Of these, 1548 were found to be conserved between each tissue whereas 945 and 142 were specific to seedling and flower, respectively. Analyses into the function of nat-siRNA has shown that they can act as post-transcriptional gene regulators, like miRNAs, by directing the RISC to sequence-specific mRNA targets, usually in *cis* [26, 205]. Degradome data provides experimental support that increases confidence with sRNA target prediction and the NATpare pipeline includes PAREsnip2 for target prediction and functional analysis of reported nat-siRNA candidates. To illustrate the usefulness of combining prediction with functional analysis, we performed analysis using NATpare on the D3 dataset, which consists of two synonymous *A. thaliana* sRNA and degradome biological replicates obtained from each flower, leaf, root and seedling.

For this analysis, and similar to the analysis performed in a previous study [228], we configured NATpare to report both *cis* and *trans*-nat-siRNAs. Similar to our previous analysis, we removed any sRNAs that were not conserved between both replicates and also removed predictions that aligned to any known miRNA, rRNA or tRNA sequences using the UEA sRNA Workbench. After performing analysis on the filtered sRNAs (Appendix C Table 5), we further processed the results to remove any predicted nat-siRNAs that were reported to originate from multiple transcripts. In total, there were 2962, 1505, 2701, 3562 nat-siRNAs candidates reported in flower, leaf, root and seedling, respectively. We then investigated the overlap between the nat-siRNAs reported from each tissue and found that 613

nat-siRNAs (9.6% of all reported sequences) were conserved between each of the tissues. The tissue with the largest number of uniquely reported nat-siRNAs was seedling, with 1438 (22.6% of all reported sequences), and the tissue with the fewest uniquely reported sequences was leaf with just 272 (4.3% of total reads). These results are consistent with those reported by Yuan *et al.* [228], where it was also found that seedling tissue produces the largest number and leaf tissue produces the smallest number of unique nat-siRNAs. A Venn diagram, created by InteractiVenn [86], showing the overlap between all tissues within the D3 dataset can be found in Figure 5.4. Further analysis into the nat-siRNA candidates found that 96.5%, 98.5%, 98.1% and 97.6%, of nat-siRNAs identified in flower, leaf, root and seedling, respectively, were uniquely reported in this study, when compared to those previously reported [228].

To identify the possible function of the captured nat-siRNAs, we performed target prediction with PAREsnip2, using default targeting criteria but without additional filtering (Appendix C Table 6 and Table 3.2), on the dataset D3 degradome libraries. The sRNA input for degradome analysis on each tissue were the captured nat-siRNAs that passed all filtering methods described above. The results of each analysis can be found within Appendix C Table 7. The time taken to perform target prediction on each dataset was 5 minutes with a peak memory usage of 4GB. After performing analysis on each dataset, we extracted the reported targets that were conserved between each of the replicates. This resulted in 6 targets from 4 nat-siRNAs captured in flower, 29 targets from 8 nat-siRNAs captured in leaf, 63 targets from 29 nat-siRNAs captured in root and 35 targets from 9 nat-siRNAs captured in seedling. To exemplify the use of degradome data for functional analysis of the predicted nat-siRNAs, we further investigated the targets reported by the root nat-siRNAs. We found that out of the 63 reported targets, 31, 12 and 1 were also found in seedling, leaf and flower, respectively, suggesting that nat-siRNAs may play both tissue-specific and wide-spread roles.

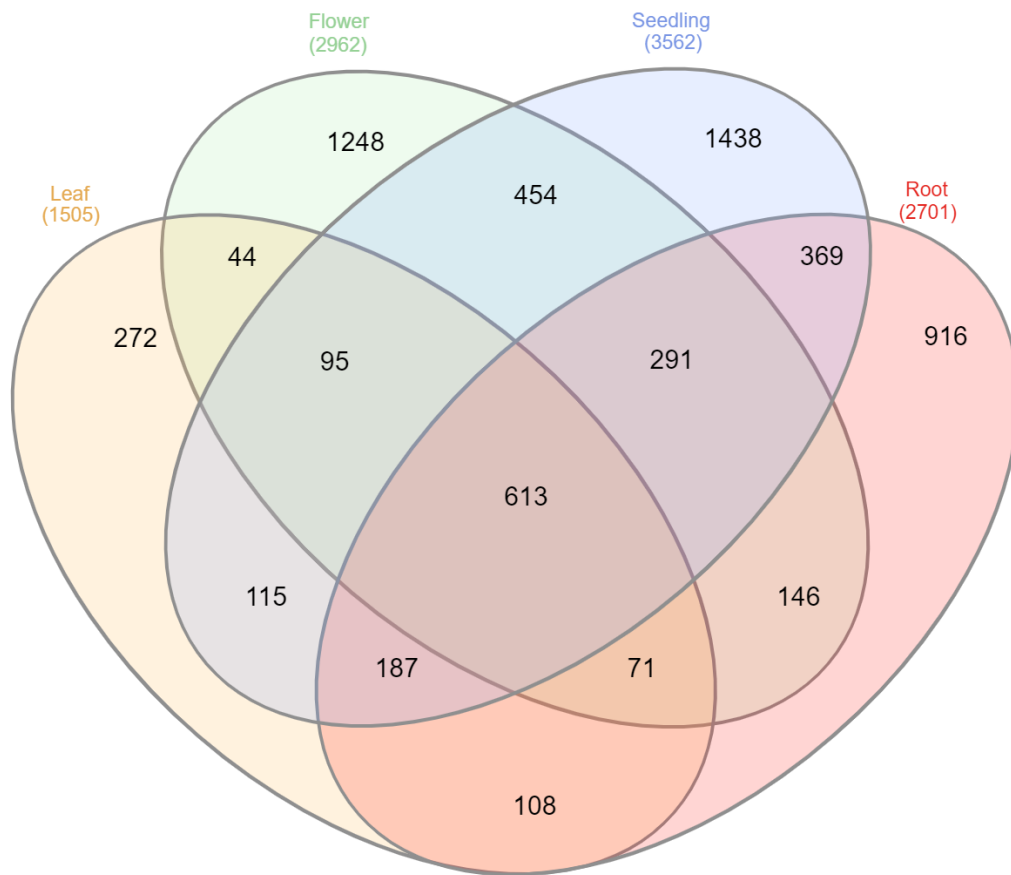


Fig. 5.4 Venn diagram showing the overlap of nat-siRNA predictions between all tissues within the D3 dataset.

5.5 Discussion

Small RNAs that originate from endogenous RNA transcripts that share sequence complementary to other RNA transcript sequences are termed nat-siRNAs, and like miRNAs, they have been shown to regulate the translation of specific mRNAs through mRNA cleavage [26]. Recently, there has been increase in the amount of research focused on classifying this type of sRNA and investigating their possible function. Even so, bioinformatics tools designed to identify nat-siRNAs from high-throughput sequencing data are limited in both number and function.

The founding examples of nat-siRNAs were in *A. thaliana* seedling, where a pair of *cis*-NATs, SRO5 and P5CDH, were shown to be involved in salt tolerance through the RNAi pathway [26]. We demonstrated the use of NATpare by performing analysis on a publicly available *A. thaliana* seedling dataset, consisting of control and salt stress libraries, followed by a DE analysis on the reported nat-siRNAs. Intriguingly, NATpare did not capture the same salt stress responsive nat-siRNAs as reported in a previous study [26] and further investigation showed that the previously found sequences were not present within the more recent salt stress dataset that we analysed. However, we did identify a number of DE nat-siRNAs in salt stress treated *A. thaliana* seedling whose originating transcripts are currently annotated as either potential natural antisense genes, unknown protein or simply described as other RNA. These results suggest that more work is required into the role of these sRNA in salt stress and also additional work into whether nat-siRNAs are specific to salt stress or indeed play a responsive role in plants under various stress conditions. However, based on previous findings [26], the function of these up-regulated nat-siRNAs may be to ensure the down-regulation of the corresponding protein coding transcripts contained within the NAT pair. Additionally, the identification of nat-siRNAs originating from transcripts where the annotation is currently unknown, for example AT3G41762.1, may enable additional annotation information to be included, similar to AT1G05562.1, which is labelled as a potential natural antisense gene in the current annotation.

In plants, post-transcriptional regulation by sRNAs usually result in mRNA cleavage and subsequent degradation. Degradome data is a useful resource for identifying the potential function of a sRNA as it captures the uncapped 5' ends of cleaved mRNAs for sequencing, which can then be aligned back to the reference transcripts and used to identify causal sRNA(s). We used a combination of NATpare and PAREsnip2 on the *A. thaliana* D3 dataset to predict and identify the possible targets of nat-siRNAs that were conserved between two biological replicates in

flower, leaf, root and seedling tissues. In this analysis, we identified a number of interactions, conserved between replicates, which were found to be either tissue-specific or present within multiple of the analysed tissues. However, as these results are based solely on predictions, without further experimental validation it is difficult to determine the exact role or function that these nat-siRNAs play. Nonetheless, bioinformatics approaches to identify possible targets from sequencing data and subsequent validation is a vital step in understanding the function of a sRNA. Thus, we hope that the development of NATpare will lead to further understanding of the origin and function of nat-siRNAs in all manner of experimental contexts.

5.6 Conclusion

In this chapter, we describe a new software tool and pipeline, called NATpare, which is able to perform analyses on recent sRNA sequencing datasets within a reasonable time frame for the very first time. When compared against the only available tool for this type of analysis, NATpare achieved a speed up of 227x (1 day, 18 hours and 34 minutes compared to just 11 minutes and 15 seconds) when benchmarked on a simulated *A. thaliana* dataset. In addition, NATpare was able to complete all analyses of the simulated non-model organism datasets, including *T. aestivum* which took just 13 hours and 2 minutes, whereas NATpipe was unable to complete any non-model organism analysis within the 10 day cut-off. Prediction performance benchmarking of NATpare demonstrated its ability capture a larger number of previously reported nat-siRNAs in *G. max* when compared with NATpipe.

In the next chapter, we exemplify the use of PAREsnip2 by performing degradome analyses on sequencing data obtained from *S. lycopersicum* to better understand the mechanisms by which the plant dies from *Cucumber mosaic virus* D-satellite RNA infection.

Chapter 6

Functional analysis of necrogenic CMV D-satRNA derived sRNA in *Solanum lycopersicum*

6.1 Summary

In Chapter 3, we introduced a software tool developed for analysing sRNA and degradome sequencing data called PAREsnip2. In this chapter, we use PAREsnip2 to perform degradome analyses on data obtained from *Solanum lycopersicum* infected with *Cucumber mosaic virus* (CMV) and D-satellite RNA (D-satRNA) to identify the possible function of necrogenic D-satRNA derived sRNA. We start by introducing CMV and satRNA and then discuss the impact that they have on host plants. We then outline the conditions in which the plants were grown, how they were infected and how the libraries were prepared for sequencing. This is followed by a description of how we pre-processed the sequencing data ready for analysis. We then explain the steps we performed to both analyse the data and to experimentally validate the results obtained. Next, we present the results

obtained from the degradome analyses followed by a comparison of the identified *S. lycopersicum* target sites to that of other species known to survive infection. Finally, we discuss the results from experimental validation of one of the more promising candidate targets.

This work produced in this chapter is from collaboration between the UEA Computational Biology group and Dr. Ping Xu's research group at Shanghai Normal University.

6.2 Background

Plant diseases pose a serious threat to global food security by reducing global food production by more than 10% [193]. Plant viruses are one type of disease-causing pathogen (others include bacteria, fungi and parasitic plants) that can affect the normal development of host plants and sometimes cause rapid plant death, often resulting in complete crop loss [102]. Lethal viral infection has occurred in many important crops such as *Zea. mays* (maize) [177], *Glycine. max* (soybean) [142] and *S. lycopersicum* (tomato) [208, 102]. Tomato is the second most important fruit or vegetable crop, behind potato, and is cultivated for its fresh fruit and processed food products [166]. Currently, China produces the largest quantity of tomato worldwide but viral infection is one of the major limiting factors in its tomato production [217].

Cucumber mosaic virus (CMV) is a plus-strand RNA virus with three RNA genomes, RNA1, RNA2 and RNA3 [96]. These three RNAs encode five proteins, 1a, 2a, 2b, movement protein (MP) and coat protein (CP). While proteins 1a and 2a are responsible for the replication of the virus, protein 2b interferes with the host RNAi pathway [154]. The function of the MP and CP is to allow movement from one infected cell to another [96]. CMV has the largest host range of any known

plant virus, infecting more than 1000 species [57]. Symptoms of CMV can vary depending on the species of plant infected and the environmental conditions but include mosaic pattern on the leaves, stunted growth, and malformation of leaves or other growing points (Figure 6.1).

CMV can also harbour small, linear RNA molecules known as satellite RNAs (satRNAs) [221]. These satRNAs are dependent on CMV for their replication but are not necessary for the survival of the helper virus. They can attenuate or worsen the symptoms induced by CMV in specific plant hosts. For example, B-satRNA and WL1-satRNA induce chlorosis and attenuate symptoms, respectively, in CMV infected *S. lycopersicum* [72]. D-satRNA is another strain of CMV satRNA that induces a lethal systemic necrosis in *S. lycopersicum*, presented in Figure 6.1, and its close relatives, which has been reported as an epidemic in France, Italy, and Spain [102]. In *Nicotiana tabacum*, however, D-satRNA attenuates the symptoms of CMV infection [72]. The mechanisms to which D-satRNA induces necrosis in CMV infected *S. lycopersicum* remain unknown, however the specific nucleotides of D-satRNA responsible for necrosis have been determined [185, 184]. Mutations at these positions, 285 (G to A), 290 (T to G) and 292 (C to T), result in the plant surviving infection and having reduced symptoms of CMV [220].

Recently, a study has shown that a sRNA derived from Y-satRNA reduces the expression of a chlorophyll synthesis related gene through the RNAi pathway, which results in leaf yellowing symptoms in CMV infected *N. tabacum* [182]. Further analyses in *Arabidopsis thaliana* and *S. lycopersicum* infected with CMV and Y-satRNA demonstrated that the leaf yellowing symptoms failed to develop, suggesting a specific interaction between Y-satRNA and *N. tabacum*. However, modification of the Y-satRNA sequence to enable complementarity to the *A. thaliana* and *S. lycopersicum* homologous genes resulted in the development of leaf yellowing symptoms [182].

In this chapter, we perform degradome analyses on sequencing data obtained from CMV and D-satRNA infected *S. lycopersicum* to investigate if D-satRNA derived sRNA may be targeting specific host genes that contribute towards plant death.

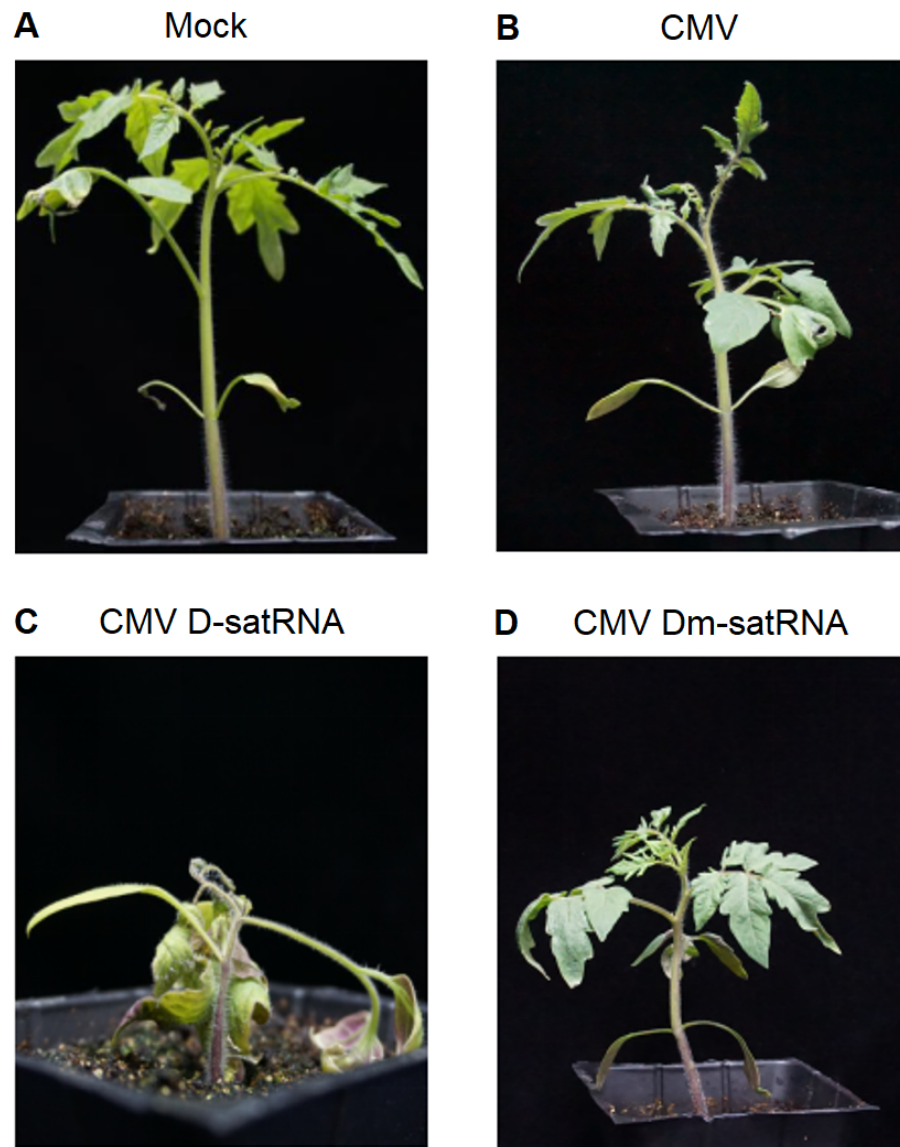


Fig. 6.1 Symptoms of CMV and D-satRNA infection in *S. lycopersicum*. Panel **A** is the control (Mock), **B** is CMV infected, **C** is CMV D-satRNA infected and **D** is CMV D-satRNA infected *S. lycopersicum*. CMV D-satRNA infection induces necrosis while Dm-satRNA attenuates some symptoms of CMV infection.

6.3 Methods

6.3.1 Plant materials and growth

Wild-type tomato (*S. lycopersicum* L. cv. Rutgers) and *Nicotiana benthamiana* were grown with 16 hours of light at 25-28°C and 60%-70% humidity, and 8 hours at night at 15-18°C and 60%-70% humidity. In vitro transcription products of CMV, D-satRNA and Dm-satRNA were obtained using methods previously described [221, 93]. CMV, CMV D-satRNA and CMV Dm-satRNA were inoculated using the friction inoculation method. The plants inoculated with phosphate buffered saline were set as the control group.

6.3.2 RNA extraction, library construction and sequencing

When the second immature stem segment close to the shoot tip of *S. lycopersicum* seedlings infected with CMV D-satRNA showed minimal death spots, the shoot tip, including the stem above the 2nd node, were collected from the plants inoculated with mock, CMV, CMV D-satRNA and CMV Dm-satRNA. The total RNA was extracted using Tri-reagent (Sigma) according to the method provided by the manufacturer. Samples for each group were extracted from 5-10 infected plants or virus-free plant controls and each group consisted of three biological replicates.

For all samples, 10 μ g of total RNA was used for the construction of each sRNA library and 100 μ g of total RNA was used for each degradome library. The sRNA libraries were constructed with HD-adapters using the previously published protocol [21]. The degradome libraries were constructed with the modified 5' RNA adapter following the previously published protocol [230]. The libraries were sequenced using HiSeq2500 platform by Berry Genomics in Beijing, China.

6.3.3 Pre-processing of sequencing libraries

FASTQ files were processed by the UEA Small RNA Workbench [192]. Adapters were trimmed using the Adapter Trimming tool and, as these libraries were constructed using HD adapters to reduce ligation bias [219], the additional random nucleotides at the 5' (PARE and sRNA) and 3' (sRNA) end of the sequences were also removed. Trimmed sequences shorter than 15 nt were discarded and sequences mapping to rRNA or tRNA sequences were also removed. After trimming of the adapter sequences, the FASTQ files were converted into FASTA format. Next, we performed sequence alignment to both the host and viral genome allowing 0 mismatches and 0 gaps using PatMaN [165]. The host genome and transcriptome versions used were SL3.0, and ITAG3.2, respectively, and were obtained from the International Tomato Genome Sequencing Project [44]. The CMV genome (strain Fny) and D-satRNA sequences were obtained from NCBI [14]. The mutated D-satRNA sequence (Dm-satRNA) was obtained using the mutated nucleotides identified in a previous study [93]. A full description of the data used in this study can be found in Appendix D Table 1.

6.3.4 Identification of necrogenic sRNA

As discussed above, the necrogenic nucleotides of D-satRNA have been identified as positions 285, 290 and 292 [93]. To investigate whether D-satRNA derived sRNA could be targeting specific host genes leading to plant death, we first identified sRNA reads that aligned to either strand of the D-satRNA sequence and contained at least one of the necrogenic nucleotides within each of our D-satRNA libraries. These reads were then extracted and used to create the necrogenic sRNA datasets for each CMV D-satRNA infected library.

6.3.5 Target prediction

Target prediction was performed with PAREsnip2 using the D-satRNA derived sRNA and corresponding degradome libraries as input. To maximise the number of predictions for further analyses, we removed the core region 2x score multiplier from the PAREsnip2 targeting criteria and lowered the MFE ratio cut-off to 0.65. We excluded filtering the results based on their *p*-value, however the value was still reported. For this analysis and to increase confidence, we configured PAREsnip2 to discard Category-3 and 4 interactions. We then post-processed the results to remove any interactions that had a cleavage signal with an abundance, raw or weighted, less than 5 as it is difficult to distinguish between true sRNA cleavage products and random degradation at such low abundance. Where multiple sRNAs were predicted to target a single site, we selected only one to present in the results. Targets that were identified in at least 2 out of the three replicated were kept for further analyses. Target plots (t-plots) were generated using the T-plot tool within the UEA sRNA Workbench.

6.3.6 Target validation

To validate the targeting sites of the candidate genes and to determine if this results in reduced expression, we used a green fluorescent protein (GFP) reporter system in CMV D-satRNA and Dm-satRNA infected *N. benthamiana*. By attaching a reporter gene to the targeting sequences from our candidates of interest, if the RISC is successfully guided to the recombinant GFP mRNA by D-satRNA derived sRNA for cleavage, we should see a reduction in the amount of expressed GFP or its fluorescence intensity. If we see a reduction in the amount of expressed GFP in CMV D-satRNA infection compared with Dm-satRNA, we can deduce that the

D-satRNA derived sRNA is targeting the candidate gene at the predicted site for degradation.

6.4 Results

6.4.1 Sequencing data

After performing the pre-processing steps on the sequencing data we aligned the sRNA and PARE libraries to the reference sequences. Summary statistics regarding each sRNA and PARE library can be found in Table 6.1 and Table 6.2, respectively. The number of sRNAs aligning to each of the virus reference sequences can be found in Table 6.3. Sequence length distribution for the sRNA and PARE libraries can be found in Appendix D Figures 1-4 and 5-8, respectively. Sequence length distribution for the sRNA and PARE libraries confirm that the predominant read lengths were within the expected range (21-24nt for sRNA and 19-21nt for PARE). The results of the sRNA host genome alignment, presented in Table 6.1, show that in each library ~80% of reads successfully aligned. The proportion of aligned reads are slightly lower in the virus infected samples and this may be a result of the additional virus-derived sRNA present within the library. With the exception of CMV1, similarly large read counts and proportion of aligned reads were also observed for the PARE libraries, presented in Table 6.2. The results obtained for the CMV1 PARE library suggest that there were issues with either the library construction or the sequencing experiment and therefore we decided to exclude it from any further analysis. In addition, there was an issue with the library construction of the M1 degradome library and as such, it was also excluded from further analyses. The virus reference sRNA alignment shows that there may be minor sample contamination, such as D-satRNA aligned sRNAs in CMV1 and

M3, however the number of aligned reads is considerably lower than that of the corresponding libraries.

sRNA library	Total	Unique	# aligned	% aligned
M1	20 881 841	11 639 845	9 641 632	82.83%
M2	27 451 115	15 193 577	12 608 442	82.99%
M3	20 089 549	12 098 427	10 095 620	83.45%
CMV1	25 136 965	12 431 862	9 922 251	79.81%
CMV2	27 354 231	13 159 140	10 509 188	79.86%
CMV3	19 544 307	10 533 855	8 473 386	80.44%
D1	24 487 617	9 870 311	7 881 618	79.85%
D2	18 115 577	7 463 839	5 981 753	80.14%
D3	19 454 258	8 730 499	7 090 078	81.21%
Dm1	18 432 744	8 924 919	7 220 300	80.90%
Dm2	20 489 244	10 122 811	8 255 131	81.55%
Dm3	27 031 049	13 509 139	11 042 624	81.74%

Table 6.1 The number of redundant, non-redundant and *S. lycopersicum* genome aligned reads in each of the sRNA libraries.

PARE library	Total	Unique	# aligned	% aligned
M2	26 179 791	7 761 293	6 419 156	82.71%
M3	23 031 781	6 561 469	5 362 282	81.72%
CMV1	20 576	18 819	15 951	84.76%
CMV2	24 706 048	8 149 176	6 729 119	82.57%
CMV3	23 955 530	7 660 919	6 269 994	81.84%
D1	21 055 183	7 177 706	5 920 495	82.48%
D2	18 165 314	5 386 997	4 411 624	81.89%
D3	50 400 671	11 202 525	8 956 440	79.95%
Dm1	22 485 867	7 349 439	5 930 472	80.69%
Dm2	15 606 177	3 876 219	3 136 670	80.92%
Dm3	24 942 614	7 269 483	5 847 360	80.44%

Table 6.2 The number of redundant, non-redundant and transcriptome aligned reads (positive direction only) in each of the PARE libraries.

sRNA library	# total reads	# unique reads	CMV RNA 1	CMV RNA 2	CMV RNA 3	D-satRNA	Dm-satRNA
C1	25 136 965	12 431 862	37 327	41 861	34 012	1 228	1 159
C2	27 354 231	13 159 140	39 411	44 005	37 399	508	498
C3	19 544 307	10 533 855	35 008	39 161	32 749	397	386
D1	24 487 617	9 870 311	19 594	17 797	28 301	6 760	-
D2	18 115 577	7 463 839	16 006	15 489	27 720	7 603	-
D3	19 454 258	8 730 499	16 576	15 362	26 375	6 625	-
Dm1	18 432 744	8 924 919	16 802	14 608	28 467	-	7 030
Dm2	20 489 244	10 122 811	16 766	14 877	28 358	-	6 917
Dm3	27 031 049	13 509 139	16 751	14 355	27 945	-	6 727
M1	20 881 841	11 639 845	864	1 220	2 059	464	454
M2	27 451 115	15 193 577	508	696	1 867	815	827
M3	20 089 549	12 098 427	1 035	992	4 556	1 746	1 930

Table 6.3 The number of redundant, non-redundant and virus aligned reads in each of the sRNA libraries.

6.4.2 Necrogenic D-satRNA derived sRNA

We identified potential necrogenic sRNA by aligning reads within each of the D-satRNA sRNA libraries to the D-satRNA sequence (plus or minus strand) and extracted those that contained at least one of the necrogenic nucleotides. The results of these can be found in Table 6.4 and the overlap between each replicate can be found in Figure 6.2. The results show a considerable overlap in sequences containing necrogenic nucleotides between replicates. We now further investigate the function these sequences by performing target prediction using PAREsnp2.

sRNA library	Total aligned reads	Unique aligned Reads
D1	267 719	690
D2	187 209	799
D3	222 345	646

Table 6.4 The number of redundant and non-redundant reads that align to the D-satRNA sequence and contain at least one of the necrogenic nucleotides in each of the CMV D-satRNA infected sRNA libraries.

6.4.3 Identification of host mRNA targets

Using as input the potential necrogenic D-satRNA derived sRNA (described above), we performed target prediction with PAREsnp2 using the criteria described in Section 6.3.5. Before performing the conservation and cleavage signal abundance filtering, the number of predicted mRNA target sites for each category are shown in Table 6.5. We then further filtered the results using the previously defined criteria (Section 6.3.5) and the number of unique reported target sites meeting these criteria are shown in Table 6.6. The filtering process removed 22, 21 and 38 lower confidence (Category-2) interactions in D1, D2 and D3, respectively, but kept all of the high confidence (Category-0) interactions in each dataset.

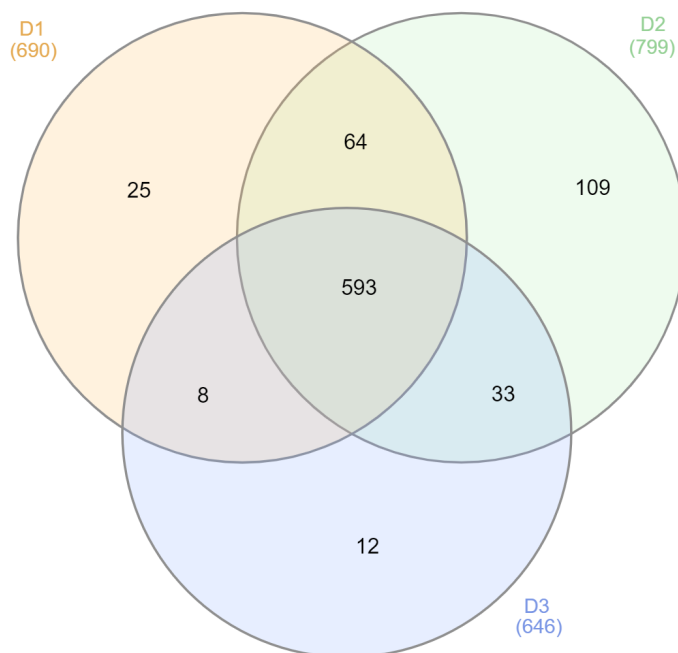


Fig. 6.2 Venn diagram showing the overlap of unique sRNA containing at least one of the necrogenic nucleotides in each of the CMV D-satRNA infected sRNA libraries.

sRNA library	Category-0	Category-1	Category-2
D1	4	0	54
D2	2	0	59
D3	2	1	86

Table 6.5 The number of unique target sites and their categories reported by PAREs-nip2 when analysing the potential necrogenic sRNAs and corresponding degradome in each of the CMV D-satRNA infected libraries.

sRNA library	Category-0	Category-1	Category-2
D1	4	0	32
D2	2	0	28
D3	2	0	48

Table 6.6 The number of unique target sites and their categories reported by PAREs-nip2 that are also conserved between at least two replicates and have a cleavage signal with abundance ≥ 5 .

We now focus on the 3 Category-0 target candidates, predicted in D1, that are conserved between each of the CMV D-satRNA infected libraries. These candidates are:

1. Solyc02g093935.1.1 Sister chromatid cohesion 1 protein 2 (SCC1P2)
2. Solyc07g065660.3.1 Cellulose synthase family protein (CSFP)
3. Solyc07g053740.1.1 Ethylene Response Factor F.4 (ERF4)

Below, we discuss each of these target interactions in more detail.

6.4.3.1 SCC1P2

The function of SCC1P2 in *S. lycopersicum* is not well understood but the homologous gene in *A. thaliana*, RAD21, has been studied. RAD21 acts as part of the sister chromatid cohesion process in dividing cells and also genomic DNA break repair in other cells [46, 47]. It exhibits higher expression in shoot apex but its expression is low in other tissues. Many chromosome breaks were found in germ cells of *A. thaliana* RAD21 mutants and, under UV or other stresses, plants would die from defective DNA break repair. Hence, RAD21 is essential for the normal development of *A. thaliana* [46, 47]. In CMV D-satRNA infected *S. lycopersicum*, primary cell death occurs in specific cells near the shoot apex. It also occurs in infected cells that are differentiating right after cell division, where DNA fragmentation occurs in the nucleus [52, 114]. Therefore, if the function of SCC1P2 in *S. lycopersicum* is similar to that of RAD21 in *A. thaliana* and its expression is reduced by D-satRNA derived sRNA, this may result in plant death.

The target site for SCC1P2 is conserved between each replicate and has a Category-0 signal in each. The T-plots for the interaction are shown in Figures 6.3, 6.4 and 6.5 for D1, D2 and D3, respectively. The sRNA predicted to target

SCC1P2 originates from the positive strand of D-satRNA, is 21nt in length and contains all three of the necrogenic nucleotides (positions 5, 10 and 12). All of these positions are within the sRNA core region (positions 2-13), which are considered to be important for sRNA-induced mRNA cleavage [3]. As such, mutations at these positions would greatly reduce the complementarity between the sRNA and mRNA, potentially abolishing the ability to locate and induce cleavage of the target mRNA. We then examined the degradation signals on this transcript in the other libraries and found that no other treatments showed signals of cleavage at this position, as demonstrated with Dm2 in Appendix D Figure 9.

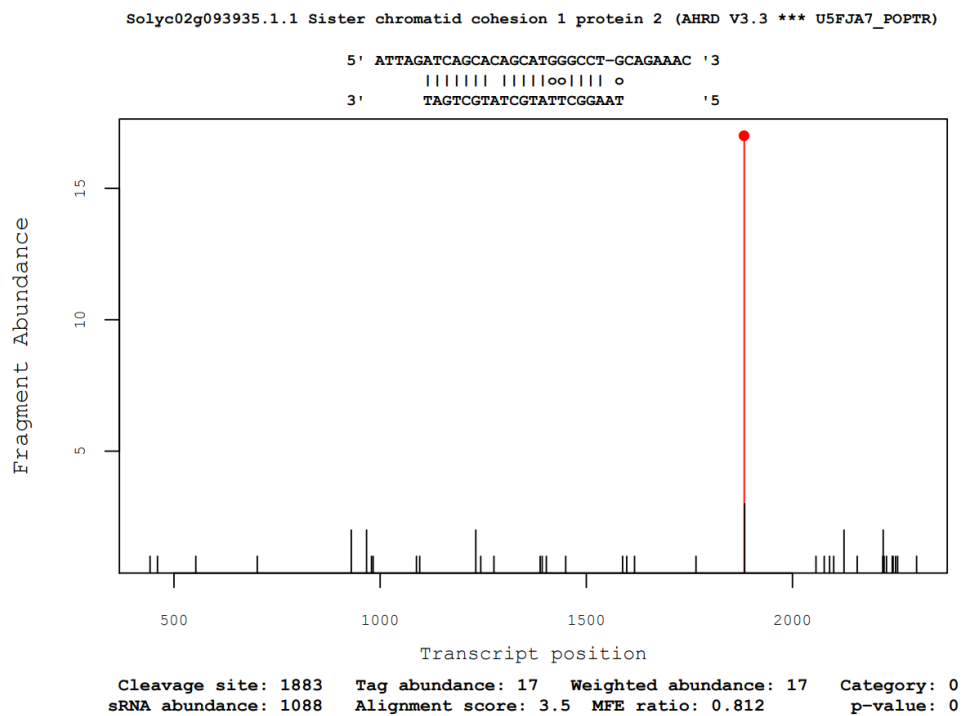


Fig. 6.3 T-plot showing the degradation activity for SCC1P2 in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a *p*-value of 0.

6.4.3.2 ERF4

As with SCC1P2, the functions of ERF4 in *S. lycopersicum* is not fully understood. However, the homologous gene in *A. thaliana* has been studied. ERF4 belongs to the AP2/ERF family of transcription factors and is the terminal regulatory gene

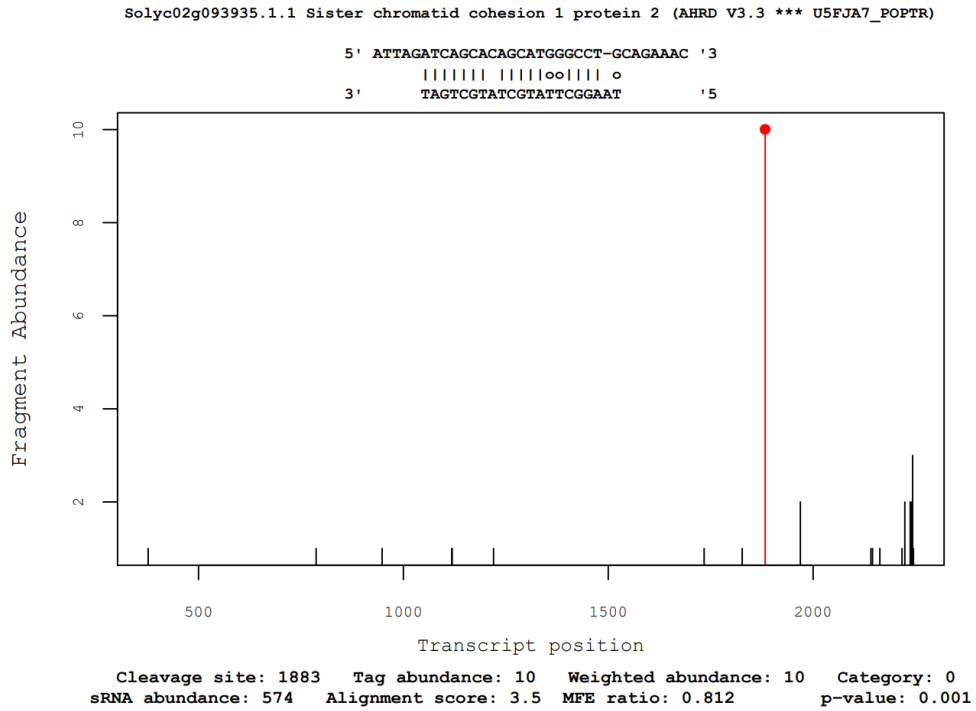


Fig. 6.4 T-plot showing the degradation activity for SCC1P2 in the D2 degradome dataset. The cleavage signal is Category-0 and the interaction has a p -value of 0.001.

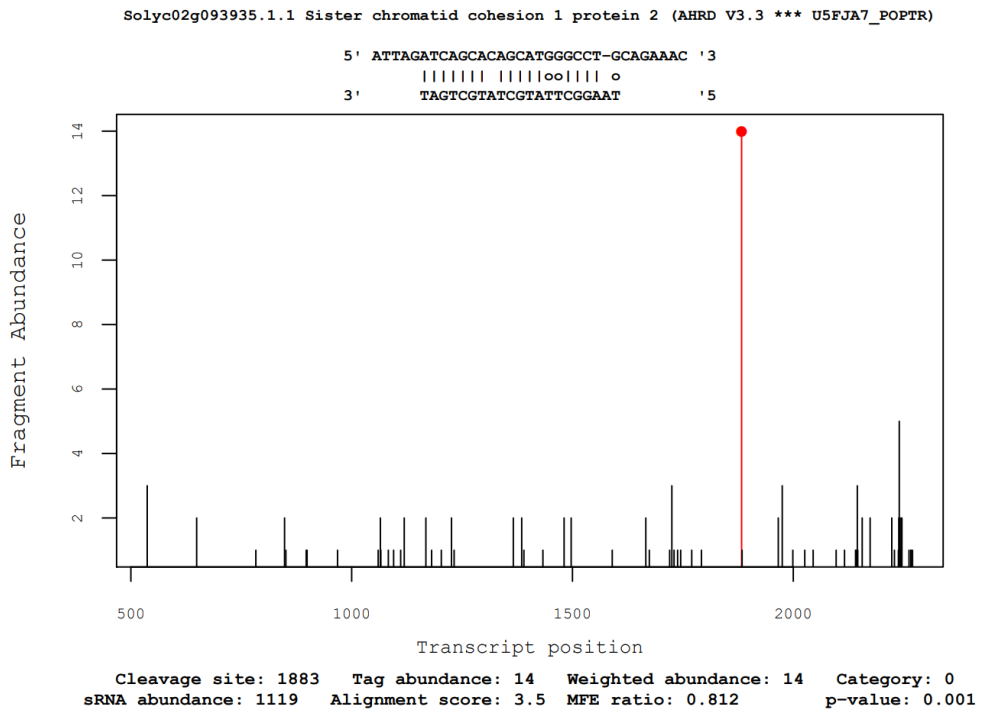


Fig. 6.5 T-plot showing the degradation activity for SCC1P2 in the D3 degradome dataset. The cleavage signal is Category-0 and the interaction has a p -value of 0.001

of the ethylene signal transduction pathway. The ERFs of *S. lycopersicum* play a role in fruit development and stress response, however the specific role of each ERF is not clear [109]. ERF4 in *A. thaliana* regulates the expression of *CATALASE* through differential splicing after transcription, which regulates the accumulation of reactive oxygen species (ROS) in cells and the process of cell senescence [171]. It has been shown that the ethylene signal transduction pathway is activated in CMV D-satRNA infected *S. lycopersicum* [52], which participates in the burst of ROS and the occurrence of secondary cell death in tissues. Thus, the decrease of ERF4 expression by D-satRNA derived sRNA may affect the ERF4-regulated ROS accumulation that is correlated with secondary cell death.

The target site for ERF4 is conserved between each replicate. The interactions are Category-0 in D1 and Category-2 in D2 and D3. The T-plots for the interaction are shown in Figures 6.6, 6.7 and 6.8 for D1, D2 and D3, respectively. The sRNA predicted to target ERF4 originates from the negative strand of D-satRNA, is 22nt in length and contains one of the necrogenic nucleotides at position 5. Upon further investigation, there also exists a highly abundant 21mer that could target this site within each CMV D-satRNA sRNA library. However, the sRNA-mRNA duplex containing the 21nt sRNA does not meet the MFE ratio cut-off value of 0.65. As with SCC1P2, the necrogenic nucleotide is contained within the sRNA core region. As such, a mutation at this position may reduce the complementarity such that the sRNA can no longer induce cleavage of the mRNA, especially with how close this would be to the two mispaired bases at positions 7 and 8. We then examined the degradation signals on this transcript in the other libraries and found that, although signals were found, as demonstrated with Dm2 in Appendix D Figure 10, the signal at the target site was considerably lower than that of the D-satRNA infected libraries.

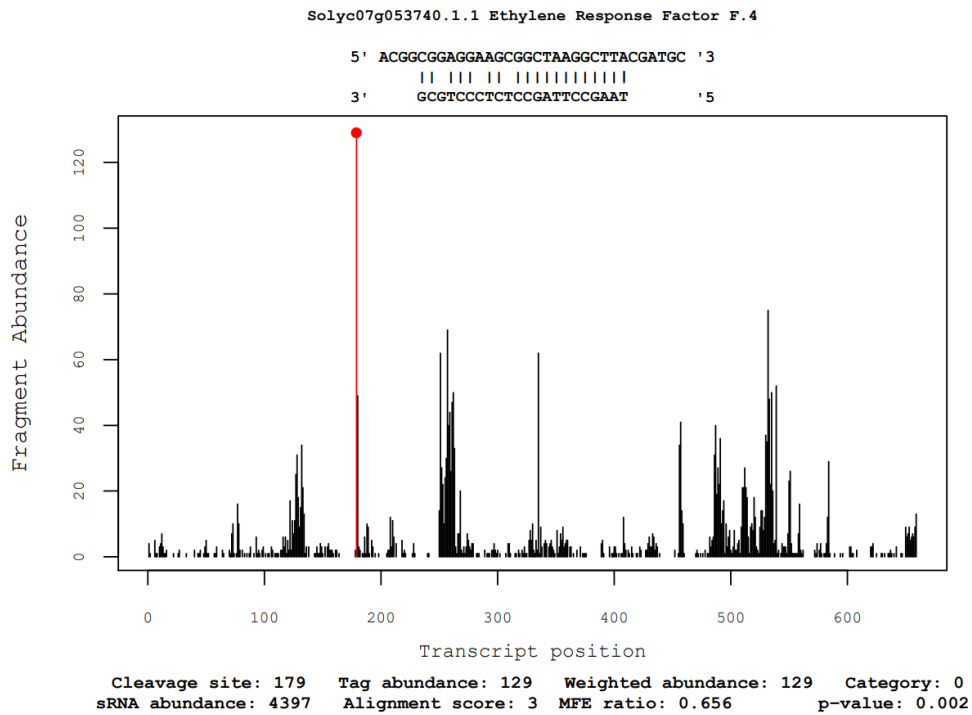


Fig. 6.6 T-plot showing the degradation activity for ERF4 in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a p -value of 0.002

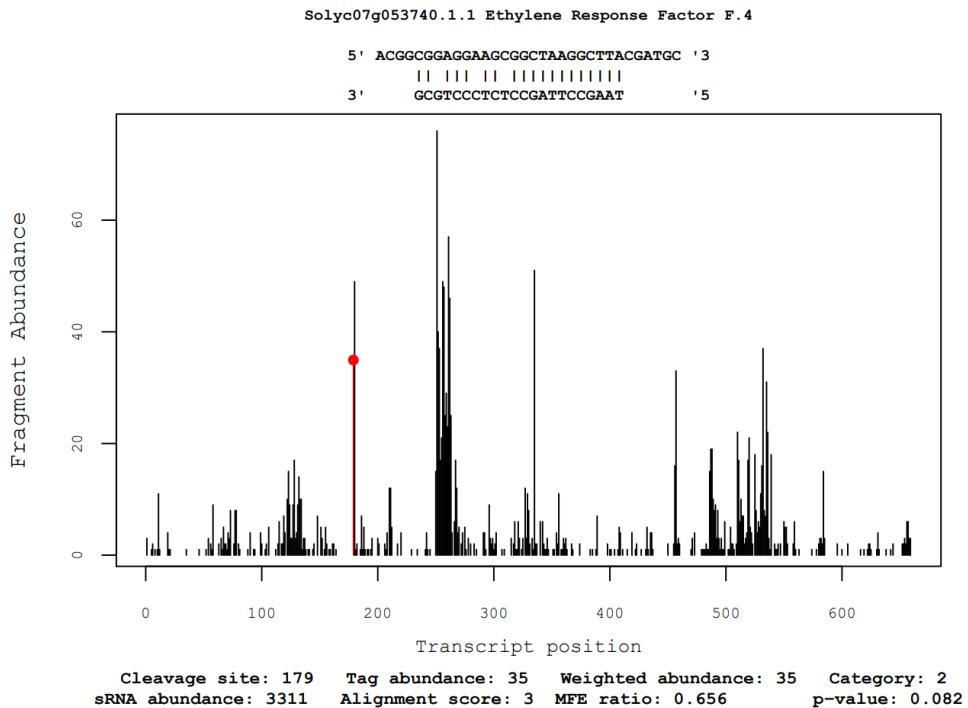


Fig. 6.7 T-plot showing the degradation activity for ERF4 in the D2 degradome dataset. The cleavage signal is Category-2 and the interaction has a p -value of 0.082.

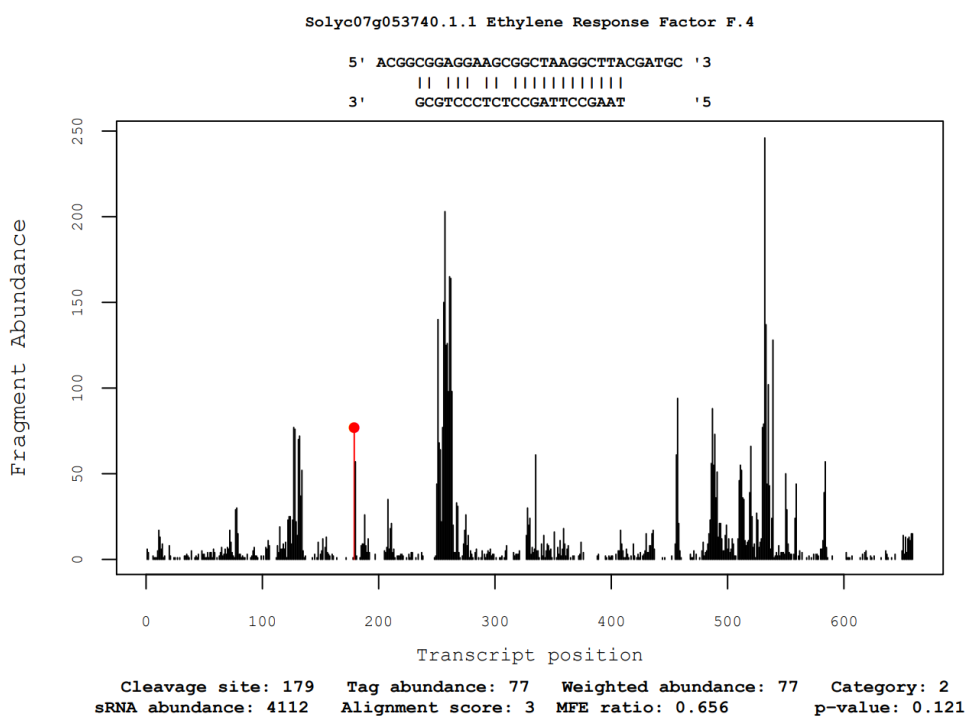


Fig. 6.8 T-plot showing the degradation activity for ERF4 in the D3 degradome dataset. The cleavage signal is Category-2 and the interaction has a p -value of 0.121

6.4.3.3 CSFP

CSFP plays a vital role in the biosynthesis of cellulose and hemicellulose in *S. lycopersicum* [189]. The target site for CSFP is conserved between each replicate. The interactions are Category-0 in D1 and D2 and Category-2 in D3. The T-plots for the interaction are shown in Figures 6.9, 6.10 and 6.11 for D1, D2 and D3, respectively. The sRNA predicted to target CSFP originates from the positive strand of D-satRNA, is 20nt in length and contains two of the necrogenic nucleotides (positions 2 and 4). Upon further investigation, there also exists a highly abundant 21mer that could target this site within each CMV D-satRNA sRNA library. However, this sRNA contains an additional mismatch at the 5' end and so does not meet the employed PAREsnip2 targeting criteria. Mutations at positions 2 and 4 greatly reduce the complementarity between the sRNA and mRNA, especially with how close these would be to the two mispaired bases at positions 5 and 6. Thus, additional mismatches may reduce or abolish the ability to locate and induce

cleavage of the target mRNA. We then examined the degradation signals on this transcript in the other libraries and found that, although signals were found, as demonstrated with Dm2 in Appendix D Figure 11, the signal at the target site was considerably lower than that of the D-satRNA infected libraries.

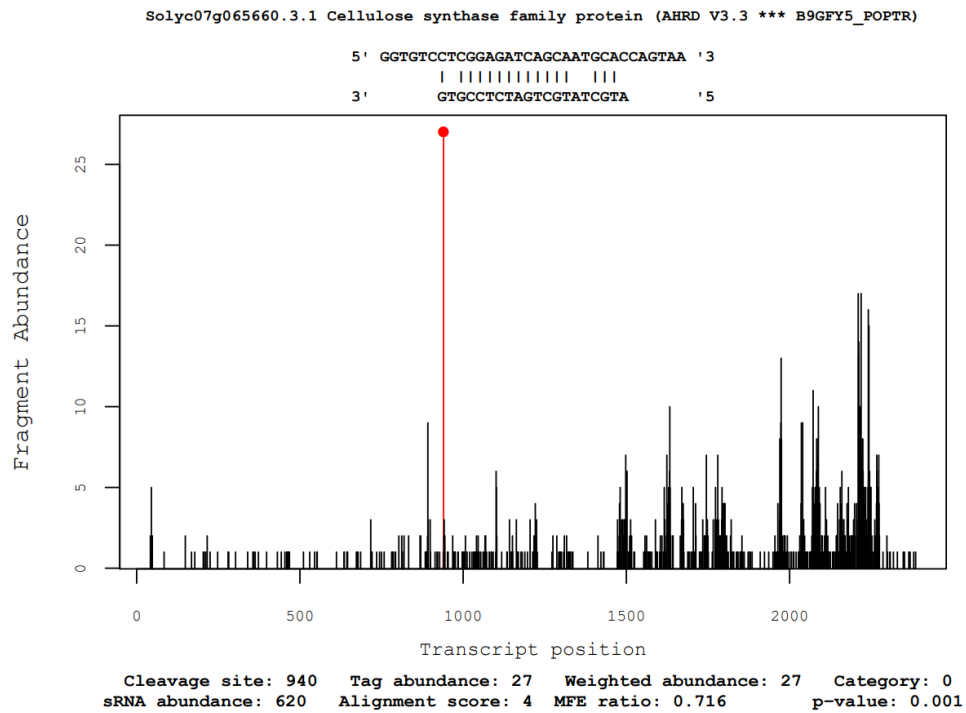


Fig. 6.9 T-plot showing the degradation activity for CSFP in the D1 degradome dataset. The cleavage signal is Category-0 and the interaction has a p -value of 0.001

6.4.4 Target site conservation in lethal and non-lethal infection

CMV D-satRNA infection is also lethal in some close relatives of *S. lycopersicum*, for example in *Solanum pennelli* and *Solanum habrochaitis*, but attenuates symptoms in other species, such as *N. tabacum* and *Solanum tuberosum*. Additional work by Dr. Ping Xu isolated a surviving line of *S. habrochaitis*, where the infected plants do not show lethal systemic necrosis.

We now investigate whether the target sites of our candidates are conserved between homologous genes in these species that are known to survive or die from CMV D-satRNA infection. The reference sequences for other species were obtained

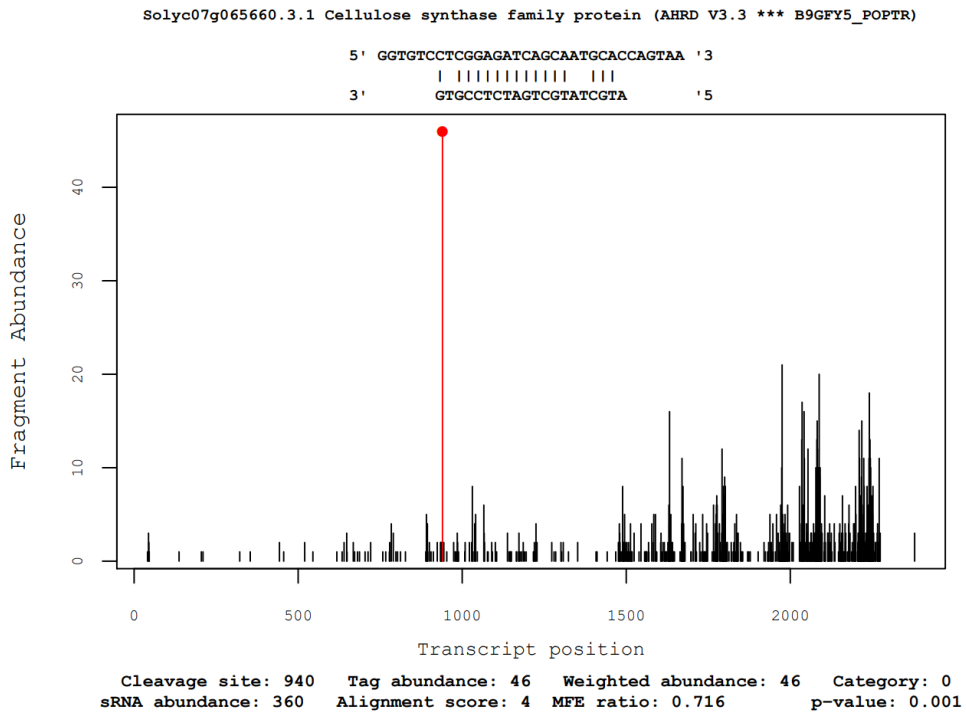


Fig. 6.10 T-plot showing the degradation activity for CSFP in the D2 degradome dataset. The cleavage signal is Category-0 and the interaction has a p -value of 0.001.

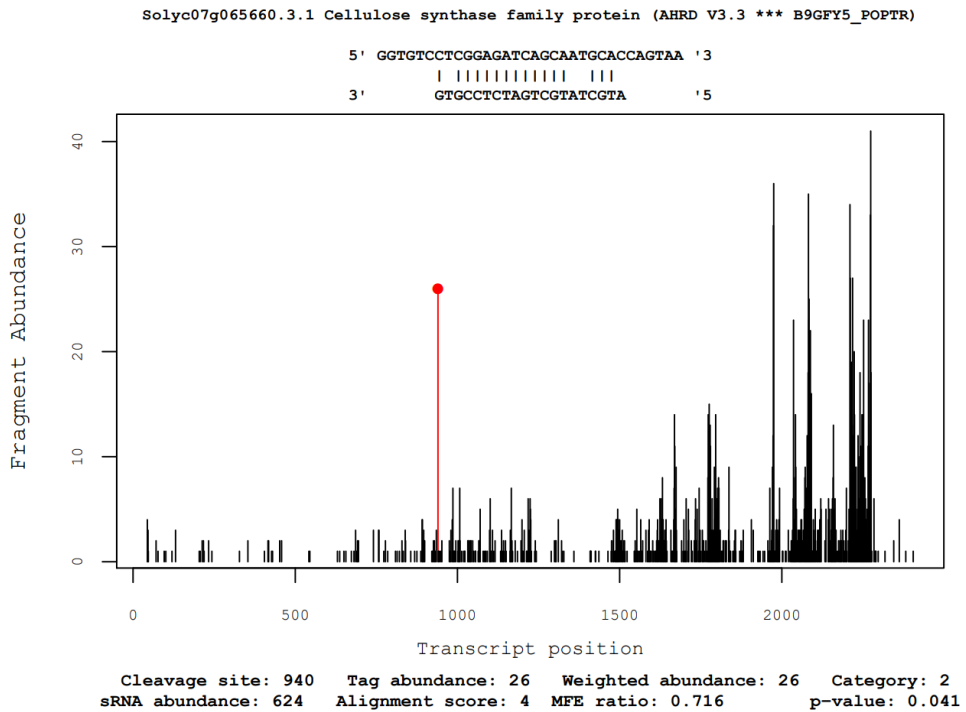


Fig. 6.11 T-plot showing the degradation activity for CSFP in the D3 degradome dataset. The cleavage signal is Category-2 and the interaction has a p -value of 0.041

through a BLAST [14] search using the NCBI website (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). The target sequence from the surviving line of *S. habrochaitis* was identified by sequencing the corresponding genomic regions.

6.4.4.1 SCC1P2

We compared the sequences of SCC1P2 in *S. lycopersicum* to its homologous genes in *S. pennelli*, *S. habrochaitis*, *N. tabacum* and *S. tuberosum*. The gene sequences are highly conserved among the above species, but with mutations at the targeting site, as shown in Table 6.7. For the plants with lethal response to the infection, the mutation is a substitution at the 13th position of the target site, with respect to the sRNA sequence, replacing an A-T pair with a G-T wobble pair. This mutation does not reduce the complementarity between the sRNA and mRNA too much and thus, the mRNA may still be silenced in these species. The SCC1P2 homologous gene in the surviving *S. habrochaitis* line was cloned. Its target site contained an additional mutation at position 16. This mutation replaced a C with a T, which removed the G-C pair at the target site.

Analysis of the *N. tabacum* homologous gene shows that it contains four mutations. These are found in the 11th, 13th, 14th and 17th position of the target site.

Analysis of the *S. tuberosum* homologous gene shows that it contains two mutations. These are found at in the 13th and 14th position of the target site.

Additional mismatches at the target site in homologous genes may reduce the complementarity between the sRNA and the target mRNA enough that the sRNA fails to recognize and/or induce cleavage. In particular, when these additional mismatches are at the core or central region of the sRNA.

Species	Target site
<i>S. lycopersicum</i> (D)	5' ATCAGCACAGCATGGGCCTGC 3'
<i>S. habrochaites</i> (D)	5' ATCAGCAC G GCATGGGCCTGC 3'
<i>S. pennelli</i> (D)	5' ATCAGCAC G GCATGGGCCTGC 3'
<i>S. habrochaites</i> (S)	5' ATCAG TAC G GCATGGGCCTGC 3'
<i>N. tabacum</i> (S)	5' ATCA ACA AGGG ATGGGCCTGC 3'
<i>S. tuberosum</i> (S)	5' ATCAGCA AG GCATGGGCCTGC 3'

Table 6.7 The mutations at the target site of the conserved SCC1P2 homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the *S. lycopersicum* target site are highlighted. **S** = survives infection and **D** = dies from infection.

6.4.4.2 ERF4

We compared the sequences of ERF4 in *S. lycopersicum* to its homologous genes in *S. pennelli*, *S. habrochaitis*, *N. tabacum* and *S. tuberosum*. The gene sequences are highly conserved among the above species, but with mutations at the targeting site in the species known to survive infection, as shown in Table 6.8. The results from the BLAST search on two closely related species that die from infection, *S. pennelli* and *S. habrochaitis*, showed that the target site of their homologous genes are identical to that of ERF4 in *S. lycopersicum*.

Analysis of the *N. tabacum* homologous gene shows that it contains three mutations. These are found in the 3rd, 14th and 17th position of the target site.

Analysis of the *S. tuberosum* homologous gene shows that it contains a single mutation. This is found in the 9th position of the target site.

Additional mismatches in homologous genes may reduce the complementarity between the sRNA and the target mRNA enough that the sRNA fails to recognize and/or cleave the target site. Particularly if they are found within the core or central region of the sRNA.

Species	Target site
<i>S. lycopersicum</i> (D)	5' GGAGGAAGCGGCTAAGGCTTA 3'
<i>S. habrochaites</i> (D)	5' GGAGGAAGCGGCTAAGGCTTA 3'
<i>S. pennelli</i> (D)	5' GGAGGAAGCGGCTAAGGCTTA 3'
<i>N. tabacum</i> (S)	5' GGAAGAGGCGGCTAAGGCGTA 3'
<i>S. tuberosum</i> (S)	5' GGAGGAAGCGGC GAAGGCTTA 3'

Table 6.8 The mutations at the target site of the conserved ERF4 homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the *S. lycopersicum* target site are highlighted. **S** = survives infection and **D** = dies from infection.

6.4.4.3 CSFP

We compared the sequences of CSFP in *S. lycopersicum* to its homologous genes in *S. pennelli*, *S. habrochaitis*, *N. tabacum* and *S. tuberosum*. The gene sequences are highly conserved among the above species, but with mutations at the targeting site in the species known to survive infection, as shown in Table 6.9. The results from the BLAST search on two closely related species that die from infection, *S. pennelli* and *S. habrochaitis*, showed that the target site of their homologous genes are identical to that of CSFP in *S. lycopersicum*.

Analysis of the *S. tuberosum* and *N. tabacum* homologous gene shows that they contains two mutations. These are found at in the 1st and 17th position of the target site. These additional mismatches reduce the complementarity between the sRNA and the target mRNA and may result in the sRNA failing to recognize and/or cleave the target site.

Species	Target site
<i>S. lycopersicum</i> (D)	5' CTCGGAGATCAGCAATGCAC 3'
<i>S. habrochaites</i> (D)	5' CTCGGAGATCAGCAATGCAC 3'
<i>S. pennelli</i> (D)	5' CTCGGAGATCAGCAATGCAC 3'
<i>N. tabacum</i> (S)	5' CTCAGAGATCAGCAATGCCGC 3'
<i>S. tuberosum</i> (S)	5' CTCAGAGATCAGCAATGCCGC 3'

Table 6.9 The mutations at the target site of the conserved CSFP homologous genes in species known to survive or die from CMV D-satRNA infection. Mutations relative to the *S. lycopersicum* target site are highlighted. S = survives infection and D = dies from infection.

6.4.5 Validation of targets

We now present the results from target validation for two of the candidate genes, CSFP and SCC1P2. Experiments to validate the ERF4 target site are currently in process.

CSFP The results from experimental validation of CSFP, presented in Figure 6.12, do not show a clear reduction in the intensity of GFP in D-satRNA infected *N. benthamiana* compared with Dm-satRNA.

SCC1P2 The results from experimental validation of SCC1P2, presented in Figure 6.13, show a clear reduction in the intensity of GFP in D-satRNA infected *N. benthamiana* compared with Dm-satRNA. These results suggest that this would also be the case in *S. lycopersicum*, where CMV D-satRNA infection leads to plant death.

6.5 Conclusion

CMV is one of the most widespread plant viruses, infecting a large number of plant species worldwide. CMV can also harbour satRNAs which usually attenuate virus symptoms, however D-satRNA in *S. lycopersicum* is an exception that eventually leads to plant death. The necrogenicity of D-satRNA has been identified as positions 285, 290 and 292, and mutations at these positions result in *S. lycopersicum* surviving infection and having reduced symptoms.

In this chapter, we exemplified the use of PAREsnip2 to identify potential host mRNA targets of necrogenic D-satRNA derived sRNA. Employing a slightly relaxed set of targeting criteria, we identified multiple mRNA targets for these sRNA evidenced through the degradome. Three of these candidate, SCC1P2, ERF4 and CSFP, had conserved cleavage signals between three biological replicates, with at least one of these signals being Category-0. In addition, previous work into the impact that down regulation of the homologous ERF4 and SCC1P2 genes in *A. thaliana*, suggest that there may be correlation with CMV D-satRNA induced necrosis.

Investigation into sequence variation at the target site of homologous genes in species known to survive CMV D-satRNA infection show mutations that reduce complementarity to the sRNA, further supporting our hypothesis that down regulation of these genes may be involved in plant death. Preliminary experimental work has confirmed the down regulation of one of these targets, SCC1P2, by D-satRNA derived sRNA.

Without further experimental validation, it is difficult to determine whether the down regulation of SCC1P2 contributes towards plant death. However, additional experimental work is now being prepared to confirm if it does play a role in plant death and these are outlined in the future work section of this thesis.

In the next chapter, we detail possible future directions and extensions to this work.

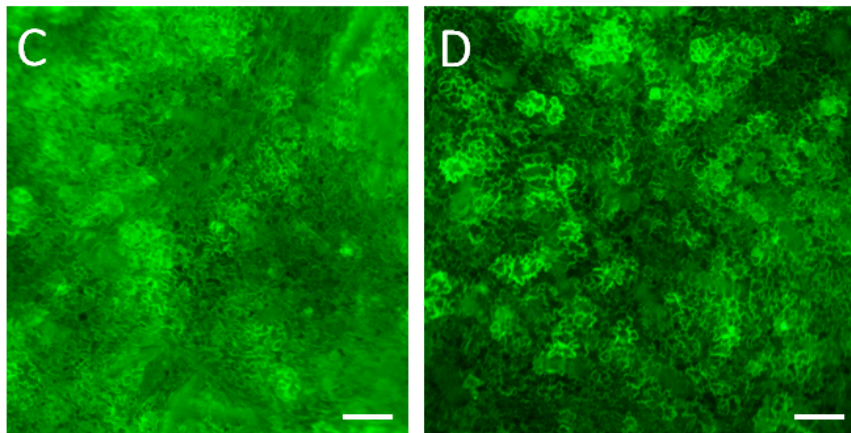
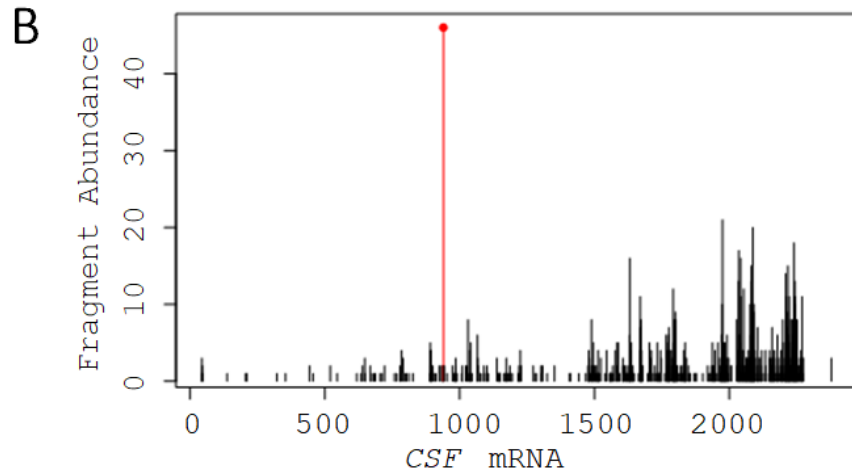
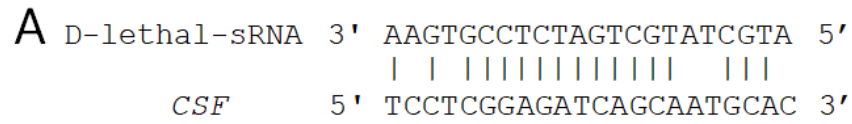


Fig. 6.12 Target validation results for CSFP in CMV D-satRNA and Dm-satRNA infected *N. benthamiana*. Panel A is the predicted target site, panel B is the cleavage signal, panel C is fluorescent intensity in D-satRNA and panel D is fluorescent intensity in Dm-satRNA.

A D-lethal-sRNA 3' CTAGTCGTATCGTATTCGGAAT ' 5
 ||||| ||||| oo ||| o
SCC1P2 5' GATCAGCACAGCATGGGCCT-G ' 3

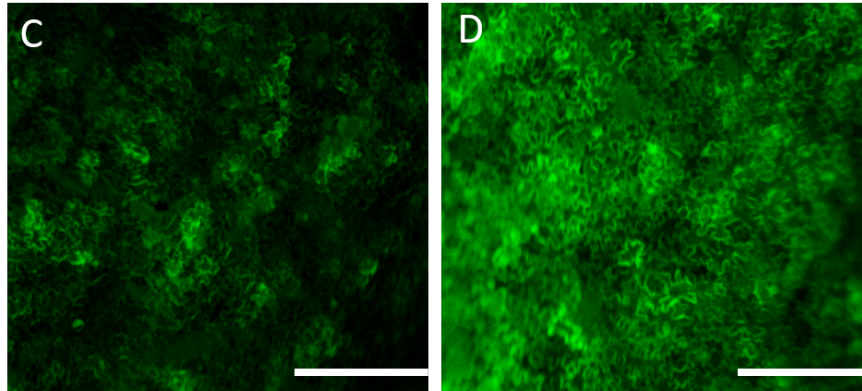
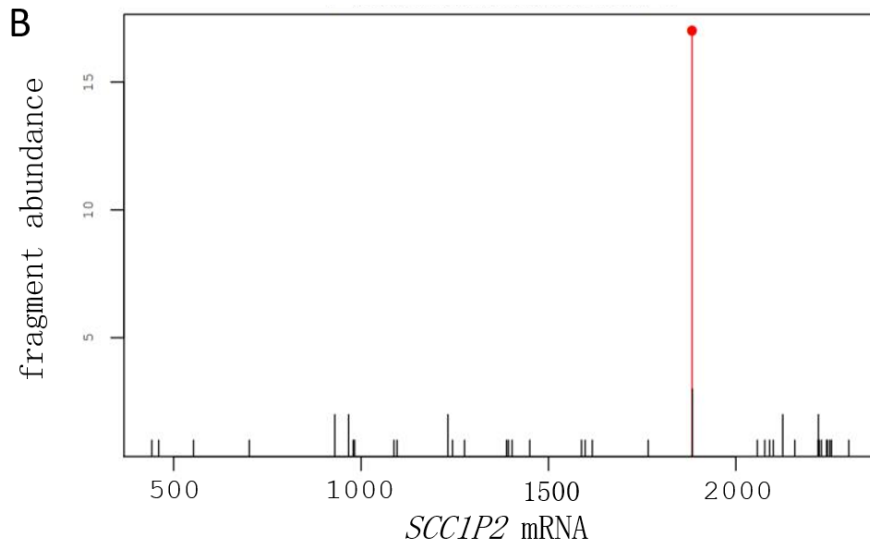


Fig. 6.13 Target validation results for *SCC1P2* in CMV D-satRNA and Dm-satRNA infected *N. benthamiana*. Panel A is the predicted target site, panel B is the cleavage signal, panel C is fluorescent intensity in D-satRNA and panel D is fluorescent intensity in Dm-satRNA. The results from this experiment show a clear reduction in the fluorescent intensity in D-satRNA when compared to Dm-satRNA.

Chapter 7

Future work and thesis conclusion

7.1 Summary

In this thesis, we have provided an introduction into sRNA biology and an overview of the computational methods used for analysing sRNA and degradome sequencing data. We presented three new tools that we developed for this type of analysis and benchmarked them against other publicly available methods, demonstrating clear improvements in computation time and/or prediction accuracy. In Chapter 6, we exemplified the use of PAREsnip2 by performing degradome analyses to better understand the role of D-satRNA derived necrogenic sRNA. In this chapter, we discuss possible extensions to this work followed by a conclusion of this thesis.

7.2 Future work

7.2.1 Combining existing workflows

There exists many individual analysis tools within the UEA sRNA Workbench, ranging from data pre-processing and quality control to sRNA classification and

functional analyses. Currently, each of the tools we have presented in this thesis are standalone applications, with all functionality required to perform analysis built-in. While each of these tools have some data pre-processing and filtering methods available, such as read length and abundance filtering, these methods are not as comprehensive as those contained in the quality checking, normalization, and differential expression tool [17]. Through combining our new tools with this existing tool, it would enable a standardised filtering, normalization and quality checking process. Consequently, this may result in more confidence in the predictions reported and may provide a more complete overview of the information contained within the sequencing data.

Moreover, the inclusion of the differential expression analysis tool may provide greater indication into the function of specific sRNAs when combined with classification or functional analysis using NATpare or PAREsnip2. This could also be coupled with mRNA differential expression analysis using RNAseq data and FiRePat [151] to further support prediction of sRNA function. For example, if a sRNA is determined to be up-regulated in a given treatment, a target for this sRNA is identified using PAREsnip2 and the target mRNA is determined to be down regulated by FiRePat, this may give a more complete picture of the regulatory processes at work.

7.2.2 Further work into sRNA targeting criteria

In Chapter 4, we introduced a software pipeline, called PAREameters, for inference of plant miRNA targeting criteria using degradome data. In that chapter, we performed a comparison of criteria inferred between species and also between tissues. This work could be extended by performing experimental validation of these miRNA targets to confirm the reported discrepancies between species or tissue specific targeting criteria. In addition, and as we have seen in Chapter 5,

miRNAs are not the only class of sRNA that induce mRNA cleavage and subsequent degradation. In Chapter 6, we performed target prediction using virus derived sRNA and a more flexible set of targeting criteria with one of these targets being confirmed experimentally.

These results suggest that an analysis into the targeting criteria employed by different classes of sRNA may lead to the identification of previously unknown sRNA-mRNA interactions. Furthermore, analyses of these interactions may lead to the identification of previously unknown differences between how each class of sRNA recognise their mRNA target(s). One such example of these differences may be whether the core region (positions 2-13) is equally important across all classes of sRNA.

7.2.3 Further work into necrogenic sRNA

In Chapter 6, we identified three target candidates for CMV D-satRNA derived sRNA that were conserved between three biological replicates and had a Category-0 signal in at least one of them. Experimental validation confirmed that one of the targets, SCC1P2, was down regulated by D-satRNA derived sRNA. However, CSFP was shown to have little to no visual change in expression in Dm-satRNA infected plants. Experiments to confirm the cleavage and down regulation of ERF4 are currently being performed.

To investigate whether the down regulation of these genes is actually involved in plant death, there are two experiments planned. The first is to express the D-satRNA derived sRNA in healthy *S. lycopersicum*. If the down regulation of this gene contributes towards plant death and the sRNA works in the same way as when it is expressed by CMV D-satRNA infection then, in principle, this would also lead to plant death. The second approach is to over-express the candidate genes with mutations at the target sites similar to that of surviving species. This should result

in reduced or complete removal of the sRNAs ability to induce cleavage of the target mRNAs. If these genes are involved in plant death, we would expect that the transgenic plants survive the infection.

7.2.4 Impact of CMV infection on host gene expression

For our analyses in Chapter 6, we focused exclusively on the sRNA originating from the necrogenic region of D-satRNA when performing degradome analysis using PAREsnip2. This meant that we ignored a large proportion of our sRNA data, potentially missing the identification of sRNA-mRNA interactions that may contribute towards CMV symptoms. Future work on this project will focus on three classes of sRNA: miRNA, ta-siRNA and virus-derived sRNAs.

With focus first on miRNAs, we will use miRCat2 to identify potential novel miRNAs in each of our treatments. Next, we will perform differential expression analyses to identify if any known or novel miRNAs are up regulated in specific treatments. These miRNAs will then be subject to target prediction using PAREsnip2 and the corresponding degradome libraries to identify possible mRNA targets. Candidates will then be investigated further based on their biological function and some selected for experimental validation. Further experiments will then be performed to determine if the down regulation of these genes play a part in virus defence or response.

Second, we will focus on ta-siRNA (described in Section 2.3.2.1), as it has been previously shown that failure of TAS3 derived ta-siRNA to regulate ARF3 and ARF4 results in wiry leaf syndrome [223] in tomato, a common symptom of CMV infection. We will first perform sequence alignment of the sRNAs to the known TAS genes. Next, and similar to the miRNAs, we will perform differential expression analyses between treatments using the identified ta-siRNAs. We will then perform target prediction using the differentially expressed ta-siRNAs to first,

determine if these can target ARF3 or ARF4 for regulation, and second, to identify any other targets that may contribute towards virus symptoms.

Finally, we will investigate if any other virus-derived sRNA, originating from CMV, D-satRNA or Dm-satRNA, could target host genes for degradation. This will be done by first aligning the sRNA sequences to the viral reference sequences. We will then perform target prediction using PAREsnip2 and the corresponding degradome libraries for each dataset of virus-derived sRNA. Candidates will then be investigated further based on their biological function and some will be selected for experimental validation. If these candidates are confirmed, further experiments will then be performed to determine if the down regulation of these genes by viral sRNA contribute towards virus symptoms.

7.3 Thesis conclusion

Research into the role of non-coding RNAs is moving away from typical model organisms and samples are now being collected from a wide range of species, tissues and conditions, ready for sequencing and computational analyses. Recent advances in high throughput sequencing technologies has resulted in larger, more complex genomes being sequenced. Moreover, not only are larger genomes being sequenced, sequencing datasets in general are growing ever larger in size and read count, with a typical sequencing experiment now containing millions of distinct reads in a single sample. In addition, the need for multiple samples and replicates is becoming the de facto standard for biological experiments, further adding to this sequence-data deluge.

The development of our new bioinformatics tools will enable processing of recent sRNA and degradome sequencing data obtained from both model and non-model organisms, something that was not previously possible without considerable

time or resource constraints. This may open new avenues of sRNA research, in particular, in the context of nat-siRNAs, which are a class of sRNA that have not yet been extensively studied. Perhaps one reason for this may be to do with the lack of available computational methods for identification and prediction of their function. Previously, nat-siRNAs have been shown to play a role in response to salt stress in *A. thaliana* [26] and we also identified some differentially expressed nat-siRNAs in the same organism and condition (Chapter 5). One possible use for our new tool NATpare would be to investigate if nat-siRNAs are also involved in salt stress response in other species. If proven to be the case and then combined with further experimental verification, this may lead to increased research interest into nat-siRNAs and therefore further understanding of the regulatory roles they play.

As demonstrated in Chapters 4 and 6, the mechanisms in which sRNAs identify their target mRNAs is not fully understood. In Chapter 4, we demonstrated that by using the degradome to infer targeting criteria, we were able to increase the sensitivity of miRNA target prediction when compared to the Allen *et al.* [3] rules. In Chapter 6, we demonstrated that a virus-derived sRNA was able to cleave its target mRNA despite having 3 mispaired bases within the sRNA core region, giving it an alignment score of 6 using the previously defined model [3]. The development of PAREamters, combined with the configurability of PAREsnp2, will hopefully contribute towards better understanding of sRNA-mRNA interactions, especially when combined with additional data, such as mRNA expression profiles, or experimental validation. Improving our understanding of the way sRNAs identify their targets may allow us to discover new regulatory interactions or networks, some of which may play critical roles in important biological pathways yet to be discovered.

Appendix A

Some of the tables referenced within Chapter 3 contain a large number predicted targets and are not practical to include within this thesis. However, for completeness, a brief description of each table is provided below and the actual data is freely available to download from Nucleic Acids Research Online at the following url: <https://doi.org/10.1093/nar/gky609>. We have also included these tables as supplementary information with this thesis.

Appendix A Table 2 We collected a set of experimentally validated *A. thaliana* interactions by combining those previously published in the literature [68, 191, 53] and those contained within miRTarBase [40] with any duplicates being removed. In total, we collected 616 validated interactions comprising 135 miRNAs. Out of these 135 miRNAs, 90 of them had unique sequences and were involved in 387 distinct miRNA–mRNA interactions.

Appendix A Table 3 contains the results of the degradome analysis of dataset D2 using the sRNA–mRNA target interaction rules as described by Allen *et al.* [3]. For this analysis, we used the default stringent parameters, which discards category-4 signals and permits a minimum sRNA abundance of 5 reads. Additionally, the built-in conservation filter was used to increase confidence in the reported interactions. In total, PAREsnip2 captured 2008 sRNA–mRNA interactions, which comprised 960 category-0, 79 category-1, 511 category-2 and 458 category-3 interactions.

Description	File/Accession	File source	Publication	Chapter abbreviation
<i>A. thaliana</i> transcriptome	^a	TAIR10	N/A	N/A
<i>A. thaliana</i> mature leaf degradome dataset used for performance evaluation	GSM1330562	GEO	[197]	D1
<i>A. thaliana</i> wild-type leaf sRNA and PARE triplicates	GSE90771 (sRNA) and GSE113958 (PARE)	GEO	[158, 198]	D2A, B and C
<i>T. aestivum</i> sRNA and PARE dataset obtained from anthers under control conditions	GSM903669 (sRNA) and GSM911924 (PARE)	GEO	[196]	N/A
<i>T. aestivum</i> transcriptome	^b	Ensembl Plants (Release 38)	N/A	N/A

Appendix A Table 1 The datasets used in Chapter 3.

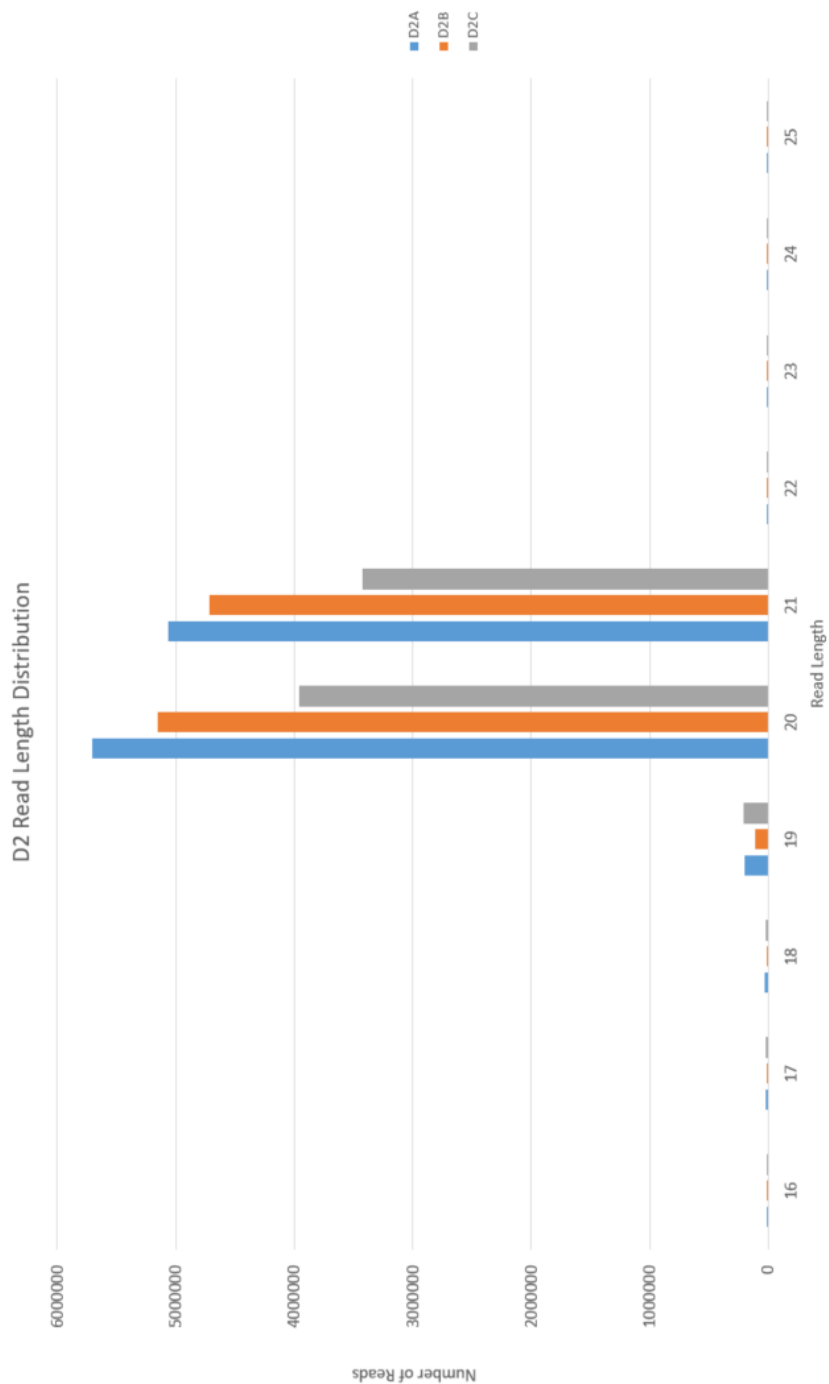
^ahttp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/TAIR10_cdna_20110103_representative_gene_model_updated

^bhttp://ftp.ensemblgenomes.org/pub/plants/release-38/fasta/triticum_aestivum/cdna/Triticum_aestivum.TGACv1.cdna.all.fa.gz

Appendix A Table 4 contains the results of the degradome analysis of dataset D2 using the Fahlgren and Carrington targeting rules, which allow mismatch and G:U wobble pairs at positions 10 and 11. Additionally, the built-in conservation filter was used to increase confidence in the reported interactions. This analysis identified 1072 category-0, 91 category 1, 611 category 2 and 529 category 3, making a total of 2303 interactions

Appendix A Table 5 contains the results of a genome-wide degradome analysis of the *T. aestivum* dataset (GEO accession GSE36867), which comprised a degradome of 4 306 082 non-redundant sequences and a corresponding sRNAome of 14 133 641 non-redundant sequences. The default stringent parameters identified 25 063 interactions, which comprised 12 120 category-0, 1026 category-1, 5576 category-2 and 6341 category-3 interactions and completed in just 31 minutes and 29 s.

Appendix A Table 6 contains the results of a genome-wide degradome analysis of the *T. aestivum* dataset (GEO accession GSE36867), which comprised a degradome of 4 306 082 non-redundant sequences and a corresponding sRNAome of 14 133 641 non-redundant sequences. The default flexible parameters identified 389 238 interactions, which comprised 83 409 category-0, 13 943 category-1, 79 935 category-2, 95 783 category-3 and 116 168 category-4 interactions with a run time of 19 h and 39 min.



Appendix A Figure 1 The sequence length distribution of the D2 degradome libraries. Majority of reads are either 20 and 21nt in length, as you would expect from data generated using the PARE protocol.

Appendix B

Dataset	# miRNAs	# V	Allen V	Inf. V	Allen NV	Inf. NV	Allen Se	Inf. Se	Allen PPV	Inf. PPV	Se gain	PPV difference
D1A	37	129	105	122	9	20	81.40%	94.57%	92.11%	85.92%	13.18%	-6.19%
D1B	38	131	109	126	11	25	83.21%	96.18%	90.83%	83.44%	12.98%	-7.39%
D1C	35	121	95	115	12	25	78.51%	95.04%	88.79%	82.14%	16.53%	-6.64%
D2A	40	140	117	133	14	43	83.57%	95.00%	89.31%	75.57%	11.43%	-13.75%
D2B	38	137	113	129	13	43	82.48%	94.16%	89.68%	75.00%	11.68%	-14.68%
D2C	40	144	117	138	3	11	81.25%	95.83%	97.50%	92.62%	14.58%	-4.88%
D3A	32	79	64	77	4	10	81.01%	97.47%	94.12%	88.51%	16.46%	-5.61%
D3B	29	70	57	66	11	17	81.43%	94.29%	83.82%	79.520%	12.86%	-4.31%
D3C	36	111	84	106	6	12	75.68%	95.50%	93.33%	89.83%	19.82%	-3.50%
D3D	35	104	88	101	3	7	84.62%	97.12%	96.70%	93.52%	12.50%	-3.19%

Appendix B Table 1 Sensitivity and precision values for the Allen *et al.* and manually inferred criteria over all the *A. thaliana* datasets. Allen = Allen *et al.* rules, Inf. = manually inferred criteria, V = validated, NV = non-validated, NV = non-validated, Se = sensitivity and PPV = precision. An increase in the achieved Se is observed for the inferred criteria; the decrease in PPV may be due to the lack of low-throughput validations for a subset of interactions, for which a clear signal is observed in the PARE datasets.

Dataset	# sRNAs	# PARE seqs	Run-time (hh:mm:ss)	Memory (GB)
D1A	134 284 6	111 145 49	00:18:51	6
D1B	109 334 4	101 036 90	00:17:41	6
D1C	110 622 2	771 525 1	00:17:11	6
D2A	193 502 5	199 306 92	00:31:18	8
D2B	908 368	194 704 87	00:38:09	8
D2C	568 633	727 512 3	01:04:24	8
D3A	379 756 1	246 325 1	00:18:15	5
D3B	163 373 0	230 054 1	00:16:52	5
D3C	283 730 4	397 528 0	00:20:30	5
D3D	117 642 4	903 209 3	00:17:59	5
D4A	176 589 3	430 500 9	01:30:15	6
D4B	456 068 4	399 261 8	01:34:12	6
D5	237 030 0	770 447 4	02:00:26	7
D6	199 194 2	250 552 3	00:41:43	5
D7	517 858 7	14363576	34:37:10	10

Appendix B Table 2 The timing and memory usage results for the PAREameters analysis of all datasets. The size of the input data (in terms of unique sequences) and complexity of the underlying genome are the main drivers for both run-time and resource usage.

Description	File/Accession	File source	Reference	Chapter abbreviation
<i>A. thaliana</i> wild-type leaf sRNA and PARE triplicates	GSE90771 (sRNA) and GSE113958 (PARE)	GEO	[158, 198]	D1A, B and C
<i>A. thaliana</i> wild-type leaves in three growth stages: young, mature and early senescence	GSE55151	GEO	[197]	D2A, B and C
<i>A. thaliana</i> wild-type flower, leaf, root and seedling of plants grown at 15°C	BioProject PRJNA407271	NCBI	[83]	D3A, B, C and D
<i>A. thaliana</i> genome	^a	TAIR10	N/A	N/A
<i>A. thaliana</i> transcriptome	^b	TAIR10	N/A	N/A
<i>A. trichopoda</i> sRNA and PARE libraries obtained from leaf and opened female flower	GSE41811	GEO	N/A	D4A and B
<i>G. max</i> sRNA and PARE obtained from leaf (Mock)	GSE76636	GEO	[36]	D5
<i>O. sativa</i> sRNA and PARE obtained from inflorescence	GSE18251	GEO	[214]	D6
<i>T. aestivum</i> sRNA and PARE dataset obtained from anthers under control conditions	GSM903669 (sRNA) and GSM911924 (PARE)	GEO	[196]	D7
Genome and transcript for all species besides <i>A. thaliana</i>	^c	Plant Ensembl (Release 43)	N/A	N/A

Appendix B Table 3 The datasets used in Chapter 4.

^ahttps://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas

^bftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/TAIR10_cdna_20110103_representative_gene_model_updated

^c<ftp://ftp.ensemblgenomes.org/pub/plants/release-43/fasta>

Dataset	Raw reads	NR reads	Genome matched	% matched	# miRNAs
D1A	6 664 998	1 342 846	1 006 022	74.92%	230
D1B	4 492 236	1 093 344	818 760	74.89%	213
D1C	5 148 552	1 106 222	837 919	75.75%	230
D2A	27 870 710	1 935 025	1 091 177	56.40%	252
D2B	26 211 889	908 368	525 503	57.90%	239
D2C	28 700 595	568 633	240 613	42.31%	209
D3A	18 460 973	3 797 561	2 534 869	66.75%	200
D3B	10 408 796	1 633 730	1 101 861	67.44%	186
D3C	19 946 757	2 837 304	1 647 286	58.06%	211
D3D	6 645 127	1 176 424	836 656	71.12%	141
D4A	18 092 450	1 765 893	656 968	37.20%	153
D4B	25 781 233	4 560 684	2 685 317	58.88%	161
D5	33 230 948	2 370 300	1 641 914	69.27%	309
D6	4 029 462	1 991 942	1 991 720	99.99%	227
D7	12 541 386	5 178 587	4 325 043	83.51%	168

Appendix B Table 4 Summary statistics of the sRNA sequencing data analysed within this chapter. The number of raw (redundant), unique (NR), genome matched reads and the number of miRNAs in the dataset. In order for an annotated miRNA to be considered present, it must have had an abundance ≥ 5 within the dataset.

Dataset	Raw reads	NR reads	Transcriptome matched	% matched
D1A	44 871 978	11 114 549	8 802 080	79.19%
D1B	34 315 808	10 103 690	8 176 679	80.93%
D1C	25 588 818	7 715 251	6 251 721	81.03%
D2A	115 224 802	19 930 692	10 565 361	53.01%
D2B	107 999 423	19 470 487	11 187 427	57.46%
D2C	191 294 550	7 275 123	1 734 835	23.85%
D3A	5 263 291	2 463 251	2 131 493	86.53%
D3B	4 809 175	2 300 541	1 998 764	86.88%
D3C	12 666 325	3 975 280	3 454 112	86.89%
D3D	53 840 936	9 032 093	6 824 080	75.56%
D4A	19 990 216	4 305 009	1 940 976	45.09%
D4B	12 609 502	3 992 618	1 679 932	42.08%
D5	26 251 057	7 704 474	6 230 026	80.86%
D6	4 426 044	2 505 523	2 268 297	90.53%
D7	35 477 509	14 363 576	10 366 761	72.17%

Appendix B Table 5 Summary statistics of the degradome sequencing data analysed in this chapter. Table presents the number of raw (redundant), unique (NR) and transcriptome (positive strand only) matching reads within each degradome dataset.

Retain rate	D1A Sensitivity	D1A Precision	D1B Sensitivity	D1B Precision	D1C Sensitivity	D1C Precision
0.5	23.3%	100.0%	22.9%	100.0%	22.3%	96.4%
0.55	24.8%	100.0%	26.7%	100.0%	23.1%	96.6%
0.6	26.4%	100.0%	27.5%	100.0%	25.6%	93.9%
0.65	27.1%	100.0%	32.8%	100.0%	30.6%	90.2%
0.7	73.6%	95.0%	72.5%	95.0%	76.9%	92.1%
0.75	75.2%	95.1%	74.8%	95.2%	78.5%	92.2%
0.8	83.7%	93.9%	84.7%	93.3%	86.0%	88.9%
0.85	86.8%	93.3%	88.6%	92.1%	88.4%	88.4%
0.9	87.6%	90.4%	89.3%	89.3%	88.4%	85.6%
0.95	96.1%	85.5%	96.2%	83.4%	94.2%	81.4%
1	99.2%	82.1%	99.2%	78.3%	99.2%	79.0%

Appendix B Table 6 Analysis of the retain rate parameter using the *A. thaliana* D1 dataset. The dual optimization problem for maximizing the increase in sensitivity (Se) and minimizing the loss in precision (PPV) is solved using the Se/PPV ratio, which for this dataset achieves its value at 0.85.

Transition	D1A			D1B			D1C		
	Se gain	PPV difference	Ratio	Se gain	PPV difference	Ratio	Se gain	PPV difference	Ratio
0.50-0.55	1.5%	0.0%	-	3.8%	0.0%	-	0.8%	0.2%	4.0
0.55-0.60	1.6%	0.0%	-	0.8%	0.0%	-	2.5%	-2.7%	0.9
0.60-0.65	0.7%	0.0%	-	5.3%	0.0%	-	5.0%	-3.7%	1.4
0.65-0.70	46.5%	-5.0%	9.3	39.7%	-5.0%	7.9	46.3%	1.9%	24.4
0.70-0.75	1.6%	0.1%	16.0	2.3%	0.2%	11.5	1.6%	0.1%	16.0
0.75-0.80	8.5%	-1.2%	7.1	9.9%	-1.9%	5.2	7.5%	-3.3%	2.3
0.80-0.85	3.1%	-0.6%	5.2	3.9%	-1.2%	3.3	2.4%	-0.5%	4.8
0.85-0.90	0.8%	-2.9%	0.3	0.7%	-2.8%	0.3	0.0%	-2.8%	0.0
0.90-0.95	8.5%	-4.9%	1.7	6.9%	-5.9%	1.2	5.8%	-4.2%	1.4
0.95-1	3.1%	-3.5%	0.9	3.0%	-5.1%	0.6	5.0%	-2.4%	2.1

Appendix B Table 7 Sensitivity (Se) gain vs precision (PPV) loss and the absolute ratio between them for values of the retain rate parameter in the 0.50-1.00 range, over three *A. thaliana* leaf replicates (D1 dataset). The sensitivity and precision were evaluated on the experimentally validated miRNA-mRNA interactions in *A. thaliana* that had corresponding HC transcript peaks within the degradome dataset. The ratio is defined as the absolute value of Se/PPV. The optimal value is obtained for the last transition that results in a Se/PPV ratio greater than or equal to 1 (i.e. there is an equal or greater gain in sensitivity with respect to the loss in precision) and for the D1 dataset, this value was obtained at the 0.80-0.85 transition.

Appendix B Table 8 A subset of the experimentally validated miRNA targets (Appendix A Table 1) containing only conserved miRNAs, which comprises 201 miRNA–mRNA interactions from 42 unique miRNA sequences. Full table can be downloaded from NAR online: <https://doi.org/10.1093/nar/gkz1234>. It is also included as supplementary information with this thesis.

Appendix B Table 9 A subset of the experimentally validated miRNA targets (Appendix A Table 1) containing miRNAs specific to the Brassicaceae family, which comprises 184 interactions from 47 unique miRNA sequences. Full table can be downloaded from NAR online: <https://doi.org/10.1093/nar/gkz1234>. It is also included as supplementary information with this thesis.

Parameter	Conserved	Species-specific
Allow MM at pos 10	No	Yes
Allow MM at pos 11	Yes	Yes
Max # adj mm in CR	0	0
Max # MM in CR	1	1
Max score	4	4
Max # MM	3	2
Max # G:U	2	2
Max # adj MM	1	1
MFE ratio cut-off	0.75	0.68

Appendix B Table 10 The PAREameters inferred parameters for the conserved and species-specific miRNA targets in *A. thaliana*.

	Flower tissue				Leaf tissue			
	<i>A. thaliana</i> (D3A)	<i>A. trichopoda</i> (D4B)	<i>O. sativa</i>	<i>T. aestivum</i>	<i>A. thaliana</i> (D1A)	<i>A. thaliana</i> (D2B)	<i>A. trichopoda</i> (D4A)	<i>G. max</i>
Allow MM at position 10	Yes	No	No	Yes	Yes	Yes	No	Yes
Allow MM at position 11	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Max # adjacent MM in CR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max # MM in CR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Max score	4.50	4.50	4.00	5.00	4.50	5.00	4.50	4.50
Max # MM	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Max # G:U	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Max # adjacent MM	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
MFE ratio cut-off	0.72	0.72	0.70	0.66	0.69	0.65	0.71	0.69

Appendix B Table 11 Overview of data-inferred thresholds inferred using PAREameters on model and non-model organisms in flower and leaf tissue. Differences in reported thresholds are observed both between organisms (e.g. monocots versus dicots) and between tissues

Species	<i>A. trichopoda</i>				<i>O. sativa</i>				<i>T. aestivum</i>			
	χ^2	MM	GU	Gap	χ^2	MM	GU	Gap	χ^2	MM	GU	Gap
1	0.740	0.170	1.000	1.000	0.007	0.0140	0.010	1.000	0.998	1.000	1.000	1.000
2	1.000	1.000	0.849	1.000	0.125	0.932	0.704	0.602	0.042	1.000	0.681	0.017
3	1.000	1.000	1.000	1.000	0.955	1.000	1.000	1.000	0.385	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	0.929	0.932	1.000	1.000	0.992	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	0.846	0.956	0.704	1.000	0.992	1.000	1.000	1.000
6	1.000	1.000	1.000	1.000	0.685	1.000	0.498	1.000	0.992	1.000	1.000	1.000
7	1.000	1.000	0.849	1.000	0.493	0.932	0.306	1.000	0.971	1.000	1.000	1.000
8	1.000	1.000	0.849	1.000	0.955	1.000	1.000	1.000	0.992	1.000	0.681	1.000
9	1.000	1.000	0.849	1.000	0.001	1.000	$p < 0.001$	1.000	0.579	1.000	1.000	1.000
10	1.000	1.000	1.000	1.000	0.889	0.932	1.000	1.000	0.992	1.000	1.000	1.000
11	1.000	1.000	1.000	1.000	0.685	0.635	1.000	1.000	0.992	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000	0.955	1.000	0.804	1.000	0.579	0.573	1.000	1.000
13	1.000	1.000	0.849	1.000	0.846	0.854	1.000	1.000	0.992	1.000	1.000	1.000
14	1.000	1.000	0.849	1.000	$p < 0.001$	$p < 0.001$	1.000	1.000	0.579	0.488	1.000	1.000
15	1.000	1.000	1.000	1.000	0.685	0.854	0.498	1.000	0.971	1.000	1.000	1.000
16	1.000	0.540	0.849	1.000	0.697	0.635	1.000	1.000	0.971	1.000	0.885	1.000
17	1.000	1.000	1.000	1.000	$p < 0.001$	0.854	$p < 0.001$	1.000	0.971	1.000	0.829	1.000
18	1.000	1.000	0.849	1.000	0.383	0.562	0.498	1.000	0.579	1.000	1.000	1.000
19	1.000	1.000	0.849	1.000	0.846	1.000	0.498	1.000	0.971	1.000	1.000	1.000
20	1.000	1.000	0.849	1.000	$p < 0.001$	$p < 0.001$	0.498	0.860	0.306	0.283	1.000	1.000
21	1.000	1.000	0.849	1.000	0.671	0.555	0.762	1.000	0.992	1.000	1.000	1.000

Appendix B Table 12 χ^2 and Fisher's exact test significance results on the position-specific properties for non-model organisms versus *A. thaliana* in flower tissue. Values below the significance threshold (0.05) are highlighted in bold. Any extreme p -values (i.e. $p < 0.001$) are reported as $p < 0.001$.

Species	<i>A. thaliana</i>	<i>A. trichopoda</i>	<i>O. sativa</i>	<i>T. aestivum</i>
<i>A. thaliana</i>	1	0.331	2.54x10 ⁻⁴	6.89x10 ⁻⁶
<i>A. trichopoda</i>		1	1.93x10 ⁻⁷	1.38x10 ⁻⁸
<i>O. sativa</i>			1	0.166
<i>T. aestivum</i>				1

Appendix B Table 13 Results of the Kolmogorov-Smirnov test when evaluating the differences between MFE ratio distributions of HC miRNA-mRNA interactions found in flower tissue in model and non-model organisms. The results highlight the significant differences observed between dicots (*A. thaliana* and *A. trichopoda*) and monocots (*O. sativa* and *T. aestivum*).

miRNA position	χ^2	MM	G:U	Gap
1	0.991	1.000	1.000	1.000
2	0.991	1.000	1.000	1.000
3	0.991	1.000	1.000	1.000
4	0.991	1.000	1.000	1.000
5	0.991	1.000	1.000	1.000
6	0.991	1.000	1.000	1.000
7	0.991	1.000	1.000	1.000
8	0.991	1.000	1.000	1.000
9	0.991	1.000	1.000	1.000
10	0.991	1.000	1.000	1.000
11	0.991	1.000	1.000	1.000
12	0.991	1.000	1.000	1.000
13	0.991	1.000	1.000	1.000
14	0.991	1.000	1.000	1.000
15	0.991	1.000	1.000	1.000
16	0.991	1.000	1.000	1.000
17	0.991	1.000	1.000	1.000
18	0.991	1.000	1.000	1.000
19	0.991	1.000	1.000	1.000
20	0.991	1.000	1.000	1.000
21	0.991	1.000	1.000	1.000

Appendix B Table 14 χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in *A. trichopoda* leaf.

miRNA position	χ^2	MM	G:U	Gap
1	0.390	1.000	0.085	1.000
2	0.390	1.000	0.085	1.000
3	0.954	1.000	1.000	1.000
4	0.954	1.000	1.000	1.000
5	0.516	1.000	0.129	1.000
6	0.848	1.000	1.000	1.000
7	0.954	1.000	1.000	1.000
8	0.921	1.000	0.741	1.000
9	0.516	0.734	1.000	1.000
10	0.954	1.000	1.000	1.000
11	0.921	1.000	1.000	1.000
12	0.954	1.000	1.000	1.000
13	0.954	1.000	1.000	1.000
14	0.921	1.000	0.589	1.000
15	0.954	1.000	1.000	1.000
16	0.954	1.000	1.000	1.000
17	0.921	1.000	1.000	1.000
18	0.954	1.000	1.000	1.000
19	0.954	1.000	1.000	1.000
20	0.921	1.000	1.000	1.000
21	0.921	1.000	0.741	1.000

Appendix B Table 15 χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in *A. trichopoda* flower.

miRNA position	χ^2	MM	G:U	Gap
1	0.981	1.000	1.000	1.000
2	0.981	1.000	1.000	1.000
3	0.981	1.000	1.000	1.000
4	0.981	1.000	1.000	1.000
5	0.981	1.000	1.000	1.000
6	0.981	1.000	1.000	1.000
7	0.981	1.000	1.000	1.000
8	0.981	1.000	1.000	1.000
9	0.981	1.000	1.000	1.000
10	0.981	1.000	1.000	1.000
11	0.981	1.000	1.000	1.000
12	0.981	1.000	1.000	1.000
13	0.981	1.000	1.000	1.000
14	0.981	1.000	1.000	1.000
15	0.981	1.000	1.000	1.000
16	0.981	1.000	1.000	1.000
17	0.981	1.000	1.000	1.000
18	0.981	1.000	1.000	1.000
19	0.981	1.000	1.000	1.000
20	0.981	1.000	1.000	1.000
21	0.981	1.000	1.000	1.000

Appendix B Table 16 χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in *G. max* leaf.

miRNA position	χ^2	MM	G:U	Gap
1	0.326	0.249	1.000	1.000
2	0.143	0.249	1.000	0.840
3	0.808	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000
5	0.880	0.747	1.000	1.000
6	0.045	0.249	0.052	1.000
7	0.580	0.249	1.000	1.000
8	0.588	0.602	1.000	1.000
9	0.204	0.059	1.000	1.000
10	0.880	1.000	1.000	1.000
11	0.880	1.000	1.000	1.000
12	0.128	0.147	0.936	1.000
13	0.980	1.000	1.000	1.000
14	0.045	0.045	1.000	1.000
15	0.174	0.228	1.000	1.000
16	0.880	1.000	1.000	1.000
17	0.880	0.602	1.000	1.000
18	0.880	1.000	1.000	1.000
19	0.880	0.602	1.000	1.000
20	0.143	0.059	1.000	1.000
21	$p < 0.001$	$p < 0.001$	0.004	1.000

Appendix B Table 17 χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in *O. sativa* inflorescence. Values below the significance threshold (0.05) are highlighted in bold. Any extreme p -values (i.e. $p < 0.001$) were reported as $p < 0.001$.

miRNA position	χ^2	MM	G:U	Gap
1	$p < 0.001$	0.516	$p < 0.001$	1.000
2	0.040	0.757	0.505	0.098
3	0.045	0.342	0.052	1.000
4	0.445	0.342	1.000	1.000
5	0.521	0.516	0.767	1.000
6	0.213	0.171	0.505	1.000
7	0.651	0.757	0.602	1.000
8	0.145	0.725	0.147	1.000
9	0.651	0.615	0.865	1.000
10	0.651	0.342	1.000	1.000
11	0.791	1.000	0.602	1.000
12	0.005	0.003	0.586	1.000
13	0.261	0.342	0.602	1.000
14	0.005	0.002	0.667	1.000
15	0.014	0.011	0.505	1.000
16	0.068	0.011	0.865	1.000
17	0.129	0.227	0.147	1.000
18	0.001	0.003	0.147	1.000
19	0.521	0.584	0.767	1.000
20	0.145	0.120	0.752	1.000
21	0.272	0.171	0.505	1.000

Appendix B Table 18 χ^2 and Fisher's exact test significance results on the position-specific properties for conserved and species-specific miRNA-mRNA interactions in *T. aestivum* spikes. Any extreme p -values (i.e. $p < 0.001$) were reported as $p < 0.001$.

Dataset	Allen <i>et al.</i> miRNAs	Allen <i>et al.</i> interactions	Inferred miRNAs	Inferred interactions
D4A	70 (3)	203 (9)	72 (5)	210 (16)
D4B	66 (2)	174 (4)	68(4)	182 (12)
D5	143 (6)	2118 (64)	143 (6)	2243 (189)
D6	42 (9)	149 (34)	33 (0)	115 (0)
D7	91 (2)	1257 (50)	99 (10)	1417 (210)

Appendix B Table 19 Intersection analysis of interactions predicted using the Allen *et al.* rules and the PAREameters inferred rules on various datasets. The number in brackets represents the miRNAs and interactions specific to the criteria used. The exact sensitivity and precision values cannot be computed on non-model organisms due to the lack of a large enough set of validated interactions

Dataset	Allen <i>et al.</i> miRNAs	Allen <i>et al.</i> interactions	Inferred miRNAs	Inferred interactions
D4A	70	203	87	272
D4B	66	174	79	208
D5	143	2118	190	2842
D6	42	149	46	161
D7	91	1257	193	2040

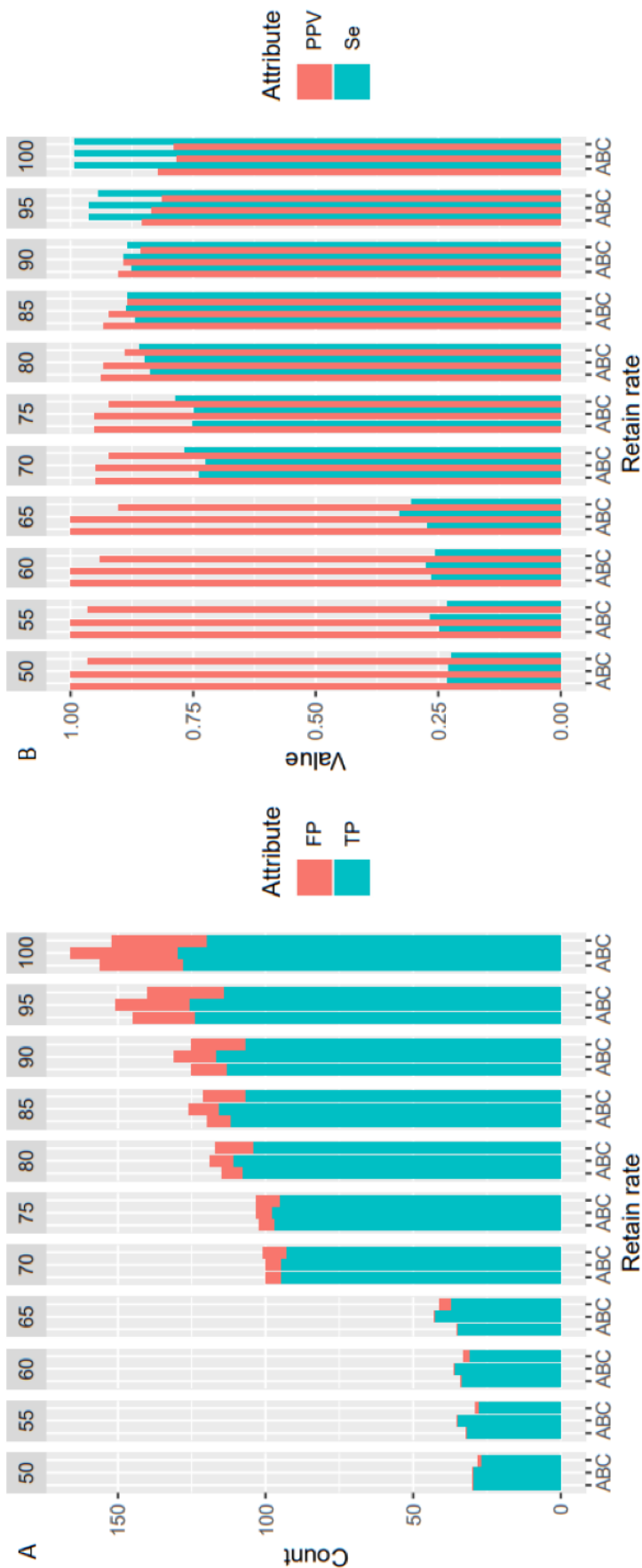
Appendix B Table 20 Intersection analysis of interactions predicted using the Allen *et al.* rules and the PAREameters inferred rules when using a retain rate of 1 on various datasets. All of the Allen *et al.* reported interactions are a subset of the inferred criteria reported interactions when using a retain rate of 1.

Dataset	D1B				D1C			
	χ^2	MM	GU	Gap	χ^2	MM	GU	Gap
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

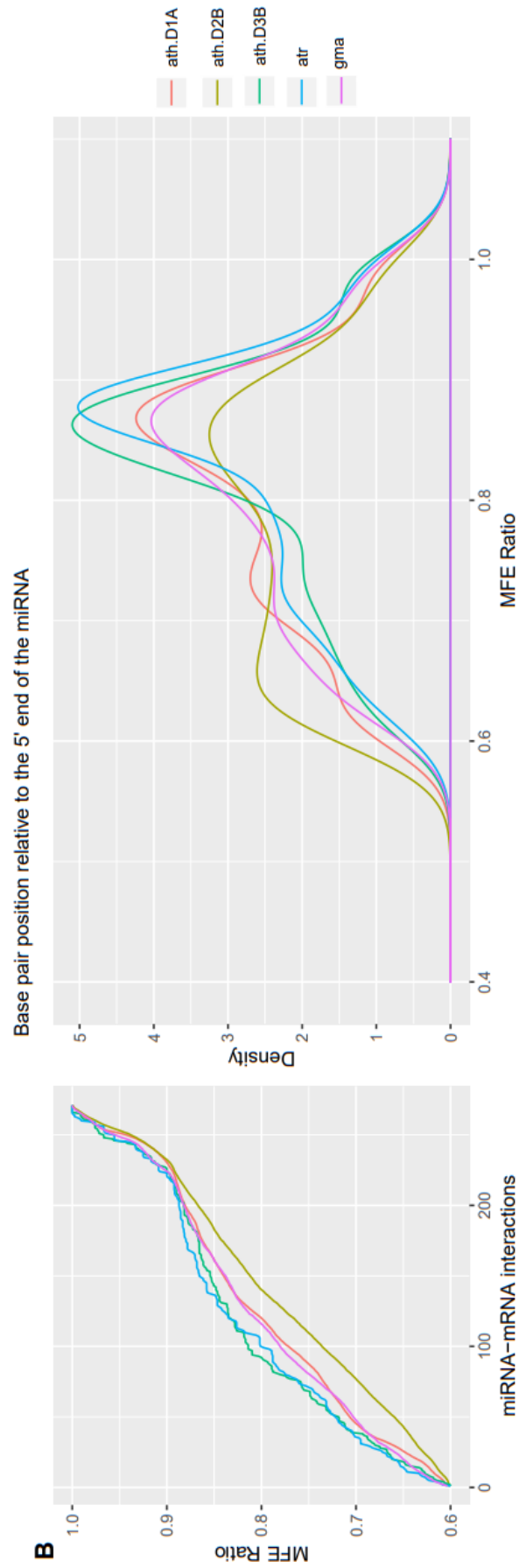
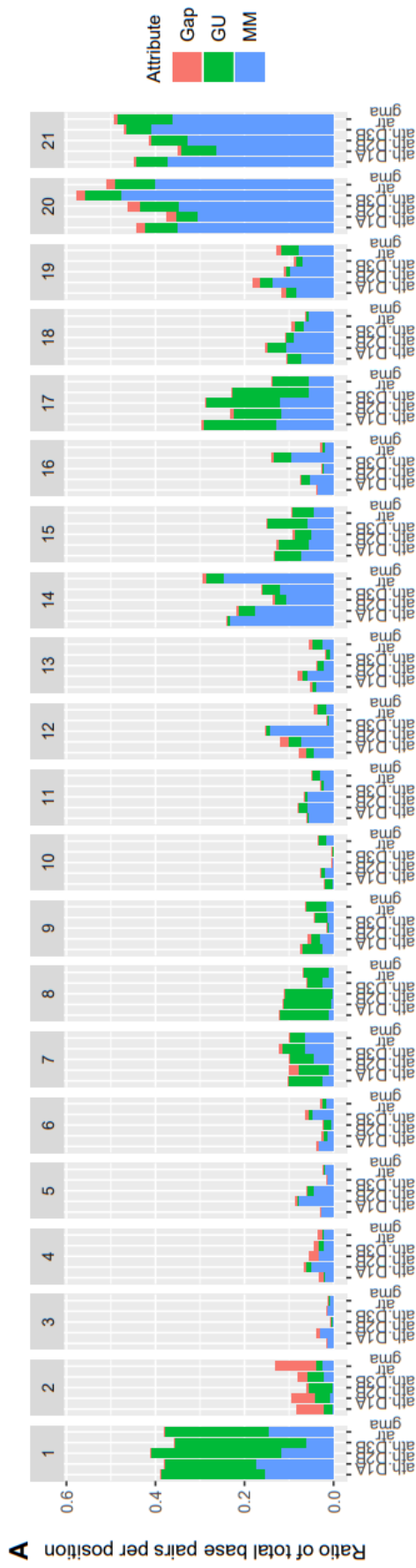
Appendix B Table 21 χ^2 and Fisher's exact test significance results on the position-specific properties for HC interactions in D1B and D1C versus *A. thaliana* D1A dataset. The contribution of specific properties such as MM, Gaps and G:U is assessed using Fisher exact tests.

Dataset	D1B				D1C			
	χ^2	MM	GU	Gap	χ^2	MM	GU	Gap
1	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
13	0.987	1.000	1.000	1.000	0.739	1.000	0.114	1.000
14	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
21	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000

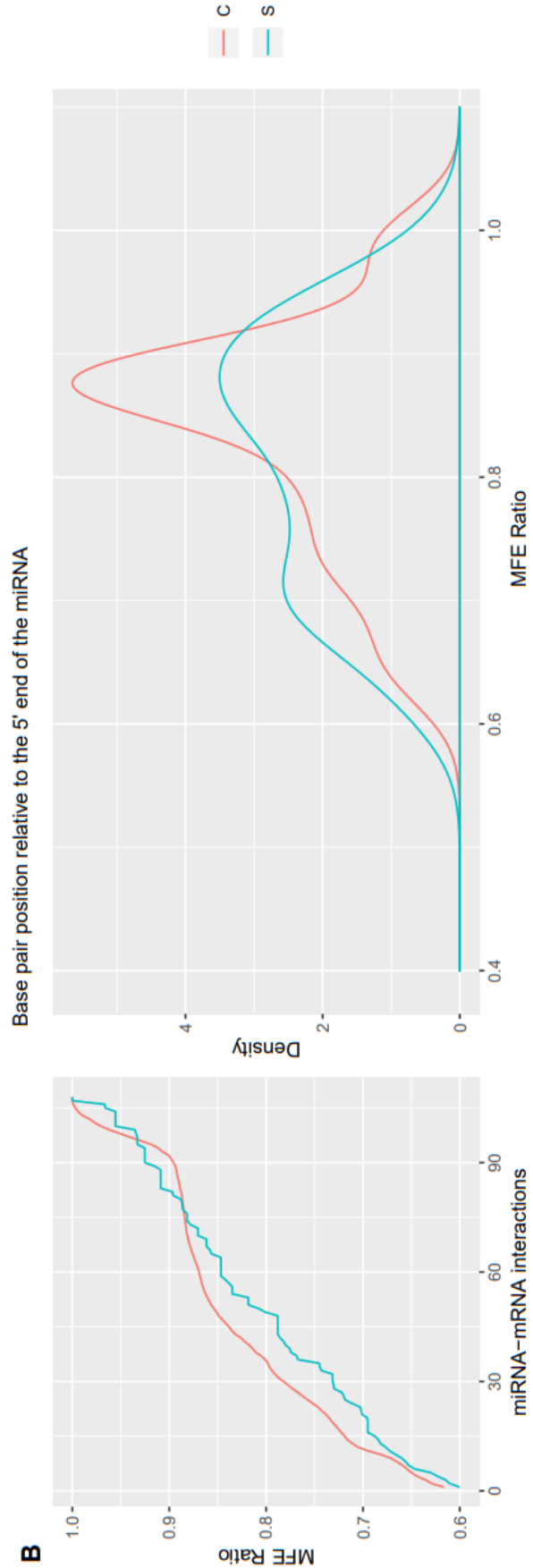
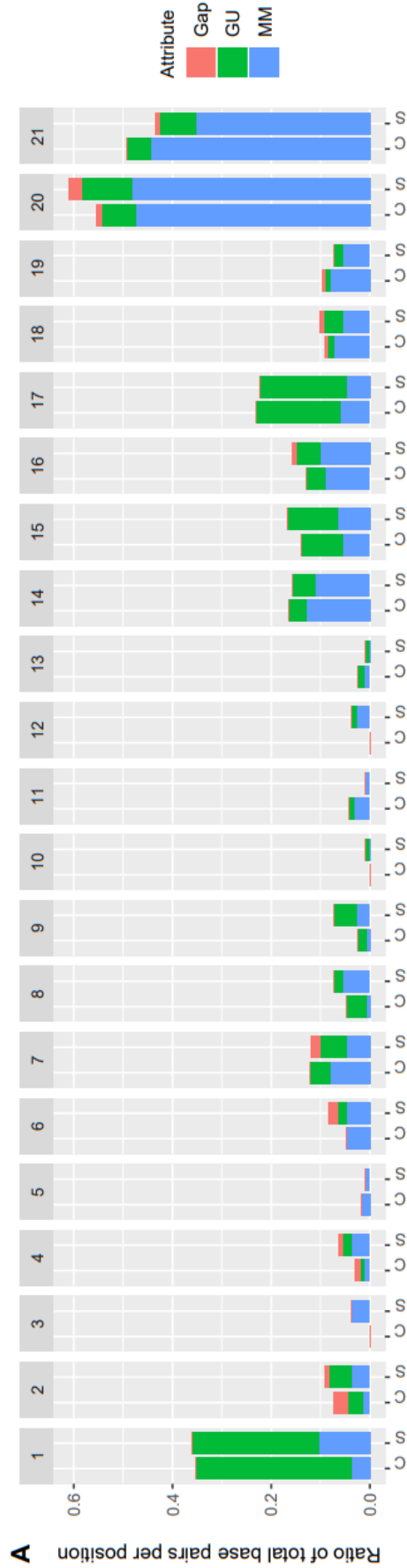
Appendix B Table 22 χ^2 and Fisher's exact test significance results on the position-specific properties for LC interactions in D1B and D1C versus *A. thaliana* D1A dataset. The contribution of specific properties such as MM, Gaps and G:U is assessed using Fisher exact tests.



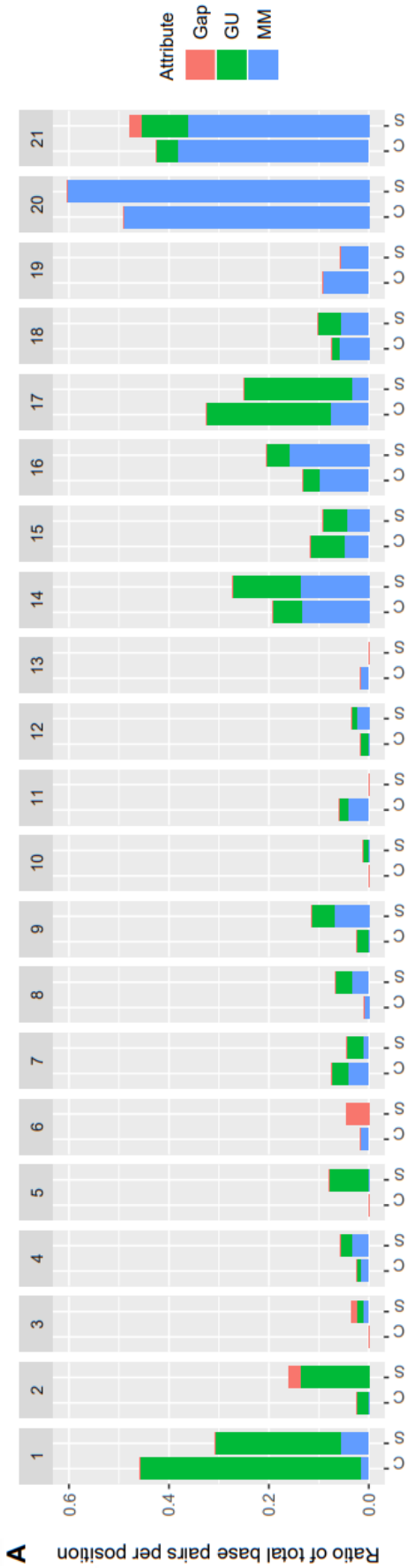
Appendix B Figure 1 Accuracy (sensitivity and precision) dependence on the retain rate parameter. Clustered frequency histograms of predicted interactions vs validated/non-validated ones (panel A) and the variation in sensitivity and precision values for increasing values of the retain rate parameter on three *A. thaliana* leaf replicates, the D1 dataset (panel B) highlight the existence of a data-driven optimum for the retain rate parameter. For this particular dataset the optimum on the Se/PPV ratio is achieved for 0.85. The data-driven optimal value for this parameter is suggested based on the input; however, it still remains a user-configurable parameter. Se = sensitivity, PPV = precision, FP = false positive, TP = true positive, A = D1A, B = D1B and C = DIC.



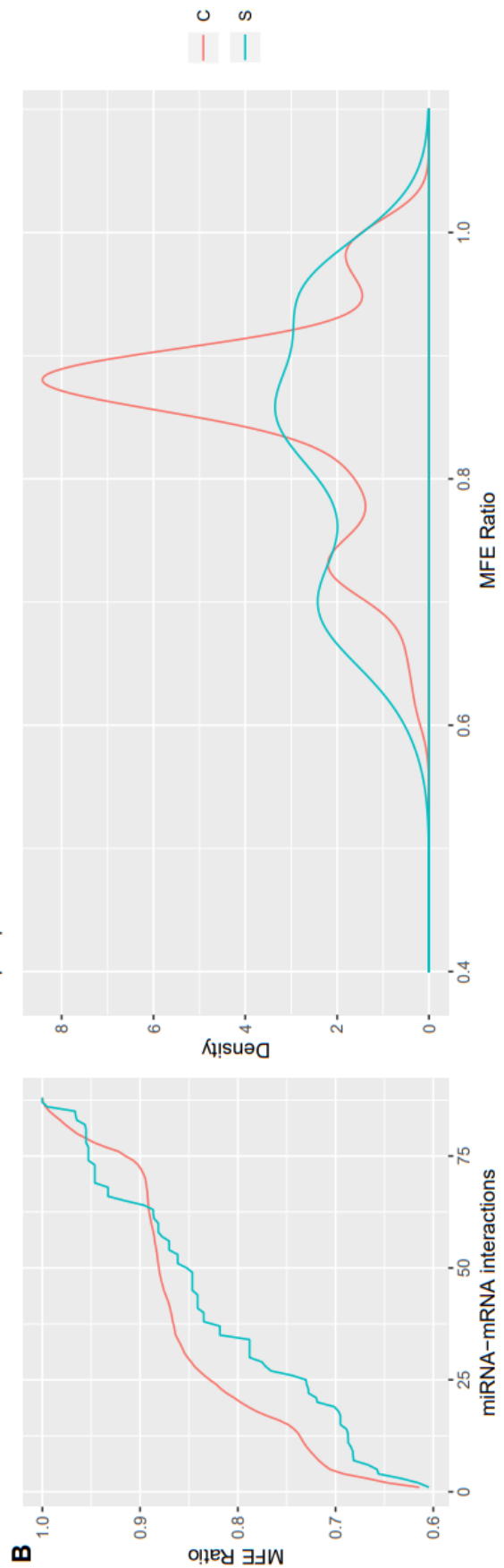
Appendix B Figure 2 Side-by-side comparison of leaf miRNA-mRNA interaction property distributions in different species and datasets. The position-specific properties (panel A) and MFE ratio distribution (panel B) of miRNA-mRNA interactions from leaf tissues in *A. thaliana*, *A. trichopoda* and *G. max*.



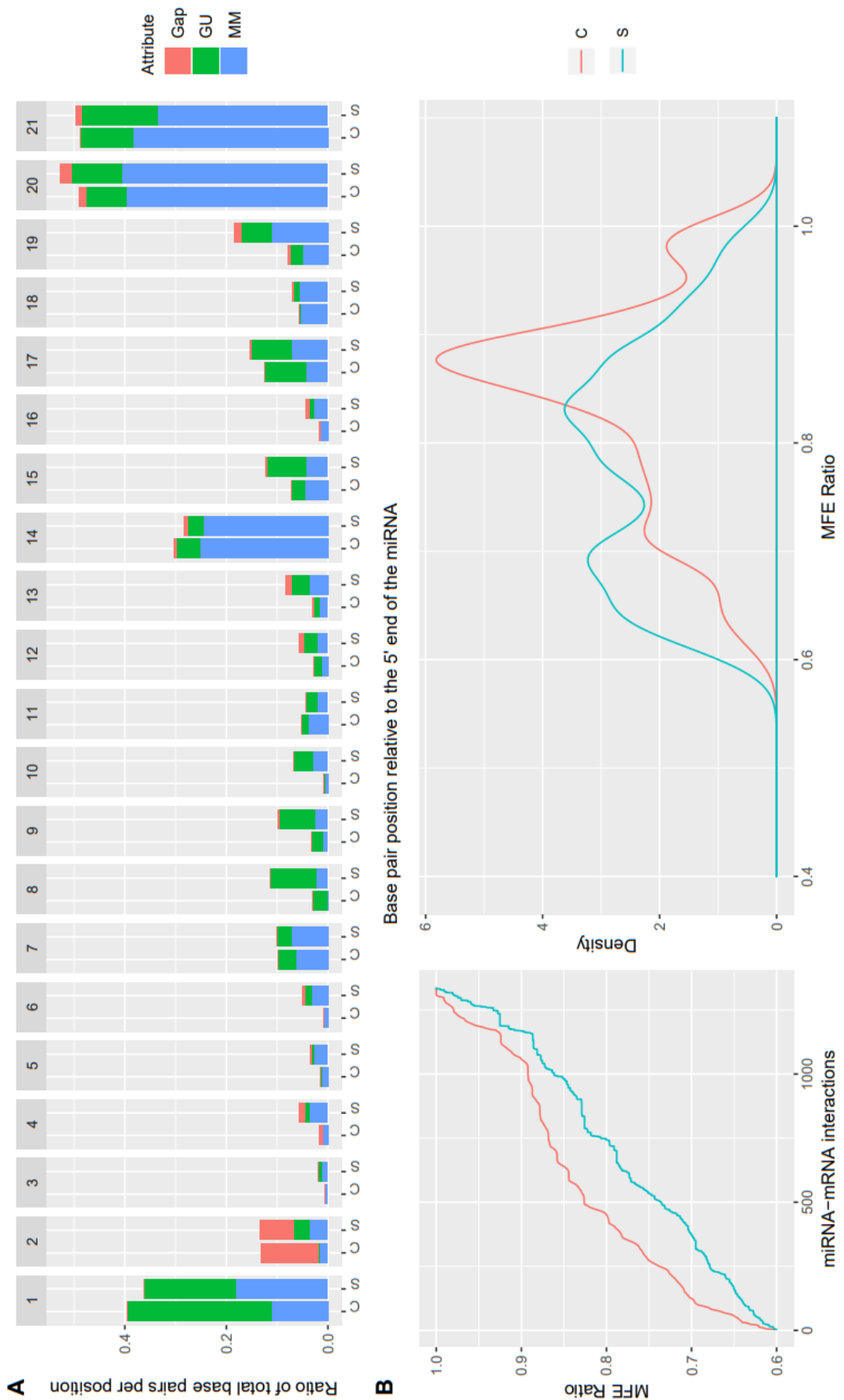
Appendix B Figure 3 Side-by-side comparison of property distributions for predicted HC interactions by conserved and species-specific miRNAs in *A. trichopoda* leaf. Using PAREamters HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of 0.1376412.



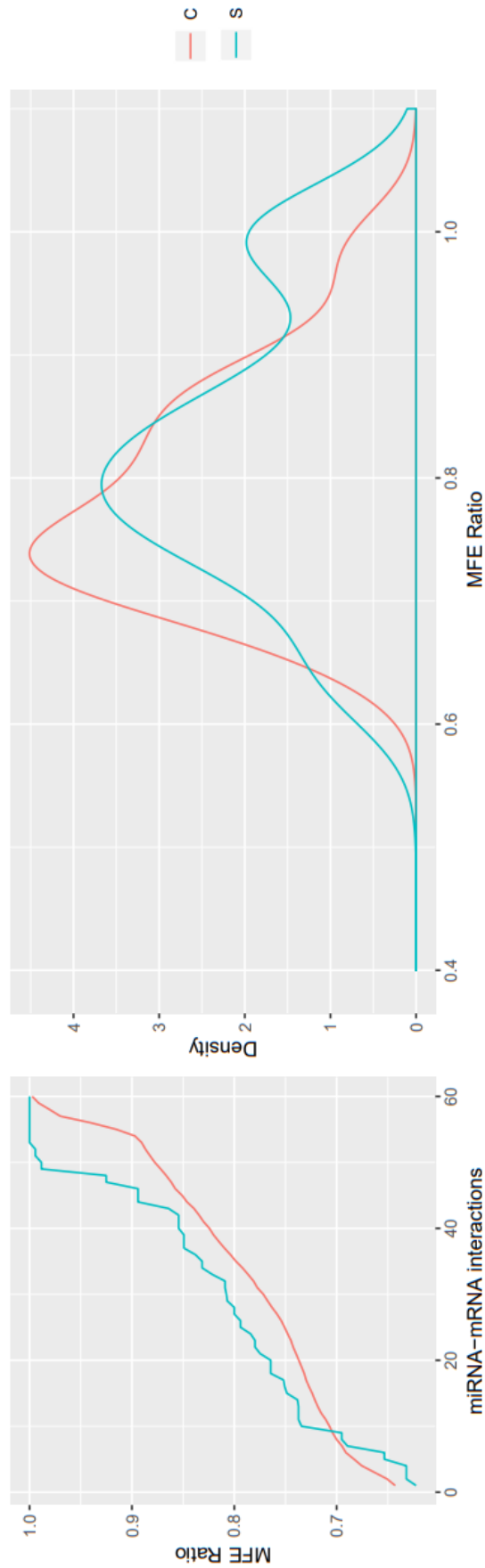
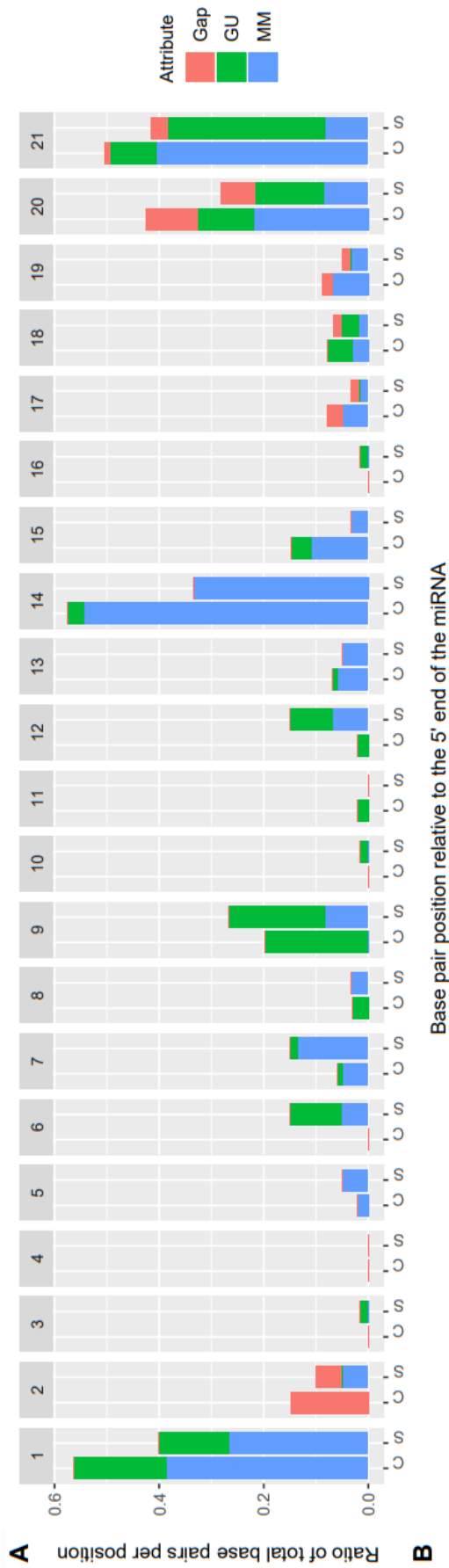
Base pair position relative to the 5' end of the miRNA



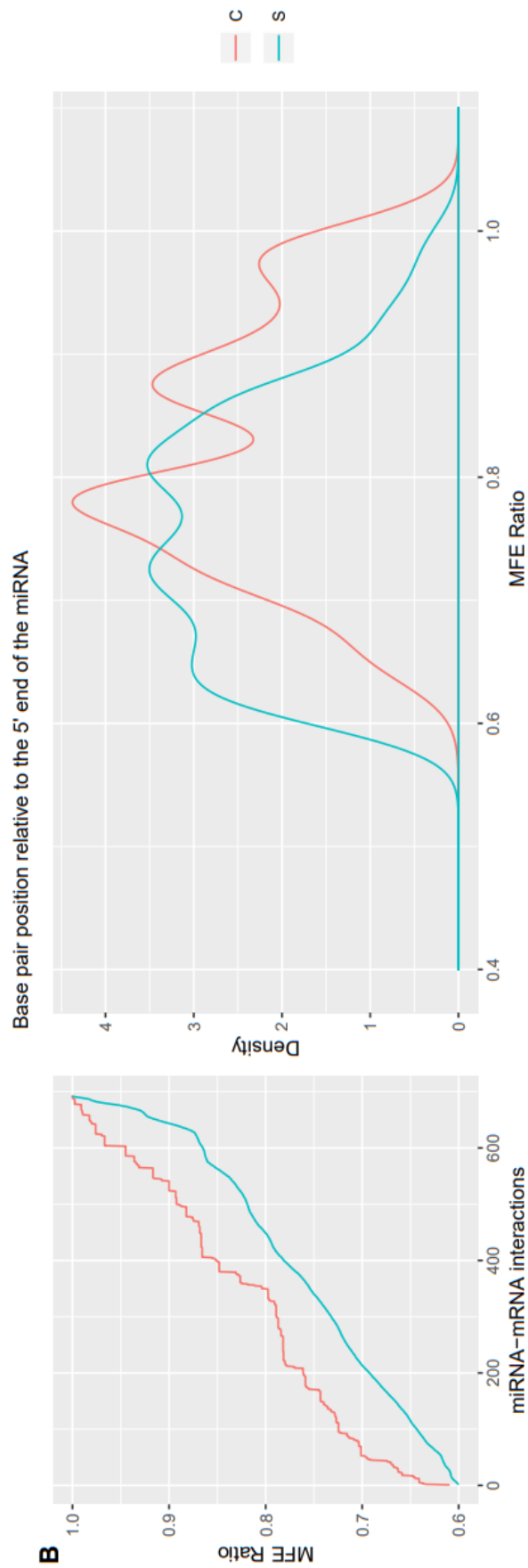
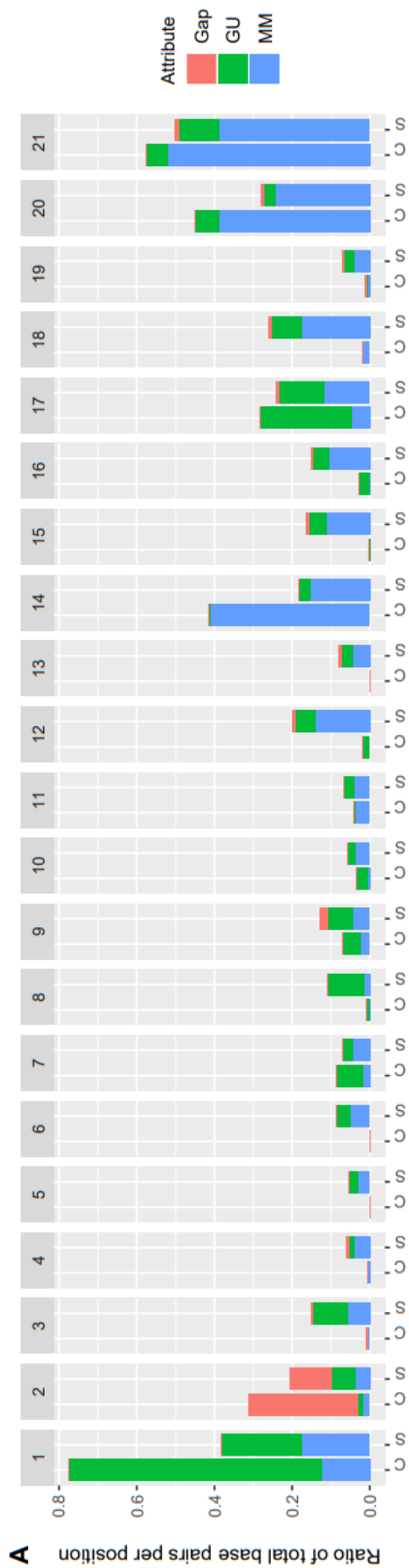
Appendix B Figure 4 Side-by-side comparison of property distributions for predicted HC interactions by conserved and species-specific miRNAs in *A. trichopoda* flower. Using PAREamters HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of 0.01332362.



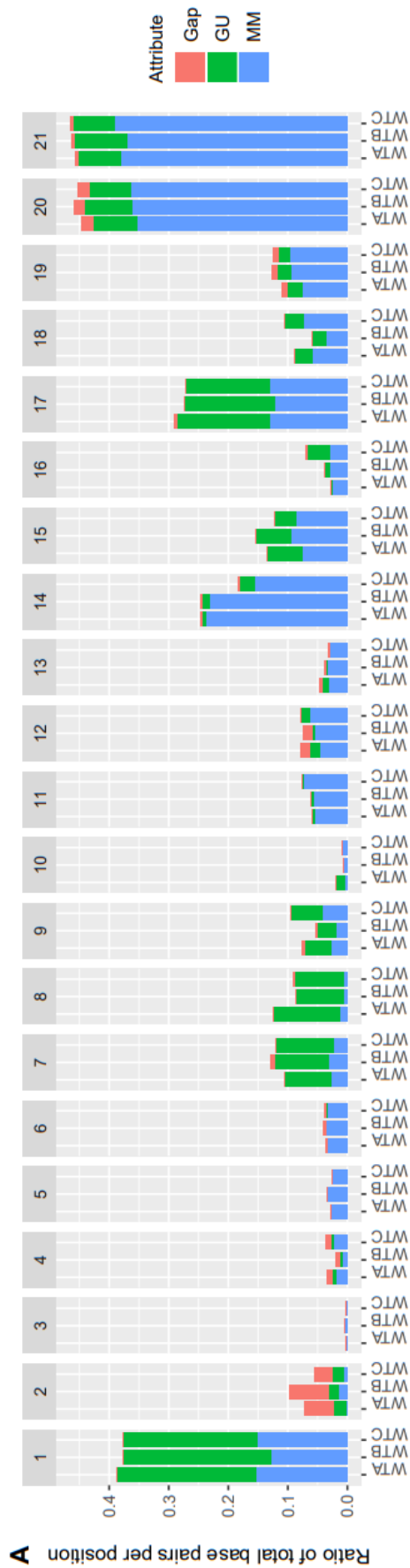
Appendix B Figure 5 Side-by-side comparison of property distributions for predicted HC interactions by conserved and species-specific miRNAs in *G. max* leaf. Using PAREamers HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of $p < 0.001$.



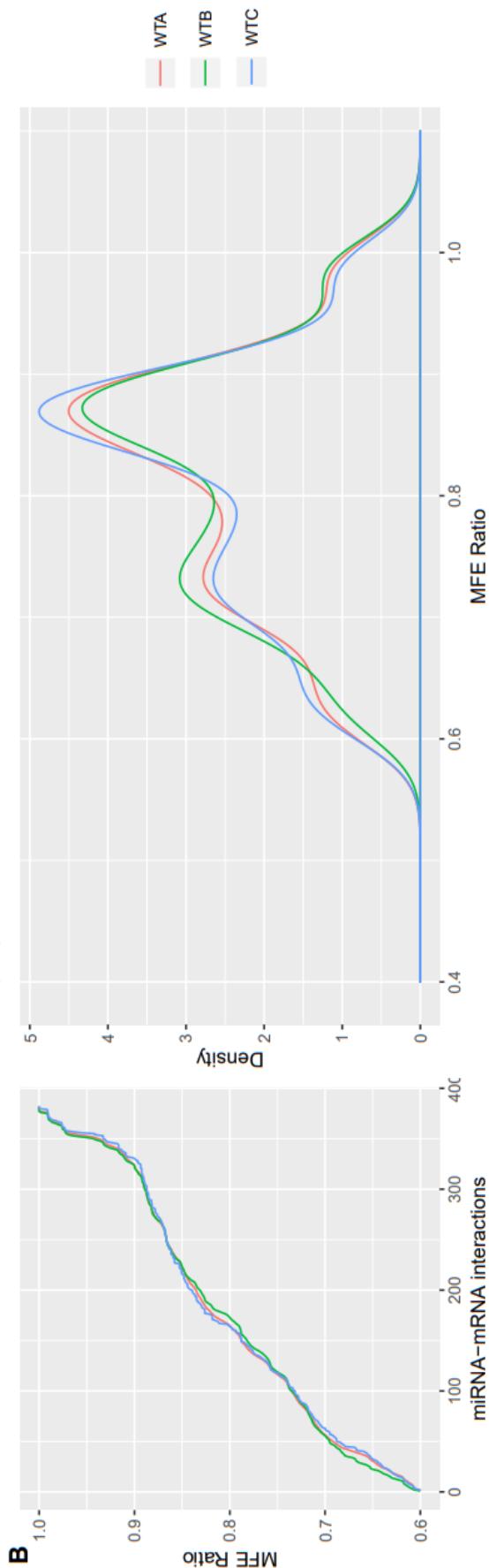
Appendix B Figure 6 Side-by-side comparison of property distributions for predicted HC interactions by conserved and species-specific miRNAs in *O. sativa* inflorescence. Using PAREamters HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of 0.2655691.



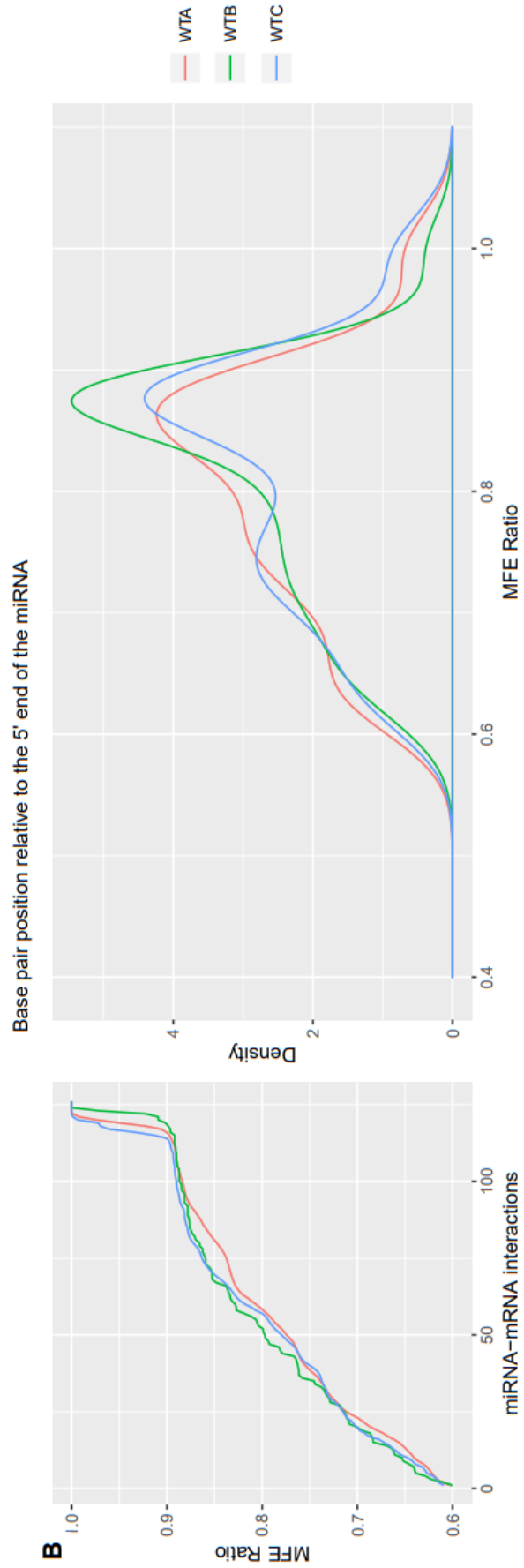
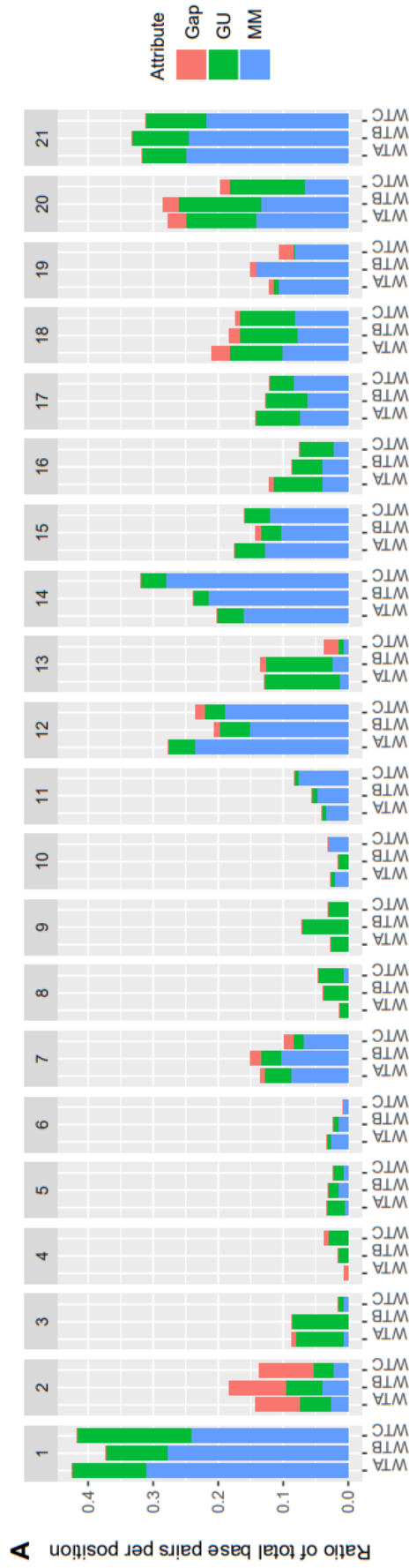
Appendix B Figure 7 Side-by-side comparison of property distributions for predicted HC interactions by conserved and species-specific miRNAs in *T. aestivum* spikes. Using PAREamters HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of $p < 0.001$.



Base pair position relative to the 5' end of the miRNA



Appendix B Figure 8 Side-by-side comparison of property distributions for predicted HC interactions in D1A (WTA), D1B (WTB) and DIC (WTC). Using PAREamers HC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a p -value of 0.9994413, 0.9973924 and 0.9596333 for D1A vs D1B, D1A vs DIC and D1B vs DIC, respectively.



Appendix B Figure 9 Side-by-side comparison of property distributions for predicted LC interactions in D1A (WTA), D1B (WTB) and DIC (WTC). Using PAREamers LC predicted miRNA-mRNA interactions as input, we calculated the position-specific properties (panel A) and the MFE ratio distribution (panel B) for the conserved and species-specific miRNA-mRNA interactions. The similarities in the distributions of the MFE ratios were evaluated using the Kolmogorov-Smirnov test, which reported a *p*-value of 0.418178, 0.617195 and 0.7229624 for D1A vs D1B, D1A vs DIC and D1B vs DIC, respectively.

Appendix C

Some of the tables referenced within Chapter 5 contain a large number predicted nat-siRNAs or their targets and are not practical to include within this thesis. However, for completeness, a brief description of each table is provided below and the actual data is provided as supplementary information included with the thesis.

Appendix C Table 2 The number of transcripts and *cis*-NATs, based on the genome annotations, in the plant species used for the computational benchmarking.

Appendix C Table 3 A comparison between the *G. max cis*-nat-siRNAs reported by Zheng *et al.* [235] and the prediction results from NATpare and NATpipe.

Appendix C Table 4 The results from the NATpare analyses on the *A. thaliana* control and salt stress tissues.

Appendix C Table 5 The nat-siRNAs, as predicted by NATpare, when performing analyses on the *A. thaliana* D3 dataset comprising of flower, leaf, root and seeding datasets.

Appendix C Table 6 The PAREsnip2 parameters used to predict targets for the reported nat-siRNAs.

Appendix C Table 7 The targets, predicted by PAREsnip2, when performing analyses on the nat-siRNAs identified by NATpare in the D3 dataset comprising of flower, leaf, root and seeding tissues.

Description	File/Accession	File source	Reference	Chapter abbreviation
Genome and transcriptome for all species used for computational benchmarking	^a	Ensembl Plant (Release 43)	N/A	N/A
<i>G. max</i> sRNA and PARE library used in the prediction performance benchmarking	GSE33380	GEO	[235]	D1
<i>G. max</i> transcriptome used in the prediction performance benchmarking	^b	Phytozome (Version 12)	N/A	N/A
<i>A. thaliana</i> sRNA triplicates obtained from seedling under salt stress	GSE66599	GEO	[13]	D2
<i>A. thaliana</i> wild-type flower, leaf, root and seedling of plants grown at 21°C	BioProject PRJNA407271	NCBI	[83]	D3

Appendix C Table 1 The datasets used in Chapter 5.

^a<ftp://ftp.ensemblgenomes.org/pub/plants/release-43/fasta>

^b<https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Gmax>

Appendix D

Description	File/Accession	File source	Reference
<i>S. lycopersicum</i> genome	^a	Sol Genomics	[44]
<i>S. lycopersicum</i> transcriptome	^b	Sol Genomics	[44]
CMV genome (strain Fny)	^c	NCBI	N/A
D-sat RNA	^d	NCBI	N/A

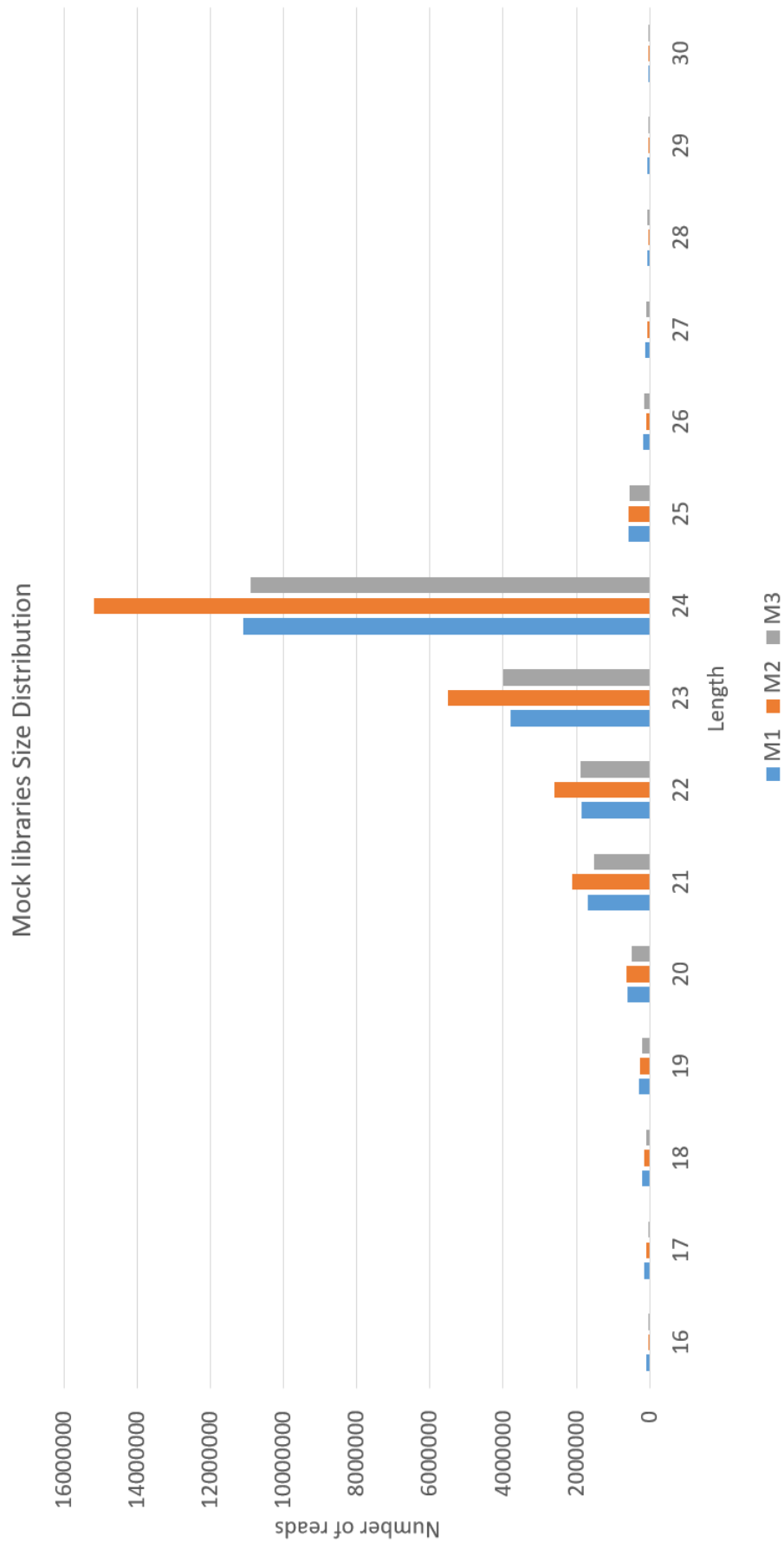
Appendix D Table 1 The datasets used in Chapter 6.

^aftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/assembly/build_3.00/S_lycopersicum_chromosomes.3.00.fasta

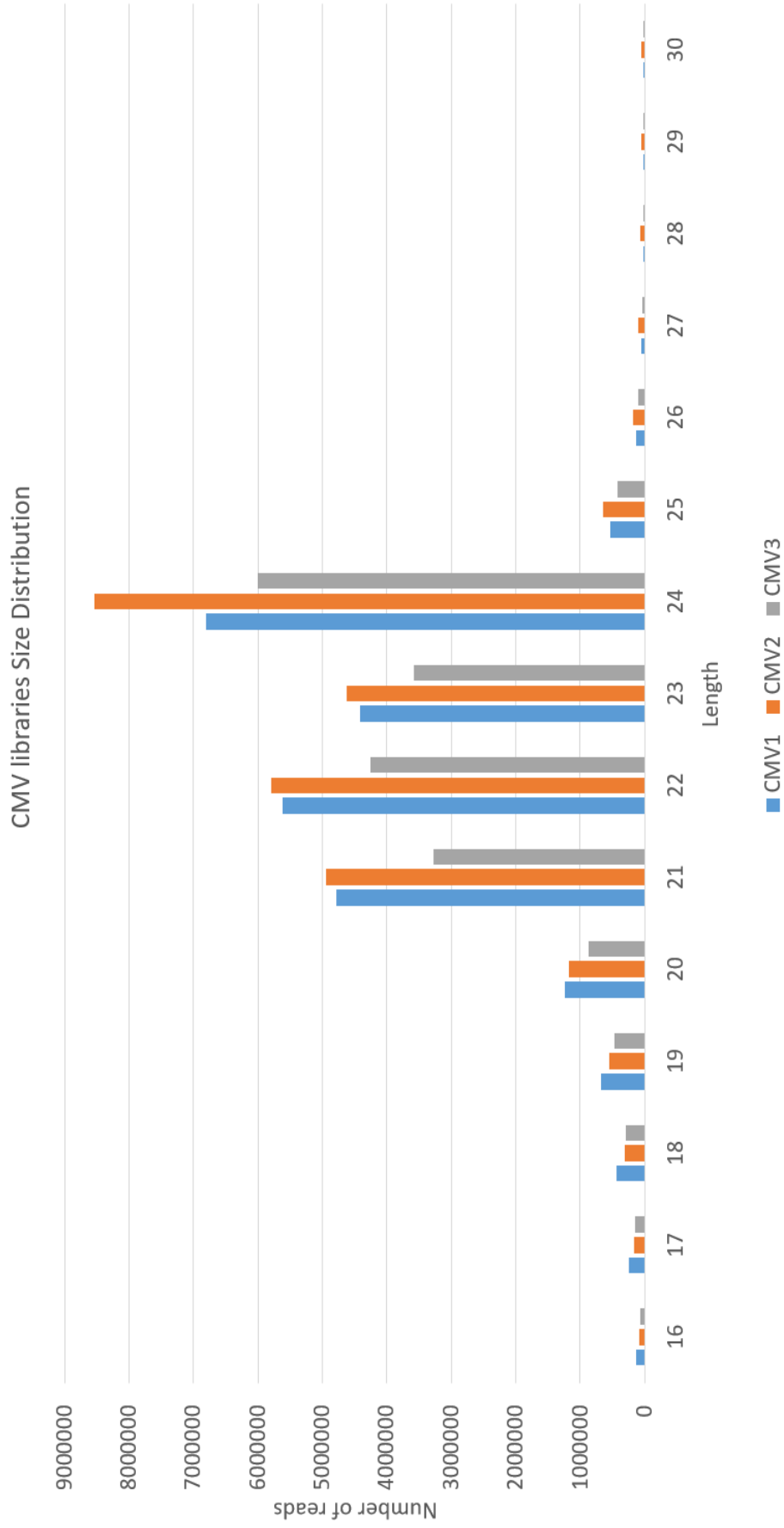
^bftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG3.2_release/ITAG3.2_cDNA.fasta

^chttps://www.ncbi.nlm.nih.gov/nuccore/9632334_9632336_9626472

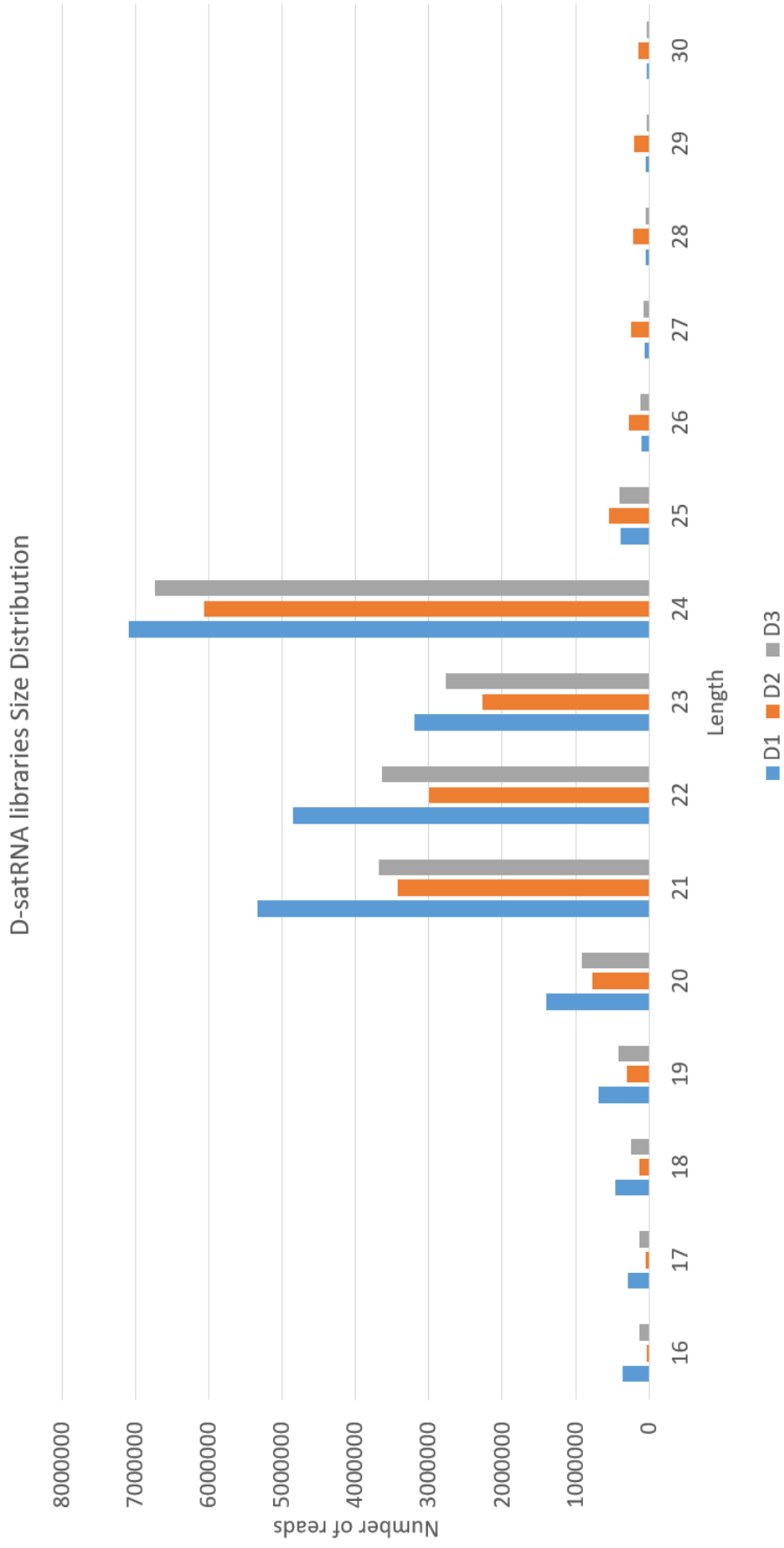
^d<https://www.ncbi.nlm.nih.gov/nuccore/M30584>



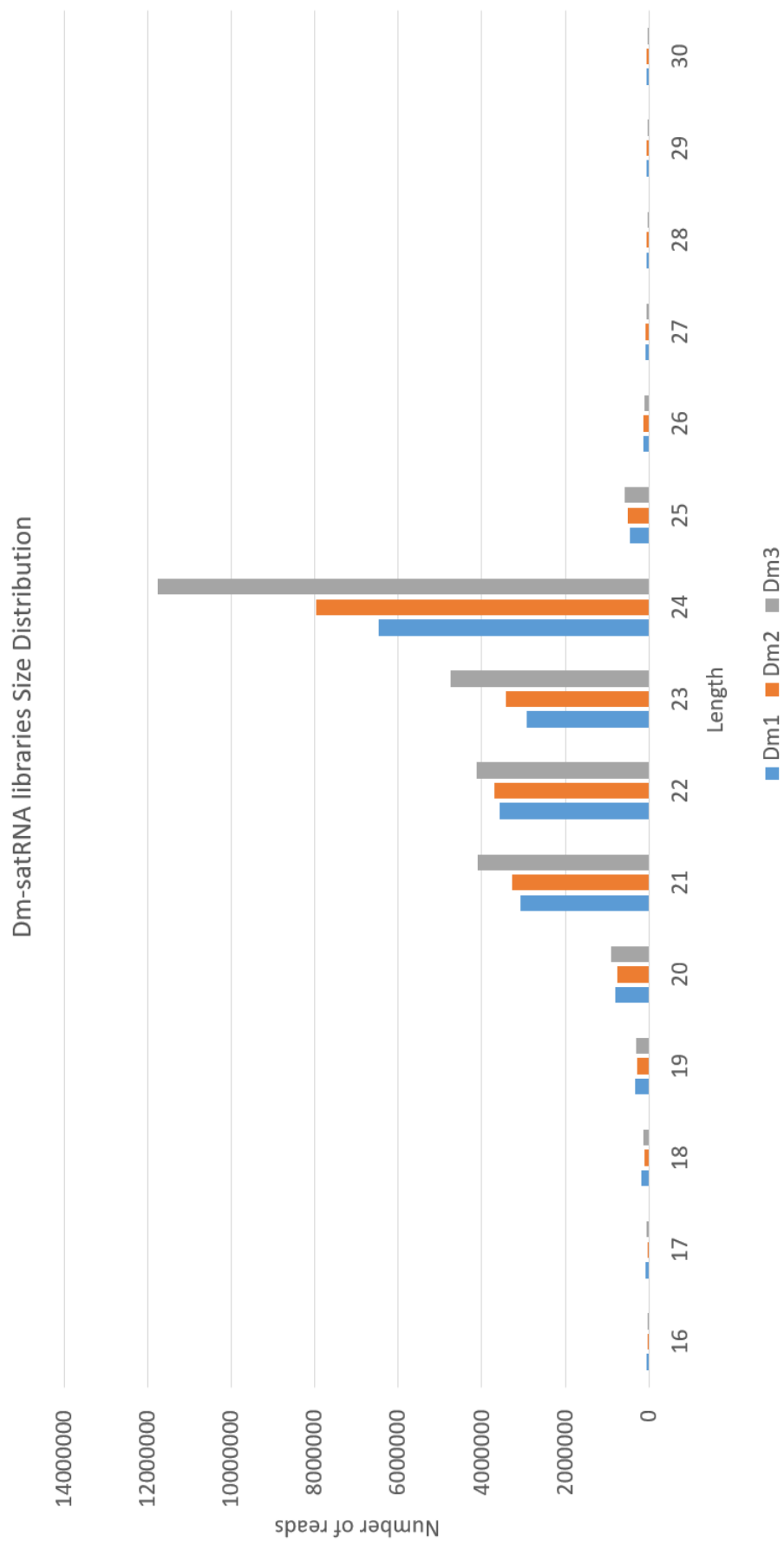
Appendix D Figure 1 Size distribution for the mock sRNA libraries. Majority of reads fall between the expected 21-24nt range.



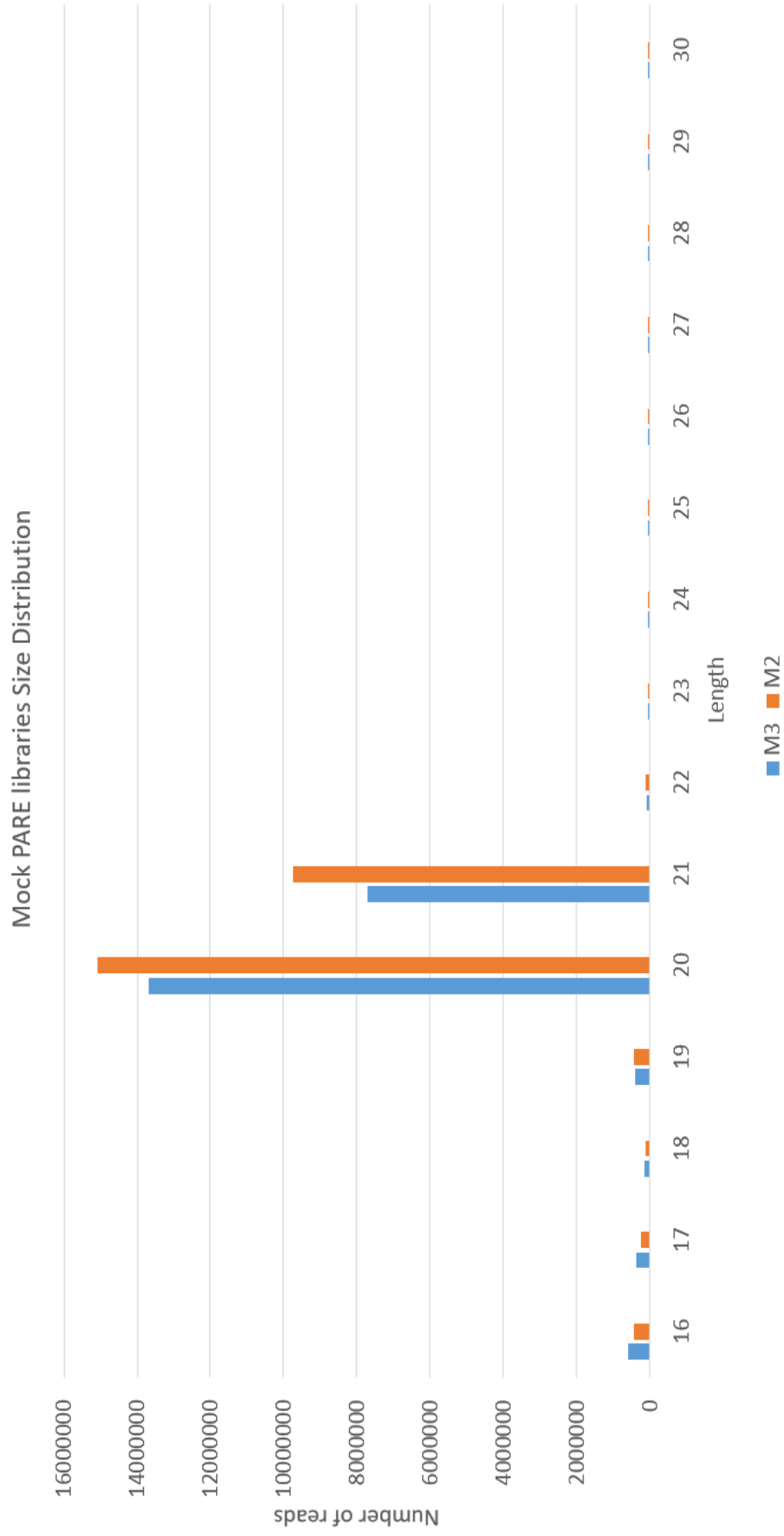
Appendix D Figure 2 Size distribution for the CMV infected sRNA libraries. Majority of reads fall between the expected 21-24nt range.



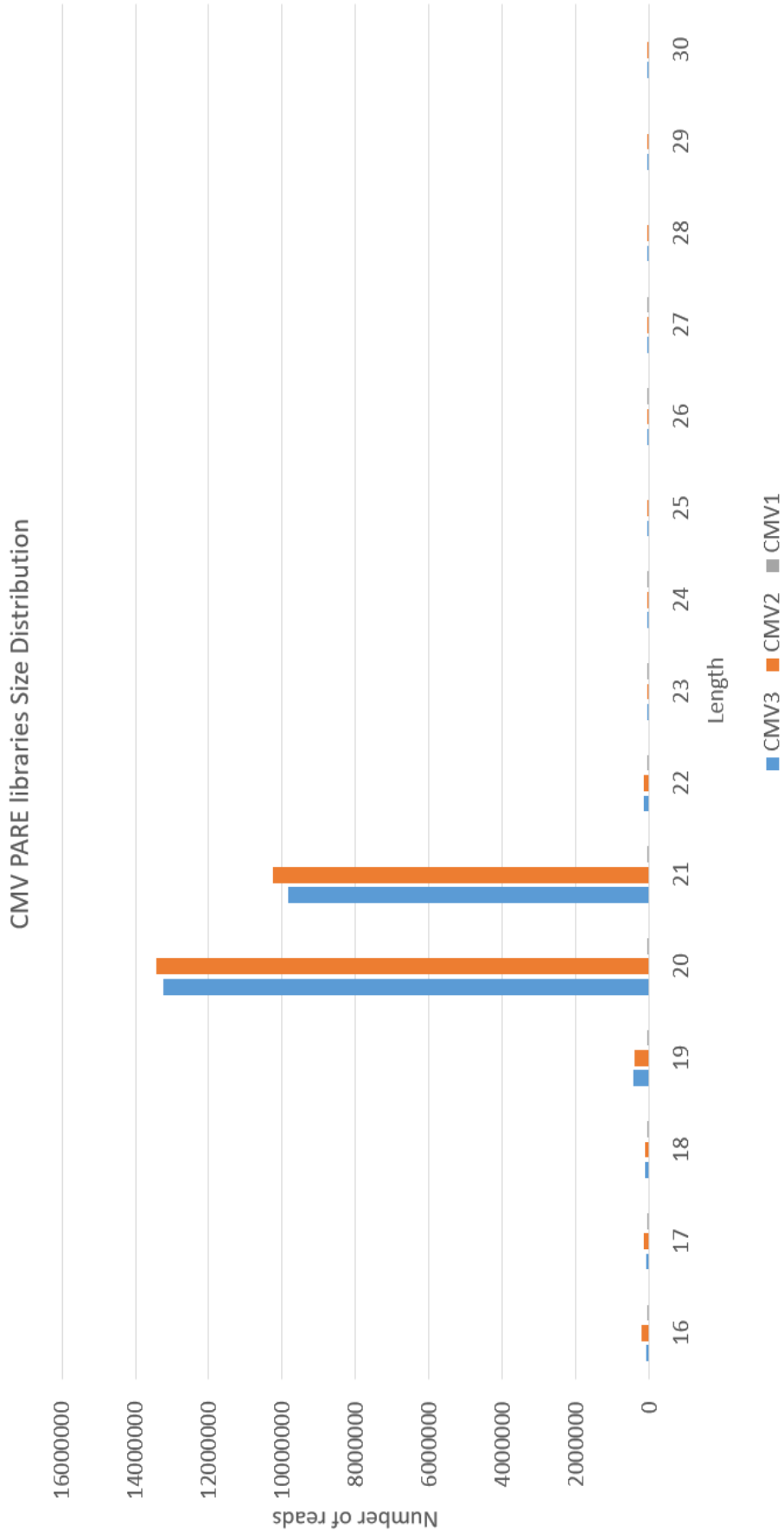
Appendix D Figure 3 Size distribution for the CMV D-satRNA infected sRNA libraries. Majority of reads fall between the expected 21-24nt range.



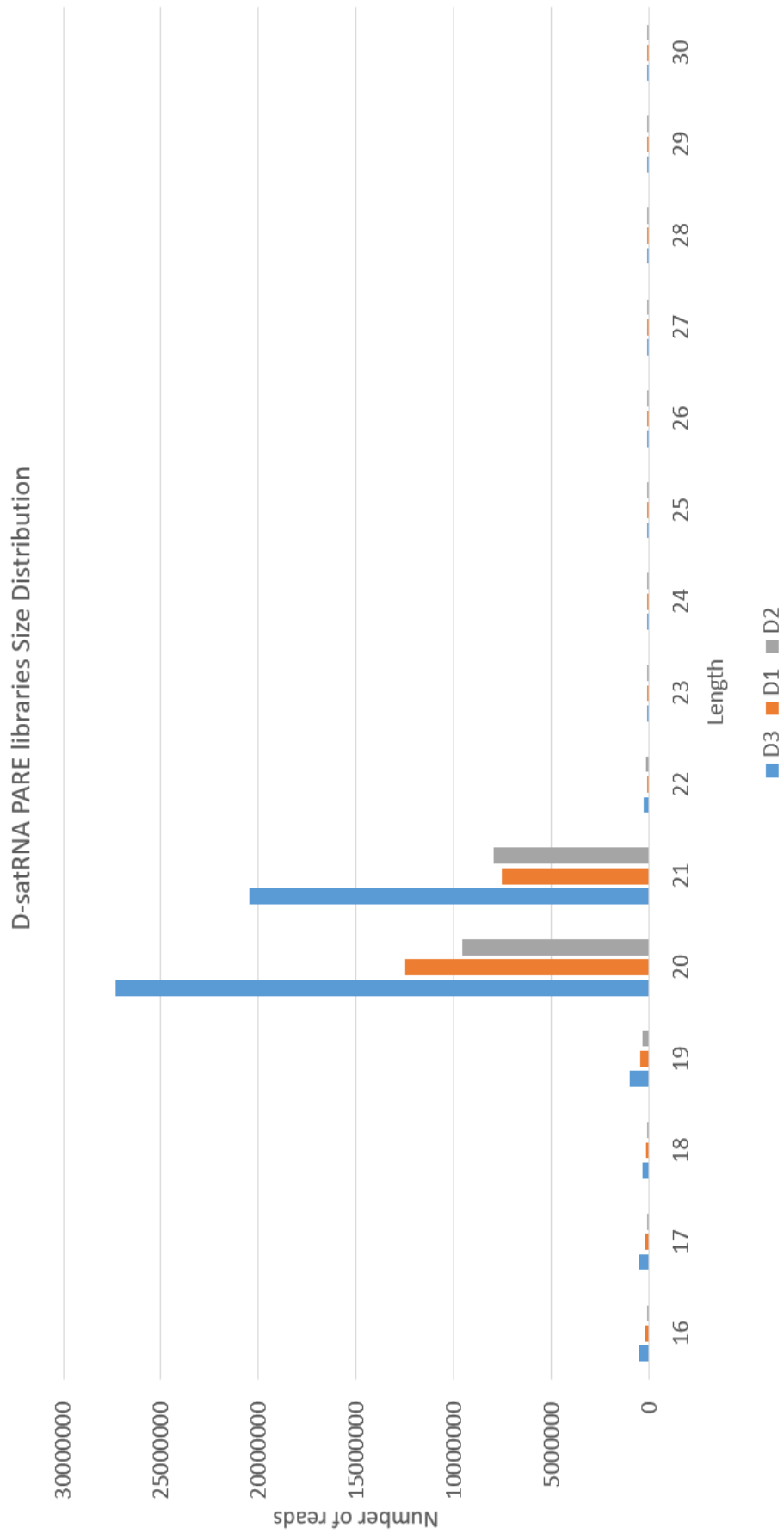
Appendix D Figure 4 Size distribution for the CMV Dm-satRNA infected sRNA libraries. Majority of reads fall between the expected 21-24nt range.



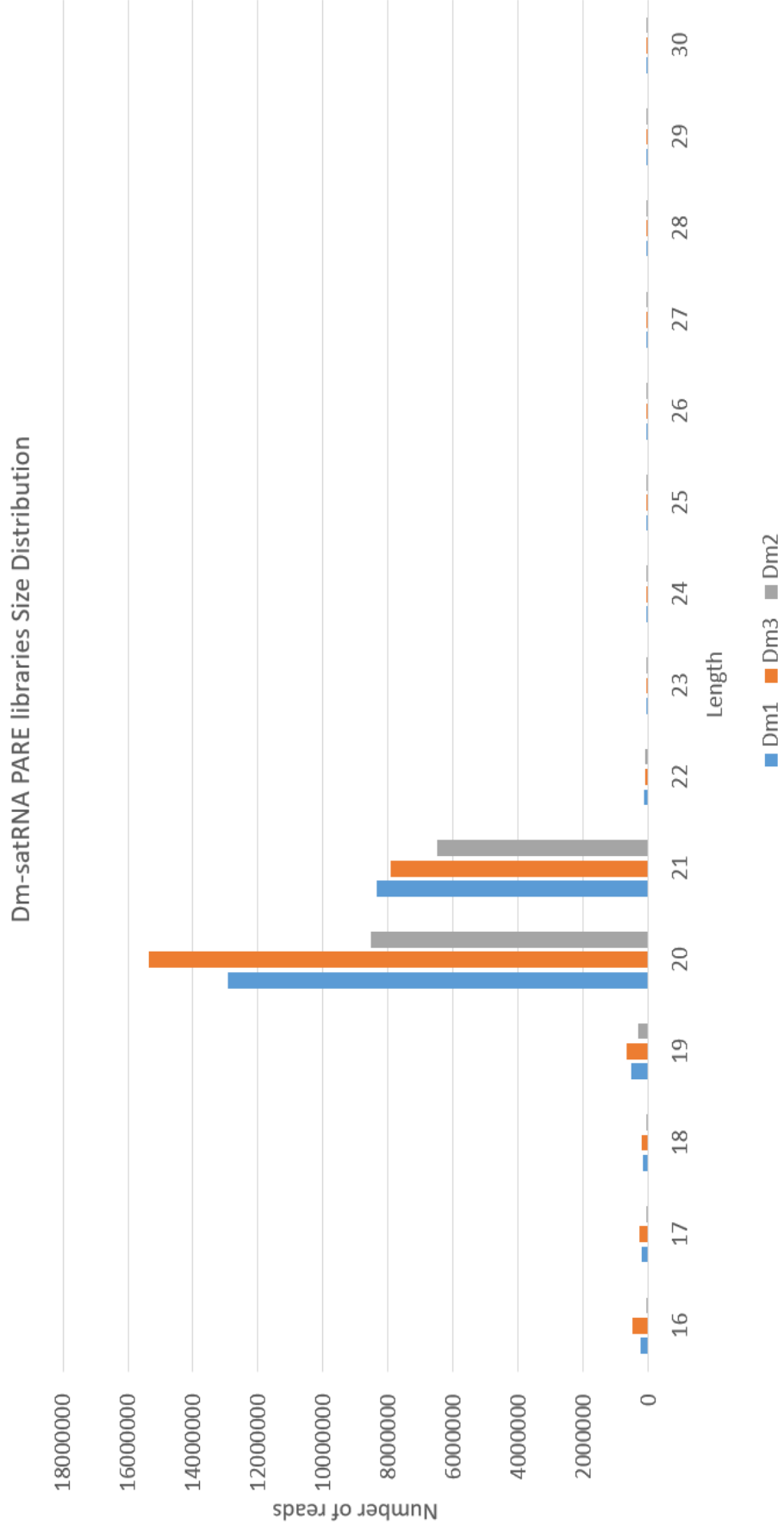
Appendix D Figure 5 Size distribution for the Mock PARE libraries. Majority of reads fall between the expected 19-21nt range.



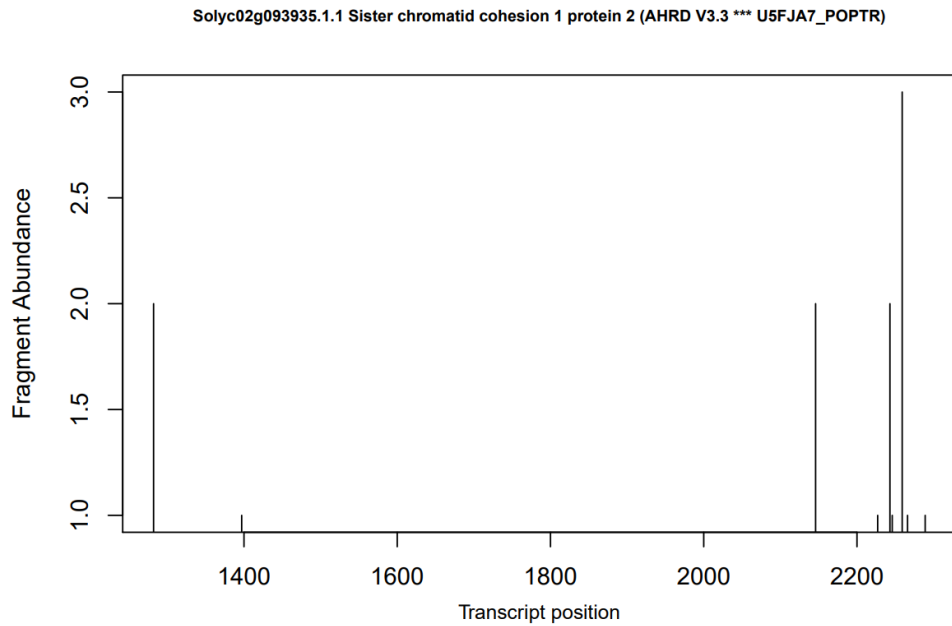
Appendix D Figure 6 Size distribution for the CMV infected PARE libraries. Majority of reads fall between the expected 19-21nt range.



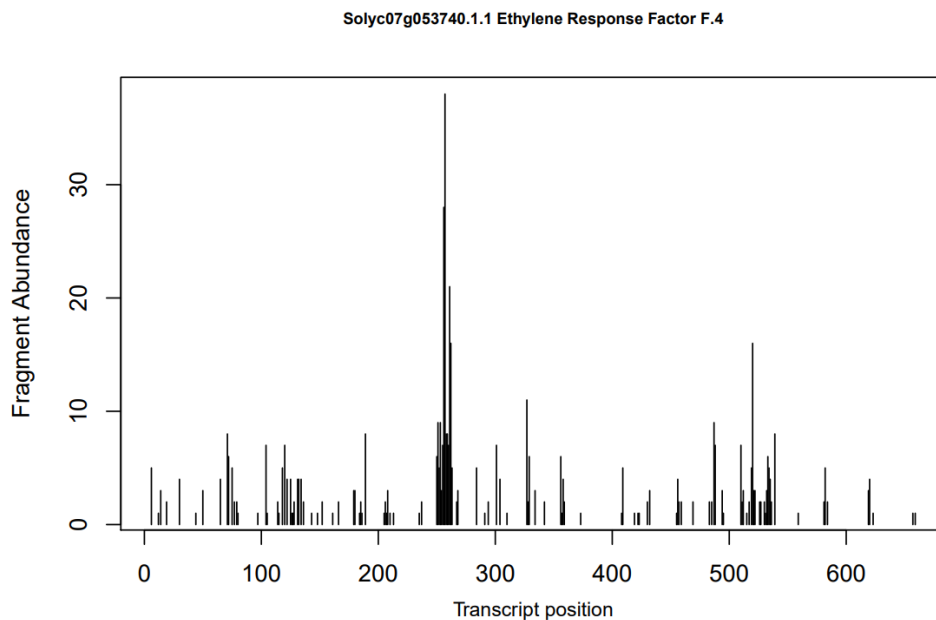
Appendix D Figure 7 Size distribution for the CMV D-satRNA infected PARE libraries. Majority of reads fall between the expected 19-21nt range.



Appendix D Figure 8 Size distribution for the CMV Dm-satRNA infected PARE libraries. Majority of reads fall between the expected 19-21nt range.

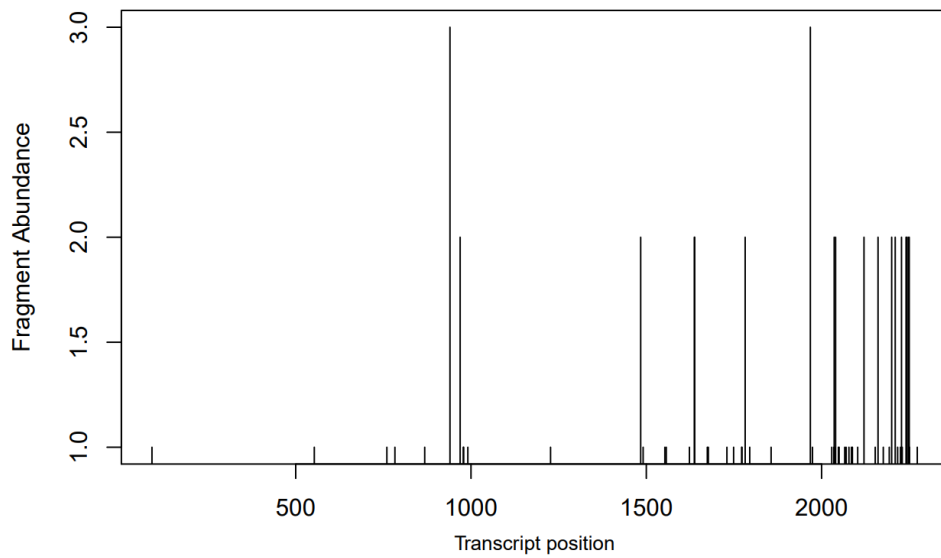


Appendix D Figure 9 The degradation activity for SCC1P2 in the Dm2 degradome dataset. There exists no evidence of mRNA cleavage at the predicted SCC1P2 target site (position 1883).



Appendix D Figure 10 The degradation activity for ERF4 in the Dm2 degradome dataset. There exists some evidence of mRNA degradation around the predicted target site (position 179), but this is considerably lower than in the D-satRNA libraries.

Solyc07g065660.3.1 Cellulose synthase family protein (AHRD V3.3 *** B9GFY5_POPTR)



Appendix D Figure 11 The degradation activity for CSFP in the Dm2 degradome dataset. There exists some evidence of mRNA degradation around the predicted target site (position 940), but this is considerably lower than in the D-satRNA libraries.

References

- [1] Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009a). Cleaveland: a pipeline for using degradome data to find cleaved small rna targets. *Bioinformatics*, 25(1):130–131.
- [2] Addo-Quaye, C., Snyder, J. A., Park, Y. B., Li, Y.-F., Sunkar, R., and Axtell, M. J. (2009b). Sliced microrna targets and precise loop-first processing of mir319 hairpins revealed by analysis of the *Physcomitrella patens* degradome. *Rna*, 15(12):2112–2121.
- [3] Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microrna-directed phasing during trans-acting sirna biogenesis in plants. *Cell*, 121(2):207–221.
- [4] Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microrna annotation. *Rna*, 9(3):277–279.
- [5] An, J., Lai, J., Sajjanhar, A., Lehman, M. L., and Nelson, C. C. (2014). mirplant: an integrated tool for identification of plant mirna from rna sequencing data. *BMC bioinformatics*, 15(1):275.
- [6] Arif, M. A., Fattash, I., Ma, Z., Cho, S. H., Beike, A. K., Reski, R., Axtell, M. J., and Frank, W. (2012). Dicer-like3 activity in *Physcomitrella patens* dicer-like4 mutants causes severe developmental dysfunction and sterility. *Molecular plant*, 5(6):1281–1294.
- [7] Arikiti, S., Zhai, J., and Meyers, B. C. (2013). Biogenesis and function of rice small rnas from non-coding rna precursors. *Current opinion in plant biology*, 16(2):170–179.
- [8] Aukerman, M. J. and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a microrna and its *apetala2*-like target genes. *The Plant Cell*, 15(11):2730–2741.
- [9] Axtell, M. J. (2013a). Classification and comparison of small rnas from plants. *Annual review of plant biology*, 64:137–159.
- [10] Axtell, M. J. (2013b). Shortstack: comprehensive annotation and quantification of small rna genes. *Rna*, 19(6):740–751.
- [11] Axtell, M. J., Jan, C., Rajagopalan, R., and Bartel, D. P. (2006). A two-hit trigger for sirna biogenesis in plants. *Cell*, 127(3):565–577.
- [12] Axtell, M. J. and Meyers, B. C. (2018). Revisiting criteria for plant microrna annotation in the era of big data. *The Plant Cell*, 30(2):272–284.
- [13] Barciszewska-Pacak, M., Milanowska, K., Knop, K., Bielewicz, D., Nuc, P., Plewka, P., Pacak, A. M., Vazquez, F., Karlowski, W., Jarmolowski, A., et al. (2015). Arabidopsis microrna expression regulation in a wide range of abiotic stress responses. *Frontiers in plant science*, 6:410.

- [14] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995.
- [15] Bartel, D. P. (2009). Micromnas: target recognition and regulatory functions. *cell*, 136(2):215–233.
- [16] Baulcombe, D. (2004). RNA silencing in plants. *Nature*, 431(7006):356–363.
- [17] Beckers, M., Mohorianu, I., Stocks, M., Applegate, C., Dalmay, T., and Moulton, V. (2017). Comprehensive processing of high-throughput small rna sequencing data including quality checking, normalization, and differential expression analysis using the uea srna workbench. *RNA*, 23(6):823–835.
- [18] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- [19] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). Genbank. *Nucleic acids research*, 39(Database issue):D32.
- [20] Bernstein, E., Caudy, A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366.
- [21] Billmeier, M. and Xu, P. (2017). Small rna profiling by next-generation sequencing using high-definition adapters. In *MicroRNA Detection and Target Identification*, pages 45–57. Springer.
- [22] Bolser, D., Staines, D. M., Pritchard, E., and Kersey, P. (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In *Plant bioinformatics*, pages 115–140. Springer.
- [23] Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). Tapir, a web server for the prediction of plant microrna targets, including target mimics. *Bioinformatics*, 26(12):1566–1568.
- [24] Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004). Evidence that microrna precursors, unlike other non-coding rnas, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917.
- [25] Borges, F. and Martienssen, R. A. (2015). The expanding world of small rnas in plants. *Nature Reviews Molecular Cell Biology*.
- [26] Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous sirnas derived from a pair of natural cis-antisense transcripts regulate salt tolerance in arabidopsis. *Cell*, 123(7):1279–1291.
- [27] Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., and Görlach, J. (2001). Growth stage-based phenotypic analysis of arabidopsis: a model for high throughput functional genomics in plants. *The Plant Cell*, 13(7):1499–1510.
- [28] Brodersen, P. and Voinnet, O. (2009). Revisiting the principles of microrna target recognition and mode of action. *Nature reviews Molecular cell biology*, 10(2):141–148.
- [29] Brousse, C., Liu, Q., Beauclair, L., Deremetz, A., Axtell, M. J., and Bouché, N. (2014). A non-canonical plant microrna target site. *Nucleic acids research*, 42(8):5270–5279.

- [30] Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm.
- [31] Carmichael, G. G. (2003). Antisense starts making more sense. *Nature biotechnology*, 21(4):371–372.
- [32] Carrington, J. C. and Ambros, V. (2003). Role of micrnas in plant and animal development. *Science*, 301(5631):336–338.
- [33] Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of mirnas and sirnas. *Cell*, 136(4):642–655.
- [34] Chapman, E. J. and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nature reviews. Genetics*, 8(11):884–896.
- [35] Chen, D., Yuan, C., Zhang, J., Zhang, Z., Bai, L., Meng, Y., Chen, L.-L., and Chen, M. (2012). Plantnatsdb: a comprehensive database of plant natural antisense transcripts. *Nucleic acids research*, 40(D1):D1187–D1193.
- [36] Chen, H., Arsovski, A. A., Yu, K., and Wang, A. (2016). Genome-wide investigation using srna-seq, degradome-seq and transcriptome-seq reveals regulatory networks of micrnas and their target genes in soybean during soybean mosaic virus infection. *PLoS One*, 11(3).
- [37] Chen, H.-M., Li, Y.-H., and Wu, S.-H. (2007). Bioinformatic prediction and experimental validation of a microrna-directed tandem trans-acting sirna cascade in arabidopsis. *Proceedings of the National Academy of Sciences*, 104(9):3318–3323.
- [38] Chen, X. (2005). Microrna biogenesis and function in plants. *FEBS letters*, 579(26):5923–5931.
- [39] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O’Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050.
- [40] Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., et al. (2016). mirtarbase 2016: updates to the experimentally validated mirna-target interactions database. *Nucleic acids research*, 44(D1):D239–D247.
- [41] Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771.
- [42] Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290.
- [43] Consortium, I. W. G. S. et al. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*triticum aestivum*) genome. *Science*, 345(6194):1251788.
- [44] Consortium, T. G. et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635.
- [45] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

- [46] da Costa-Nunes, J., Bhatt, A., O'shea, S., West, C., Bray, C., Grossniklaus, U., and Dickinson, H. (2006). Characterization of the three arabidopsis thaliana rad21 cohesins reveals differential responses to ionizing radiation. *Journal of experimental botany*, 57(4):971–983.
- [47] da Costa-Nunes, J. A., Capitão, C., Kozak, J., Costa-Nunes, P., Ducasa, G. M., Pontes, O., and Angelis, K. J. (2014). The atrad21. 1 and atrad21. 3 arabidopsis cohesins play a synergistic role in somatic dna double strand break damage repair. *BMC plant biology*, 14(1):353.
- [48] Dai, X. and Zhao, P. X. (2008). pssnaminer: a plant short small rna regulatory cascade analysis server. *Nucleic acids research*, 36(suppl_2):W114–W118.
- [49] Dai, X. and Zhao, P. X. (2011). psrnatarget: a plant small rna target analysis server. *Nucleic acids research*, 39(suppl 2):W155–W159.
- [50] Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psrnatarget: a plant small rna target analysis server (2017 release). *Nucleic acids research*, 46(W1):W49–W54.
- [51] Darmon, S. K. and Lutz, C. S. (2012). Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. *RNA biology*, 9(10):1255–1265.
- [52] Devic, M., Jaegle, M., and Baulcombe, D. (1989). Symptom production on tobacco and tomato is determined by two distinct domains of the satellite rna of cucumber mosaic virus (strain y). *Journal of general virology*, 70(10):2765–2774.
- [53] Ding, J., Li, D., Ohler, U., Guan, J., and Zhou, S. (2012). Genome-wide search for mirna-target interactions in arabidopsis thaliana with an integrated approach. In *BMC genomics*, volume 13, page S3. BioMed Central.
- [54] Duan, C.-G., Wang, C.-H., and Guo, H.-S. (2012). Application of rna silencing to plant disease resistance. *Silence*, 3(1):5.
- [55] Eddy, S. R. (2001). Non-coding rna genes and the modern rna world. *Nature Reviews Genetics*, 2(12):919–929.
- [56] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.
- [57] Edwardson, J. and Christie, R. (1991). Cucumoviruses. *CRC handbook of viruses infecting legumes*, pages 293–319.
- [58] Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.
- [59] Elbashir, S. M., Lendeckel, W., and Tuschl, T. (2001). Rna interference is mediated by 21-and 22-nucleotide rnas. *Genes & development*, 15(2):188–200.
- [60] Faghihi, M. A. and Wahlestedt, C. (2009). Regulatory roles of natural anti-sense transcripts. *Nature reviews Molecular cell biology*, 10(9):637–643.
- [61] Fahlgren, N. and Carrington, J. C. (2010). mirna target prediction in plants. *Plant MicroRNAs: Methods and Protocols*, pages 51–57.
- [62] Fahlgren, N., Montgomery, T. A., Howell, M. D., Allen, E., Dvorak, S. K., Alexander, A. L., and Carrington, J. C. (2006). Regulation of auxin response factor3 by tas3 ta-sirna affects developmental timing and patterning in arabidopsis. *Current biology*, 16(9):939–944.

- [63] Fei, Q., Xia, R., and Meyers, B. C. (2013). Phased, secondary, small interfering rnas in posttranscriptional regulatory networks. *The Plant Cell*, 25(7):2400–2415.
- [64] Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE.
- [65] Ferragina, P. and Manzini, G. (2001). An experimental study of an opportunistic index. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 269–278. Society for Industrial and Applied Mathematics.
- [66] Fire, A., Albertson, D., Harrison, S. W., and Moerman, D. (1991). Production of antisense rna leads to effective and specific inhibition of gene expression in *c. elegans* muscle. *Development*, 113(2):503–514.
- [67] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded rna in *caenorhabditis elegans*. *nature*, 391(6669):806–811.
- [68] Folkes, L., Moxon, S., Woolfenden, H. C., Stocks, M. B., Szittyá, G., Dalmay, T., and Moulton, V. (2012). PAREsnip: A tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Research*, 40(13).
- [69] Fourounjian, P., Tang, J., Tanyolac, B., Feng, Y., Gelfand, B., Kakrana, A., Tu, M., Wakim, C., Meyers, B. C., Ma, J., et al. (2019). Post-transcriptional adaptation of the aquatic plant *spirodela polyrhiza* under stress and hormonal stimuli. *The Plant Journal*, 98(6):1120–1133.
- [70] Frohman, M. A. et al. (1990). Race: rapid amplification of cDNA ends. *PCR protocols: A guide to methods and applications*, 28.
- [71] Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., et al. (2009). The challenges of sequencing by synthesis. *Nature biotechnology*, 27(11):1013.
- [72] Garcia-Arenal, F. and Palukaitis, P. (1999). Structure and functional relationships of satellite rnas of cucumber mosaic virus. In *Satellites and defective viral RNAs*, pages 37–63. Springer.
- [73] Ge, S. X., Son, E. W., and Yao, R. (2018). idep: an integrated web application for differential expression and pathway analysis of rna-seq data. *BMC bioinformatics*, 19(1):534.
- [74] German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., et al. (2008). Global identification of microRNA–target rna pairs by parallel analysis of rna ends. *Nature biotechnology*, 26(8):941–946.
- [75] Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (*oryza sativa* l. ssp. *japonica*). *Science*, 296(5565):92–100.
- [76] Gregory, B. D., O’Malley, R. C., Lister, R., Urich, M. A., Tonti-Filippini, J., Chen, H., Millar, A. H., and Ecker, J. R. (2008). A link between rna metabolism and silencing affecting *arabidopsis* development. *Developmental cell*, 14(6):854–866.

- [77] Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., and Enright, A. J. (2006). mirbase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl 1):D140–D144.
- [78] Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The vienna rna websuite. *Nucleic acids research*, 36(suppl_2):W70–W74.
- [79] Gupta, O. P., Meena, N. L., Sharma, I., and Sharma, P. (2014). Differential regulation of microRNAs in response to osmotic, salt and cold stresses in wheat. *Molecular Biology Reports*, 41(7):4623–4629.
- [80] Gupta, V., Markmann, K., Pedersen, C. N., Stougaard, J., and Andersen, S. U. (2012). shortran: a pipeline for small rna-seq data analysis. *Bioinformatics*, 28(20):2698–2700.
- [81] Gutierrez, L., Bussell, J. D., Păcurar, D. I., Schwambach, J., Păcurar, M., and Bellini, C. (2009). Phenotypic plasticity of adventitious rooting in arabidopsis is controlled by complex regulation of auxin response factor transcripts and microRNA abundance. *The Plant Cell*, 21(10):3119–3132.
- [82] Gužvić, M. (2013). The history of dna sequencing/istorijat sekvenciranja dnk. *Journal of medical biochemistry*, 32(4):301–312.
- [83] Gyula, P., Baksa, I., Tóth, T., Mohorianu, I., Dalmay, T., and Szittyá, G. (2018). Ambient temperature regulates the expression of a small set of srnas influencing plant development through *nf-ya2* and *yuc2*. *Plant, cell & environment*, 41(10):2404–2417.
- [84] Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. *Nature genetics*, 41(10):1088.
- [85] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.
- [86] Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015). Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics*, 16(1):169.
- [87] Held, M. A., Penning, B., Brandt, A. S., Kessans, S. A., Yong, W., Scofield, S. R., and Carpita, N. C. (2008). Small-interfering rnas from natural antisense transcripts derived from a cellulose synthase gene modulate cell wall biosynthesis in barley. *Proceedings of the National Academy of Sciences*, 105(51):20534–20539.
- [88] Hofacker, I. L. (2003). Vienna rna secondary structure server. *Nucleic acids research*, 31(13):3429–3431.
- [89] Hofacker, I. L. (2009). Rna secondary structure analysis using the vienna rna package. *Current protocols in bioinformatics*, 26(1):12–2.
- [90] Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel dna pyrosequencing. *Genome biology*, 8(7):R143.
- [91] Hutvagner, G. (2005). Small rna asymmetry in *RNAi*: function in RISC assembly and gene regulation. *FEBS letters*, 579(26):5850–5857.

- [92] Initiative, A. G. et al. (2000). Analysis of the genome sequence of the flowering plant *arabidopsis thaliana*. *nature*, 408(6814):796.
- [93] Irian, S., Xu, P., Dai, X., Zhao, P. X., and Roossinck, M. J. (2007). Regulation of a virus-induced lethal disease in tomato revealed by long-range analysis. *Molecular plant-microbe interactions*, 20(12):1477–1488.
- [94] Ivashuta, S., Banks, I. R., Wiggins, B. E., Zhang, Y., Ziegler, T. E., Roberts, J. K., and Heck, G. R. (2011). Regulation of gene expression in plants through miRNA inactivation. *PLoS one*, 6(6).
- [95] Iwakawa, H.-o. and Tomari, Y. (2013). Molecular insights into miRNA-mediated translational repression in plants. *Molecular cell*, 52(4):591–601.
- [96] Jacquemond, M. (2012). Cucumber mosaic virus. In *Advances in virus research*, volume 84, pages 439–504. Elsevier.
- [97] Jin, H., Vacic, V., Girke, T., Lonardi, S., and Zhu, J.-K. (2008). Small RNAs and the regulation of cis-natural antisense transcripts in *arabidopsis*. *BMC molecular biology*, 9(1):6.
- [98] Johnson, C., Kasprzewska, A., Tennessen, K., Fernandes, J., Nan, G.-L., Walbot, V., Sundaresan, V., Vance, V., and Bowman, L. H. (2009). Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome research*, 19(8):1429–1440.
- [99] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S., and Madden, T. L. (2008). Ncbi blast: a better web interface. *Nucleic acids research*, 36(suppl_2):W5–W9.
- [100] Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant miRNAs and their targets, including a stress-induced miRNA. *Molecular cell*, 14(6):787–799.
- [101] Kakrana, A., Hammond, R., Patel, P., Nakano, M., and Meyers, B. C. (2014). sparta: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic acids research*, page gku693.
- [102] Kaper, J. and Waterworth, H. (1977). Cucumber mosaic virus associated RNA 5: causal agent for tomato necrosis. *Science*, 196(4288):429–431.
- [103] Karlova, R., van Haarst, J. C., Maliepaard, C., van de Geest, H., Bovy, A. G., Lammers, M., Angenent, G. C., and de Maagd, R. A. (2013). Identification of miRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis. *Journal of experimental botany*, 64(7):1863–1878.
- [104] Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Zhu, J.-K., Staskawicz, B. J., and Jin, H. (2006). A pathogen-inducible endogenous siRNA in plant immunity. *Proceedings of the National Academy of Sciences*, 103(47):18002–18007.
- [105] Kent, W. J. (2002). Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664.
- [106] Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., et al. (2016). Ensembl genomes 2016: more genomes, more complexity. *Nucleic acids research*, 44(D1):D574–D580.

- [107] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–1284.
- [108] Kincaid, R. P. and Sullivan, C. S. (2012). Virus-encoded microRNAs: an overview and a look to the future. *PLoS Pathog*, 8(12):e1003018.
- [109] Klee, H. J. and Giovannoni, J. J. (2011). Genetics and control of tomato fruit ripening and quality attributes. *Annual review of genetics*, 45:41–59.
- [110] Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). mirbase: from microRNA sequences to function. *Nucleic acids research*, 47(D1):D155–D162.
- [111] Kozomara, A. and Griffiths-Jones, S. (2014). mirbase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73.
- [112] Krüger, J. and Rehmsmeier, M. (2006). Rnahybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research*, 34(suppl 2):W451–W454.
- [113] Kulski, J. K. (2016). Next-generation sequencing—an overview of the history, tools, and “omic” applications. *Next Generation Sequencing—Advances, Applications and Challenges*, pages 3–60.
- [114] Kuwata, S., Masuta, C., and Takanami, Y. (1991). Reciprocal phenotype alterations between two satellite RNAs of cucumber mosaic virus. *Journal of general virology*, 72(10):2385–2389.
- [115] Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210.
- [116] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.
- [117] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.
- [118] Lapidot, M. and Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO reports*, 7(12):1216–1222.
- [119] Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854.
- [120] Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060.
- [121] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2010a). The European Nucleotide Archive. *Nucleic acids research*, 39(suppl_1):D28–D31.
- [122] Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. N. S. D. (2010b). The Sequence Read Archive. *Nucleic acids research*, 39(suppl_1):D19–D21.

- [123] Li, B., Qin, Y., Duan, H., Yin, W., and Xia, X. (2011). Genome-wide characterization of new and drought stress responsive microRNAs in *populus euphratica*. *Journal of experimental botany*, 62(11):3765–3779.
- [124] Li, F., Orban, R., and Baker, B. (2012). Somart: a web server for plant miRNA, ta-siRNA and target gene analysis. *The Plant Journal*, 70(5):891–901.
- [125] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.
- [126] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.
- [127] Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483.
- [128] Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858.
- [129] Li, J., Reichel, M., and Millar, A. A. (2014). Determinants beyond both complementarity and cleavage govern miR159 efficacy in *arabidopsis*. *PLoS genetics*, 10(3).
- [130] Li, L., Xu, J., Yang, D., Tan, X., and Wang, H. (2010a). Computational approaches for miRNA studies: a review. *Mammalian Genome*, 21(1-2):1–12.
- [131] Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714.
- [132] Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009b). Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967.
- [133] Li, S., Liberman, L. M., Mukherjee, N., Benfey, P. N., and Ohler, U. (2013). Integrated detection of natural antisense transcripts using strand-specific rna sequencing data. *Genome research*, 23(10):1730–1739.
- [134] Li, Y.-F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M. J., Zhang, W., and Sunkar, R. (2010b). Transcriptome-wide identification of miRNA targets in rice. *The Plant Journal*, 62(5):742–759.
- [135] Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.
- [136] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed research international*, 2012.
- [137] Liu, Q., Wang, F., and Axtell, M. J. (2014). Analysis of complementarity requirements for plant miRNA targeting using a *nicotiana benthamiana* quantitative transient assay. *The Plant Cell*, 26(2):741–753.
- [138] Liu, S., Li, J.-H., Wu, J., Zhou, K.-R., Zhou, H., Yang, J.-H., and Qu, L.-H. (2015). Starscan: a web server for scanning small rna targets from degradome sequencing data. *Nucleic acids research*, 43(W1):W480–W486.
- [139] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of scarecrow-like mrna targets directed by a class of *arabidopsis* miRNA. *Science*, 297(5589):2053–2056.

- [140] Lopez-Gomollon, S., Mohorianu, I., Szittyá, G., Moulton, V., and Dalmay, T. (2012). Diverse correlation patterns between miRNAs and their targets during tomato fruit development indicates different modes of miRNA actions. *Planta*, 236(6):1875–1887.
- [141] Lu, C., Jeong, D.-H., Kulkarni, K., Pillay, M., Nobuta, K., German, R., Thatcher, S. R., Maher, C., Zhang, L., Ware, D., et al. (2008). Genome-wide analysis for discovery of rice miRNAs reveals natural antisense miRNAs (nat-miRNAs). *Proceedings of the National Academy of Sciences*, 105(12):4951–4956.
- [142] Ma, G., Chen, P., Buss, G., and Tolin, S. (2003). Genetic study of a lethal necrosis to soybean mosaic virus in pi 507389 soybean. *Journal of Heredity*, 94(3):205–211.
- [143] Mallory, A. C. and Vaucheret, H. (2006). Functions of miRNAs and related small RNAs in plants. *Nature genetics*, 38(6):S31–S36.
- [144] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.
- [145] Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.
- [146] Markowitz, F. (2017). All biology is computational biology. *PLoS biology*, 15(3):e2002050.
- [147] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564.
- [148] May, P., Liao, W., Wu, Y., Shuai, B., McCombie, W. R., Zhang, M. Q., and Liu, Q. A. (2013). The effects of carbon dioxide and temperature on miRNA expression in Arabidopsis development. *Nature communications*, 4(1):1–11.
- [149] Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431(7006):343–349.
- [150] Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., Cao, X., Carrington, J. C., Chen, X., Green, P. J., et al. (2008). Criteria for annotation of plant miRNAs. *The Plant Cell*, 20(12):3186–3190.
- [151] Mohorianu, I., Lopez-Gomollon, S., Schwach, F., Dalmay, T., and Moulton, V. (2012). Firepat—finding regulatory patterns between sRNAs and genes. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3):273–284.
- [152] Mohorianu, I., Stocks, M. B., Wood, J., Dalmay, T., and Moulton, V. (2013). Colide: a bioinformatics tool for co-expression based small RNA loci identification using high-throughput sequencing data. *RNA biology*, 10(7):1221–1230.
- [153] Morgado, L. and Johannes, F. (2019). Computational tools for plant small RNA detection and categorization. *Briefings in bioinformatics*, 20(4):1181–1192.
- [154] Moury, B. (2004). Differential selection of genes of cucumber mosaic virus subgroups. *Molecular Biology and Evolution*, 21(8):1602–1611.
- [155] Moxon, S., Schwach, F., Dalmay, T., MacLean, D., Studholme, D. J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(19):2252–2253.

- [156] Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans.
- [157] Oh, M., Rhee, S., Moon, J. H., Chae, H., Lee, S., Kang, J., and Kim, S. (2017). Literature-based condition-specific mirna-mrna target prediction. *PLoS one*, 12(3).
- [158] Paicu, C., Mohorianu, I., Stocks, M., Xu, P., Coince, A., Billmeier, M., Dalmay, T., Moulton, V., and Moxon, S. (2017). mircat2: accurate prediction of plant and animal micrnas from next-generation sequencing datasets. *Bioinformatics*, 33(16):2446–2454.
- [159] Pan, C., Ye, L., Zheng, Y., Wang, Y., Yang, D., Liu, X., Chen, L., Zhang, Y., Fei, Z., and Lu, G. (2017). Identification and expression profiling of micrnas involved in the stigma exertion under high-temperature stress in tomato. *BMC genomics*, 18(1):843.
- [160] Pantaleo, V., Szittyá, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., and Burgyan, J. (2010). Identification of grapevine micrnas and their targets using high-throughput sequencing and degradome analysis. *The Plant Journal*, 62(6):960–976.
- [161] Pasquinelli, A. E. (2012). Micrnas and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282.
- [162] Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. *Genomics*, 11(3):635–650.
- [163] Pearson, W. R. (2016). Finding protein and nucleotide similarities with fasta. *Current protocols in bioinformatics*, 53(1):3–9.
- [164] Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H. L., and Poethig, R. S. (2004). Sgs3 and sgs2/sde1/rdr6 are required for juvenile development and the production of trans-acting sirnas in arabidopsis. *Genes & development*, 18(19):2368–2379.
- [165] Prüfer, K., Stenzel, U., Dannemann, M., Green, R. E., Lachmann, M., and Kelso, J. (2008). Patman: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531.
- [166] Quinet, M., Angosto, T., Yuste-Lisbona, F. J., Blanchard-Gros, R., Bigot, S., Martinez, J.-P., and Lutts, S. (2019). Tomato fruit development and metabolism. *Frontiers in Plant Science*, 10.
- [167] Quintero, A., Pérez-Quintero, A. L., and López, C. (2013). 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.
- [168] Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of micrnas in arabidopsis thaliana. *Genes & development*, 20(24):3407–3425.
- [169] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597.
- [170] Reuter, J. S. and Mathews, D. H. (2010). Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129.

- [171] Riester, L., Köster-Hofmann, S., Doll, J., Berendzen, K. W., and Zentgraf, U. (2019). Impact of alternatively polyadenylated isoforms of ethylene response factor4 with activator and repressor function on senescence in arabidopsis thaliana l. *Genes*, 10(2):91.
- [172] Ron, M., Saez, M. A., Williams, L. E., Fletcher, J. C., and McCormick, S. (2010). Proper regulation of a sperm-specific cis-nat-sirna is essential for double fertilization in arabidopsis. *Genes & Development*, 24(10):1010–1021.
- [173] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352.
- [174] Ruiz-Ferrer, V. and Voinnet, O. (2009). Roles of plant small rnas in biotic stress responses. *Annual review of plant biology*, 60:485–510.
- [175] Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441IN19447–446IN20448.
- [176] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.
- [177] Scheets, K. (1998). Maize chlorotic mottle machlomovirus and wheat streak mosaic rymovirus concentrations increase in the synergistic disease corn lethal necrosis. *Virology*, 242(1):28–38.
- [178] Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *nature*, 463(7278):178–183.
- [179] Schon, M. A., Kellner, M. J., Plotnikova, A., Hofmann, F., and Nodine, M. D. (2018). Nanopare: parallel analysis of rna 5' ends from low-input rna. *Genome research*, 28(12):1931–1942.
- [180] Schroeder, J. I., Delhaize, E., Frommer, W. B., Guerinot, M. L., Harrison, M. J., Herrera-Estrella, L., Horie, T., Kochian, L. V., Munns, R., Nishizawa, N. K., et al. (2013). Using membrane transporters to improve crops for sustainable food production. *Nature*, 497(7447):60–66.
- [181] Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135.
- [182] Shimura, H., Pantaleo, V., Ishihara, T., Myojo, N., Inaba, J.-i., Sueda, K., Burgyán, J., and Masuta, C. (2011). A viral satellite rna induces yellow symptoms on tobacco by targeting a gene involved in chlorophyll biosynthesis using the rna silencing machinery. *PLoS pathogens*, 7(5).
- [183] Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., and Bartel, D. P. (2010). Expanding the microrna targeting code: functional sites with centered pairing. *Molecular cell*, 38(6):789–802.
- [184] Sleat, D., Zhang, L., and Palukaitis, P. (1994). Mapping determinants within cucumber mosaic virus and its satellite rna for the induction of necrosis in tomato plants. *Molecular plant-microbe interactions: MPMI*, 7(2):189–195.
- [185] Sleat, D. E. and Palukaitis, P. (1990). Site-directed mutagenesis of a plant viral satellite rna changes its phenotype from ameliorative to necrogenic. *Proceedings of the National Academy of Sciences*, 87(8):2946–2950.

- [186] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.
- [187] Song, Q.-X., Liu, Y.-F., Hu, X.-Y., Zhang, W.-K., Ma, B., Chen, S.-Y., and Zhang, J.-S. (2011). Identification of mirnas and their target genes in developing soybean seeds by deep sequencing. *BMC plant biology*, 11(1):5.
- [188] Song, X., Li, P., Zhai, J., Zhou, M., Ma, L., Liu, B., Jeong, D.-H., Nakano, M., Cao, S., Liu, C., et al. (2012). Roles of dcl4 and dcl3b in rice phased small rna biogenesis. *The Plant Journal*, 69(3):462–474.
- [189] Song, X., Xu, L., Yu, J., Tian, P., Hu, X., Wang, Q., and Pan, Y. (2019). Genome-wide characterization of the cellulose synthase gene superfamily in *solanum lycopersicum*. *Gene*, 688:71–83.
- [190] Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012). Reducing ligation bias of small rnas in libraries for next generation sequencing. *Silence*, 3(1):4.
- [191] Srivastava, P. K., Moturu, T. R., Pandey, P., Baldwin, I. T., and Pandey, S. P. (2014). A comparison of performance of plant mirna target prediction tools and the characterization of features for genome-wide target prediction. *BMC genomics*, 15(1):348.
- [192] Stocks, M. B., Mohorianu, I., Beckers, M., Paicu, C., Moxon, S., Thody, J., Dalmay, T., and Moulton, V. (2018). The uea srna workbench (version 4.4): a comprehensive suite of tools for analyzing mirnas and srnas. *Bioinformatics*, 34(19):3382–3384.
- [193] Strange, R. N. and Scott, P. R. (2005). Plant disease: a threat to global food security. *Annual review of phytopathology*, 43.
- [194] Sunkar, R., Chinnusamy, V., Zhu, J., and Zhu, J.-K. (2007). Small rnas as big players in plant abiotic stress responses and nutrient deprivation. *Trends in plant science*, 12(7):301–309.
- [195] Tafer, H. and Hofacker, I. L. (2008). Rnaplex: a fast tool for rna–rna interaction search. *Bioinformatics*, 24(22):2657–2663.
- [196] Tang, Z., Zhang, L., Xu, C., Yuan, S., Zhang, F., Zheng, Y., and Zhao, C. (2012). 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.
- [197] Thatcher, S. R., Burd, S., Wright, C., Lers, A., and Green, P. J. (2015). Differential expression of mirnas and their target genes in senescing leaves and siliques: insights from deep sequencing of small rnas and cleaved target rnas. *Plant, cell & environment*, 38(1):188–200.
- [198] Thody, J., Folkes, L., Medina-Calzada, Z., Xu, P., Dalmay, T., and Moulton, V. (2018). Paresnip2: a tool for high-throughput prediction of small rna targets from degradome sequencing data using configurable targeting rules. *Nucleic acids research*, 46(17):8730–8739.
- [199] Thody, J., Folkes, L., and Moulton, V. (2020a). NATpare: a pipeline for high-throughput prediction and functional analysis of nat-siRNAs. *Nucleic Acids Research*. gkaa448.

- [200] Thody, J., Moulton, V., and Mohorianu, I. (2020b). Pareameters: a tool for computational inference of plant miRNA-mRNA targeting rules using small RNA and degradome sequencing data. *Nucleic Acids Research*.
- [201] Turner, D. H., Sugimoto, N., and Freier, S. M. (1988). RNA structure prediction. *Annual review of biophysics and biophysical chemistry*, 17(1):167–192.
- [202] Umu, S. U. and Gardner, P. P. (2017). A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life. *Bioinformatics*, 33(7):988–996.
- [203] Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.
- [204] Vanh ee-Brossollet, C. and Vaquero, C. (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene*, 211(1):1–9.
- [205] Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes & development*, 20(7):759–771.
- [206] Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gascioli, V., Mallory, A. C., Hilbert, J.-L., Bartel, D. P., and Cr et e, P. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Molecular cell*, 16(1):69–79.
- [207] Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell*, 136(4):669–687.
- [208] Vuittenez, A. and Putz, C. (1972). Catastrophe pour les producteurs alsaciens de tomates. *Alsace*, August, 20.
- [209] Watson, J. D. and Crick, F. H. (1953a). The structure of DNA. In *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press.
- [210] Watson, J. D. and Crick, F. H. C. (1953b). Molecular structure of nucleic acids.
- [211] Williams, L., Carles, C. C., Osmont, K. S., and Fletcher, J. C. (2005). A database analysis method identifies an endogenous trans-acting short-interfering RNA that targets the Arabidopsis *arf2*, *arf3*, and *arf4* genes. *Proceedings of the National Academy of Sciences*, 102(27):9703–9708.
- [212] Won, J.-I., Lee, J., Lee, H., Shin, J., Yoon, J., and Jeong, D.-H. (2019). Webpord: a web-based pipeline of RNA degradome. *International Journal of Data Mining and Bioinformatics*, 22(3):216–230.
- [213] Wu, G., Park, M. Y., Conway, S. R., Wang, J.-W., Weigel, D., and Pothig, R. S. (2009a). The sequential action of *mir156* and *mir172* regulates developmental timing in Arabidopsis. *Cell*, 138(4):750–759.
- [214] Wu, L., Zhang, Q., Zhou, H., Ni, F., Wu, X., and Qi, Y. (2009b). Rice microRNA effector complexes and targets. *The Plant Cell*, 21(11):3421–3435.
- [215] Xie, F., Xiao, P., Chen, D., Xu, L., and Zhang, B. (2012). *mirdeepfinder*: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant molecular biology*, 80(1):75–84.
- [216] Xie, F. and Zhang, B. (2010). Target-align: a tool for plant microRNA target identification. *Bioinformatics*, 26(23):3002–3003.

- [217] Xu, C., Sun, X., Taylor, A., Jiao, C., Xu, Y., Cai, X., Wang, X., Ge, C., Pan, G., Wang, Q., et al. (2017a). Diversity, distribution, and evolution of tomato viruses in china uncovered by small rna sequencing. *Journal of virology*, 91(11):e00173–17.
- [218] Xu, J., Wang, Q., Freeling, M., Zhang, X., Xu, Y., Mao, Y., Tang, X., Wu, F., Lan, H., Cao, M., et al. (2017b). Natural antisense transcripts are significantly involved in regulation of drought stress in maize. *Nucleic acids research*, 45(9):5126–5141.
- [219] Xu, P., Billmeier, M., Mohorianu, I.-I., Green, D., Fraser, W., and Dalmay, T. (2015). An improved protocol for small rna library construction using high definition adapters. *Methods in next generation sequencing*, 2(1).
- [220] Xu, P., Blancaflor, E. B., and Roossinck, M. J. (2003). In spite of induced multiple defense responses, tomato plants infected with cucumber mosaic virus and d satellite rna succumb to systemic necrosis. *Molecular plant-microbe interactions*, 16(6):467–476.
- [221] Xu, P. and Roossinck, M. J. (2000). Cucumber mosaic virus d satellite rna-induced programmed cell death in tomato. *The Plant Cell*, 12(7):1079–1092.
- [222] Yang, X. and Li, L. (2011). mirdeep-p: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, 27(18):2614–2615.
- [223] Yifhar, T., Pekker, I., Peled, D., Friedlander, G., Pistunov, A., Sabban, M., Wachsman, G., Alvarez, J. P., Amsellem, Z., and Eshed, Y. (2012). Failure of the tomato trans-acting short interfering rna program to regulate auxin response factor3 and arf4 underlies the wiry leaf syndrome. *The Plant Cell*, 24(9):3575–3589.
- [224] Yu, D., Meng, Y., Zuo, Z., Xue, J., and Wang, H. (2016). Natpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (nats) and phase-distributed nat-sirnas from de novo assembled transcriptomes. *Scientific reports*, 6(1):1–6.
- [225] Yu, D., Xu, M., Ito, H., Shao, W., Ma, X., Wang, H., and Meng, Y. (2018). Tracking microRNA processing signals by degradome sequencing data analysis. *Frontiers in genetics*, 9:546.
- [226] Yu, H. and Kumar, P. P. (2003). Post-transcriptional gene silencing in plants by RNA.
- [227] Yu, X., Yang, J., Li, X., Liu, X., Sun, C., Wu, F., and He, Y. (2013). Global analysis of cis-natural antisense transcripts and their heat-responsive nat-sirnas in brassica rapa. *BMC plant biology*, 13(1):208.
- [228] Yuan, C., Wang, J., Harrison, A. P., Meng, X., Chen, D., and Chen, M. (2015). Genome-wide view of natural antisense transcripts in arabidopsis thaliana. *DNA research*, 22(3):233–243.
- [229] Yue, D., Liu, H., and Huang, Y. (2009). Survey of computational algorithms for microRNA target prediction. *Current genomics*, 10(7):478–492.
- [230] Zhai, J., Arikait, S., Simon, S. A., Kingham, B. F., and Meyers, B. C. (2014). Rapid construction of parallel analysis of rna end (pare) libraries for illumina sequencing. *Methods*, 67(1):84–90.
- [231] Zhang, C., Li, G., Wang, J., and Fang, J. (2012a). Identification of trans-acting sirnas and their regulatory cascades in grapevine. *Bioinformatics*, 28(20):2561–2568.

- [232] Zhang, X., Lii, Y., Wu, Z., Polishko, A., Zhang, H., Chinnusamy, V., Lonardi, S., Zhu, J.-K., Liu, R., and Jin, H. (2013). Mechanisms of small rna generation from cis-nats in response to environmental and developmental cues. *Molecular plant*, 6(3):704–715.
- [233] Zhang, X., Xia, J., Lii, Y. E., Barrera-Figueroa, B. E., Zhou, X., Gao, S., Lu, L., Niu, D., Chen, Z., Leung, C., et al. (2012b). Genome-wide analysis of plant nat-sirnas reveals insights into their distribution, biogenesis and function. *Genome biology*, 13(3):R20.
- [234] Zhang, Y. (2005). miru: an automated plant mirna target prediction server. *Nucleic acids research*, 33(suppl 2):W701–W704.
- [235] Zheng, H., Qiyan, J., Zhiyong, N., and Hui, Z. (2013). Prediction and identification of natural antisense transcripts and their small rnas in soybean (glycine max). *BMC genomics*, 14(1):280.
- [236] Zheng, Y., Li, Y.-F., Sunkar, R., and Zhang, W. (2012). Seqtar: an effective method for identifying microrna guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic acids research*, 40(4):e28–e28.
- [237] Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D., Roberts, M., Wegrzyn, J. L., de Jong, P. J., et al. (2014). Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, 196(3):875–890.
- [238] Zubko, E. and Meyer, P. (2007). A natural antisense transcript of the petunia hybrida sho gene suggests a role for an antisense mechanism in cytokinin regulation. *The Plant Journal*, 52(6):1131–1139.
- [239] Zuker, M. (2000). Calculating nucleic acid secondary structure. *Current opinion in structural biology*, 10(3):303–310.
- [240] Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148.
- [241] Zuo, J., Wang, Q., Han, C., Ju, Z., Cao, D., Zhu, B., Luo, Y., and Gao, L. (2017). Srnaome and degradome sequencing analysis reveals specific regulation of srna in response to chilling injury in tomato fruit. *Physiologia plantarum*, 160(2):142–154.