Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

MicroRNA and mRNA analysis of two species of New Zealand *Pachycladon*

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in

Genetics

at Massey University, Manawatu, New Zealand.

Louise Michelle Carr

2014

Abstract

MicroRNAs (miRNAs) are small, non-coding RNAs important in post-transcriptional regulation. In this study, potential miRNAs from two New Zealand *Pachycladon* species, *P. cheesemanii* and *P. fastigiatum*, are identified and compared.

Sixteen miRNAs were differentially expressed between the species, most of which have roles in flower and leaf development. Potential targets for 15 miRNAs were located in expressed sequence tag (EST) libraries for *P. cheesemanii* and/or *P. fastigiatum*, including a new potential relationship in *P. cheesemanii* between miR825 and MYB29 (AT5G07690), a transcription factor involved in the synthesis of methionine-derived glucosinolates.

From the results of the differential expression analysis and target identification, 27 miRNAs from 21 miRNA families were chosen for pre-miRNA sequencing. Sequences of 15 *P. cheesemanii* miRNA hairpins and 13 *P. fastigiatum* miRNA hairpins were validated experimentally.

Additionally, mRNA-Seq data obtained at the same time as the miRNAs were analysed. A gene ontology analysis indicated enriched terms for defence responses and miRNAs in *P. fastigiatum*.

This study is the first investigation of the miRNAs present in *Pachycladon* and how their differential expression contributes to the adaptive divergence between the species.

Acknowledgements

Firstly, I thank my supervisor Claudia Voelckel for all your support and guidance over the past two years. Thanks to Peter Lockhart for being a wonderful officemate, and thanks to Trish McLenachan for assistance with lab work and proofreading. I could not have asked for a more caring group of people to work with, and I truly appreciate all of the help you have given me over the time we have shared.

Thanks to Lesley Collins for microRNA wisdom and helping to shape this project, and thanks to Nicole Gruenheit for helping with the bioinformatics. Thanks to everyone who assisted with sequencing and extraction of small RNA and mRNA data.

Thank you to my friends, and especially my flatmates, for all your companionship. And lastly, huge thanks to my parents, brother, and grandparents for your unerring love and support.

Table of Contents

List of Figures

List of Tables

Supplementary Tables and Files

Abbreviations

Aly	*A. lyrata*
Ath	*A. thaliana*
BLAST	basic local alignment search tool
bp	base pair
EST	Expressed Sequence Tag
GO	Gene Ontology
kb	kilobase
logFC	log fold change
miRNA	microRNA
mRNA	messenger RNA
Mya	million years ago
nt	nucleotide
PC, Pch	*P. cheesemanii*
PCR	polymerase chain reaction
PF, Pfa	*P. fastigiatum*
RNA	ribonucleic acid
rRNA	ribosomal RNA
tRNA	transfer RNA

Nucleotide Ambiguity Code

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N

Chapter One: Literature Review

1.1 Introduction

MicroRNAs (miRNAs) are short (21-22 nucleotides) non-coding RNA that have critical roles in gene regulation in plants and animals. MiRNAs are involved in development and adaptation, and act as post-transcriptional super-regulators by repressing translation or directing cleavage of the mRNA. Targets of the miRNAs have a region that is partially complementary to the miRNA. In plants, there is typically a very high complementarity between the miRNA and the binding site of the target, while in animals, there is less complementarity (1).

Along with identifying miRNAs and the processes they are involved in, miRNAs are also interesting from an evolutionary perspective - how do miRNAs change over the course of evolution, and how does this affect the phenotype and adaptation of the species?

1.2 MiRNA Biogenesis and Function

1.2.1 Biogenesis of miRNAs in plants

1.2.1.2 Processing in the nucleus

Transcription of miRNA genes by RNA Polymerase II generates primary miRNA transcripts, the pri-miRNAs that form secondary structures that include a hairpin loop (2). These pri-miRNAs have 5’ caps and poly-A tails (3). Dicer-like endonucleases in plants (predominately DCL1) (4) cleave the pri-miRNAs to produce the pre-mRNA (which consists only of the hairpin) and then cleave the hairpin to produce the mature miRNA duplex. This duplex consists of the guide miRNA and the passenger miRNA (denoted with *). The duplex is not perfectly complementary and possesses two-nucleotide 3' overhangs.

1.2.2.2 Processing in the cytosol

The duplex is exported to the cytosol where the guide miRNA is then loaded into the ARGONAUTE (AGO) protein of the RNA induced silencing complex (RISC) and the miRNA* is degraded. The guide miRNA is usually the miRNA strand with the less stable

5' base pair in the duplex (5), but the other strand is sometimes incorporated instead (6). The miRNA guides the RISC to the target mRNA, and the AGO protein (which is typically AGO1 in *A. thaliana*) catalyses cleavage of the target mRNA.

1.2.2 Functions of miRNAs

1.2.2.1 Roles of miRNAs in stress responses

MiRNAs have been shown to be involved in various biotic and abiotic stress responses. MiR395, miR398, and miR399 have roles in nutrient homeostasis, regulating the uptake of sulfate, copper, and phosphate respectively by being induced when the specific nutrient is in low quantities. The miRNA then targets the mRNA of specific proteins that use the nutrient (7-9). This allows the nutrient to be utilised by other essential processes. MiRNAs are also involved in thermotolerance (10) and protection against oxidative stress (7), bacteria (11), and other pathogens (reviewed in (12)).

1.2.2.2 MiRNAs and development

Mutations in genes involved in miRNA processing have been shown to affect normal development (13-15) and altered expression of individual miRNAs also show developmental defects. For example, overexpression of miR165 in *Arabidopsis* affects establishment of organ polarity, vascular development, and apical meristem formation (16). MiR156 and miR172 have roles in the initiation of flowering and maintaining the vegetative phase (17) and miR159 is involved in seed germination (18).

1. 3 Approaches to Studying miRNAs

1.3.1 Identifying miRNA genes

To identify miRNAs present in a species, small RNA libraries are typically constructed (19, 20). Depending on the objectives of the study, small RNAs may be extracted from different tissues or from samples under various environmental conditions. Alternatively, previously identified miRNAs from other species can be used to identify conserved miRNAs.

To identify which of these sequences are miRNAs, the sequences are mapped against a reference for the organism. If the organism has its genome sequenced, the small RNAs

can be mapped against it with no mismatches, and then nucleotides from either end of the match are extracted, and software such as MFOLD (21) is used to predict whether the sequence can form the characteristic hairpin structure of a pre-miRNA with the mature sequence on the stem of the hairpin. MiRNAs can be found in intragenic regions (untranslated regions, exons, and introns of coding regions) and intergenic regions (22). MiRNAs derived from introns are known as mirtrons, and are the result of spliced introns forming hairpins (1).

If the genome sequence is unavailable, a similar technique is used, except miRNA libraries from the species of interest or other species are mapped to expressed sequence tags (ESTs, mRNA sequences) or genome survey sequences (GSS, sequenced fragments of genome). This method has been used to identify miRNAs in species such as citrus plants (23), where *Arabidopsis* miRNA sequences were used as a query against citrus ESTs, and 13 conserved miRNAs were found to match to hairpin-forming ESTs. This method has also been used in potato, where miRNA repertoires from nine plant species were mapped against potato ESTs, finding 48 potential miRNAs (24).

A combinatory technique was used in red algae (*Porphyra yezoensis*), where a small RNA library was sequenced, which was mapped against miRBase (25) to identify conserved miRNAs, and the remaining sequences were mapped against ESTs from the red alga to identify new miRNAs (19).

MiRNA genes can also be directly predicted from genomic sequences using characteristics of pre-miRNAs (26). Typical pre-miRNAs consist of non-exact stems which are regions of perfect base pairing separated by shorter regions of no base pairing (bulges). The stems are typically symmetrical, with the same number of bulges on each side of the stem (26). A typical pre-miRNA hairpin is shown in Figure 1.

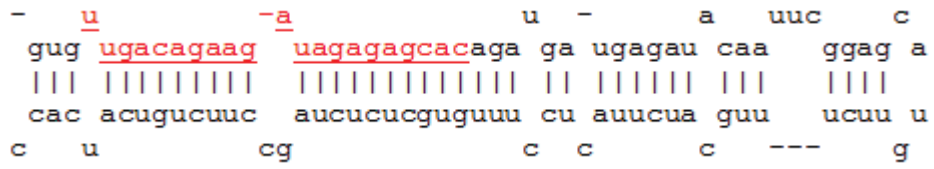

Figure 1: The pre-miRNA hairpin of miR157a in *A. thaliana*, a typical hairpin structure. MiRNA hairpins are typically symmetrical, and are characterised by regions of high complementarity interspersed with loops. The mature sequence is underlined.

1.3.2 Experimental and computational approaches to find miRNA targets

To understand the function of the discovered miRNAs, the target genes are predicted. To computationally predict targets, miRNAs are mapped to the available reference for the organism (the genomic sequence or EST libraries) using programs such as psRNAtarget (27) or miRanda (28).

Another technique is to use the degradome approach (29) to identify miRNA targets, as described in Figure 2. This screens degraded mRNAs for those whose degradation was mediated by miRNAs. By mapping the sequenced mRNA fragments to miRNA sequences, it can be predicted if the mRNA is a target; it is more likely if the last nucleotide of the degraded mRNA matches to the tenth nucleotide of the miRNA as this is where cleavage typically occurs (30).

Using this approach, miR398 targets in rice were validated via gene-specific 5' RACE, which showed one target to be a copper chaperone protein CCS1 (31).

An experiment to further validate this interaction was carried out in *Agrobacterium tumefaciens* (32). The bacteria were transformed with *Arabidopsis thaliana* miR398 primary transcript and *Arabidopsis* CCS1 and then infected into *Nicotiana benthamiana* leaves for co-expression analysis. There was no expression of CCS1 when miR398 was present. Interestingly, the miRNA had five mismatches with the target – this shows the importance of experimental validation as computational predictions often have a cut-off of three mismatches (31).

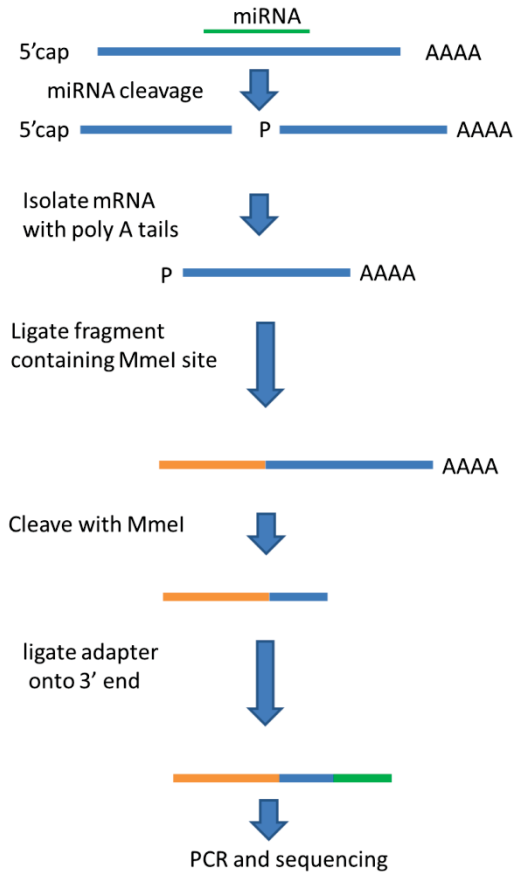

Figure 2: MiRNA target identification using degradome sequencing. RNAs that contain polyA tails are selected, then adapters are ligated to only fragments that have been degraded (The 5' cap blocks ligation). Treating with MmeI cleaves the mRNA downstream of the site, leaving a small identifier of the mRNA attached to the MmeI adapter. Another adapter is ligated to the 3' end and this is used for PCR amplification.

1.4 MiRNAs and Evolution

1.4.1 MiRNA families

MiRNAs that are widely conserved, for example, conserved amongst angiosperms, are typically coded by multiple loci. These are referred to as miRNA families and are grouped by identical or near-identical mature miRNA sequences. The regions of the pre-miRNA that do not contain the mature miRNA sequence are under less selection pressure, so the pre-miRNAs of family members exhibit variation in sequence and length.

The expansion of miRNA families are the result of whole genome, large-scale (segmental), and tandem duplications (33), akin to the origins of protein-coding gene families.

MiRNA family members can have different expression patterns, for example, the three members of the miR164 family in *Arabidopsis* are expressed in different tissue combinations, with only slight overlap between them (34). Older miRNAs (miRNAs that are more conserved) may also have more targets than young miRNAs (35).

1.4.2 MiRNA loss and gain

MiRNA analysis can also be used for evolutionary studies. Comparing miRNAs between species shows the change in regulation that has occurred; the loss or gain of a miRNA gives clues to how the species have changed during development and how this ultimately affects the adult phenotype. Such an analysis has recently been performed using 11 plant species and algae (36). It showed that there were no miRNAs shared between green algae and land plants and that many of the ancient land plant miRNAs were lost in present day flowering plants.

Highly conserved miRNAs have more important functions than less conserved (young) miRNAs, thus when losing the former there may be a large change in the phenotype of the organism; loss of these miRNAs may be an important mechanism for speciation (37). In contrast, miRNA genes may arise sporadically and may be under no selection pressure, so they diverge, and are subsequently lost, at a faster rate. However, it is possible for young miRNAs to become incorporated into regulatory networks, for example the *Brassicaceae*-specific miR824 regulates stomata development (38).

The expression of young miRNAs may be considerably lower than that of older miRNAs (37). While older miRNAs typically come in gene families (duplicated loci with similar or identical mature sequences), miRNAs with recent origin are likely found in only one locus in the genome.

1.4.3 MiRNAs and polyploidy

Polyploidisation results in a large increase in gene expression (a “genomic shock”) due to an increase in the number of copies of each gene, and organisms must be able to

quickly control this increase in gene expression to survive. Polyploids are of interest to study as polyploidisation may precede speciation and phenotypic diversification in plants (39). The diversification is suggested to be the result of different changes in gene expression to reduce the effect of the genomic shock, and such changes may be mediated by changes in miRNA expression.

Changes in miRNA expression have been implicated in polyploidisation of *Arabidopsis* plants. The levels of miRNA expression in natural and artificially generated *Arabidopsis* polyploids were analysed, and 40-50% of miRNAs were found to be differentially expressed between the parental *Arabidopsis* and the polyploids, as well as amongst the different polyploids (40).

1.5 *Pachycladon*

Pachycladon is a relatively new genus of eleven species of rockcress that is closely related to the model plant *Arabidopsis thaliana*. *Pachycladon* was formed after an allopolyploidisation event approximately 2 Mya followed by rapid diversification (41). The species are morphologically diverse and inhabit different altitudes and soil types (42). These characteristics make *Pachycladon* an appropriate model system for studying the genetics underlying adaptive diversification. As one of the parental lineages of *Pachycladon* has been a close relative of the model plant *Arabidopsis thaliana,* the large amounts of genetic and molecular resources for the latter are transferable to the former to a large extent.

Molecular and ecological evidence have suggested an adaptive radiation event has occurred in *Pachycladon* (43). Analysis of 10 nuclear genes of 52 individuals from the 11 *Pachycladon* species revealed distinct clustering by species and high diversification rates. The species occupy specific niches with ecological differences higher than would be expected from gradual selection after reproductive isolation (43).

Despite this high ecological divergence that would suggest adaptive speciation, no species-specific molecular adaptations are known in *Pachycladon* apart from differential expression of glucosinolate genes in *P. enysii, P. exile, P. fastigiatum*, and *P.*

novae-zelandiae (44, 45). The differential expression causes accumulation of different types of glucosinolates and their hydrolysis products, chemicals that are used for defence against herbivores, and this proposes a potential role for biotic factors in species divergence. Ample genetic diversity in glucosinolate hydrolysis genes has recently been identified in three species of *Pachycladon* (46).

1.6 Project Overview

1.6.1 Background Information

Considering the fast rate of divergence of *Pachycladon*, changes in miRNA diversity and expression may have contributed to species diversification within the genus. To investigate the nature of these changes, two species with very different ecological niches, *P. cheesemanii* and *P. fastigiatum,* were chosen for a comparative miRNA analysis.

P. cheesemanii is a generalist, inhabiting a wide range of altitudes and soil types, and found over a larger area than the other *Pachycladon* species (42). It is predicted to have a morphology the most similar to the ancestor of *Pachycladon* (47), and, along with *P. exile,* forms a sister group to all other *Pachycladon* species (Figure 3). The leaves of *P. cheesemanii* exhibit leaf heterophylly (i.e. different shaped leaves on the same plant) and may be broad elliptic to ovate in shape and lobed or serrate (42).

P. fastigiatum is a specialist, and only grows on greywacke at mid-altitudes (~1500 m above sea level). *P. fastigiatum* has narrow serrate leaves that are elliptic to lanceolate in shape (42).

Small RNA libraries had been sequenced from greenhouse-grown *P. cheesemanii* and *P. fastigiatum*, with mRNA sampled at the same time to construct EST libraries and perform differential expression analysis.

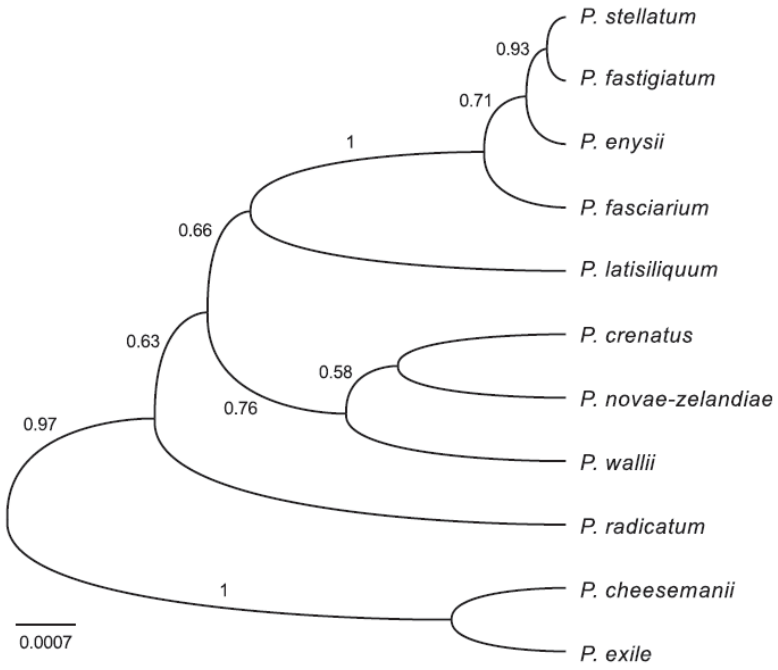

Figure 3: Phylogenetic tree of *Pachycladon*. Adapted from Joly, 2013.

1.6.2 Aims

The aims of this study are to identify and characterise miRNAs in *P. cheesemanii* and *P. fastigiatum* and to predict phenotypic differences between the species from differential miRNA and mRNA expression. These aims were split into four objectives:

1. Identify and quantify miRNAs present in both *Pachycladon* species (Chapter 2)
2. Identify potential targets of identified miRNAs in both species (Chapter 3)
3. Experimentally validate the sequence of selected miRNAs (Chapter 4)
4. Identify genes and GO terms differentially expressed between *Pachycladon* species (Chapter 5)

Chapter Two: Identification and Differential Expression Analysis of miRNAs

2.1 Introduction

MiRNA analyses in species with genome sequences typically involve mapping small RNA sequences to the genomes, but as there are no *Pachycladon* genome sequences, alternate approaches to identifying miRNAs must be devised. In this study, two methods are used to identify miRNAs. The first is to map the *Pachycladon* small RNAs against *A. thaliana* miRNAs, allowing mismatches to identify miRNAs with diverged sequences (similar to the study in *Porphyra yezoensis* (19). The second method is to search the *Pachycladon* EST libraries for pre-miRNAs using *A. thaliana* miRNA sequences as probes (as per (23)).

As *Pachycladon* is polyploid, there is added complexity in identifying miRNAs. As some miRNAs come in gene families, there is a possibility for sequence divergence not only in the gene families, but also between homeologs (homologs from the two parental genomes). For example, miR157 has three loci in *A. thaliana*: miR157a, miR157b, and miR157c. If both ancestors of *Pachycladon* also had these three loci, then the *Pachycladon* species potentially have six loci for miR157 (Figure 4).

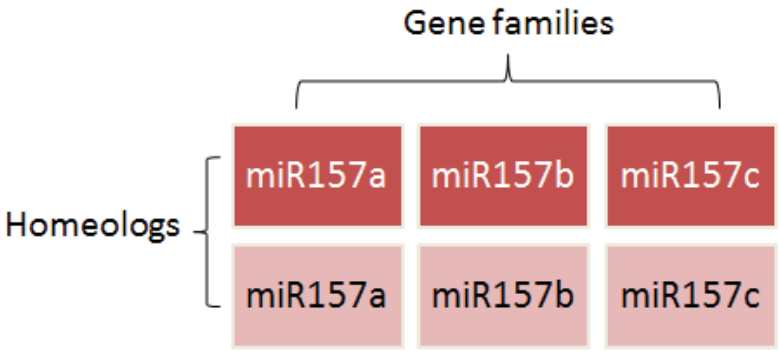

Figure 4: Potential miR157 loci in *Pachycladon*.

2.2 Methods

2.2.1 Small RNA sequencing

Total RNA was extracted from rosette leaves of glasshouse-grown *P. cheesemanii* and *P. fastigiatum* using the RNeasy kit (Qiagen). Small RNA was extracted from total RNA and small RNA libraries were prepared and sequenced on the Illumina Genome Analyzer GAII using Illumina's small RNA sample preparation kit and the 35 cycle sequencing kit. Due to budget considerations at the time, only one replicate of each library was sequenced.

2.2.2 Sequence trimming and sorting

The adapter used for sequencing was removed via the FASTQ/A Clipper from the FASTX Toolkit software (20). This also removed reads containing ambiguous nucleotides (labelled as "N") and reads shorter than five nucleotides. The DynamicTrim and LengthSort algorithms, implemented in the SolexaQA package (48) were used to quality trim and filter the data. Reads were trimmed using DynamicTrim to remove low-quality nucleotides from the ends of reads that had a p-value greater than 0.01. LengthSort was used to remove sequences shorter than 17 nucleotides and longer than 26 nucleotides.

2.2.3 rRNA and tRNA filtering

The remaining small RNA sequences were aligned to *Arabidopsis thaliana* tRNA sequences from the genomic tRNA database (49) and rRNA sequences from Rfam (50) using Bowtie, and any that matched perfectly were removed.

2.2.4 Identification of miRNAs using small RNA data

Bowtie (51) was used to map the filtered reads to all *Arabidopsis thaliana* miRNAs from the miRNA database miRBase (25), with the number of mismatches varied from 0-2 to test the effect mismatches have on the numbers and expression levels of identified miRNAs.

A. thaliana miRNAs from miRBase have gene family members reported separately, for example miR157 had three members named miR157a, miR157b, and miR157c. Mature miRNAs in the same gene family often have the same sequence, and as there is no way

to determine which gene the mature sequences come from in this case, the miRNA set was condensed down to unique sequences. For example, as miR157 family members have identical mature sequences, the three miRNAs were replaced by a single miRNA named miR157abc. MiRNA gene families with multiple sequences were grouped according to sequence.

Tablet (52) was used for visualisation of the Bowtie output in this chapter and Chapter 3.

2.2.5 Mining EST libraries of *Pachycladon* for miRNAs

The EST libraries for *P. cheesemanii* and *P. fastigiatum* were generated from leaves of plants that were cultivated for five months (53). ESTs were assembled for 13,284 unique genes for *P. fastigiatum* and 8,890 genes for *P. cheesemanii*, 5,684 of which were common to both species. The genes in the EST libraries are annotated according to their *A. thaliana* homolog, with unique identifiers added to distinguish homeologs.

To identify miRNA genes in the EST libraries, *A. thaliana* miRNAs were mapped to the *Pachycladon* EST libraries using Bowtie, with all perfect matches reported.

To account for changes in miRNA sequence between *Arabidopsis* and *Pachycladon* or for short ESTs that fail to include the mature region, the accession numbers from the *Pachycladon* EST libraries were compared to a list of *A. thaliana* miRNA accession numbers from TAIR (54) using the intersect function of R (55).

The pre-miRNAs of any identified ESTs were folded using the MFOLD (21). Alignments were done using ClustalX (56).

2.2.6 Differential expression analysis

The package edgeR (57), designed for differential expression analysis of digital data within the R environment, was used to perform differential expression analyses of miRNAs. Since there were no replicates in the miRNA data, the dispersion parameter was borrowed from the mRNA analysis in Chapter 4. A strict p-value of 0.001 was used to account for no replicates in the data and a log fold change of log(2)2 were used. In all figures, negative log fold change denotes up-regulation in *P. fastigiatum*, while

positive values are up-regulated in *P. cheesemanii*. The script used is given in Supplementary File 1.

2.3 Results

2.3.1 Identification of miRNAs

2.3.1.1 Data Pre-processing

Between 21 and 35 million reads were sequenced for each of *P. cheesemanii* and *P. fastigiatum* (Supplementary Table 1). After removing adapters and sequences smaller than 17 nucleotides, the number of reads decreased to approximately 12 million. rRNA and tRNA comprised a large proportion of these reads; 52% of the *P. cheesemanii* reads, and 40% of the *P. fastigiatum* reads matched perfectly to the rRNA sequences from Rfam and tRNA from the genomic tRNA database (Supplementary Table 1).

Of the 6-7 million reads that remained, approximately 300,000 in *P. cheesemanii*, and 200,000 in *P. fastigiatum* mapped to *A. thaliana* miRNAs using zero mismatches. The mapping was repeated with one and two mismatches. Allowing one mismatch increased the number of reads by approximately 10,000 reads in each species, while the second mismatch increased the number of reads matching by 500 reads.

These reads corresponded to a total of 52 miRNAs using 0 mismatches, 45 of which were present in both species (Figure 5). Six additional miRNAs were discovered using one mismatch, bringing the total to 58 miRNAs, 47 of which were in both *Pachycladon* species. Using two mismatches, 65 miRNAs were identified, 50 of which were shared between the species. The names and sequences of the identified miRNAs are given in Supplementary Table 2. Of the unique miRNAs, only miR852 in *P. cheesemanii* and miR161.1 in *P. fastigiatum* had more than ten reads mapping to the miRNA.

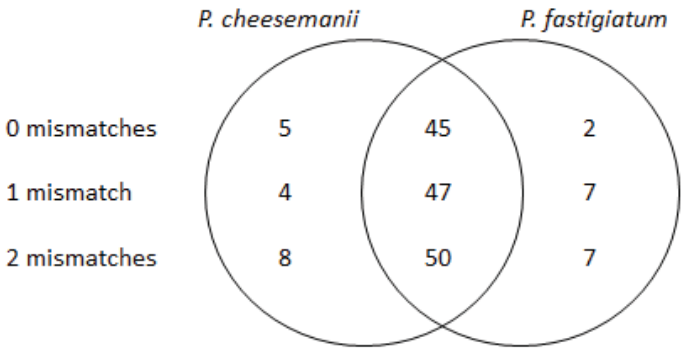

Figure 5: Comparison of the number of miRNAs identified in *P. cheesemanii* and *P. fastigiatum* with varying number of mismatches.

2.3.1.2 Identification of potential homeologs

To identify potential homeologs of the miRNAs, the sequence variants of the reads mapped to each identified miRNA were analysed. Graphs were made showing the major sequence variants as a proportion of the total number of reads for each miRNA (only variants with at least 10% of the reads for that miRNA in at least one species were considered). 10 miRNAs were identified as having possible sequence variants (Figure 6, Supplementary Table 3).

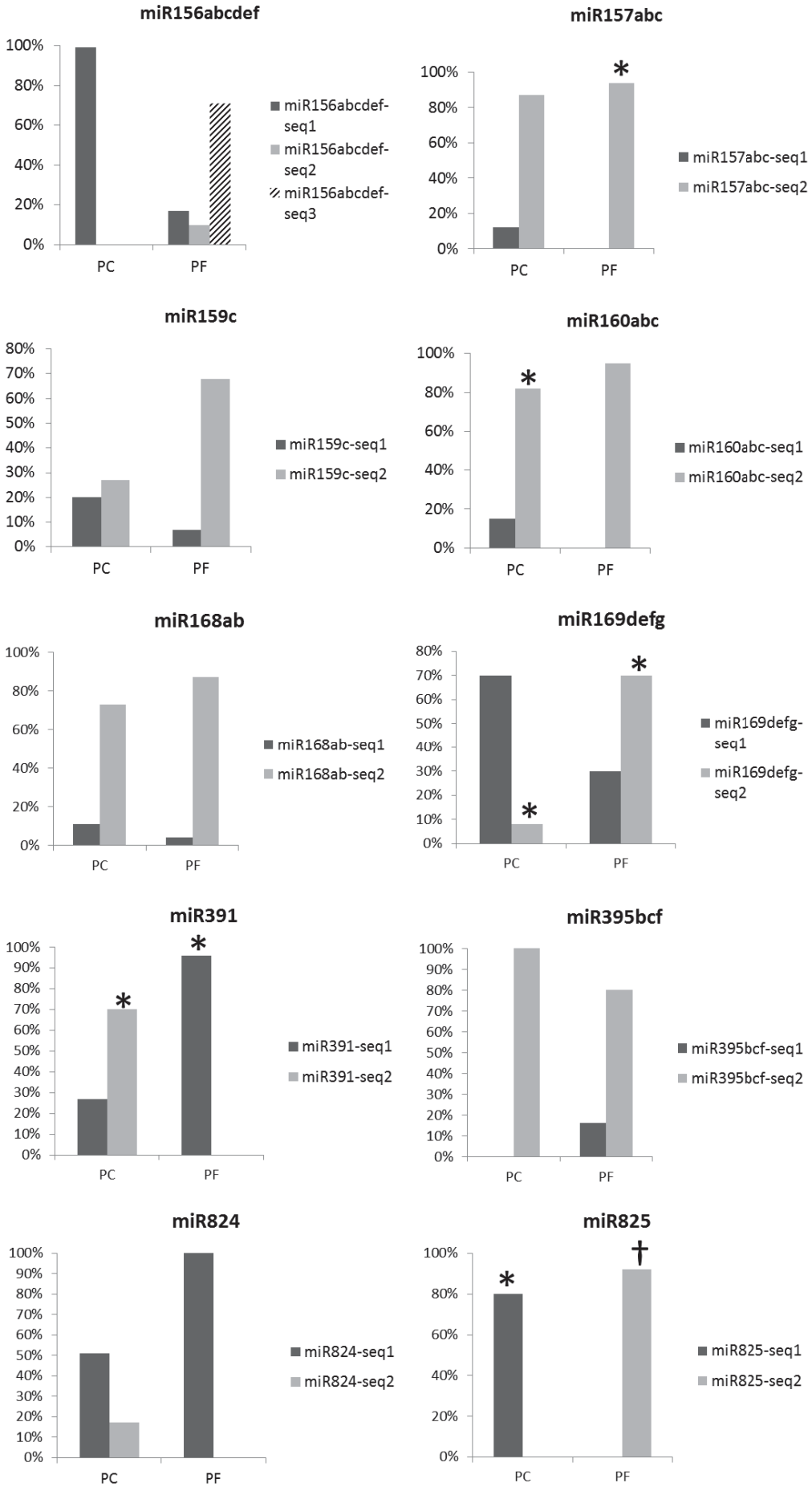
miR156abcdef
100%
80%
60%
40%
20%
0%
PC
PF
miR156abcdef-seq1
miR156abcdef-seq2
miR156abcdef-seq3
miR157abc
*
miR157abc-seq1
miR157abc-seq2
miR159c
miR159c-seq1
miR159c-seq2
miR160abc
*
miR160abc-seq1
miR160abc-seq2
miR168ab
miR168ab-seq1
miR168ab-seq2
miR169defg
*
*
miR169defg-seq1
miR169defg-seq2
miR391
*
*
miR391-seq1
miR391-seq2
miR395bcf
miR395bcf-seq1
miR395bcf-seq2
miR824
miR824-seq1
miR824-seq2
miR825
*
†
miR825-seq1
miR825-seq2

Figure 6: Potential homeologs of miRNA genes. The abundance of each sequence variant is given as a percentage of the total reads for the particular miRNA. Confirmed sequences are marked with an asterisk; the partially confirmed sequence of Pfa-miR825 is marked with a cross. Sequences for each variant are given in Supplementary Table 3.

2.3.2 Mining *Pachycladon* EST libraries for miRNAs

None of the *A. thaliana* miRNAs mapped perfectly to the EST libraries of *P. cheesemanii and P. fastigiatum.*

Searching the EST libraries for miRNA accession numbers of *A. thaliana* identified one miRNA precursor, miR414 (AT1G67195), in both species. This gene was present in two copies in the *P. cheesemanii* EST library and one copy in *P. fastigiatum* (Figure 7). The gene appears poorly conserved between the species, with deletions present in the mature region (highlighted in black) in all three *Pachycladon* ESTs. The predicted hairpins are in general very weak, in particular the predicted hairpin for *P. fastigiatum* which has many bulges and very few stem regions. The *P. cheesemanii* copy 2 hairpin resembled the *A. thaliana* hairpin the closest. Mature regions were predicted for both copies of *P. cheesemanii* (Figure 8).

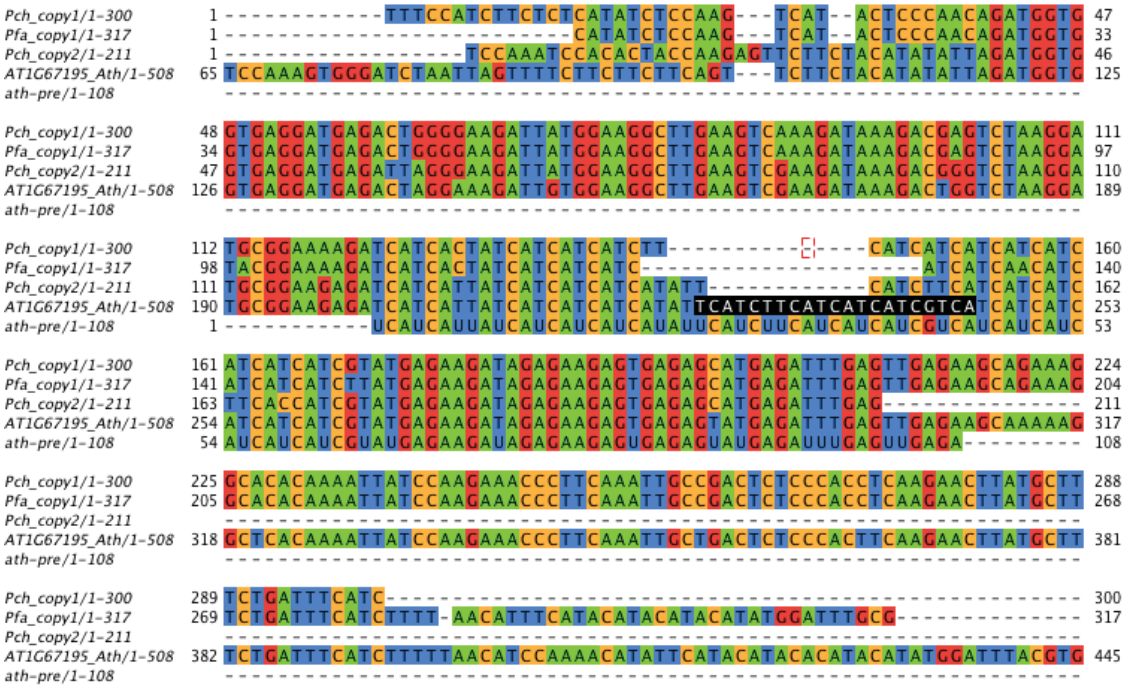

Figure 7: ClustalX alignment of miR414 (AT1G67195). Sequences included are from *A. thaliana* ("AT1G67195_Ath"), *P. fastigiatum* ("Pfa_copy1"), and two copies of *P. cheesemanii* ("Pch_copy1" and "Pch_copy2"). The pre-miRNA of *A. thaliana* is also included ("ath-pre") with the mature miRNA highlighted in black with white text.

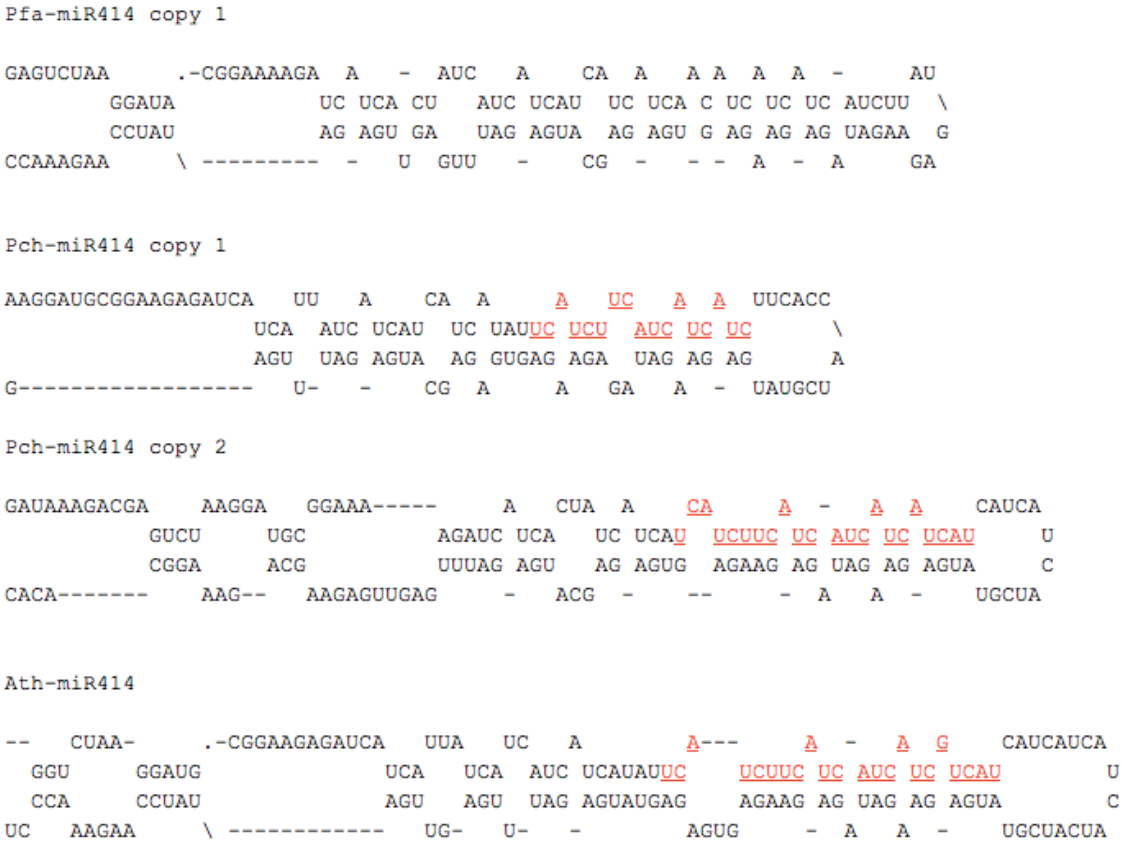

Figure 8: Hairpins of miR414 in *Pachycladon*. The hairpin of miR414 from *A. thaliana* is also shown. The mature region of *A. thaliana* is underlined, as well as potential mature regions for the *Pachycladon* sequences.

2.3.3 Differential expression analysis

2.3.3.1 Changing the number of mismatches

The differential expression analysis was performed on the datasets generated from mapping with zero, one and two mismatches using parameters log(2)2 and p =0.001 (Table 1). Twelve miRNAs were differentially expressed no matter the number of mismatches. Two miRNAs (miR156 and miR171a) were identified as differentially expressed only with zero mismatches. Four miRNAs were only differentially expressed when mismatches were allowed (miR825, miR472, miR169, miR848). Interestingly, one of these (miR825) was only differentially expressed when one mismatch was allowed, and the log FC was 10.5, which was the greatest difference in expression in these results. MiR852 and miR161.1, which were identified in the small RNA library of only one species, were also differentially expressed.

Table 1: The effect of changing mismatches on the identification of differentially expressed miRNAs.

	0	1	2
ath-miR159b	1.36	1.39	1.39
ath-miR164c	1.61	1.50	1.50
ath-miR852 [a]	8.24	8.29	8.29
ath-miR156abcdef	3.23		
ath-miR395bcf	-2.14	-2.39	-2.39
ath-miR166abcdefg	-1.94	-1.86	-1.86
ath-miR165ab	-1.59	-1.58	-1.58
ath-miR398bc	-1.83	-1.75	-1.75
ath-miR396b	-2.22	-2.20	-2.20
ath-miR171a	-1.18		
ath-miR391	-4.71	-2.97	-2.96
ath-miR394ab	-1.99	-2.02	-2.02
ath-miR157abc	-1.56	-1.52	-1.35
ath-miR408	-2.16	-2.15	-2.15
ath-miR825		-10.51	
ath-miR472		-1.53	-1.54
ath-miR169hijklmn		-2.65	-2.65
ath-miR848		-12.17	-8.75
ath-miR161.1 [a]			-8.56

Note: Numbers refer to the respective logFC. Positive logFC are miRNAs up in *P. cheesemanii*, and negative logFC are up in *P. fastigiatum*. [a] miRNA only present in the small RNA library of the respective species.

2.3.3.3 Variant Analysis

To test if the unusual patterns of differential expression that arise with changing the mismatches are due to variants, the differential expression analysis was repeated with the identified variants from section 1.3.1.2 added (Table 2).

Five miRNA families had more than one of the predicted sequence variants differentially expressed; miR825 being one of them, with a unique sequence up-regulated in each species. MiR157abc, miR391, and miR395bcf also had the previously identified sequence variants differentially expressed. MiR156abcdef had two of the three predicted variants differentially expressed.

MiR159c, miR160abc, and miR169defg, which were not differentially expressed when the variants were counted together, each had one sequence variant differentially expressed when the variants were considered individually.

Table 2: Differentially expressed sequence variants.

miRNA name	Sequence	logFC	p-value
miRNAs with one variant up-regulated in each species			
miR157abc	Seq1	7.60	1.67E-05
	Seq2	-1.50	1.52E-05
miR391	Seq1	-4.80	2.01E-33
	Seq2	10.14	1.17E-20
miR825	Seq1	9.98	2.25E-19
	Seq2	-11.40	8.32E-34
miRNAs with two variants up-regulated in one species			
miR395bcf	Seq1	-8.03	1.48E-07
	Seq2	-2.16	1.04E-07
miRNA with one variant not differentially expressed			
miR156abcdef	Seq1	3.26	2.42E-14
	Seq3	-9.94	8.47E-20
	Seq2	-2.04	1.58E-02
miR159c	Seq1	2.31	8.98E-04
	Seq2	-0.38	4.16E-01
miR160abc	Seq1	10.62	8.75E-24
	Seq2	-1.00	2.19E-03
miR169defg	Seq1	2.32	1.40E-05
	Seq2	-1.13	6.82E-04

Note: Sequences can be found in Supplementary Table 3. Variants not differentially expressed are shaded grey.

2.4 Discussion

2.4.1 Data Pre-processing

Even after tRNA and rRNA filtering, most of the reads did not map to miRNAs. A large proportion of the remaining reads are likely to still be small pieces of rRNA or tRNA – no mismatches were used in the rRNA and tRNA filtering steps, so any reads with sequence changes between *Arabidopsis* and *Pachycladon* would not have been filtered out. The variable regions in the rRNA in particular would not have been removed. Small interfering RNAs (siRNAs) would also be present in the data. SiRNAs are involved in antiviral defence and chromatin modification, and were not the focus of this study.

2.4.2 The effect of mismatches

A typical miRNA analysis would involve finding sequences that perfectly match miRNA sequences from other species. The disadvantage of this is that it does not account for sequence divergence, and miRNAs present may not be identified.

Conserved miRNAs typically have the same sequence across species, while genus or family-specific miRNAs exhibit variation. These miRNAs are likely of more recent origin and are not as essential as the conserved miRNAs. They may also be involved in speciation or adaptation and their presence or absence may give evolutionary information.

Allowing mismatches in this study enabled the identification of more miRNAs including miR472 and miR825. MiRBase entries for miR472 are present for *Citrus sinensis* and *Populus trichocarpa* as well as *A. thaliana* and *A. lyrata*, but the mature and pre-miRNA sequences are largely different in sequence. The sequence for miR472 shared between the two *Pachycladon* species was different to the previously reported sequences for miR472 and thus would not have been identified if mismatches had not been allowed.

MiR825 presents a similar case: the miRNA is presently only identified in *A. thaliana* and *A. lyrata*. The two *Arabidopsis* species had different sequences; the *A. lyrata* sequence was the same sequence as identified in *P. fastigiatum*, while *P. cheesemanii* had a unique sequence.

The method used for counting reads for certain miRNAs has implications in differential expression analyses. The different datasets created by allowing certain numbers of mismatches had different miRNAs differentially expressed. MiR825 represents an interesting case, as the sequence differences between the *Pachycladon* species, and the differences to the *A. thaliana* reference, resulted in the miRNA only being differentially expressed using one mismatch.

2.4.3 Variant Analysis

By separating the reads that map to a miRNA by sequence, potential homeologs were identified for 10 miRNAs (156abcdef, 157abc, 159c, 160abc, 168ab, 169defg, 391, 395bcf, 824, 825). In the case of 156, 157, 160, 168, 169, and 395, these variants could be either family members or homeologs. In *A. thaliana,* these loci have the same mature sequence (and so were grouped together in this analysis), but this may not be the same for *Pachycladon*. For the remaining miRNAs with only one locus but with sequence variants, it is possible these may represent the two homeologous copies of the locus.

Giving these sequence variants unique identifiers resolved the patterns shown in the differential expression analysis, in which miRNAs were only differentially expressed when a certain number of mismatches were allowed. Analysing sequence variants, especially when dealing with miRNAs that have multiple loci, gives additional information and should be an important step in miRNA identification.

2.4.4 Mining EST libraries for miRNAs

The attempt to find miRNAs in the EST library was largely unfruitful, with the exception of miR414. One read mapped to this miRNA in the small RNA library of *P. cheesemanii* using two mismatches (Supplementary Table 2) but this sequence did not match to the potential mature sequences of either *P. cheesemanii* copy.

The absence of other miRNA transcripts in the EST libraries could be attributed to low expression levels or high turnover rates of the miRNA transcripts. The usefulness of EST libraries is also dependent on the characteristics of the library –depth of coverage, tissues and developmental stages mRNA has been obtained from. The mRNA used to make the EST library used for this study was sampled at the same time as the miRNAs.

This is a positive if comparisons are to be made between the expression of mRNA and the target miRNAs, but to obtain sequences of the pre-miRNAs, the EST library must be more comprehensive, as it has been previously demonstrated that it is possible to obtain pre-miRNAs from EST libraries (19, 23, 24).

2.4.5 MiRNAs of interest from this section

Eighteen miRNAs were chosen to validate experimentally due to their patterns of differential expression (miR157abc, miR159b, miR160abc, miR161.1, miR164c, miR165ab, miR166abcdefg, miR169hijklmn, miR391, miR394ab, miR395bcf, miR396b, miR398bc, miR408, miR472, miR825, miR848 and miR852) (see Chapter 3). mir156, miR159, miR164, and miR852 were up-regulated in *P. cheesemanii*, and the rest of the selected miRNAs are up in *P. fastigiatum*. To assess their particular role in the divergence of the two *Pachycladon* species, the literature was searched for experimental and other evidence of their function.

2.4.5.1 MiRNAs up-regulated in *P. cheesemanii*

MiR156 targets SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) factors involved in the initiation of flowering in response to cold (58). MiR159 targets MYB transcription factors that are involved in the regulation of leaf, flower, and seed development (59, 60).

MiR164 regulates NAC-domain proteins involved in root, shoot and flower development, including CUP-SHAPED COTYLEDON (CUC) proteins (61), which are regulators of leaf serration. Interestingly, reduced expression of miR164c causes serrated leaves in *A. thaliana* (61), and miR164c was up-regulated in *P. cheesemanii*. *P. fastigiatum* has serrated leaves while *P. cheesemanii* does not (42), thus differential expression of miR164 may have a role in leaf diversification in *Pachycladon*.

The function of miR852 has not yet been reported.

2.4.5.2 MiRNAs up-regulated in *P. fastigiatum*

MiR157, miR159, and miR160 regulate plant development. MiR157 is a family member of miR156 and also targets SPL factors (62). MiR160 is involved in vegetative and

flower development (63). MiR161.1 targets a superfamily of pentatricopeptide repeat (PPR) proteins, which have roles in organelle biogenesis (64).

MiR165 and miR166 are members of the same miRNA family, and regulate meristem formation and leaf polarity (65, 66). MiR169 targets the transcription factor NFYA5, which promotes drought resistance (67). MiR171 targets SCARECROW-LIKE (SCL) transcription factors involved in development of root and meristem development (68) and is also involved in drought response in *Solanum tuberosum* (69).

MiR396 is a regulator of cell proliferation and pistil development (70). Overexpression of miR396 causes narrow leaves in *A. thaliana* (71), a trait possessed by *P. fastigiatum* which had miR396 up-regulated when compared with *P. cheesemanii*, a species with broad leaves (42). However, reduced expression of miR394, or a miRNA-resistant version of its target, LEAF CURLING RESPONSIVENESS, LCR, also causes narrow leaves (72). MiR394 was up-regulated in *P. fastigiatum*.

MiR398 and miR408 regulate copper homeostasis and the response to oxidative stress by targeting Cu/Zn superoxide dismutases, and miR408 also promotes vegetative growth (7, 73, 74).

There is less functional information on the remaining miRNAs. MiR472 in *P. trichocarpa* is predicted to target a number of disease resistance proteins (75) and is up-regulated when infected with tomato mosaic virus (76). MiR391 and miR825 are both preferentially expressed in rosette-stage leaves (77) and while miR391 expression increases under hypoxia, miR825 expression was shown to decrease after bacterial infection (78). MiR848 has no function currently identified.

Taken together, interesting correlations between differential miRNA expression and leaf morphology were observed (miR164, miR396) as well as up-regulation of miRNAs involved in initiation of flowering (miR156) and abiotic stress responses (miR169, miR171, miR398, miR408). The correlations between miRNA expression and *Pachycladon* phenotype have to be interpreted with caution as inferences of differential expression were made based on a qualitative, unreplicated analysis (see methods). For example, both miR394 and miR396 were up-regulated in *P. fastigiatum*

but have opposing effects on leaf width. While these contradictory expression patterns may be real they might also be an artefact of the qualitative analysis. Nonetheless, these results hint at particular traits being important in species diversification, such as initiation of flowering and response to drought and osmotic stress.

Chapter Three: Identification of Potential miRNA Targets

3.1 Introduction

As demonstrated in the literature review of the differentially expressed miRNAs in section 2.5, miRNAs can be involved in multiple processes, targeting multiple genes. MiRNAs that are conserved typically have multiple targets while less conserved miRNAs are likely to have fewer targets (35, 79). Changes in miRNA sequences, as well as changes in the miRNA binding site of the target, affect the strength of the miRNA binding. As a result, miRNA-target relationships can be lost, and this would change the regulation of the targets. Similarly, miRNA binding sites in genes can also be gained via mutations.

Pachycladon may have gained or lost miRNA targets after polyploidisation. Potential gains and losses of miRNA targets have been investigated in rice (80), which had a whole genome duplication ~70 million years ago. Gains or losses of targets were observed between homeologs, with only one copy of the target retaining a miRNA binding site.

In this chapter, targets of the identified *Pachycladon* miRNAs are predicted, and the miRNA-target relationships are compared with *A. thaliana* miRNA-target relationships to determine if interactions have been gained or lost.

3.2 Methods

Two programs were used to identify potential targets of the *Pachycladon* miRNAs: Bowtie (allowing three mismatches, (51)) and psRNAtarget (using default parameters, (27)), with the respective *Pachycladon* EST library used as a reference. Both programs rely on the high complementarity between the miRNA and its target binding sites. Bowtie simply reports all ESTs matching to miRNAs with up to three mismatches, while psRNATarget takes factors such as target accessibility and statistical significance into consideration.

For the miRNAs that mapped to at least one EST, the targets were compared between the two *Pachycladon* species, and these were compared with the targets confirmed in *A. thaliana*. *Arabidopsis thaliana* miRNA-target relationships were obtained from TarBase (81).

Only miRNAs that mapped to one EST or more were included in this comparison. The other identified miRNAs may have targets present, but they did not map with either program.

The miRNA-EST relationships were split into three categories:

1. Similar to *A. thaliana*: MiRNA-target relationships confirmed in *A. thaliana* were also identified in at least one of the *Pachycladon* species.
2. Potential loss of target: The miRNA-target relationship is confirmed in *A. thaliana*, but could not be established in either *Pachycladon* species despite the targets being present in the EST libraries.
3. Potential gain of target: The *Pachycladon* species has a potentially new miRNA-target relationship suggested by computational analysis that has not yet been reported in *A. thaliana*.

ClustalX alignments were created for the cases where a miRNA-target relationship was present in one *Pachycladon* species but did not occur for the other species. ClustalX alignments were also generated for the cases where the miRNAs failed to map to a predicted target.

Due to the polyploid nature of *Pachycladon*, the two homeologs of the genes were often present in the EST library. If present, both copies of the ESTs are included in the analyses in this chapter.

3.3 Results

A total of 15 miRNAs mapped to at least one EST from either *Pachycladon* species, with a total of 111 miRNA-target relationships identified.

3.3.1 miRNA-EST relationships that are the same in *A. thaliana* and *Pachycladon*

Twenty-two miRNA-EST relationships in one of both of the *Pachycladon* species that are the same as in *A. thaliana* were identified (Figure 9). Six of these were present and mapped in both, nine were present in both but mapped in only one species. Six miRNA-target relationships were predicted in one species but the EST was absent in the other species. For each of the cases where ESTs were present in both species, but mapped in only one, ClustalX alignments were produced to visualize why the mapping only occurred in one species. In all of these cases, the mRNA was too short in the other species, and did not cover the binding region of the miRNA.

miR156	AT2G42200	AT2G33810
PC	BT PS	
PF	BT PS	BT PS

miR157	AT2G33810
PC	
PF	PS

miR159	AT3G11440
PC	NP
PF	BT PS

miR164	AT1G56010	AT5G39610
PC	BT PS	
PF	BT PS	PS

miR168	AT1G48410
PC	BT PS
PF	NP

miR169	AT3G20910	AT5G12840	AT1G54160
PC	PS		BT PS
PF	PS	PS	BT PS

	Target present and maps in *Pachycladon*
	EST too short
BT	Maps with Bowtie
PS	Maps with psRNATarget
NP	Not present in EST library

Figure 9: MiRNA-target relationships confirmed in *A. thaliana* that are present in *Pachycladon*. PC = *P. cheesemanii*; PF = *P. fastigiatum*. *A. thaliana* locus identifiers are given as IDs for target genes.

miR171	AT4G00150
PC	NP
PF	BT PS

miR172	AT4G36920	AT2G28550	AT2G39250
PC			PS
PF	PS	BT PS	PS

miR394	AT1G27340
PC	BT PS
PF	BT PS

miR395	AT5G10180	AT5G43780
PC	BT PS	NP
PF		BT PS

miR396	AT1G10120	AT2G36400
PC	NP	NP
PF	PS	PS

miR398	AT1G08830
PC	NP
PF	BT PS

miR403	AT1G31280
PC	
PF	BT

miR408	AT2G02850
PC	BT PS
PF	

Figure 9 (continued)

3.3.2 Potential loss of target in *Pachycladon*

In total, there were 26 predicted targets of the miRNAs present in one or both of the *Pachycladon* EST libraries that did not map to their respective miRNAs (Figure 10). The lack of mapping could be due to one of three scenarios:

1. Neither program has the potential for accurate target prediction i.e. there are more than three mismatches in the event of Bowtie or the interaction is outside of the psRNATarget parameters.
2. The EST is too short and does not cover the binding region.
3. The miRNA has lost the target in *Pachycladon.*

To test the first scenario, the mappings using psRNAtarget and Bowtie were repeated using *A. thaliana* miRNAs and cDNA sequences. Neither program mapped eleven of the 26 targets to their corresponding miRNA in *A. thaliana*. These are the false negatives.

For the remaining 15 targets that did map in *A. thaliana*, ClustalX alignments were produced with the *A. thaliana* cDNA and the corresponding *Pachycladon* EST(s). Seven of the ESTS were present in *P. cheesemanii*, four of which did not map because of length, and 14 of the ESTs were present in *P. fastigiatum*, eight of which did not map because of length.

In all the remaining three *P. cheesemanii* and six *P. fastigiatum* ESTs, except for one, the sequence where the miRNA binds was identical to the binding sequence in *A. thaliana*, but the surrounding sequences were altered –either because of deletions, insertions, or nucleotide changes (Figure 10). In all cases, there were no changes in miRNA sequence between *A. thaliana* and *Pachycladon*.

The one case where the binding site had changed was the miR396-AT4G25210 pair (Figure 11). There were differences in the binding site, along with multiple nucleotide changes up and downstream of the binding site. The EST was only expressed in the *P. fastigiatum* library, and both copies were present. One copy had a one-nucleotide change in the binding site, and the other copy had three changes. This copy also had two regions of 20 base pair length insertions, so the product may have been non-functional. AT4G25210 codes for a transcriptional regulator with an unknown role.

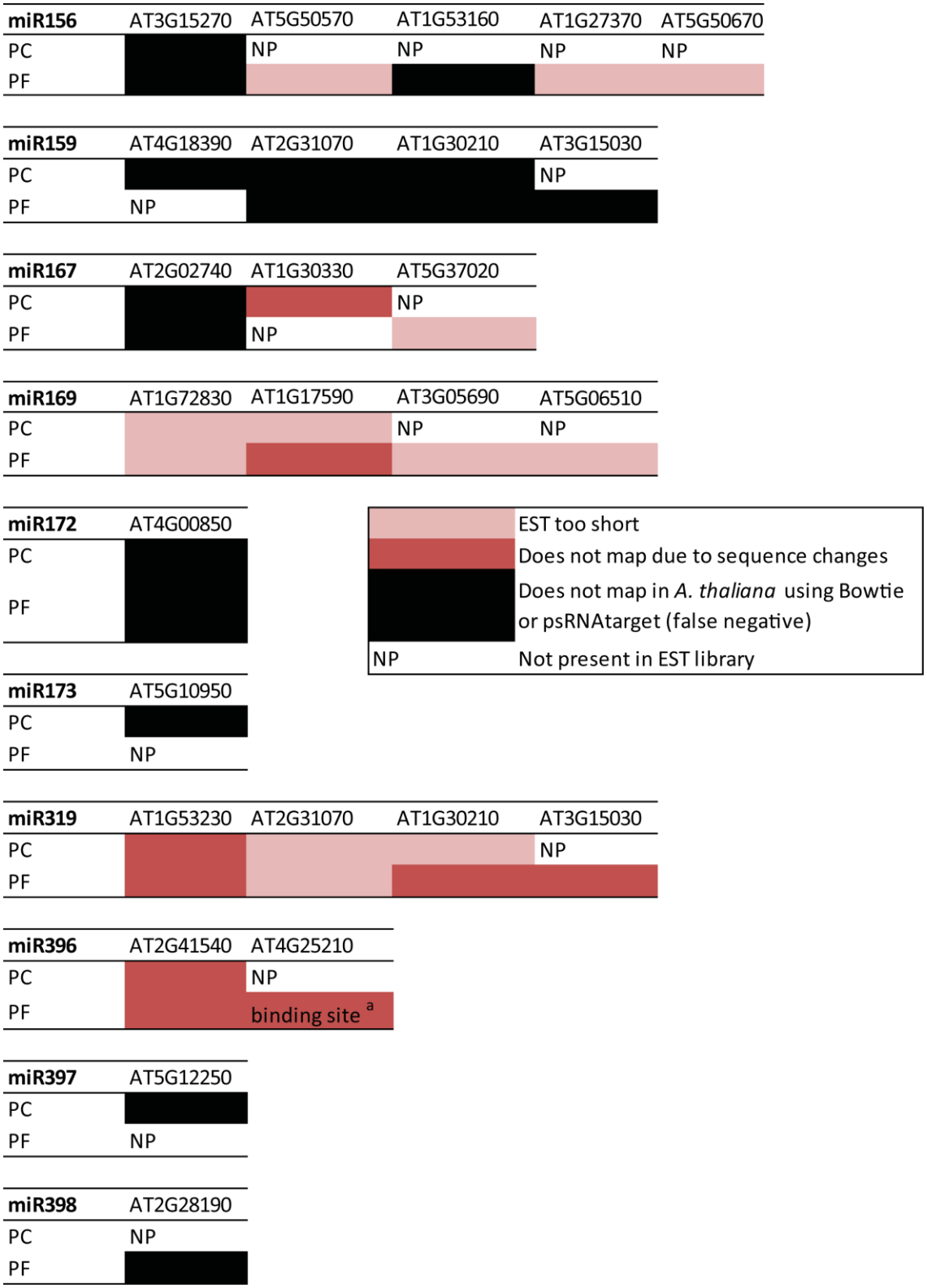

Figure 10: Predicted miRNA targets present in *Pachycladon* that do not map. [a] Sequence changes are present in the miRNA-target binding site.

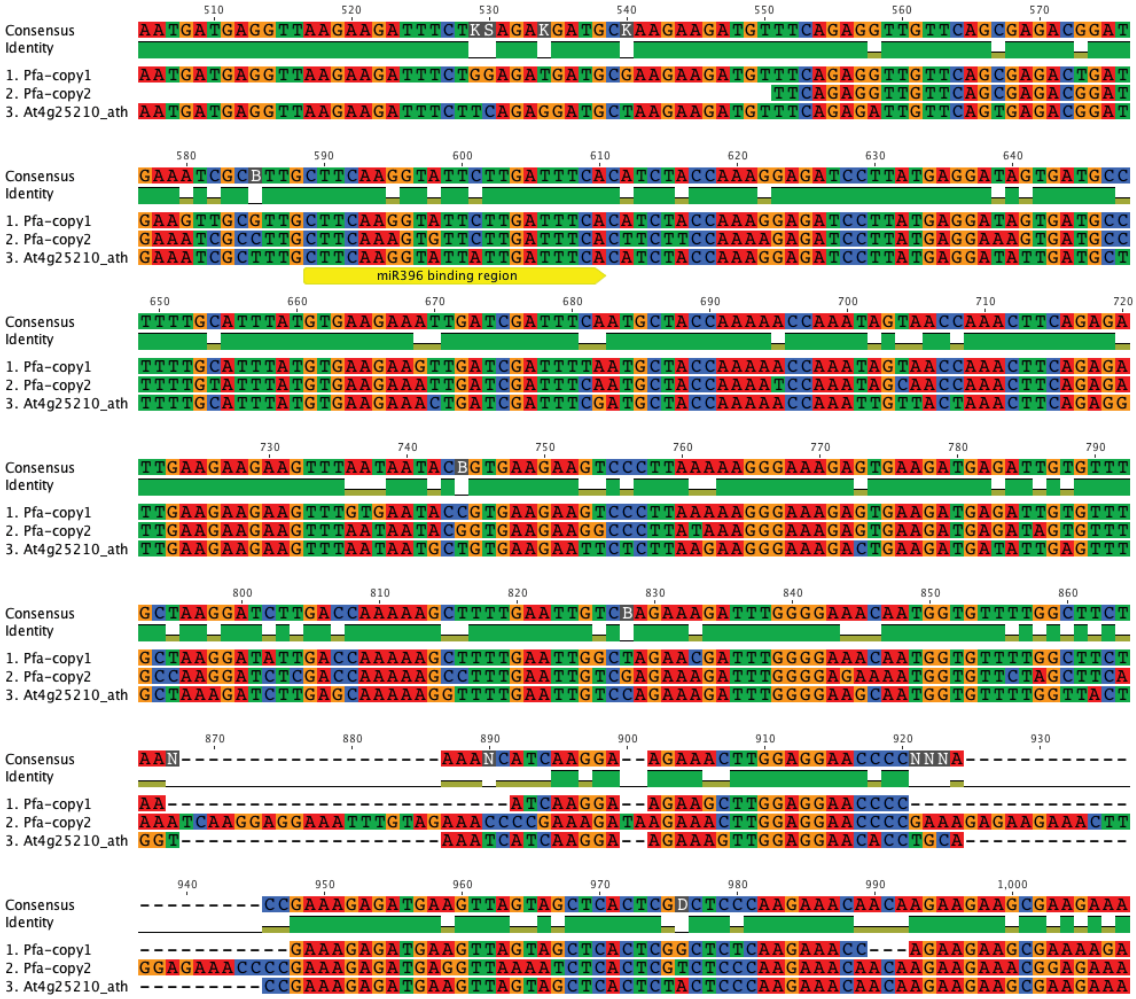

Figure 11: ClustalX alignment of AT4G25210 in *A. thaliana* and *P. fastigiatum*, a potential loss of target in *Pachycladon*. The binding site of miR396 in *A. thaliana* (AT4G25210_ath) is annotated. Both copies of AT4G25210 were present in *the P. fastigiatum* EST library ("Pfa_copy1" and "Pfa_copy2") and both failed to map to miR396.

3.3.3 Potential gain of target

Additionally, 63 miRNA-target relationships were identified that have not been reported in *A. thaliana* (Figure 12). As in section 3.3.2, the target prediction was repeated using *A. thaliana* sequences, and 19 of these relationships were identified in *A. thaliana*, i.e. they were false positives.

For the remaining 44 interactions, 10 of these interactions were identified in both species, 19 of the interactions were only present in one species due to the EST being present in only one of the libraries, and five were present in both but did not map due

to the EST being too short in one species. Six of the remaining nine mapped in one, but did not map in the other due to sequence differences in the predicted binding region. The remaining three miRNA-target relationships were due to changes in the miRNA sequence itself.

One relationship (miR157-AT1G75860) was specific to *P. fastigiatum*, while the other two (miR825 with AT2G19310 and AT5G07690) were specific to *P. cheesemanii*.

miR156	AT1G20100	AT5G41810	AT5G61380
PC	PS	NP	NP
PF	NP	PS	

miR157	AT2G42200	AT2G45990	AT5G54290	AT1G20100	AT5G12950	AT1G75860
PC			NP	PS	PS	PS
PF			BT	NP	1MM	miRNA

miR159	AT5G10480	AT1G53430	AT1G80230	AT4G26840	AT2G15220
PC	BT PS			PS	
PF	2MM			PS	PS

miR162	AT1G48430	AT3G63180
PC	BT PS	PS
PF	BT PS	3MM

miR164	AT1G26820	AT1G07400
PC	BT PS	
PF	BT	NP

miR167	AT1G63160	AT5G53620
PC	PS	NP
PF	PS	PS

	MiRNA maps to EST
	Does not map due to sequence changes
	EST too short
	Maps in *A. thaliana* (false positive)
BT	Maps with Bowtie
PS	Maps with psRNAtarget
NP	Not present in EST library
xMM	x mismatches between species
miRNA	Difference in miRNA sequence

miR169	AT3G14020	AT3G55110	AT2G34070	AT3G55940	AT2G41180	AT5G24680	AT3G56690	AT1G15110
PC	BT PS	BT PS	BT	BT	PS	NP	NP	NP
PF		NP	NP	BT		PS	PS	PS

miR171	AT4G27000	AT4G36710	AT5G19430
PC	BT PS	NP	2MM
PF	BT PS	BT PS	PS

miR172	AT2G05830	AT4G27830	AT1G31800	AT2G16500	AT3G54440	AT3G47360	AT1G59990
PC	BT	PS	2MM	NP	NP	NP	PS
PF	BT	1MM			BT		PS

Figure 12: New potential miRNA-target relationships identified in Pachycladon.

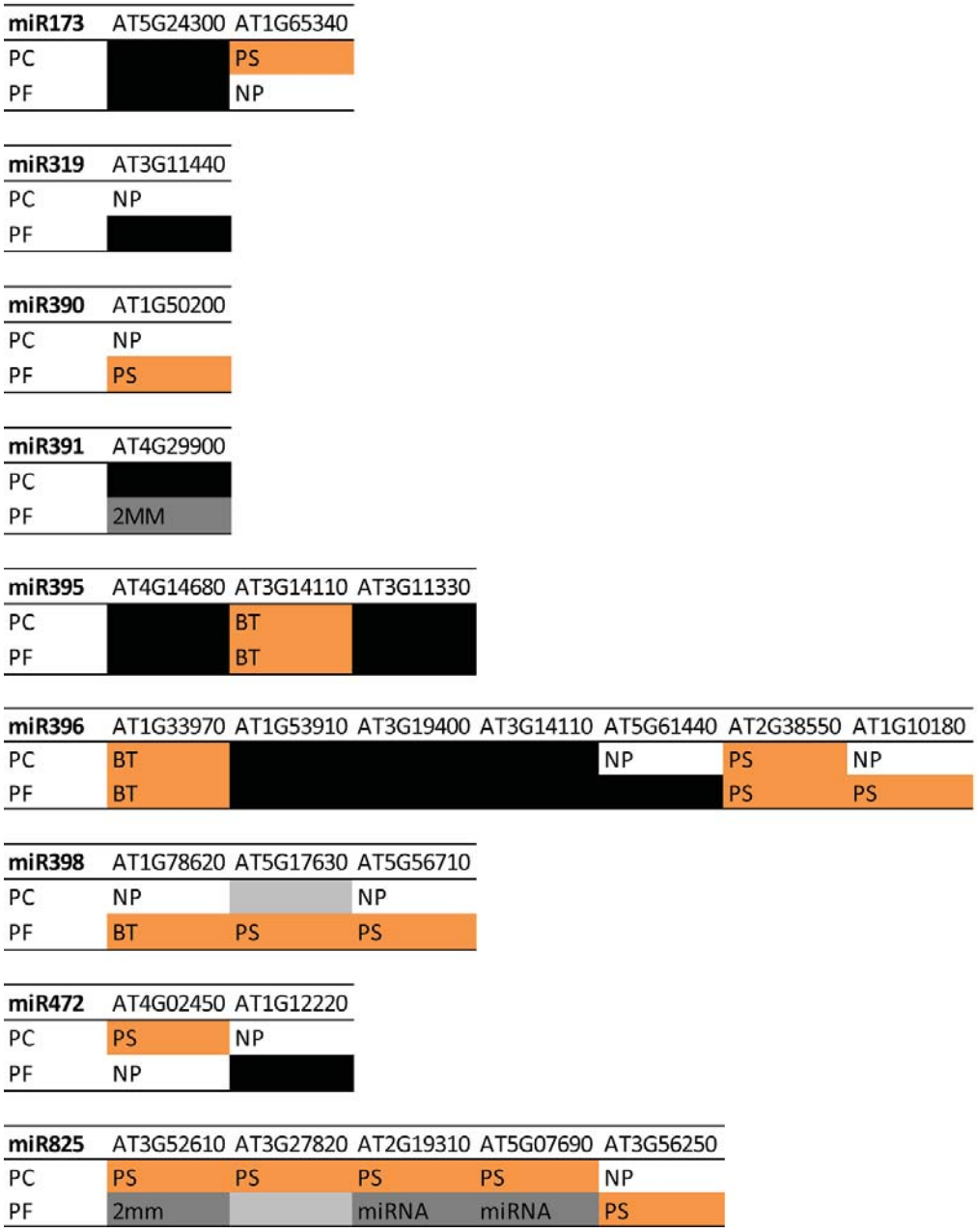

Figure 12 (continued)

3.4 Discussion

3.4.1 Bowtie vs psRNAtarget

The two methods used for target prediction have their advantages and disadvantages. Bowtie is able to map the miRNA in the forward and reverse directions, but does not simulate the binding of miRNA to target, and only allows up to three mismatches between miRNA and target. Typically, miRNAs bind to their targets with high complementarity, but weaker interactions would not be identified. PsRNAtarget takes

into consideration binding energy, target accessibility, and where the mismatches are located in the miRNA-target duplex. It does not consider binding in the reverse direction, however. The EST libraries used often had the sequences of the reverse complement, an antiquity of the library assembly, and psRNAtarget failed to map the miRNAs to them. The mapping can, however, be repeated using the reverse complement of the ESTs.

The two programs often disagreed in the targets predicted. Overall, only 25% (17 out of 69) of the total miRNA-target relationships predicted were predicted by both psRNAtarget and Bowtie. For the miRNA-target relationships confirmed in *A. thaliana*, the two programs agreed for 57% (8/14) of the relationships. Conversely, for the new miRNA-target relationships, only 16% (9/55) of the relationships were predicted by both programs. Many of the new miRNA-target relationships were only predicted by psRNAtarget (22/55, 60%).

3.4.2 Potential losses and gains of targets in *Pachycladon*

Of the 109 miRNA-target interactions predicted computationally in this study, over half were previously unreported miRNA-target interactions in *A. thaliana*. These potentially new miRNA-target interactions may have contributed to the evolution of *Pachycladon*. One example in particular is the interaction between miR825 and AT5G07690 in *P. cheesemanii*. AT5G07690 codes for MYB29, a transcription factor involved in the synthesis of methionine-derived glucosinolates. The change in the sequence of miR825 causes this new interaction. Repeating the mapping of miR825 to AT5G07690 using *A. thaliana* sequences of the gene and miRNA gave no result, so it is less likely to be the result of a false positive. An alignment of these genes with the new binding site in *P. cheesemanii* is shown in Figure 13.

The two programs have a rather high error rate, as seen in Figure 10; half of the miRNA-target interactions that were confirmed in *A. thaliana* but failed in *Pachycladon* also failed in *A. thaliana* when the same parameters were used (false negatives). There was also a high rate of false positives (Figure 12).

Most of the potential losses of targets in *Pachycladon* (Figure 10) were due to changes in the regions surrounding the binding site. The surrounding sequences are important for correct miRNA-target binding as they affect the secondary structure of the mRNA, which in turn affects the accessibility of the RISC proteins to the binding site (82).

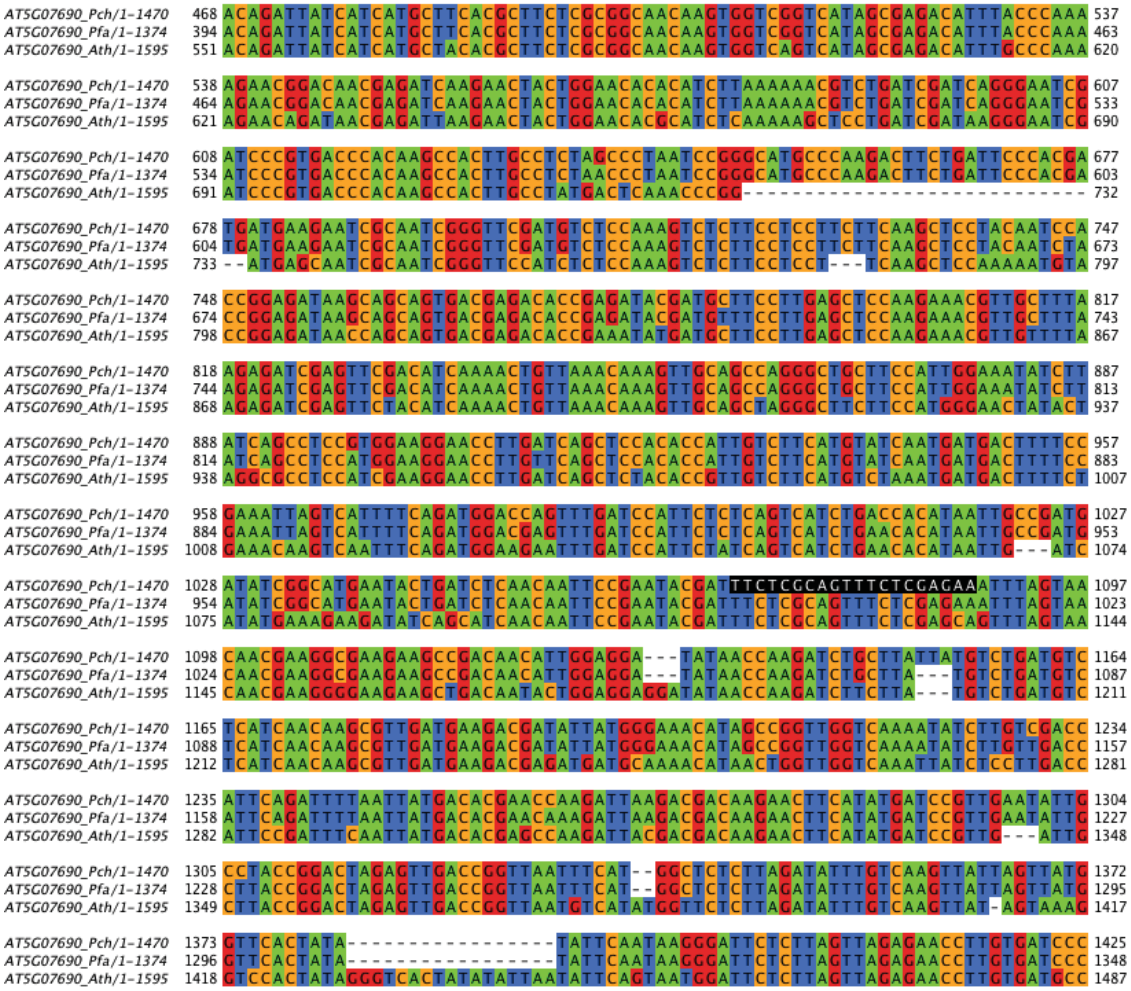

Figure 13: ClustalX alignment of AT5G07690, a potential gain in target in *Pachycladon*. This target is new in *P. cheesemanii* due to a change in miRNA sequence. The binding site of miR825 in *P. cheesemanii* is highlighted in black.

3.4.3 Differences between the species

In most cases, the differences in targets between the two *Pachycladon* species were due to ESTs being too short and not covering the miRNA binding site, or the ESTs being absent from one of the libraries. The cases where there were differences in miRNA-target relationships between the species were in the potentially new target relationships.

In some cases, this was due to the polyploid nature of *Pachycladon* – the two copies of the gene varied in miRNA binding sequence, so that one maps and the other does not. One species expressed the copy of the target that maps while the other species expressed only the other copy. For example, AT5G10480 was predicted to map to miR159 in *P. cheesemanii* and not in *P. fastigiatum* (Figure 12). Two copies of AT5G10480 were present in the EST library of *P. cheesemanii* ("copy1" and "copy2", Figure 14), while only copy1 was present in *P. fastigiatum*. MiR159 only mapped to copy2, accounting for the apparent difference in targets between the species.

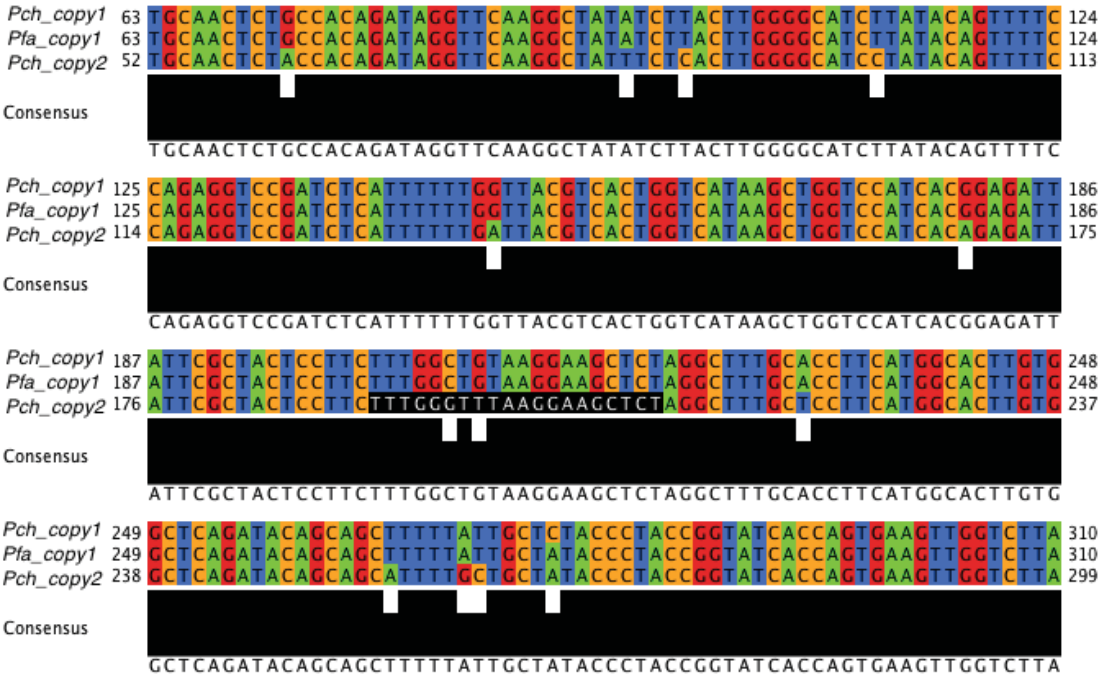

Figure 14: ClustalX alignment of AT5G10480 in *P. fastigiatum* and *P. cheesemanii*, an example of a different miRNA-target relationship between the species. In this case, the different relationship was due to the presence of different homeologs in the EST libraries ("copy1" and "copy2"). The binding site of miR159 in copy2 is highlighted in black.

In three cases, the difference between the *Pachycladon* species was due to a potential change in the miRNA sequence. In *P. fastigiatum*, a sequence variant of miR157 absent from *P. cheesemanii* mapped to AT1G75860. The other two cases were in *P. cheesemanii* between miR825 and two targets (AT2G19310 and AT5G07690). *P. cheesemanii* and *P. fastigiatum* have different predicted sequences for miR825.

All of the new miRNA-target relationships in *Pachycladon* cannot be verified without experimental work such as knockouts or luciferase assays, but these interactions that change due to miRNA sequence can be investigated further by sequencing the miRNA gene(s).

3.4.4 MiRNAs with interesting target profiles

MiR157, miR159, miR162, miR171, miR172, miR391, and miR825 had different target profiles between the two *Pachycladon* species. The functions of these targets are given in Table 3. For miR157 and miR825, some of these changes were due to changes in miRNA sequence. In the other cases, the changes were due to sequence differences in the regions surrounding the predicted miRNA binding site.

Table 3: Functions of targets predicted for only one *Pachycladon* species due to sequence difference between the species

miRNA	Target	Description
P. cheesemanii targets		
miR157	AT5G12950	Protein with catalytic activity containing putative glycosyl hydrolase of unknown function (DUF1680)
	AT1G75860	unknown protein
miR159	AT5G10480	Protein tyrosine phosphatase-like involved in cell division and differentiation
miR162	AT3G63180	TIC-like (TKL)
miR172	AT4G27830	Encodes a beta-glucosidase that may be responsible for acyl-glucose-dependent anthocyanin glucosyltransferase activity in *Arabidopsis*
miR391	AT4G29900	ACA10, one of the type IIB calcium pump isoforms. encodes an autoinhibited Ca(2+)-ATPase
miR825	AT3G52610	unknown protein
	AT2G19310	HSP20-like chaperones superfamily protein
	AT5G07690	Encodes a putative transcription factor (MYB29).
P. fastigiatum targets		
miR171	AT5G19430	RING/U-box superfamily protein that functions in zinc ion binding;
miR172	AT1G31800	CYTOCHROME P450, Encodes a protein with β-ring carotenoid hydroxylase activity

Chapter Four: Verification of miRNAs

4.1 Introduction

As no genome sequences are available for *Pachycladon*, to validate the presence of candidate miRNAs, the genomes of *A. thaliana* and *A. lyrata* were used to design primers. The regions surrounding the pre-miRNAs were obtained for each species and aligned, and primers were designed in conserved regions between the two species. If these regions are conserved between *Arabidopsis* species, they are more likely to be conserved in *Pachycladon*, and so the primers are more likely to amplify the target region in *Pachycladon*. However, as *Pachycladon* is polyploid, there is the possibility the primers will amplify both homeologous regions.

4.2 Methods

4.2.1 MiRNA candidate selection

From Chapter Two, miR156abcdef, miR157abc, miR159b and miR159c, miR160abc, miR161.1, miR164c, miR165ab, miR166abcdefg, miR169defg and miR169hijklmn, miR171a, miR391, miR394ab, miR395bcf, miR396b, miR398bc, miR408, miR472, miR825, miR848, and miR852 were chosen as interesting miRNAs to sequence due to being differentially expressed between the two *Pachycladon* species. Of these, miR157, miR159, miR162, miR171a, miR391, and miR825 were also of interest due to species-specific target profiles (see Chapter 3), along with miR162 and miR172, resulting in a total of 22 miRNAs selected for experimental validation.

For miR157, miR159 and miR160, which all have three family members, all family members were selected for sequencing. MiR157 and miR160 had a potential sequence variant in *P. cheesemanii*, so all three loci were chosen. MiR159 is a well-conserved miRNA, but a potential sequence variant was identified in *P. fastigiatum*, so all three loci were chosen to compare sequence conservation.

For miR169 and miR395, in which many loci were present, one representative of each of the two sequence variants were chosen: miR169g and miR169h, and miR395b and miR395d. Two loci were also chosen from the large miR156 and miR166 families:

miR156a and miR156b, and miR166a and miR166g. MiR162a and miR162b were both chosen for sequencing, and for the remaining miRNAs, one locus was chosen.

4.2.2 Primer design

A. thaliana and *A. lyrata* sequences for each miRNA precursor and 2kb of surrounding sequences on each side of the precursor were obtained from PlantGDB (83) and JGI (84) respectively, and aligned using a Geneious alignment. Geneious, which implements the Primer3 program (85), was used for primer design. An example of primer design is given in Figure 15.

Primers were manually designed for miR395d (both directions), and the reverse primers for miR165a and miR160c. No primers could be created for miR159a, miR171a, and miR395b as there was not enough complementarity between *A. thaliana* and *A. lyrata* or the GC content was too low. If possible, more than two primers were designed for each miRNA in case the primers failed to work using the *Pachycladon* DNA templates (Supplementary Table 4).

The predicted amplicons range in size from 300 – 2300 nt and average 1,200 nt.

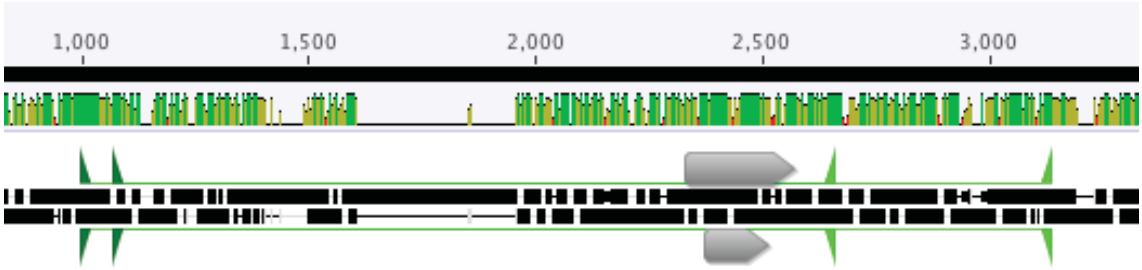

Figure 15: Example alignment for primer design. *A. thaliana* (top) and *A. lyrata* (bottom) sequences surrounding miR162a were aligned. The pre-miRNAs of miR162a are represented by grey arrows, and primers are shown as green triangles.

4.2.3 Experimental procedure

Amplification was attempted from *P. cheesemanii* and *P. fastigiatum,* using *A. thaliana* as a positive control. PCR optimisation included variations of annealing temperature and additions of PCR enhancers. The final amplification conditions were: 3 min of denaturation at 95°C; 40 cycles at 94°C for 1 min, 48°C for 1 min, and 72°C for 1 min 30 s, followed by a 5 min extension at 72°C.

A 20 µL reaction volume was used containing 1X PCR buffer (10mM Tris-HCl, 50mM KCl, 1.5 mM $MgCl_2$ pH8.3; Roche), 0.5mM $MgCl_2$, 0.25mM dNTPs, 1 M betaine, 10 pmol each of forward and reverse primers, 0.2 µL (1 U) Taq polymerase, and 1 µL DNA (5 – 50 ng). A negative control that contained no DNA was also included to test for contamination.

5 µL of 2X loading buffer was added to 5 µL of each PCR product and these were run on a 2% (w/v) agarose gel with a 1 kB+ ladder (Invitrogen®). The gel was visualised with SYBR® Safe gel stain (Invitrogen®) to confirm the size of the amplification products and to verify the presence of a unique PCR product (see Figure 16 for an example).

If a single product was present, SAP/EXO (2 µL shrimp alkaline phosphatase and 1 µL exonuclease; USB®) was used to clean the product for sequencing. If multiple products were present, bands were excised from the gels and purified using the Zymoclean™ Gel DNA Recovery Kit.

Templates were mixed with a single primer in the appropriate concentration (http://www.massey.ac.nz/massey/learning/departments/centres-research/genome/massey-genome-service-home.cfm) and submitted to the Massey Genome Service for sequencing. Electrophoretograms were edited using Geneious.

4.2.4 Hairpin prediction

Pachycladon sequences were added into the alignments of *A. thaliana* and *A. lyrata* that had been used for primer design, to confirm that the corresponding region in *Pachycladon* had been amplified. The secondary structures of the *Pachycladon* sequences were predicted by the web server MFOLD (21). *Pachycladon* pre-miRNA sequences were submitted to Genbank. Accession numbers for the sequences are given in Supplementary File 2.

4.3 Results and Discussion

4.3.1 PCR amplification

Out of the 27 primer sets, 22 and 17 amplified products in *P. cheesemanii* and *P. fastigiatum,* respectively (Table 4). Example gel photographs are shown in Figure 16. Of these products, the sequences of 15 *P. cheesemanii* and 13 *P. fastigiatum* miRNA genes were obtained. MiRNA sequences were not obtained from the remaining products due to amplification of products of the same size, or the amplification of incorrect regions.

Five primer pairs in *P. cheesemanii* and two in *P. fastigiatum* amplified both copies of the miRNA gene. In all but two of these, the PCR products were of identical sizes leading to inconclusive sequencing results. The PCR of miR852 in both *Pachycladon* species gave two products that differed in size by approximately 130 base pairs and were able to be separated by gel electrophoresis (Figure 16).

The primers for miR157c, miR161.1 and miR848 amplified the incorrect regions due to non-specific binding. The PCR program used had lowered annealing temperatures to allow for sequence divergence between *Arabidopsis* and *Pachycladon.*

Lack of sequence conservation in the surrounding regions of certain miRNAs between *A. thaliana* and *A. lyrata* impeded optimal primer design and degenerate primers were unsuccessful. Of the three miRNAs that had one or both of the primers manually designed (miR160c, miR165a, and miR395d), only miR165a worked in *P. cheesemanii*. Sometimes primers were only successful in one of the species.

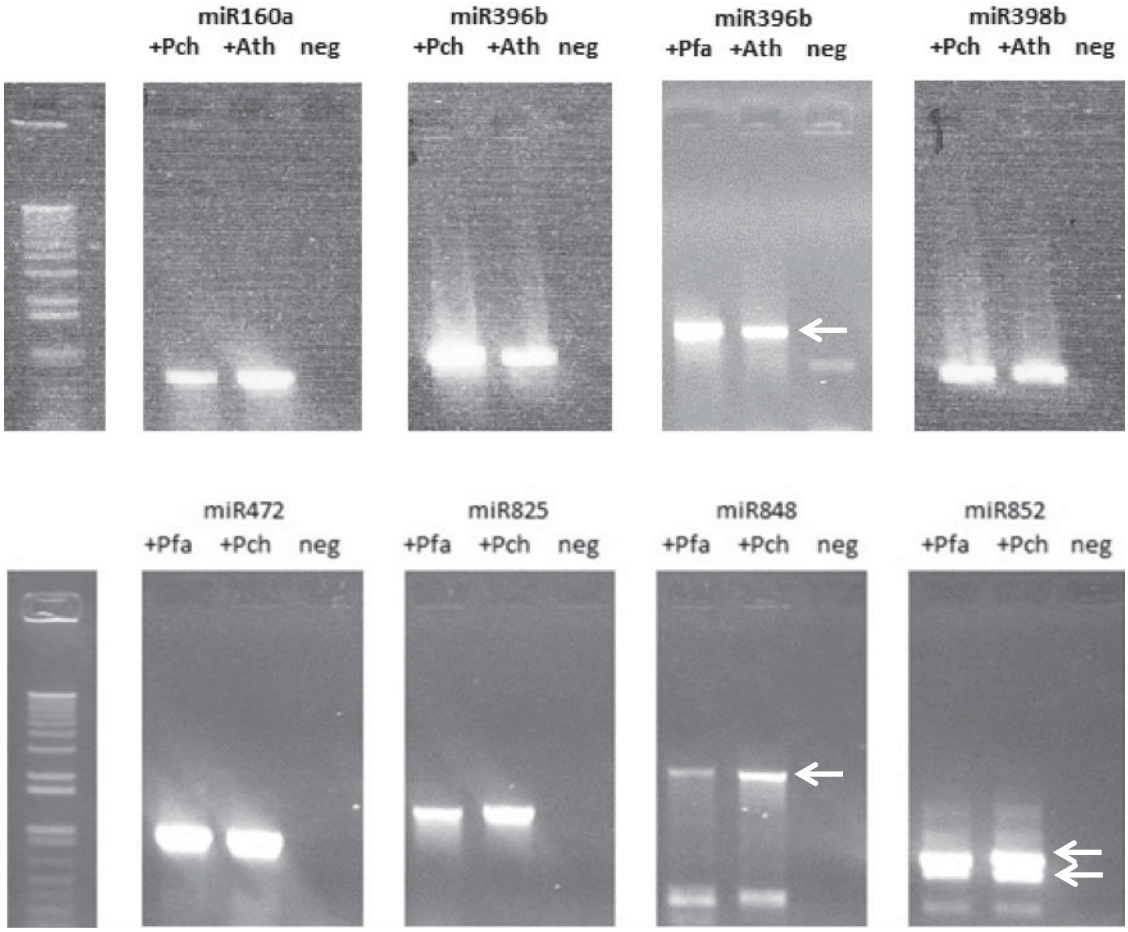

Figure 16: Amplification of miRNAs 160a, 396b, 398b, 472, 825, 848 and 852 in *Pachycladon*. +Pfa, *P. fastigiatum* DNA added; +Pch, *P. cheesemanii* DNA added; +Ath, *A. thaliana* DNA added; neg, no DNA added. For lanes with multiple bands, arrows indicate the bands that were excised.

Table 4: PCR and sequencing results of pre-miRNAs.

	Length of hypothetical (*Arabidopsis*) or real (*Pachycladon*) PCR product (bp)				Pre-miRNAs obtained?		Mature sequence same as *A. thaliana*?		Mature sequence same as miRNA sequencing?	
	Ath	**Aly**	**Pch**	**Pfa**	**Pch**	**Pfa**	**Pch**	**Pfa**	**Pch**	**Pfa**
157a	835	1323		1515		yes		yes		yes
157b	932	974	907		no [a]					
157c	1235	1296	1135	1306	no [b]	yes		yes		yes
159b	1417	1704	1415	1457	yes	yes	yes	yes	yes	yes
159c	1427	1184	551		no [a]					
160a	732	723	696		yes		yes		yes	
161.1	720	730	360	360	no [b]	no [b]				
162a	1487	1117	1192	1156	yes	yes	yes	yes	yes	yes
164c	1138	1052	1085		yes		yes		yes	
165a	413	411	495		yes		yes		yes	
166a	560	556	530	551	yes	yes	yes	yes	yes	yes
169g	1666	1616	1535	1641	yes	yes	yes	yes	yes	yes
169h	454	550	516	509	yes	yes	no	yes	n/a [c]	no
172	1760	1775	1729	1725	no [d]	no [d]				
391	836	784	805	835	yes	yes	no	yes	yes	yes
394a	374	366	374	414	yes	yes	yes	yes	yes	yes
396b	1026	956	942	958	yes	yes	yes	yes	yes	yes
398b	740	715	775	719	no [a]	yes		yes		yes
408	1760	1748	1600		yes		yes		yes	
472	787	796	828	825	yes	yes	no	no	yes	yes
825	1528	1604	1221	1225	yes	Partially [d]	no		yes	
848	1372	2246	1753	2102	no [b]	no [b]				
852	761	846	659 830	670 772	yes both copies	yes both copies	yes	n/a [e]	yes	n/a [c]

Note: Only primers that worked in at least one species are shown. Predicted sizes in *A. thaliana and A. lyrata* are given along with the *Pachycladon* products. If no length is given for one of the *Pachycladon* species, no products were obtained for that species using the primers tested. [a] the two homeologs were amplified together, resulting in an indistinguishable sequence. [b] the wrong region was amplified due to primers binding in a wrong position. [c] the miRNA was not expressed or is present in a very low copy number with no definite sequence. [d] poor sequence quality. [e] no mature sequence present in either pre-miRNA.

4.3.2 Validation of mature sequences by PCR

For all miRNAs except for miR169h, the sequences of the miRNAs identified in the small RNA sequencing matched the mature sequences of the corresponding PCR products (Table 4). The mature region for the *P. cheesemanii* copy of miR169h had two base changes and a single base pair deletion. There was no definite predicted sequence for miR169h in *P. cheesemanii* however, as only 5 reads of three variants from the small RNA sequencing matched to the *A. thaliana* miRNA sequence. For *P. fastigiatum*, the *A. thaliana* sequence was obtained, whereas a sequence with one base pair change was predicted. This miRNA was also present in low amounts (22 reads) so may represent the sequence of a different gene family member. MiR169n of *A. lyrata* has this sequence variant, so the corresponding *P. fastigiatum* sequence could potentially be the same.

For miR391, two sequence variants were predicted in *P. cheesemanii* by small RNA sequencing, one of which is present in *A. thaliana*. The sequence variant not present in *A. thaliana* was the one validated in *P. cheesemanii*. *P. fastigiatum* had only the *A. thaliana* sequence predicted, and this sequence was validated in *P. fastigiatum*.

The small RNA sequencing predicted miR472 to have a different sequence from *A. thaliana*, the same single nucleotide change in both species. This was confirmed by PCR.

Similarly, for miR825, different sequences from *A. thaliana* were predicted for *P. fastigiatum* and *P. cheesemanii*. *P. fastigiatum* was predicted to have the same mature sequence as *A. lyrata*, while *P. cheesemanii* was predicted to have a unique sequence. The sequence of *P. cheesemanii* was confirmed, but poor sequence quality prevented the full mature sequence and hairpin of *P. fastigiatum* from being obtained. The portion of the mature sequence that was verified is identical to the *A. lyrata* sequence, however (Figure 17)).

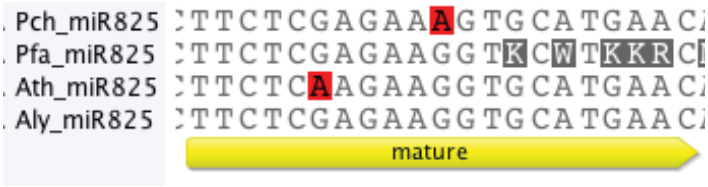

Figure 17: *Pachycladon* miR825 mature sequences obtained by validation sequencing. Also included are the *A. thaliana* and *A. lyrata* miR825 mature sequences.

MiR157 and miR160 had predicted sequence variants in *P. cheesemanii*, so all three loci for each miRNA were chosen to be sequenced. However, no new variants were observed. MiR157a failed to amplify, miR157b amplified two copies, and miR157c amplified the wrong region in *P. cheesemanii*. MiR160a amplified in *P. cheesemanii*, but the sequence was the same as *A. thaliana*, and miR160b and miR160c failed to work in both species. MiR159c had a sequence variant in *P. fastigiatum*, but the PCR failed to work.

4.3.3 Hairpin prediction

Hairpins were predicted for the 15 *P. cheesemanii* and 13 *P. fastigiatum* pre-miRNAs (Figure 18). Hairpins were only able to be predicted for one copy of miR852 in each species. The sequences of the *Pachycladon* pre-miRNAs are given in FASTA format in Supplementary File 3.

Of the 13 *P. fastigiatum* and 15 *P. cheesemanii* hairpins obtained, ten were obtained in both species (Table 4). For the miRNAs obtained in both species (miR159b, miR166a, miR169g, miR169h, miR391, miR394a, miR396b, miR472, and miR852), the hairpins were typically very similar in structure (Figure 18). Changes in miR169h affect the shape of these hairpins slightly.

Family members were present for miR157 in *P. fastigiatum,* the miR165/166 family in *P. cheesemanii*, and miR169 in both species. The hairpins of these family members are greatly different, varying in shape and sequence, which highlights the lack of selection pressure on the regions of the pre-miRNA that do not contain the mature sequence.

miR157a
PF

```
     U  G     C  C  UC   U         A-                A      A   UUU   G
5' GG UU AGAGG AU GA  GUG UGACAGAAG  UAGAGAGCACAGAUGA UGAGAU CAA   GGA C
3' CC AA UCUUU UA UU  CAC ACUGUCUUC  AUCUCUCGUGUUUCCU AUUCUA GUU  -UUU A
     U  G     -  U  UU   U         CG                C      C   UC    A
```

miR157c
PF

```
       UU     U------    UU   U  U  C     -           U      GACAUGCAAGUACAUUACAUAUAUAUCAUCACACCGCAUGUGAUGUUAAAAUAU
5'UAGGU  GAGAG       GAUG  GGU GU GA AGAAG AUAGAGAGCAC AAGGAU                                                        \
3'AUCCA  CUCUC       UUAU  CCA CA CU UCUUC UAUCUCUCGUG UUUCUA                                                        G
       UU     UUCUUCU    UU   C  -  A     A           U      AUCUCUACGUCCGAGAAAGAGAAAGAGAGGGAGAGAGAGAGAUUAAAGAAAUAU
```

miR159b
PC

```
     ACAA   A   AUUAG            GA       U       UU   ------    UG    -      G  A  C     -  C       C---       AAU    AA
5'AGA    AGG AGA     GAAGAGCUCUUU  AGUUCAA GAAGGGU  AGC      AGGG  AAGU AAAGCU CU AG UAUGG AU CCAUAAG    CUUAUCA   UCAA  \
3'UCU    UCC UCU     CUUCUCGAGGGA  UUAGGUU CUUCUCA  UCG      UCCC  UUCA UUUCGA GA UC AUACC UA GGUAUUU    GAAUAGU   AGUU  U
     CCC-   C   CUA--            AG       U       CU   GUAAUU    CA    C      G  C  U     A  -       UUUU       ---    AA
```

PF

```
     ACAA   A   -    G            GA       U       UU   ------    UC    -      G  A  C     -    CAU--           AAU    AA
5'AGA    AGG AGA AUUA GAAGAGCUCUUU  AGUUCAA GGAGGGU  AGC      AGGG  AAGU AAAGCU CU AG UAUGG AUCC     AAGCCUUAUCA   UCAA  \
5'UCU    UCC UCU UAGU CUUCUCGAGGGA  UUAGGUU CUUCUCA  UCG      UCCC  UUCA UUUCGA GA UC AUACC UAGG     UUUGGAAUAGU   AGUU  U
     CCC-   C   C    A            AG       U       CU   GUAAUU    CA    C      G  C  U     A    UAUUU           ---    AA
```

miR160a
PC

```
5'G----------      U         --------   C       CU         A   UU  CC     AG
             GUGUAU AUAUAUGUA        UGC UGGCUCC  GUAUGCCAU CGC  AG  CAUCG  \
             UACAUA UAUAUACAU        ACG ACCGAGG  UAUGCGGUA GUG  UC  GUAGC  U
3'UAAUCUCUUUA      -         ACUCCUAU   U       AG         G   CC  CA     UA
```

Figure 18: Predicted *Pachycladon* miRNA hairpin structures. Mature miRNA sequences are underlined.

miR162a

PC

```
         G    C    C        C   C  GGAAACAAAAUA
5'CGCUGGA GCAG GGUU AUCGAUCU UUC UG            \
3'GUGACCU CGUC CCAA UAGCUAGA AAG AC            A
         A    U    A        U   U  AAAAAAAAAAGA
```

PF

```
        G    CC    C        C   C  GGAAACAAAAUUU
5'GCUGGA GCA  GGUU AUCGAUCU UUC UG             A
3'UGACCU CGU  CCAA UAGCUAGA AAG AC             A
        A   CU    A        U   U  AAAACAAAAAAAA
```

miR164c

PC

```
     AU   U   UA    U  C     A  C              CACAAAUGAGAUCGGU
5'GUA  GGG GAG  ACAC UG UGGAG AG AGGGCACGUGCGAA               \
3'CAU  CCC-CUC  UGUG AC ACCUC UC UCUUGUGCACGCUU               A
     C-       UC    C  A     C  A              UUCAUACUAGUUGUGC
```

miR165a

PC

```
      A    A   UU      A            AUAUAUACGGACACA      C         G  A
5'GUUG GGGG AUG  GUCUGG UCGAGGAUAUUU               UAUAUA AUGUAUGUU AU C
3'CAAC CCCC UAC  CGGACC GGCUCCUAUGAG               AUAUGU UAUAUACGA UG A
      C    C   UU      A            AG-------------      A         G  A
```

miR166a

PC

```
         UU      CU   G   U    CUC      U  AUGUUGGAUCUUCUUCGAUCUAAUAU
5'GGGGAAUG  GUCUGG  CGA GAC CUGG   GCUCUA UC                          U
3'CCCCUUAC  CGGACC  GAU CUG GAUU   UGGGAU AG                          U
         UU      AG   G   C    -U-      U  UCUAGACUUUAGAACUCUAAGUUAAG
```

PF

```
       UUUCUCUUU   A        UU      CU   G   U    CU   UGUAUUCAUGUUGGAUCUUCUUCGAUCUAAUA
5'GGGGC         UUG GGGGAAUG  GUCUCC  CGA GAC CUGG  CCC                               U
3'CCUCG         AAC CCCCUUAC  CGGACC  GCU CUG GAUU  GGG                               U
       --UUGUU--   C        UU      AG   G   C    UU   AUUAGUCUAGACUUUAGAACUCCAAGUUAAGU
```

Figure 18 (continued)

miR169g

PC

```
     G-     G         GUCUC   UU   A  CAG        U            UA     U     U         A          AUUU      CGA   UG
5'GAU  AUGAU AUGAUGAGA     UAG  UGU CC   AGAAGUCU GCAU-GGAAAGAA  GAGAA GAGGU GAGCCAAGG UGACUUGCCG    UAGGAA   AUC  A
3'CUA  UGCUG UACUACUUU-----AUC  AUA-GG---UGUUCAGA UGUA CUUUUCUU  UUCUU CUUUA CUCGGUUCC GUUGAACGGC    GUGGUU---UAG  A
     AG     A               U-                 U    G        C-     C     U         A          CU--            UC
```

PF

```
            AAAAUAA        G-     --AAAA--     G-     G         GUCUC   UU   A   GA      U    G      U      U    UU         A          AUUU     AC     UG
5'AAGA-UCACA       CAA-GAGA  GUAGAG       AUGAU  AUGAU AUGAUGAGA     UAG  UGU CCA  GAGUCU GCAU GAAGAG AGAGAA GAGG  GAGCCAAGG UGACUUGCCG    UACCA  GAAUC  A
3'UUCU AGUGU       GUU CUUU  CGUUUC       UGCUA  UGCUG UACUACUUU-----AUC  AUA-GGU--CUCAGA UGUA CUUUUC UCUCUU CUUU  CUCGGUUCC GUUGAACGGC    GUGGU--UUUAG  A
      C     --AAA--   A    AA      AAACUAG     AG     A                U-               U    G      U      C    GU         A          -CU-            UC
```

miR169h

PC

```
5'CAACA  AA  U     U     GAGUAA     AUAAAA    AUG    C       --    CAAC    G   -     A   CU    U   --   UUUUA     U    AUCAAA    CGA
       AC  CA CCUCA AAAUA      AUCAU      GAAA   GUGA AUGAAGA  AUGA    UUGU UGG UAGUC AGG  GACU GCU  GCG     AACCA AUCU      GACU   U
       UG  GU GGGGU UUUAU      UAGUA      CUUU   CACU UACUUCU  UGCU    AACA ACU AUCGG UCC  CUGA CGG  UGC     UUGGU UGGA      CUGA   C
3'-----  CG  U     -     ------     ------    A--    U       CU    CUC-    G   U     C   U-    -   AC   UGA--     -    AUA---    UAG
```

PF

```
       -     G   C       --       C    G  -----      A    U    U     -  UUUUA     U    AUCAAA    CGA
5'AAAGA AAAUG UGA AUGAAGA  AUGAGAA UUGU UG     GUAGCC AGGA GACU GCCUG CG     AACCA AUCU      GACU   U
3'UUUCU UUUAC ACU UACUUCU  UGCUCUU AACA AC     UAUCGG UCCU CUGA CGGAC GC     UUGGU UGGA      CUGA   C
       C     -   U       CU       -    G  UUAAU      C    -    -     U  UGA--     -    AUA---    UAG
```

miR391

PC

```
       G     G   ACA   AU     U  C          G  CCC     UAAGUU
5'AUUUU AAACU CGA   AAG  UUGCU CG AGGAGAGAUA CG   UCGCC      U
3'UGGAA UUUGA GCU   UUC  AACGA GC UCCUCUCUAU GC   AGUGG      A
       A     A   ---   CU     U  A          G  ---     UGGCCA
```

PF

```
       G   ACA   AU     U  C          G  C       UAAGU
5'AAACU CGA   AAG  UUGCU CG AGGAGAGAUA CG CAUCGCC     \
3'UUUGA GCU   UUC  AACGA GC UCCUCUCUAU GC GUGGUGG     U
       A   ---   CU     U  A          G  A       CCAAU
```

Figure 18 (continued)

miR394a

PC

```
     U        G  A     UU     C  ACU   ---      UUC               A-        -    U   UG     A
5'UCA GAGGGUUU AC AAGAG  UCUUA CG   UCU   UUGGCA   UGUCCACCUCCUUCU  UACAUAUA UGCA GUG  UGUAU U
3'AGU CUCCCAAA UG UUCUU  GGGAU GU   AGA   AACCGU   AUAGGUGGAGGAAGA  GUGUGUGU GCGU UAU  AUGUG A
     U        A  C     U-     U  CU-   AGU      CAT               AG        U    U   GU     A
```

PF

```
     U        G  A     U    U   UA-   CU      UUC               AUA        U     U   UG     AUA
5'UCA GAGGGUUU AC AAGAG UUCU ACA   CUU  UUGGCA   UGUCCACCUCCUUCU   CAUAUAUA AUGCA GUG  UAUAU   U
3'AGU CUCCCAAA UG UUCUU GGGA UGU   GAG  AACCGU   AUAGGUGGAGGAAGA   GUGUGUGU UGCGU UAU  AUGUG   A
     U        A  C     U    U   CUA   AU      CAU               AG-        -     U   GU     AAU
```

miR396b

PC

```
5'UUCAGAA      GA     AA       ----     A         G         A    -CUUUUUUAUUUCCUUUUUUUAACAA
         GAAGGA  UGAUG  GAUCCUG    GUCAU UUUUUCCACA CUUUCUUGA CUUU                        \
         CUUCUU  ACUAC  UUGGGAC    CAGUA AAAAGGGUGU GAAAGAACU GAAA                        A
3'-------      AG     --       AUAA     C         G         C    AAUCGAUUCUAAAAAUCGAUAUAACU
```

PF

```
        GA     AA       ----     A                   A    CUUUUUUAUUUCCUUUUUUUACCAAAUCAAUAUAG
5'GAAGGA  UGAUG  GAUCCUG    GUCAU UUUUUCCACAGCUUUCUUGA CUUU                                  \
3'CUUCUU  ACUAC  UUGGGAC    CAGUA AAAAGGGUGUCGAAAGAACU GAAA                                  C
        AG     --       UUAA     C                   C    AAGAAACAAAGGUUUCACGAUUAAUCUCUAAAAAU
```

miR398b

PF

```
      UUU-  AA-     U---   U    A     U    A           UGCAAUC     C---     UG  GCC
5'GAUA    UG   GGUAG    GGA CUCG CAGGG UGAU UGAGAACACAUG       AACGG    UGUAA  AU   A
3'CUAU    AC   CCAUC    UCU GAGU GUCCC ACUG ACUCUUGUGUAC       UUGCU    ACAUU  UA   U
      UACU  CUG     UCUU   C    C     C    G           UU-----     CUCU     GU  AUG
```

Figure 18 (continued)

miR408
PC

```
       U   C    UGGU  U    U   ------       CAA    A      AUU   UUU  U       UAAA
5'AGAAG AGA AAAG    GA GAGA AGA      CAGGGAA   GCAG GCAUGG   GAG   AC AAAACAU    \
3'UUUUC UCU UUUC    CU CUCU UCU      GUCCCUU   CGUC CGUACC   CUC   UG UUUUGUG    C
       C   C    UUAU  -    U   CCCUCG       CUC    A      CAU   ---  -       UCAG
```

miR472
PC

```
      G-   -    --       A             C   CGCCUC     -         A  A           C  CUU--  AC  A    AACAA  AA-     UUGA     U  UG    UGUAA
5'UGGA  UCA UAUU  CUCAUCA AGAUGGAUUCGGG UUA      UGUAU GUAUGGGCG AG AGGCAAAAUCU AC     UC  GC GAUC     UA   UUUUG    AUAGA GC  GAUU     \
3'ACCU  AGU AUGA  GAGUGGU UCUACCUAAGUCC AAU      ACAUA CAUACCCGC UC UCCGUUUUAGA UG     AG  CG CUAG     AU   AAAGC    UGUUU UG  CUAG     G
      GA   C    AC       C             U   ACCCAU     C         C  A           A  AACUU  AA  C    AG---  GUG     UUG-     U  GU    UGUUG
```

PF

```
5'UUAUGC  CUAG    G-   AA--  -       A             C   CGCCUC     -        AA  A           C  CUUUCAC  A    AACAA  AA-     UUGA     U  UG    UUGUAA
        GC    UGGA  UCA    UU CUUAUCA AGAUGGAUUCGGG UUA      UGUAU GUAUGGGC  AG AGGCAAAAUCU AC       GC GAUC     UA   UUUUG    AUAGA GC  GAUU      \
        CG    ACCU  AGU    AA GAGUGGU UCUACCUAAGUCC AAU      ACAUA CAUACCCG  UC UCCGUUUUAGA UG       CG CUAG     AU   AAAGU    UGUUU UG  CUAG      G
3'U-----  AA--    GA   CAUG  C       C             U   ACCCAU     C        CC  A           A  AACUUAA  C    AG---  GUG     UUG-     U  GU    UGUUG
```

miR825
PC

```
          C     A    CAG     A           AA    U-    GAA
5'CAUCAACU GUUCA GCAC   CUCGA GAAGCGUAGCU  UUUA  UUUA   A
3'GUAGUUGA CAAGU CGUG   GAGCU CUUCGCAUCGA  AAGU  AGGU   A
          A     A    AAA     -           GA    CU    ACU
```

miR852
PC

```
                                UG-      AC    U   A  UC
5'UCAGAACUAAGGCGCUUAUCUUCUUUGAUAU   CAUGGA  AUGC UCU CU  \
3'AGUCUUGAUUCCGCGAAUAGAAGAGACUAUA   GUAUUU  UACG AGA GG  U
                              UAG      --    U   G  UC
```

PF

```
    UCAGAAC---                      UG-      AC    U   A  UC
5'GC          UAAGGCGCUUAUCUUCUUUGAUAU   CAUGGA  AUGC UCU CU  \
3'CG          AUUCCGCGAAUAGAAGAGACUAUA   GUAUUU  UACG AGA GG  U
    UAAUAUACCU                      UAG      --    U   G  UC
```

Figure 18 (continued). The incomplete mature miRNA of miR852 in *P. fastigiatum* is bolded.

The alignment of the pre-miRNA for miR472, which was sequenced from both species, is shown in Figure 19. This is representative of the typical scenario of strong conservation of pre-miRNA sequences between *Pachycladon* species, with less similarity to *Arabidopsis* but with no significant changes between the genera.

The exception to this generalisation was miR852, in which both copies of the gene were obtained in each species. Figure 20 shows the alignment of all the *Pachycladon* sequences for miR852 with the corresponding *Arabidopsis* sequences. A 120 bp deletion in one of the copies (which is referred to as copy two) in both species truncates the pre-miRNA, and completely removes the sequence of the mature miRNA. In copy one of *P. fastigiatum*, a 20 base pair deletion removed the end of the mature sequence. This is consistent with the miRNA sequencing data, as sequences for this miRNA were only obtained for *P. cheesemanii*. The deletion of the mature region in *P. fastigiatum* likely prevents expression of the miRNA. The hairpin was still predicted, however, but the deletion of part of the mature region introduces a loop not present in the hairpin of *P. cheesemanii* (Figure 17).

Pairwise percent identity scores were calculated for the miR852 sequences (Table 5). Orthologs (e.g. copy one in both species) were 95-98% similar, which are similar to values reported for orthologs in the EST libraries for *P. fastigiatum* and *P. cheesemanii* (53). The homeologs (i.e. copy one and two in one species) were more dissimilar, with scores between 70 – 74%. These scores are lower than the 85 – 95% identity reported for homeologs in the *Pachycladon* EST libraries (53). Copy 1 was the most similar to the *Arabidopsis* sequences (80% similar, compared with ~60% for copy 2) (Table 5).

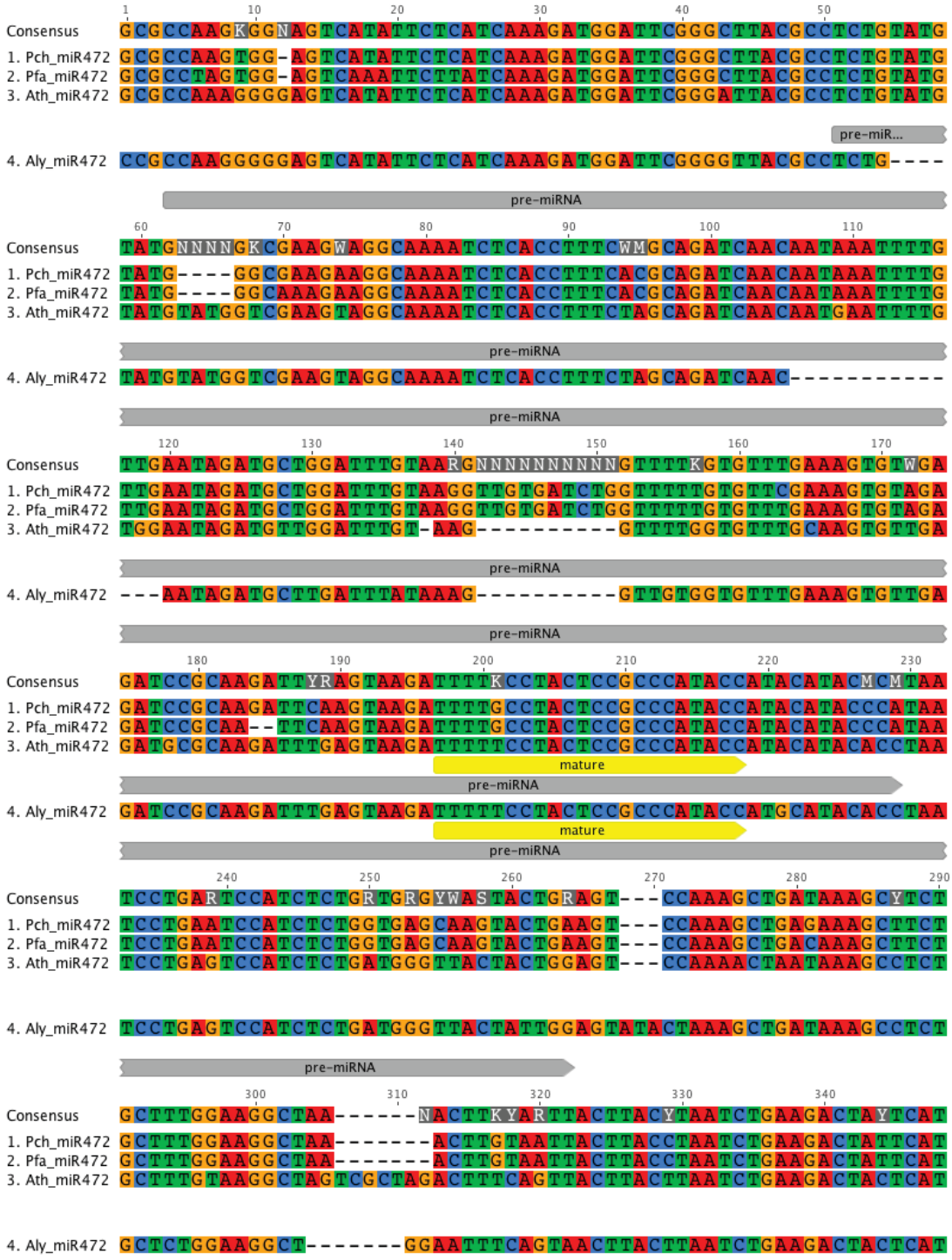

Figure 19: Alignment of miR472 in *Pachycladon* and *Arabidopsis*. *A. thaliana* ("Ath_miR472") and *A. lyrata* ("Aly_miR472") pre-miRNAs are represented by grey arrows, and mature miRNAs are represented by yellow arrows.

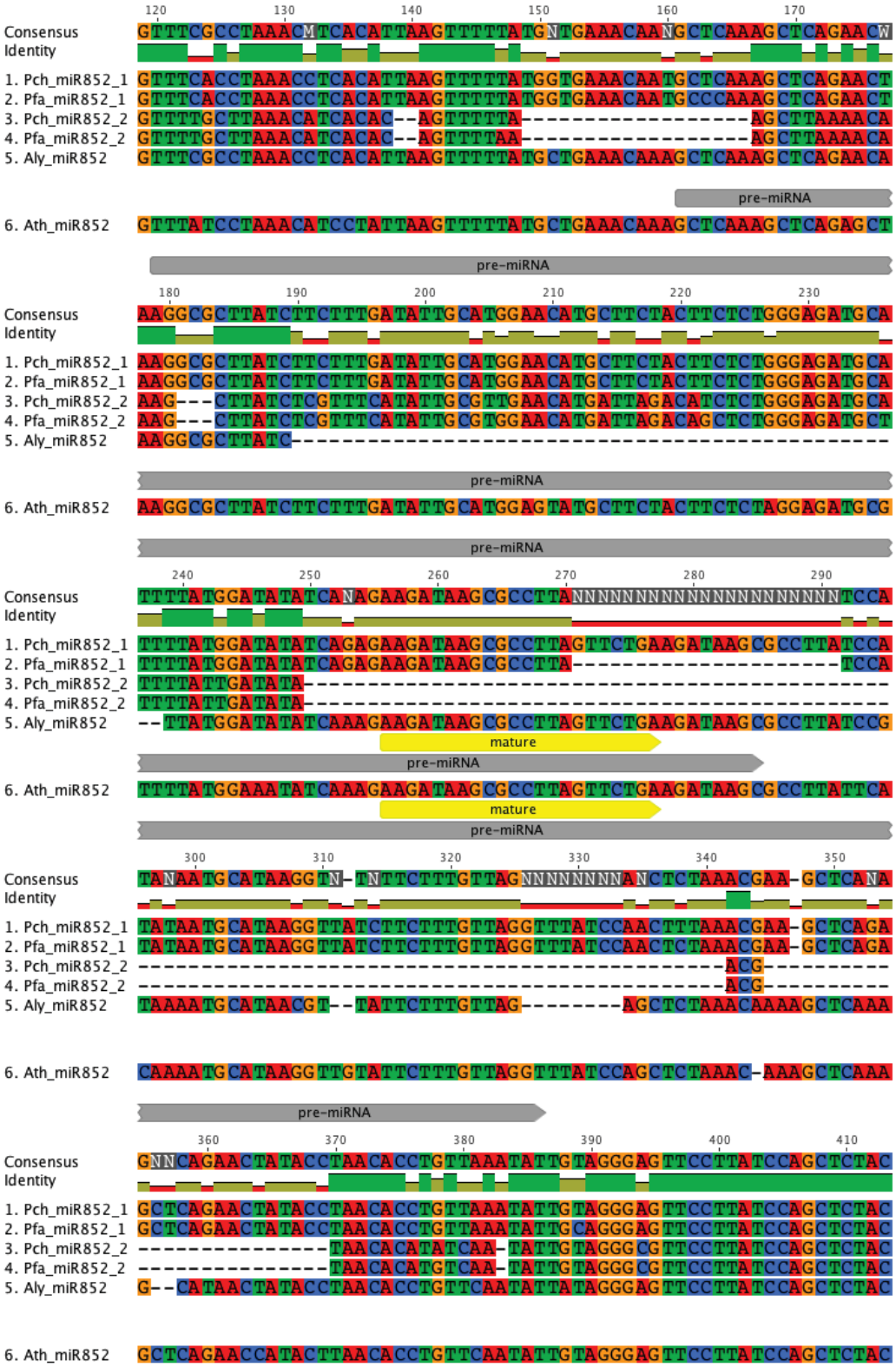

Figure 20: Alignment of miR852 pre-miRNA regions in *Pachycladon* and *Arabidopsis*.. Both copies of the *Pachycladon* miRNAs are shown (designated as miR852_1 (copy one) and miR852_2 (copy two)). Of particular interest are the large deletions in both *Pachycladon* miR852_2 copies that span the mature region.

Table 5: Pairwise percent identity for miR852 of *Pachycladon* and *Arabidopsis*.

	Pch-c1	Pfa-c1	Pch-c2	Pfa-c2	Aly	Ath
Pch-c1	100%					
Pfa-c1	95%	100%				
Pch-c2	70%	73%	100%			
Pfa-c2	70%	74%	98%	100%		
Aly	80%	79%	63%	64%	100%	
Ath	81%	80%	63%	64%	79%	100%

Note: Approximately 1000 bp was used for the comparisons. The two copies of each *Pachycladon* sequence are abbreviated as c1 and c2.

4.3.4 Future Work

With the exception of miR852, amplification of homeologs did not allow for straightforward sequencing and thus require cloning to obtain the two sequences. The amplicons that were too long for standard Sanger Sequencing and had poor sequence quality near the middle would require either increased elongation times during PCR or additional primers designed closer to the pre-miRNAs. Alternative primer combinations may allow products to be obtained for the PCRs that failed to amplify any product.

Chapter Five: mRNA-Seq Analysis

5.1 Introduction

The aim of a comparative mRNA expression analysis is to predict trait differences and ecological drivers of species divergence by comparing expression of genes between two species.

This has previously been done in *P. fastigiatum* and *P. enysii* using microarrays (45) and later confirming the results using mRNAseq (44). Differential gene expression analysis predicted genes of interest involved in defence against herbivores. It was predicted that since *P. enysii* expressed epithiospecifier protein (ESP, AT1G54040), it would produce nitriles, and because *P. fastigiatum* expressed epithiospecifier modifier 1 (ESM1, AT3G14210), it would produce isothiocyanates. These results were confirmed by analysing the glucosinolate hydrolysis products via gas chromatography – mass spectrometry (GC-MS) (45).

In this study, the mRNA analysis was done in combination with a comparative miRNA analysis (see chapters 2-4). As the mRNA and small RNAs were extracted at the same time, a comparison of the two methods and their results was possible. Firstly, differentially expressed genes were identified in *P. cheesemanii* and *P. fastigiatum*, and then enriched Gene Ontology (GO) terms were identified. Additionally, the expression levels of the miRNAs were compared with the expression of the potential targets present in the mRNA data.

5.2 Materials and Methods

5.2.1 mRNA sequencing and data pre-processing

Leaves were harvested from three plants each of greenhouse-grown *P. fastigiatum* and *P. cheesemanii* five months after germination. Total RNA for each sample was extracted using the RNeasy plant mini kit (Qiagen) and was prepared for sequencing using the mRNA-Seq sample prep kit (Illumina) according to the manufacturer's instruction. Samples were sequenced on a Genome Analyzer IIx (Illumina) for 75 cycles.

Data pre-processing, read mapping and quantification were done by N. Gruenheit (unpublished) as follows:

Primers were removed from the 75 bp reads using the FASTQ/A Clipper of the FASTX toolkit. Quality assessment of the reads was performed using DynamicTrim, and all reads less than 60 bp in length were removed.

Reference sequences were created for each *Pachycladon* species using the homologous sequences from the EST libraries of *P. fastigiatum* and *P. cheesemanii* (5,684 sequences). Reads were mapped to their conspecific reference using Bowtie. The table of counts that was generated from these mappings was used for the analyses in this chapter.

5.2.2 Differential expression analysis

The R package edgeR (57) was used to identify differentially expressed genes. A negative binomial model was fitted to the data to estimate common dispersion, and an exact test was used to determine p-values. The Benjamini-Hochberg (BH) method was used to adjust the p-values for multiple testing. Genes were classified as differentially expressed if there was at least a twofold difference in expression between the two species and an adjusted p-value of <0.05. The script used is given in Supplementary File 4.

5.2.3 GO Analysis

Two tests were used to determine enriched GO terms between the *Pachycladon* species: A gene enrichment analysis in R and the online tool AgriGO.

5.2.3.1 Gene set enrichment analysis using R

To identify enriched GO terms, a z-test was used, by comparing the average logFC of all genes in a GO term with the average logFC of all genes. To identify which genes belong to which GO terms, a binary GO matrix is used. The generic GO matrix provided had GO terms organised in columns and gene identifiers (*A. thaliana* locus identifiers) in rows, with each identifier occurring only once. A gene associated with a GO term was given a value of one, while a gene not associated with a GO term

was given a value of zero. The script used is given in Supplementary File 5.

As homeologous copies of genes are often present in *Pachycladon* expression data, unique identifiers are added to the ends of the gene names, but these are not present in a generic GO matrix. Thus for the purpose of the gene set enrichment analysis, an average logFC was calculated for both homeologous copies.

5.2.3.2 AgriGO

A gene singular enrichment analysis (SEA) was performed using AgriGO (86) as another method to determine enriched GO terms. A list of differentially expressed genes from each species was provided and enriched GO terms were calculated using a Fisher test with the 'TAIR9 *Arabidopsis* gene model' as a background, and Hochberg FDR as the multi-test adjustment method.

5.2.4. MiRNA and mRNA expression comparison

Any potential miRNA targets that were differentially expressed were identified by comparing a list of the targets to a list of the differentially expressed genes using the intersect() function of R. The logFC of these genes were compared with the logFC of their miRNAs via scatter plots to examine the relationship between miRNA and mRNA expression. MiRNA expression levels were taken from Section 2.3.3. Thus each miRNA was separated into unique sequences, and the sequence variants (potential homeologs) were considered separately.

False negatives (where the target prediction programs failed to predict a confirmed *A. thaliana* miRNA-target relationship using *A. thaliana* sequences) and false positives (where new miRNA-target relationships were predicted using *A. thaliana* sequences) were not included in the scatter plots.

5.3 Results

5.3.1 Differentially expressed genes

Out of the 5684 genes analysed, 1018 genes were identified as differentially expressed, 470 of which were up-regulated in *P. fastigiatum* and 548 were up-regulated in *P. cheesemanii* (Figure 21). The top 10 genes in *P. fastigiatum* and *P.*

cheesemanii are shown in Tables 6 and 7, respectively. All genes with absolute logFC >2 for *P. fastigiatum* (81 genes) and *P. cheesemanii* (110 genes) are shown in Supplementary Tables 5 and 6 respectively.

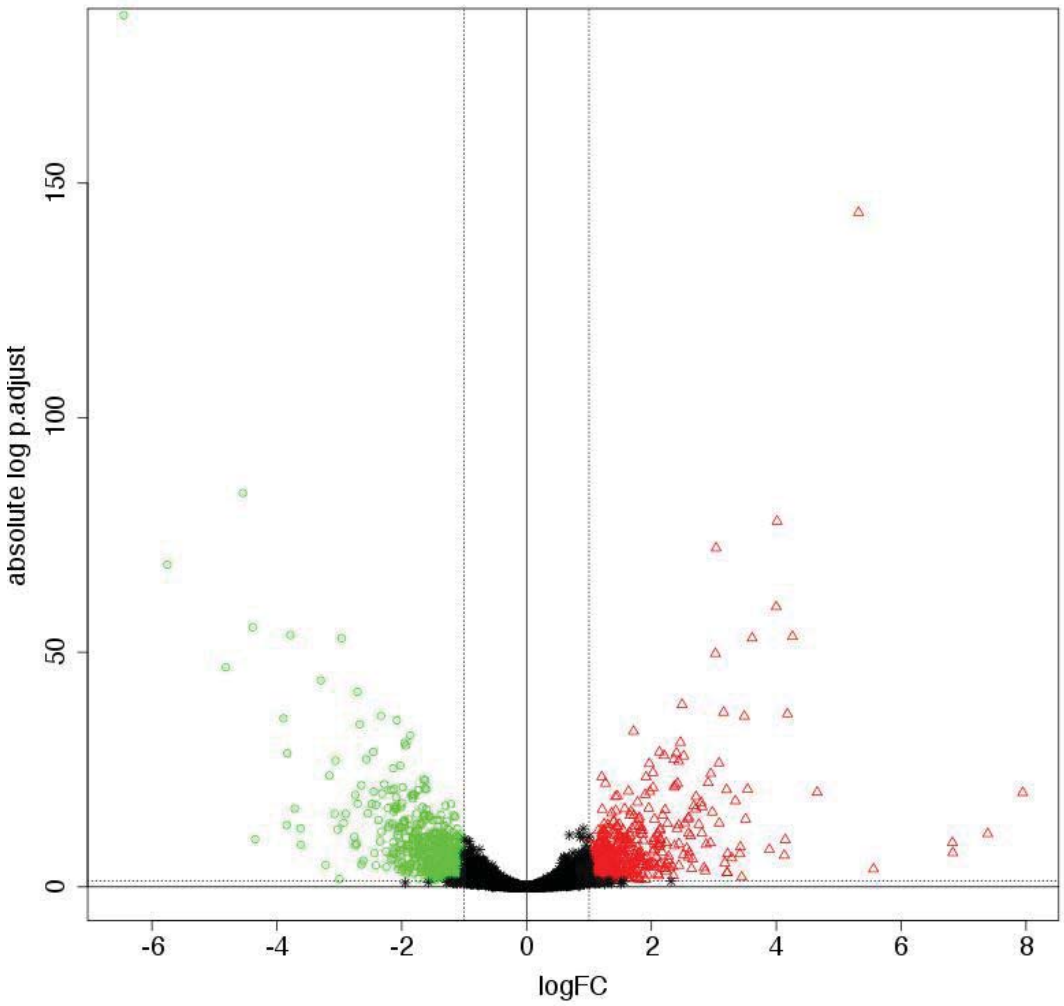

Figure 21: Log fold change vs log p-value of 5685 mRNAs expressed in both *Pachycladon* species. Circles represent genes up-regulated in *P. fastigiatum*, triangles represent genes up-regulated in *P. cheesemanii*, and stars represent genes not differentially expressed between both species. Genes were classified as up-regulated if there was at least a twofold difference in expression (beyond the -1 and +1 dashed lines) and with a maximum adjusted p-value of 0.05.

Table 6: Top 10 differentially expressed genes up-regulated in *P. fastigiatum*

	logFC	p.adjust	description
AT1G09310	-6.45	1.57E-186	Protein of unknown function, DUF538
AT2G29150	-4.54	1.12E-84	NAD(P)-binding Rossmann-fold superfamily
AT1G62510	-5.76	2.09E-69	Bifunctional inhibitor/lipid-transfer protein /seed storage 2S albumin superfamily protein
AT2G02120	-4.39	4.33E-56	Predicted to encode a PR (pathogenesis-related) protein. Belongs to the plant defensin (PDF) family
AT2G44130	-3.78	2.11E-54	Galactose oxidase/kelch repeat superfamily protein
AT2G38430	-2.97	9.35E-54	unknown protein
AT1G21550	-4.82	1.67E-47	Calcium-binding EF-hand family protein
AT2G19310	-3.30	1.00E-44	HSP20-like chaperones superfamily protein
AT2G47710	-2.71	2.51E-42	Adenine nucleotide alpha hydrolases-like superfamily protein
AT1G35720	-2.33	3.46E-37	Encodes a member of the annexin gene family

Table 7: Top 10 differentially expressed genes up-regulated in *P. cheesemanii*

	logFC	p.adjust	description
AT4G11640	5.32	1.76E-144	Serine racemase
AT2G31955	4.01	1.20E-78	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 2.
AT3G44310	3.04	6.26E-73	NITRILASE 1
AT1G45201	4.00	2.01E-60	Target of AtGRP7 regulation.
AT3G23000	4.26	4.36E-54	Encodes a serine/threonine protein kinase with similarities to CBL-interacting protein kinases, SNF1 and SOS2.
AT1G53430	3.61	9.28E-54	Leucine-rich repeat transmembrane protein kinase
AT5G25180	3.03	1.92E-50	putative cytochrome P450
AT3G56910	2.49	1.33E-39	plastid-specific 50S ribosomal protein 5 (PSRP5)
AT3G09680	3.16	6.56E-38	Ribosomal protein S12/S23 family protein
AT3G44990	4.18	1.42E-37	xyloglucan endo-transglycosylase

5.3.2 GO Analysis

5.3.2.1 Gene set enrichment analysis

A hundred and ninety-two GO terms were enriched, 89 of which were up in *P. fastigiatum* and 103 were up in *P. cheesemanii.* The ten GO terms with the greatest logFCs are shown for *P. cheesemanii* and *P. fastigiatum* in Tables 8 and 9 respectively.

Two glucosinolate metabolism GO terms were up-regulated in *P. fastigiatum*: GO:0019762 (glucosinolate catabolic process, logFC -4.39, pval = 1.15E-05) and GO:0019761 (glucosinolate biosynthetic process, logFC -3.83, pval = 0.00013).

Additionally, GO:0035195 (miRNA-mediated gene silencing) was up-regulated in *P. fastigiatum* (logFC -2.35, pval=0.022).

Table 8: Top 10 GO terms enriched in *P. cheesemanii* using gene set enrichment (sorted by p-value). Adj. pval = adjusted p-value.

	z	Adj. pval	Annotation
GO:0000154	15.76	6.19E-56	rRNA modification
GO:0042546	15.11	1.45E-51	cell wall biogenesis
GO:0006487	7.83	4.88E-15	protein amino acid N-linked glycosylation
GO:0009069	6.09	1.12E-09	serine family amino acid metabolic process
GO:0009561	5.43	5.65E-08	megagametogenesis
GO:0016311	5.36	8.06E-08	dephosphorylation
GO:0040008	5.34	8.88E-08	regulation of growth
GO:0009817	5.05	4.35E-07	defense response to fungus, incompatible interaction
GO:0048528	4.98	6.14E-07	post-embryonic root development
GO:0009664	4.93	8.11E-07	cellulose and pectin-containing cell wall organization and biogenesis

Table 9: Top 10 GO terms enriched in *P. fastigiatum* using gene set enrichment

	z	Adj .pval	Annotation
GO:0006414	-10.34	-4.71E-25	translational elongation
GO:0045333	-8.09	-5.98E-16	cellular respiration
GO:0009408	-8.03	-9.69E-16	response to heat
GO:0009960	-7.05	-1.75E-12	endosperm development
GO:0007049	-6.99	-2.75E-12	cell cycle
GO:0006952	-6.74	-1.61E-11	defense response
GO:0042542	-6.42	-1.37E-10	response to hydrogen peroxide
GO:0006114	-5.58	-2.45E-08	glycerol biosynthetic process
GO:0009585	-5.26	-1.49E-07	red, far-red light photo-transduction
GO:0006949	-5.15	-2.61E-07	syncytium formation

5.3.2.2 AgriGO

Fourty-eight GO terms were up in *P. cheesemanii*, and 39 were up in *P. fastigiatum*, 13 of which were common between the species. Tables 10 and 11 show the top 10 GO terms in *P. cheesemanii* and *P. fastigiatum.*

Table 10: Top 10 GO terms up in *P. cheesemanii* using AgriGO

GO term	Description	p-value	FDR
GO:0034641	cellular nitrogen compound metabolic process	1.20E-08	1.10E-05
GO:0044271	cellular nitrogen compound biosynthetic process	2.00E-06	0.00086
GO:0046483	heterocycle metabolic process	1.80E-05	0.0038
GO:0050896	response to stimulus	1.60E-05	0.0038
GO:0009611	response to wounding	4.30E-05	0.0068
GO:0010876	lipid localization	4.70E-05	0.0068
GO:0018130	heterocycle biosynthetic process	0.00011	0.013
GO:0009620	response to fungus	0.00014	0.015
GO:0009694	jasmonic acid metabolic process	0.0002	0.018
GO:0009628	response to abiotic stimulus	0.0002	0.018

Table 11: Top 10 GO terms up in *P. fastigiatum* using AgriGO

GO term	Description	p-value	FDR
GO:0050896	response to stimulus	2.40E-11	1.90E-08
GO:0006950	response to stress	1.40E-08	5.70E-06
GO:0042221	response to chemical stimulus	7.20E-08	1.90E-05
GO:0009628	response to abiotic stimulus	1.50E-07	3.10E-05
GO:0006979	response to oxidative stress	1.20E-06	0.00019
GO:0010876	lipid localization	2.20E-05	0.0029
GO:0006970	response to osmotic stress	5.80E-05	0.0066
GO:0006949	syncytium formation	6.60E-05	0.0066
GO:0009266	response to temperature stimulus	0.00013	0.011
GO:0009607	response to biotic stimulus	0.00014	0.011

5.3.2.3 GO terms shared between the R enrichment analysis and AgriGO results

Six GO terms were enriched in *P. cheesemanii* for both the edgeR and AgriGO analyses (Table 12), and 11 for *P. fastigiatum* (Table 13).

Table 12: GO terms enriched in *P. cheesemanii* for both the R and AgriGO analyses

X	z	adj.pval	anno
GO:0009693	4.24	2.13E-05	ethylene biosynthetic process
GO:0009611	3.65	2.46E-04	response to wounding
GO:0009058	2.21	2.32E-02	biosynthetic process
GO:0009695	2.19	2.41E-02	jasmonic acid biosynthetic process
GO:0015986	2.07	3.22E-02	ATP synthesis coupled proton transport
GO:0009414	1.87	4.94E-02	response to water deprivation

Table 13: GO terms enriched in *P. fastigiatum* for both the R and AgriGO analyses

X	z	adj.pval	anno
GO:0009651	-2.42	-1.81E-02	response to salt stress
GO:0009739	-2.75	-6.58E-03	response to gibberellin stimulus
GO:0051707	-3.37	-8.14E-04	response to other organism
GO:0006979	-4.10	-4.37E-05	response to oxidative stress
GO:0009631	-4.15	-3.40E-05	cold acclimation
GO:0006970	-4.90	-1.00E-06	response to osmotic stress
GO:0009828	-5.01	-5.58E-07	cellulose and pectin-containing cell wall loosening
GO:0009644	-5.01	-5.49E-07	response to high light intensity
GO:0006949	-5.15	-2.61E-07	syncytium formation
GO:0042542	-6.42	-1.37E-10	response to hydrogen peroxide
GO:0009408	-8.03	-9.69E-16	response to heat

5.3.3 Is miRNA up-regulation predictive of target gene down-regulation?

Having both miRNA and mRNA expression available offers an opportunity to test the predictive potential of differential miRNA profiles. To use differential expression patterns of miRNAs in a predictive way means to conclude that up-regulation of a miRNA leads to down-regulation of its targets. The logFC of all the miRNA targets that were differentially expressed were compared with the logFC of their corresponding miRNA.

Four out of eight miRNA-target relationships confirmed in *A. thaliana* that were also predicted in *Pachycladon* had inverse miRNA-target expression levels (Figure 22). Seven of 23 new miRNA-target relationships and five of nine losses of targets in *Pachycladon* also had inverse miRNA-target expression levels (Figures 23 and 24 respectively).

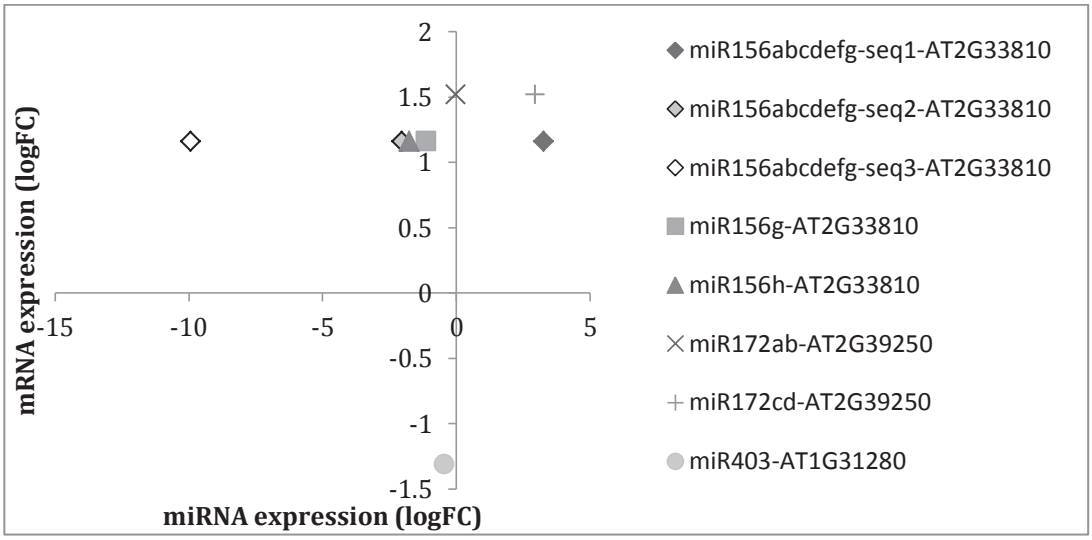

Figure 22: MiRNA expression versus target expression for *Pachycladon* miRNA-target relationships confirmed in *A. thaliana*.

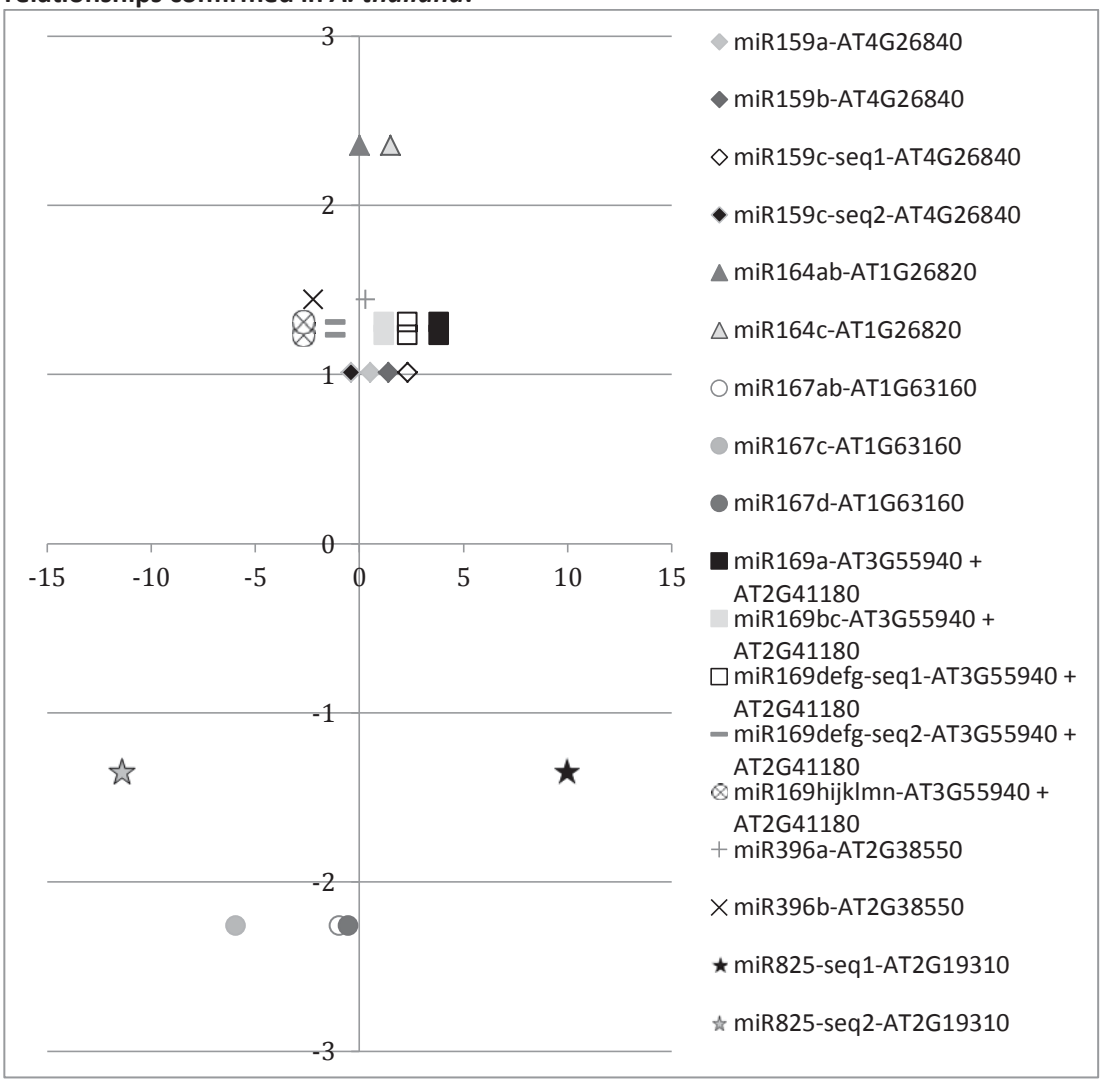

Figure 23: MiRNA expression vs target expression for new miRNA-target relationships in *Pachycladon*.

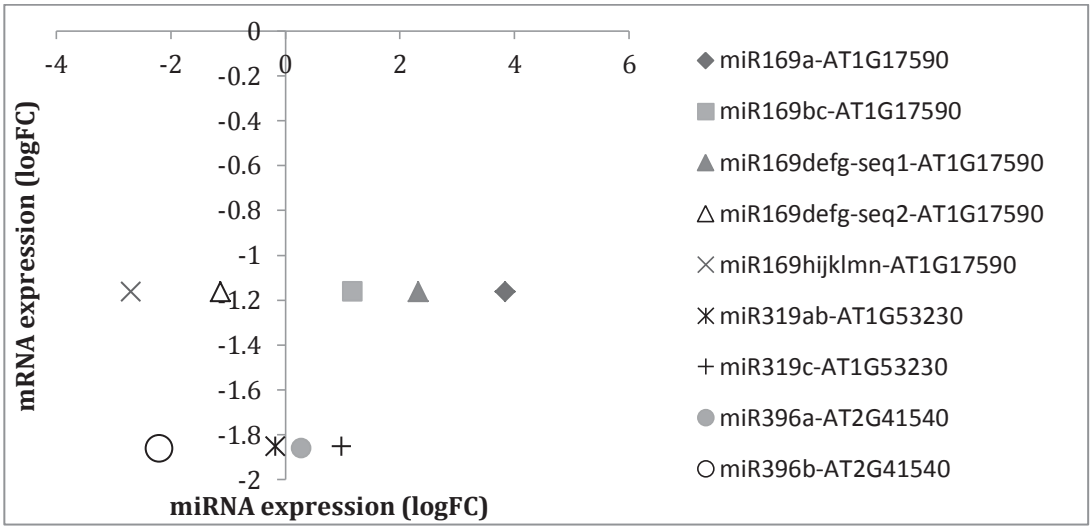

Figure 24: MiRNA expression vs target gene expression for potential loss of targets in *Pachycladon*.

5.4 Discussion

5.4.1 Approaches to identifying enriched GO terms

5.4.1.1 Averaging duplicate accession numbers

To accommodate for the presence of homeologs in the RNA-Seq data, the logFCs of the accession numbers that had both copies present were averaged so that a premade GO matrix of accession numbers and their associated GO terms could be used. The obvious solution to the problem of duplicated accession numbers is to generate a custom matrix for *Pachycladon* that has a unique name for each copy. Different homeologs were occasionally up-regulated in each species, for example AT1G05010 (logFCs = -1.66 and 1.90). If it would be worth spending time separating the homeologs and generating a custom matrix, as opposed to averaging the duplicates, remains to be seen.

5.4.1.2 AgriGO versus gene set enrichment

Two gene enrichment tests were performed to determine GO terms up-regulated in each species. The gene set enrichment test in R used a z-test to determine differentially expressed GO terms. In contrast, AgriGO compared lists of differentially expressed GO terms against an *A. thaliana* background reference, and determined enriched GO terms by comparing the percentage of GO terms present

in the *Pachycladon* set compared with the background. The former takes into account logFC of all genes annotated with a particular GO term, whereas the latter only includes genes with a significant logFC in the enrichment analysis.

A perhaps more meaningful AgriGO analysis would have been to use the corresponding *Pachycladon* EST library as a background, but the available EST libraries are unsuitable for two reasons: The first is that the size of the EST library is very small (containing 8000-12000 terms, compared with the ~37,000 terms for the *Arabidopsis* gene model TAIR9). Secondly, as the EST library was only generated from leaves, this introduces bias as leaf-specific GO terms would dominate the library.

The two methods identified different GO terms in each species; the common GO terms were very few (11 in *P. fastigiatum* and 6 in *P. cheesemanii*). However, this comparison using R's intersect function had a drawback - if two associated but different GO terms were present in the different analyses, the terms would not show up in the intersection. For example, in *P. fastigiatum* AgriGO identified GO:0019760 glucosinolate metabolic process as enriched while the R test identified GO:0019761 glucosinolate biosynthetic process as enriched, and so these were not present in the common list.

5.4.2 Differentially expressed genes and GO terms

In general, the types of enriched GO terms differed between the species. *P. fastigiatum* had more GO terms associated with defence against biotic and abiotic factors (20/39 (51%) for the AgriGO analysis, compared with 9/48 (19%) for *P. cheesemanii*) while *P. cheesemanii* had more biosynthetic-related GO terms enriched. The up-regulation of defence-related GO terms may have had a role in the evolution of the more specialised *P. fastigiatum*.

Six GO terms for *P. cheesemanii* and eleven for *P. fastigiatum* were enriched in both the AgriGO and the gene set enrichment test using R.

Eight of these GO terms in *P. fastigiatum* were involved in various stress responses;

all except one were responses to abiotic stressors. These included oxidative stress, osmotic stress, and high light intensity. Cold acclimation was also enriched for *P. fastigiatum.* This suggests *P. fastigiatum* may be more adapted to an extreme environment.

Also, *P. fastigiatum* had miRNA-mediated gene silencing (GO:0035195) enriched in the gene set enrichment test, which may also reflect its adaptation as a specialist plant.

Ethylene biosynthesis and jasmonic acid biosynthesis were amongst the up-regulated GO terms in *P. cheesemanii*. Ethylene and jasmonic acid are involved in the response to pathogens via the ethylene response factor ERF1 (87). ERF1 is induced by water deprivation (another GO term up-regulated in *P. cheesemanii*) (88).

5.4.2.1 Genes up-regulated in *P. fastigiatum*

In agreement with the enriched GO terms, a number of the top 10 genes up-regulated in *P. fastigiatum* are involved in various stress responses. AT2G29150 and AT1G62510 are up-regulated under cold stress in the wild crucifer *Thlaspi arvense* (89) and AT1G2550 and AT2G19310 are up-regulated under high light intensity and hydrogen peroxide feeding in *A. thaliana* (90). AT2G02120 encodes Pdf2.1, an antimicrobial plant defensin (91). AT2G47710 is a universal stress response protein family member.

Additionally, ESM1 was also up-regulated in *P. fastigiatum* (Supplementary Table 5), which is in accordance with previous studies (44, 45).

5.4.2.2 Genes up-regulated in *P. cheesemanii*

Genes up-regulated in *P. cheesemanii* are involved in processes such as auxin signalling (AT3G44310), molybdenum cofactor biosynthesis (AT2G31955), nitrogen assimilation, along with two ribosomal proteins and a serine racemase.

While not one of the top 10 genes, *P. cheesemanii* also had miR414 up-regulated, the miRNA found in Chapter Two by searching EST libraries for miRNA accession numbers (logFC = 1.46, p-value = 0.0043). MiR414 was found in both EST libraries,

but, interestingly, the hairpin for one of the *P. cheesemanii* copies most closely resembled the *A. thaliana* miR414 hairpin.

5.4.3 Comparing the expression of miRNAs and their targets

If there was differential regulation of a miRNA between two species, it would be expected that the targets had the opposite regulation. For example, if a miRNA was up-regulated in *P. cheesemanii*, it would have a positive logFC, and its targets would be expected to have a higher expression in *P. fastigiatum,* i.e. a negative logFC. Overall, no definite trend was observed in the scatter plots that plotted the miRNA logFCs against the logFCs of their predicted targets (Figures 22 – 24). However, it appears the inverse miRNA and mRNA expression occurs more often for potential gains and potential losses (Figures 22 and 24) than for the new target relationships (Figure 23).

However, the miRNA-target relationship of miR825 and AT2G19310 (which encodes a HSP20-like chaperones superfamily protein) presented an interesting case. This miRNA-target relationship was predicted for *P. cheesemanii* only, due to changes in the miR825 sequence between the species. While miR825 was present in similar levels in both species, AT2G19310 was up-regulated in *P. fastigiatum*. The reduced expression of this gene in *P. cheesemanii* may indicate a species-specific regulation of this gene by a change in miRNA sequence. This was the only target predicted to be different between the species that was differentially expressed in the mRNA-Seq data.

The lack of correlation between miRNA and target expression is likely a fault of the lack of replicates in the miRNA; the data is not reliable for expression-type analyses due to lack of statistical support. However, it is also possible – particularly for the newly identified targets – that not all the miRNA family members may target these genes. Additionally, there may be a time lag in miRNA up-regulation and target down-regulation that is not accounted for when analysing both profiles from a single time point.

Chapter 6.0 Discussion

6.1 MiRNA Analysis by Small RNA-Seq: Pros & Cons

6.1.1 Reliability of small RNA sequencing

In this study, unreplicated small RNA-Seq libraries were used to predict mature miRNA sequences in two species of the New Zealand alpine herb *Pachycladon*. For a selected set of miRNAs, the predicted mature sequences were validated by pre-miRNA amplification in all but one case (miR169h). However, different miRNA family members may have been expressed at the time of small RNA sampling than those selected for pre-miRNA amplification, which may explain the discrepancies between predicted and empirically determined mature sequences for miR169h.

Empirically validating predicted miRNAs by pre-miRNA amplification was a vital step as it helped explain the presence and absence of miRNAs in the small RNA-Seq data. For example, miR852 was only expressed in the *P. cheesemanii* small RNA library due to mutations in both copies of the *P. fastigiatum* miRNA gene which are likely to have caused its lack of expression in this species. For miRNA quantification, replicate measurements are preferable although the presence or absence of a miRNA may be deduced from unreplicated libraries (e.g. miR852).

6.1.2 Limitations of EST libraries in miRNA analyses

A limitation of not having genome sequences is that it is hard to infer loci numbers for miRNA families. MiRNAs may have been duplicated in *Pachycladon*, and these loci would not be detected by small RNA-Seq. Conversely, *Pachycladon* may have lost loci for certain miRNAs, particularly for larger miRNA families like miR169. The numbers of loci are interesting from an evolutionary perspective, particularly if polyploids are involved. Did the numbers of loci for certain miRNAs increase or decrease after hybridisation, and did both parental lineages have different miRNAs in different locations?

EST libraries are an alternative reference to use for small RNA-Seq. While it is possible to identify miRNA genes in EST libraries, their numbers are limited by the

size of the EST library and the tissues and developmental stages the ESTs are extracted from.

EST libraries are more useful for predicting miRNA-target relationships, but it is possible for the ESTs to be too short to cover the miRNA-binding region.

6.1.3 Using another species as a reference

MicroRNA analysis of non-model organisms has been greatly facilitated by next-generation sequencing-mediated small RNA-Seq. However, miRNA identification is often hampered by the lack of genomic resources in these organisms. The approach applied in this thesis for two species of *Pachycladon*, namely the combined analysis of small RNA-Seq and mRNA-Seq data as well as pre-miRNA amplification, was largely enabled by the presence of miRNA, EST and genome databases of the closely related model species *A. thaliana.* Mapping of filtered small RNA sequences to miRBase predicted 56 and 55 miRNAs in *P. cheesemanii* and *P. fastigiatum*, respectively, including as yet unknown sequence variants such as the *Pachycladon*-specific sequence of miR472, and unique sequences for miR825 and miR391 in *P. cheesemanii*. The majority of miRNAs had the same sequence as *A. thaliana*, as most of the miRNAs identified have the mature sequence conserved across many plant species. These are often miRNAs involved in plant development, and thus might be expected to be under purifying selection.

Allowing for mismatches between the *A. thaliana* and *Pachycladon* sequences allowed the identification of miRNAs that were less conserved. These miRNAs had roles in plant defence, and mature sequences often varied between *A. thaliana* and *A. lyrata*.

A. thaliana target information was used in the target identification step. Information of identified miRNA-target relationships for *A. thaliana* was available from TarBase, allowing the target profiles to be compared between *A. thaliana* and *Pachycladon*. The sequences of *A. thaliana* targets were compared with those from *Pachycladon* in cases where a confirmed miRNA-target relationship in *A. thaliana* failed to be predicted in *Pachycladon*.

The genome sequences of *A. thaliana* and *A. lyrata* were also used for designing primers for PCR. Once the miRNAs of interest were decided upon, 4 kb of genome sequence surrounding the miRNAs was obtained from online genome browsers and aligned using a Geneious alignment. Primers were designed in conserved regions, to increase the probability of amplification of this region in *Pachycladon*. The PCRs were largely a success, with 22/27 of the primers working in *P. cheesemanii* and 18/27 of the primers working in *P. fastigiatum*. Of these, 15 and 13 hairpins were obtained from the species respectively. *Pachycladon* has the added complication of being polyploid, so often two copies of each miRNA were obtained, but cloning can resolve the two copies if multiple products of the same size were present.

6.2 Biological Aspects

6.2.1 Differentially expressed miRNAs and target genes

A large proportion of the miRNAs with new target genes were also differentially expressed between the species. Twenty-two miRNAs in total were identified as differentially expressed once sequence variants were analysed separately. Of these, 11 were predicted to target new genes (16 miRNAs were predicted to have new targets; the remaining five were not differentially expressed).

Four miRNAs (miR167, miR169, miR319, and miR396) had potential loss of targets in *Pachycladon* due to sequence differences. Of these, two were differentially expressed (miR169 and miR396).

Seven miRNAs (miR162, miR157, miR159, miR171, miR172, miR391 and miR825) had differences in target genes between the two *Pachycladon* species due to sequence changes. Five of these miRNAs were also differentially expressed (miR157, miR159, miR171, miR391 and miR825)

For these miRNAs, the changes in targets and changes in expression level suggests a potential role in the diversification of the species. Considering 65 miRNAs in total were identified, the proportion of miRNAs with changes in target profiles that are also differentially expressed is high.

6.2.2 MiRNAs of Interest

When results were consolidated across the differential expression analysis of miRNAs, target prediction, experimental validation, and mRNA-Seq analysis, five miRNAs were selected as particularly interesting in the context of species diversification: miR825, miR852, miR396, miR391, and miR472.

For miR825, sequence differences between *P. fastigiatum*, *P. cheesemanii* and *A. thaliana* were confirmed experimentally. Two new targets were predicted for miR825 in *P. cheesemanii* due to the change in miRNA sequence, one of which (MYB29) has a role in the biosynthesis of a subset of glucosinolates, which is an area of interest in *Pachycladon* research. MYB29 is involved in the biosynthesis of short-chained, methionine-derived glucosinolates. The other predicted target of Pch-miR825 that was due to the new miRNA sequence encodes a member of the HSP20-like chaperones superfamily, and was up-regulated in *P. fastigiatum*. This scenario would be expected if only the *P. cheesemanii* miR825 targeted this mRNA. Little is known about miR825 in *A. thaliana*, except that it decreases under infection (78).

MiR852 is the only validated example of loss of miRNA in *Pachycladon*, as both copies of the miRNA were sequenced for each species. In both *Pachycladon* species, the copy less similar to the *Arabidopsis* homologs was missing the mature sequence due to a ~150 bp deletion. The copy most similar to *Arabidopsis* was also missing the mature region in *P. fastigiatum* due to a smaller deletion. No targets were predicted for this miRNA though, and its role is uncertain in *A. thaliana*, but it is expected to be still functional in *P. cheesemanii*.

MiR396b, one of two family members in the miR396 family, was up-regulated in *P. fastigiatum*. MiR396 causes narrow leaves when overexpressed, and *P. fastigiatum* has narrow leaves when compared with *P. cheesemanii*. A loss of target gene in *P. fastigiatum* was predicted, the AT4G25210 gene that encodes a DNA-binding storekeeper protein-related transcriptional regulator. This gene was only present in

the *P. fastigiatum* library, and both copies were present. Changes in the miRNA binding sites in both copies prevented the miRNA from mapping. A possible inactivation of one copy was also predicted due to insertions when compared with the other copy and the *A. thaliana* sequence. The processes this gene regulates is unknown, so the effect the loss of this gene may have in *P. fastigiatum* is unknown. Nine new targets were also predicted between the two species, but the targets that were not false positives (AT1G33970, AT2G38550, AT2G36400, AT1G10180, and AT1G10120) were mainly of unknown function.

MiR391 possessed an unusual case of sequence variants, with different miRNA sequences confirmed in the *Pachycladon* species. The *A. thaliana* sequence was predicted for both species, but a sequence variant was also predicted for *P. cheesemanii*. The sequence variant was confirmed for *P. cheesemanii*, and the *A. thaliana* sequence was confirmed in *P. fastigiatum*. With a pairwise percentage score of 90% it is possible a different copy of the miRNA was amplified in each species, but without both copies this is uncertain. Little is known about miR391, except that it is induced by hypoxia.

AT4G29900, which encodes AUTOINHIBITED CA(2+)-ATPASE 10 (ACA10), was predicted as a new target in *P. cheesemanii* and this relationship was also predicted in *A. thaliana*. The same copy of the target was present in the two species (98.8% similarity), but sequence changes were present in the miRNA binding site of *P. fastigiatum*. This would normally classify the relationship as a false positive, but as no targets have been confirmed in *A. thaliana* for miR391 it may actually be a loss of target in *P. fastigiatum*. ACA10 has roles in floral and leaf development as well as antibacterial defense.

A new sequence variant of miR472 was confirmed in both *Pachycladon* species. MiR472 was predicted to target AT4G02450 in *P. cheesemanii* (which was absent from the *P. fastigiatum* EST library). AT4G02450 encodes a member of the HSP20-like chaperones superfamily, a different protein to the one predicted to be a target for miR825.

6.2.3 Comparing information obtained from miRNA and mRNA-Seq analyses

Having both the miRNAs and mRNA from the same samples from the same time point allows the comparison of information and the identification of genes of interest that might have been missed otherwise. While in general, the expression levels of miRNAs failed to be inverse to mRNA expression of their targets, one interesting case of inverse expression was for the target relationship of miR825 and AT2G19310, a HSP20-like chaperone in *P. cheesemanii*. This relationship was only predicted in *P. cheesemanii*, due to a change in miRNA sequence, and the down-regulation of this target in *P. cheesemanii* correlates with the miRNA only targeting this gene in *P. cheesemanii*.

Moreover, a different HSP20-like chaperone protein (AT4G02450) was also predicted as a target for miR472. Six of these HSP20-like proteins were differentially expressed between the two *Pachycladon* species (although the one predicted to be targeted by miR472 was not one of them). Five HSP20-like chaperones were up-regulated in *P. fastigiatum* and one was up-regulated in *P. cheesemanii*. Along with their role in protein folding under heat stress, HSP20-like chaperones are involved in plant adaptation to various environmental stresses as well as development (92-94). The differential expression of both HSP20-like proteins and miRNAs that may regulate them (miR825 and miR472) suggests a potential role of these chaperones in the diversification of *Pachycladon* species requiring further investigation.

GO analysis of the mRNA-Seq data also showed an enrichment of many kinds of stress responses in *P. fastigiatum*; response to salt, organism, oxidative stress, osmotic stress, high light, hydrogen peroxide, and heat were identified by both AgriGO and gene set enrichment analyses, along with cold acclimation. The large number of stress response genes up-regulated in *P. fastigiatum* may be indicative of being adapted to a more extreme environment. In contrast, *P. cheesemanii* had fewer GO terms enriched, and these were predominantly processes that are involved in response to pathogens and wounding.

In comparison, most of the "interesting" miRNAs (for example, miR825 and miR852) have not been studied in detail, so their importance remains unknown. These are typically non-conserved miRNAs limited to *Arabidopsis* or *Brassicaceae*, many of which are predicted to be involved in defense responses, and may be up- or down-regulated under infection (for example, miR472 and miR825, respectively). Two conserved miRNAs up-regulated in *P. fastigiatum* are involved in response to oxidative stress (miR398, miR408), along with the *Arabidopsis*-specific miR391 which increases under hypoxia.

The other miRNAs that were differentially expressed were involved in developmental processes. The miRNAs up-regulated in *P. cheesemanii* covered a broader range of processes, including leaf, flower, and root development, while *P. fastigiatum* had miRNAs involved in leaf development and leaf serration.

The main similarity between the miRNA and mRNA analyses was that *P. fastigiatum* had more miRNAs/GO terms up-regulated than *P. cheesemanii*. These often had roles in stress responses. In general, the miRNA differential expression analysis identified processes with a broader role than the mRNA-Seq analysis; miRNAs with roles in development of flowers and leaves were up-regulated in both species. Thus, the small RNA-Seq data identified miRNA loci potentially underlying morphological differences between the species, such as leaf serration, while the mRNA-Seq data identified genes potentially underlying adaptations that are different between the species, such as to cold and light.

6.3 Conclusion

MiRNAs were identified and compared for two species of *Pachycladon*: *P. cheesemanii* and *P. fastigiatum*. 16 miRNAs were identified as differentially expressed between the species, with roles in a broad range of processes including development and stress response. Potential targets were predicted for some of the identified miRNAs, and species-specific target relationships between the species were among the new miRNA-target relationships identified.

Pre-miRNA sequences and hairpin structures were obtained for 13 *P. fastigiatum* and 15 *P. cheesemanii* miRNAs. Sequence changes in the mature miRNA sequences were confirmed by experimental validation. A difference in the presence of miR852 was observed between homologs and homeologs after both copies were able to be sequenced. Both copies of the miRNA possessed deletions in *P. fastigiatum*, and one copy remained intact in *P. cheesemanii*.

MRNA-Seq data, which was obtained at the same time as the miRNA data, were also analysed. Two gene ontology analyses identified enrichment of processes involved in various environmental stresses in *P. fastigiatum*, which correlates with the species being a specialist plant adapted to high altitudes.

Differences in miRNA sequence and expression were identified between the two *Pachycladon* species, but the effect these have on the phenotype of the plants requires further investigation.

6.3.1 Future Work

Future work could involve the cloning of the PCR products in which both copies of the miRNAs were amplified but were of similar length and hence could not be retrieved by gel extractions. Possessing both copies of more miRNAs may provide more information into the evolution of miRNAs in the plants. For example, miRNA-target relationships may be different for homeologous copies of miRNAs as well as their targets within a single *Pachycladon* species. Differences between *Pachycladon* species may reveal additional complex scenarios such as the one described for miR852 for other miRNAs.

Furthermore, predicted new miRNA-target relationships could be validated experimentally, in particular the potential regulation of both a transcription factor that stimulates glucosinolate synthesis and the HSP20-like chaperone by miR825 in *P. cheesemanii* but not *P. fastigiatum*. Experimental validation involves over- or underexpression of a candidate miRNA in combination with monitoring mRNA and/or protein expression of target genes by real-time quantitative PCR (qPCR) or Western blotting. A similar procedure has been implemented in *Arabidopsis*, with

overexpression of miR165 combined with qPCR of its target genes (16) to confirm the miRNA-target relationships.

To validate the expression differences between the two species, the small RNA sequencing could be repeated using replicates so that the expression patterns seen in this study can be statistically supported. Alternatively, differential miRNA expression suggested by the small RNA-Seq approach may be validated by quantifying mature miRNAs or their pri-miRNA precursors via real time PCR. This has been previously used to compare miRNA expression between *Arabidopsis* hybrids and polyploids (40).

References

1. Axtell MJ, Westholm JO, Lai EC (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol* 12(4):221.
2. Xie Z, *et al.* (2005) Expression of Arabidopsis MIRNA genes. *Plant Physiol* 138(4):2145-2154.
3. Zhou X, Ruan J, Wang G, Zhang W (2007) Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol* 3(3):e37.
4. Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol* 12(17):1484-1495.
5. Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115(2):209-216.
6. Guo L, Lu Z (2010) The Fate of miRNA* Strand through Evolutionary Analysis: Implication for Degradation As Merely Carrier Strand or Potential Regulatory Molecule? *PLoS One* 5(6):e11387.
7. Sunkar R, Kapoor A, Zhu JK (2006) Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18(8):2051-2065.
8. Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* 14(6):787-799.
9. Bari R, Datt Pant B, Stitt M, Scheible WR (2006) PHO2, microRNA399, and PHR1 define a phosphate-signaling pathway in plants. *Plant Physiol* 141(3):988-999.
10. Guan Q, Lu X, Zeng H, Zhang Y, Zhu J (2013) Heat stress induction of miR398 triggers a regulatory loop that is critical for thermotolerance in Arabidopsis. *Plant J* 74(5):840-851.
11. Navarro L, *et al.* (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* 312(5772):436-439.
12. Katiyar-Agarwal S, Jin H (2010) Role of small RNAs in host-microbe interactions. *Annu Rev Phytopathol* 48:225-246.
13. Han MH, Goud S, Song L, Fedoroff N (2004) The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proc Natl Acad Sci U S A* 101(4):1093-1098.

14. Vazquez F, Gasciolli V, Crete P, Vaucheret H (2004) The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Curr Biol* 14(4):346-351.
15. Boutet S*, et al.* (2003) Arabidopsis HEN1: a genetic link between endogenous miRNA controlling development and siRNA controlling transgene silencing and virus resistance. *Curr Biol* 13(10):843-848.
16. Zhou GK, Kubo M, Zhong R, Demura T, Ye ZH (2007) Overexpression of miR165 affects apical meristem formation, organ polarity establishment and vascular development in Arabidopsis. *Plant Cell Physiol* 48(3):391-404.
17. Wu G, *et al.* (2009) The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* 138(4):750-759.
18. Reyes JL, Chua NH (2007) ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. *Plant J* 49(4):592-606.
19. Liang C*, et al.* (2010) Identification of miRNA from Porphyra yezoensis by high-throughput sequencing and bioinformatics analysis. *PLoS One* 5(5):e10698.
20. Gordon A, Hannon G (2010) Fastx-toolkit. *FASTQ/A short-reads pre-processing tools (unpublished)* http://hannonlab.cshl.edu/fastx_toolkit.
21. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13):3406-3415.
22. Sun J, Zhou M, Mao Z, Li C (2012) Characterization and evolution of microRNA genes derived from repetitive elements and duplication events in plants. *PLoS One* 7(4):e34092.
23. Song C, Fang J, Li X, Liu H, Thomas Chao C (2009) Identification and characterization of 27 conserved microRNAs in citrus. *Planta* 230(4):671-685.
24. Zhang W, Luo Y, Gong X, Zeng W, Li S (2009) Computational identification of 48 potato microRNAs and their target. *Comput Biol Chem* 33(1):84-93.
25. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36(Database issue):D154-158.
26. Tempel S, Tahi F (2012) A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res* 40(11):e80.
27. Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 39(Web Server issue):W155-159.
28. Enright AJ*, et al.* (2003) MicroRNA targets in Drosophila. *Genome Biol* 5(1):R1.
29. Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol* 18(10):758-762.
30. Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 297(5589):2053-2056.
31. Li YF*, et al.* (2010) Transcriptome-wide identification of microRNA targets in rice. *Plant J* 62(5):742-759.
32. Li Y, Li C, Xia J, Jin Y (2011) Domestication of Transposable Elements into MicroRNA Genes in Plants. *PLoS One* 6(5):e19212.
33. Maher C, Stein L, Ware D (2006) Evolution of Arabidopsis microRNA families through duplication events. *Genome Res* 16(4):510-519.
34. Sieber P, Wellmer F, Gheyselinck J, Riechmann JL, Meyerowitz EM (2007) Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. *Development* 134(6):1051-1060.
35. Huang Y, Gu X (2011) A study of the evolution of human microRNAs by their apparent repression effectiveness on target genes. *PLoS One* 6(9):e25034.
36. Nozawa M, Miura S, Nei M (2012) Origins and evolution of microRNA genes in plant species. *Genome Biology and Evolution* 4(3):230-239.

37. Roux J, Gonzalez-Porta M, Robinson-Rechavi M (2012) Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. *Nucleic Acids Res* 40(13):5890-5900.
38. Kutter C, Schob H, Stadler M, Meins F, Jr., Si-Ammour A (2007) MicroRNA-mediated regulation of stomatal development in Arabidopsis. *Plant Cell* 19(8):2417-2429.
39. Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci* 16(2):108-116.
40. Ha M*, et al.* (2009) Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. *Proc Natl Acad Sci U S A* 106(42):17835-17840.
41. Joly S, Heenan PB, Lockhart PJ (2009) A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. *Mol Phylogenet Evol* 51(2):365-372.
42. Yogeeswaran K, Voelckel C, Joly S, Heenan P (2011) Pachycladon. *Wild Crop Relatives: Genomic and Breeding Resources*, ed Kole C (Springer Berlin Heidelberg), pp 227-249.
43. Joly S, Heenan PB, Lockhart PJ (2014) Species Radiation by Niche Shifts in New Zealand's Rockcresses (Pachycladon, Brassicaceae). *Syst Biol* 63(2):192-202.
44. Voelckel C, Gruenheit N, Biggs P, Deusch O, Lockhart P (2012) Chips and tags suggest plant-environment interactions differ for two alpine Pachycladon species. *BMC Genomics* 13:322.
45. Voelckel C*, et al.* (2008) Transcriptional and biochemical signatures of divergence in natural populations of two species of New Zealand alpine Pachycladon. *Mol Ecol* 17(21):4740-4753.
46. Becker M*, et al.* (2013) Hybridization may facilitate in situ survival of endemic species through periods of climate change. *Nature Climate Change* 3:1039–1043.
47. Heenan P, Mitchell A (2003) Phylogeny, biogeography and adaptive radiation of Pachycladon (Brassicaceae) in the mountains of South Island, New Zealand. *Journal of Biogeography* 30(11):1737-1749.
48. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485.
49. Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* 37(Database issue):D93-97.
50. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Res* 31(1):439-441.
51. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
52. Milne I*, et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 14(2):193-202.
53. Gruenheit N*, et al.* (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics* 13:92.
54. Lamesch P*, et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202-1210.
55. R Development Core Team (2011) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
56. Larkin MA*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.

57. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139-140.
58. Bergonzi S*, et al.* (2013) Mechanisms of age-dependent response to winter temperature in perennial flowering of Arabis alpina. *Science* 340(6136):1094-1097.
59. Allen RS*, et al.* (2007) Genetic analysis reveals functional redundancy and the major target genes of the Arabidopsis miR159 family. *Proc Natl Acad Sci U S A* 104(41):16371-16376.
60. Millar AA, Gubler F (2005) The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. *Plant Cell* 17(3):705-721.
61. Nikovics K*, et al.* (2006) The balance between the MIR164A and CUC2 genes controls leaf margin serration in Arabidopsis. *Plant Cell* 18(11):2929-2945.
62. Xing S, Salinas M, Hohmann S, Berndtgen R, Huijser P (2010) miR156-targeted and nontargeted SBP-box transcription factors act in concert to secure male fertility in Arabidopsis. *Plant Cell* 22(12):3935-3950.
63. Liu X*, et al.* (2010) The role of floral organs in carpels, an Arabidopsis loss-of-function mutation in MicroRNA160a, in organogenesis and the mechanism regulating its expression. *Plant J* 62(3):416-428.
64. Lurin C*, et al.* (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16(8):2089-2103.
65. Miyashima S*, et al.* (2013) A comprehensive expression analysis of the Arabidopsis MICRORNA165/6 gene family during embryogenesis reveals a conserved role in meristem specification and a non-cell-autonomous function. *Plant Cell Physiol* 54(3):375-384.
66. Liu Z, Jia L, Wang H, He Y (2011) HYL1 regulates the balance between adaxial and abaxial identity for leaf flattening via miRNA-mediated pathways. *J Exp Bot* 62(12):4367-4381.
67. Li WX*, et al.* (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *Plant Cell* 20(8):2238-2251.
68. Bolle C (2004) The role of GRAS proteins in plant signal transduction and development. *Planta* 218(5):683-692.
69. Hwang E-W, Shin S-J, Yu B-K, Byun M-O, Kwon H-B (2011) miR171 family members are involved in drought response in Solanum tuberosum. *Journal of Plant Biology* 54(1):43-48.
70. Rodriguez RE*, et al.* (2010) Control of cell proliferation in Arabidopsis thaliana by microRNA miR396. *Development* 137(1):103-112.
71. Liang G, He H, Li Y, Wang F, Yu D (2014) Molecular Mechanism of microRNA396 Mediating Pistil Development in Arabidopsis. *Plant Physiol* 164(1):249-258.
72. Song JB, Huang SQ, Dalmay T, Yang ZM (2012) Regulation of Leaf Morphology by MicroRNA394 and Its Target LEAF CURLING RESPONSIVENESS. *Plant Cell Physiol* 53(7):1283-1294.
73. Zhang H, Li L (2013) SQUAMOSA promoter binding protein-like7 regulated microRNA408 is required for vegetative development in Arabidopsis. *Plant J* 74(1):98-109.
74. Abdel-Ghany SE, Pilon M (2008) MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis. *J Biol Chem* 283(23):15932-15945.
75. Lu S, Sun YH, Chiang VL (2008) Stress-responsive microRNAs in Populus. *Plant J* 55(1):131-151.

76. Chen L, *et al.* (2012) Genome-wide profiling of novel and conserved Populus microRNAs involved in pathogen stress response by deep sequencing. *Planta* 235(5):873-883.
77. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* 20(24):3407-3425.
78. Khraiwesh B, Zhu JK, Zhu J (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim Biophys Acta* 1819(2):137-148.
79. Archak S, Nagaraju J (2007) Computational prediction of rice (Oryza sativa) miRNA targets. *Genomics, proteomics & bioinformatics* 5(3-4):196-206.
80. Guo X, *et al.* (2008) Selection and mutation on microRNA target sequences during rice evolution. *BMC Genomics* 9:454.
81. Vergoulis T, *et al.* (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 40(Database issue):D222-229.
82. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39(10):1278-1284.
83. Duvick J, *et al.* (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res* 36(Database issue):D959-965.
84. Grigoriev IV, *et al.* (2012) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* 40(Database issue):D26-32.
85. Untergasser A, *et al.* (2012) Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40(15):e115.
86. Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* 38(Web Server issue):W64-70.
87. Berrocal-Lobo M, Molina A, Solano R (2002) Constitutive expression of ETHYLENE-RESPONSE-FACTOR1 in Arabidopsis confers resistance to several necrotrophic fungi. *Plant J* 29(1):23-32.
88. Cheng MC, Liao PM, Kuo WW, Lin TP (2013) The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *Plant Physiol* 162(3):1566-1582.
89. Sharma N, Cram D, Huebert T, Zhou N, Parkin IA (2007) Exploiting the wild crucifer Thlaspi arvense to identify conserved and novel genes expressed during a plant's response to cold stress. *Plant Mol Biol* 63(2):171-184.
90. Yamamoto Y, *et al.* (2004) Global classification of transcriptional responses to light stress in Arabidopsis thaliana. *Endocytobiosis Cell Res* 15(2):438-452.
91. Siddique S, Wieczorek K, Szakasits D, Kreil DP, Bohlmann H (2011) The promoter of a plant defensin gene directs specific expression in nematode-induced syncytia in Arabidopsis roots. *Plant Physiol Biochem* 49(10):1100-1107.
92. Swindell WR, Huebner M, Weber AP (2007) Transcriptional profiling of Arabidopsis heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics* 8:125.
93. Kotak S, Vierling E, Baumlein H, von Koskull-Doring P (2007) A novel transcriptional cascade regulating expression of heat stress proteins during seed development of Arabidopsis. *Plant Cell* 19(1):182-195.
94. Wehmeyer N, Vierling E (2000) The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance. *Plant Physiol* 122(4):1099-1108.

Appendix

Supplementary Table 1: Data pre-processing of small RNA reads

	P. cheesemanii	*P. fastigiatum*
No. of reads sequenced	21,643,175	34,266,883
After quality filtering	14,845,042	23,009,789
Reads ≥ 17	12,269,297	12,253,277
After tRNA and rRNA filtering	5,912,660	7,551,693
Reads that map to *A. thaliana* miRNAs		
with zero mismatches	283,849	185,771
with one mismatch	292,060	194,298
with two mismatches	292,629	194,620

Supplementary Table 2: Sequences of miRNAs identified in *Pachycladon* species

miRNA	*P. cheesemanii* sequence	*P. fastigiatum* sequence
P. cheesemanii-specific miRNAs		
ath-miR169a	AGCCAAGGACAACTTGCCGA	
ath-miR169g-3p	TCGGCAAGTTGTCCTTGGCT	
ath-miR172b-5p	GCAGCACCATCGAGATTCAC	
ath-miR172cd	GAATCTTGATGATGCTGCAG	
ath-miR414	CATCTTCATCATCATCAC	
ath-miR5634	AGGGATTTTGTGAAATTAGGG	
ath-miR852	AAGATAAGCGCCTTAGTTCTG	
ath-miR858a	TTTCGTTGTCTGTTCGACCTT	
P. fastigiatum-specific miRNAs		
ath-miR161.1		TGAAAGTGACTAAACCGGGGT
ath-miR167c		TAAGCTGCCAGCATGATCTA
ath-miR2111b-3p		ATCCTCGGGATACAGATTACC
ath-miR399a		TGCCAACGGAGATTTGCCCTA
ath-miR5646		GTTCGAGGCACGTTGGGAGG
ath-miR5642ab		TTTCGCGCTTGTGCGGCTTT
ath-miR400		TATGAGAGTATTATAAGTCAC

Note: For each miRNA, the sequence of the read with the most abundance is given. Refer to Supplementary Table 3 for sequence variant analysis.

Sequences shaded grey have no more than ten reads that map to the miRNA in either species.

Supplementary Table 2 (continued)

miRNAs common to both species		
ath-miR156abcdef	TGACAGAAGAGAGTGAGCAC	TGACAGAAGAGAGTGAGCGC
ath-miR156g	CGACAGAAGAGAGTGAGCAC	
ath-miR156h	TGACAGAAGAAAGAGAGCAC	
ath-miR157abc	TTGACAGAAGATAGAGAGCAC	
ath-miR157d	TGACAGAAGATAGAGAGCAC	
ath-miR159a	TTTGGATTGAAGGGAGCTCTA	
ath-miR159b	TTTGGATTGAAGGGAGCTCTT	
ath-miR159c	TTTGGATTGAAGGGAGCTCCA	
ath-miR160abc	TGCCTGGCTCCCTGTATGCCA	
ath-miR162a	TCGATAAACCTCTGCATCCAG	
ath-miR164ab	TGGAGAAGCAGGGCACGTGCA	
ath-miR164c	TGGAGAAGCAGGGCACGTGCG	
ath-miR165ab	TCGGACCAGGCTTCATCCCCC	
ath-miR166abcdefg	TCGGACCAGGCTTCATTCCCC	
ath-miR167ab	TGAAGCTGCCAGCATGATCTA	
ath-miR167d	TGAAGCTGCCAGCATGATCTGG	
ath-miR168ab	TCGCTTGGTGCAGGTCGGGAA	
ath-miR169bc	CAGCCAAGGATGACTTGCCGG	
ath-miR169defg	TGAGCCAAGGATGACTTGCCG	
ath-miR169hijklmn	TAGCCAAGGATGACTTGCCTT	TAGCCAAAGATGACTTGCCTG
ath-miR170	TGATTGAGCCGTGTCAATATC	
ath-miR171a	TGATTGAGCCGCGCCAATATC	
ath-miR171bc	TTGAGCCGTGCCAATATCACG	
ath-miR172ab	AGAATCTTGATGATGCTGCAT	
ath-miR172e	GGAATCTTGATGATGCTGCAT	
ath-miR173-3p	TGATTCTCTGTGCAAGCAAAA	
ath-miR173-5p	TTCGCTTGCAGAGAGAAATCAC	
ath-miR2111ab-5p	TAATCTGCATCCTGAGGTTTA	
ath-miR319ab	TTGGACTGAAGGGAGCTCCCT	
ath-miR319c	TTGGACTGAAGGGAGCTCCTT	
ath-miR390ab	AAGCTCAGGAGGGATAGCGCC	
ath-miR391	TTCGCAGGAGAGATAGCGCCC	TTCGCAGGAGAGATAGCGCCA
ath-miR393ab	TCCAAAGGGATCGCATTGATCC	
ath-miR394ab	TTGGCATTCTGTCCACCTCC	
ath-miR395ade	CTGAAGTGTTTGGGGGAACTC	
ath-miR395bcf	CTGAAGTGTTTGGGGGGACTC	
ath-miR396a	TTCCACAGCTTTCTTGAACTG	
ath-miR396b	TTCCACAGCTTTCTTGAACTT	
ath-miR397a	TCATTGAGTGCAGCGTTGATG	

Note: for the miRNAs common to both species, if no sequence is given for *P. fastigiatum* it is the same as the *P. cheesemanii* sequence.

Supplementary Table 2 (continued)

ath-miR398a	TGTGTTCTCAGGTCACCCCTT	
ath-miR398bc	TGTGTTCTCAGGTCACCCCTG	
ath-miR399bc	TGCCAAAGGAGAGTTGCCCTG	
ath-miR403	TTAGATTCACGCACAAACTCG	
ath-miR408	ATGCACTGCCTCTTCCCTGGC	
ath-miR472	TTTTGCCTACTCCGCCCATACC	
ath-miR824	TAGACCATTTGTGAGAAGGGA	
ath-miR825	TTCTCGAGAAAGTGCATGAAC	TTCTCGAGAAGGTGCATGAA
ath-miR827	TTAGATGACCATCAACAAACG	
ath-miR840	ACACTGAAGGACCTAAACTAAC	
ath-miR848	TGACTTGCGACTGCCTAAGCT	TGAGATGGGACTGCCTAAGCTA

Supplementary Table 3: Sequences of potential homeologs in *Pachycladon*

miRNA	Sequence Identifier [a]	Sequence
miR156abcdef	1	TGACAGAAGAGAGTGAGCAC
	2	TGACAGAAGAGAGTGAGCAT
	3	TGACAGAAGAGAGTGAGCGC
miR157abc	1	TTGACAGAAGAGAGTGAGCAC
	2	TTGACAGAAGATAGAGAGCAC
miR159c	1	TTTGGATTGAAGGGAGCTCA
	2	TTTGGATTGAAGGGAGCTCCA
miR160abc	1	TGCCTGACTCCCTGTATGCCA
	2	TGCCTGGCTCCCTGTATGCCA
miR168ab	1	TCGCTTGGTGCAGATCGGGAC
	2	TCGCTTGGTGCAGGTCGGGAA
miR169defg	1	TGAGCCAAGGAAGACTTGCCG
	2	TGAGCCAAGGATGACTTGCCG
miR391	1	TTCGCAGGAGAGATAGCGCCA
	2	TTCGCAGGAGAGATAGCGCCC
miR395bcf	1	CTGAAGTGTTAGGGGGGACTC
	2	CTGAAGTGTTTGGGGGGACTC
miR824	1	TAGACCATTTGTGAGAAGGGA
	2	TAGACCATTTGTGAGAAGGGC
miR825	1	TTCTCGAGAAAGTGCATGAAC
	2	TTCTCGAGAAGGTGCATGAAC

Note: Sequence differences between homeologs are underlined. [a] Labeled as per Figure 6

Supplementary Table 4: Primers designed for PCR of miRNA genes.

Name	Sequence (5'-3')	Tm[a] (°C)
156aF1	TCTCCGTCAATCTTTGAACCCT	59.4
156aR1	CGCTTTCCCTAGATCGCACT	59.4
156aF2	CGACTGTTTCGCCATTAGAGTC	59.9
156bF1	CGCCAGACATCTGTTCCCTT	60.0
156bR2	AGTCCAACGACTCACTTACCC	59.9
156bF2	TGATGTTGGTCTTCCGCCAA	59.4
157aF1 [b]	GGCTTCAAGAAATCTCATCATCAT	57.4
157aR1 [b]	AACCATCAAACCTTATGGAATTCTT	57.2
157bF1 [b]	AGTGGCTTCAAGAAACTTCATCA	58.5
157bR1 [b]	ACCTTAATCTCCGCAATGAAACA	58.7
157bR2	GGGCTCTTTGAGATAGTTGAAACG	59.9
157cF1	TTTTGCCAGCCATTGCCAAA	59.8
157cR1 [b]	GCTTCTTGAAACACCTGCCA	59.0
157cF2 [b]	TGAGGACGTACGGTTGGTTG	60.0
159bF1 [b]	GAGGCATGTGACGTTTTTATGT	57.6
159bR1 [b]	TCATCCATGTTTATACACCTGCA	57.9
159bR2	CCATCATCCATGTTTATACACCTGC	60.0
159cF1 [b]	ACCCTAACCGTATCTCTCTCTAAA	58.0
159cR1 [b]	TGAAATCCGGCCGGTATACA	58.9
159cR2	GTTCAACAAACGGCAATGCC	59.1
160aF1	GTGGGTGTGTGTGAGGATGA	59.6
160aR1 [b]	GTGGTGTTCCAACTTCGTGC	60.0
160aF2 [b]	TTCATGCATGGACCAGGTGG	60.3
160bF1	GCCCACTAAAACTAAAGCAAACT	57.8
160bR1	TTGGGAGGGACAAATGTATGAA	57.3
160bF2	GAAGTGTTGCACGCTGTGTC	60.3
160bR2	TGCCTATGTTGCTTCACGACA	60.3
160cF1	CGATTCAAGCCAAGATCCACG	59.7
160cR1 [c]	ACTATCAATTYATAGTGTTCCGTG	60.3
161.1F1 [b]	CGCCGCTTTTCTCTTTCTC	57.0
161.1R1 [b]	TGATGCAATCTCAAACAAAGTACA	57.4
161.1R2	AGGTATAACTCAAAGTGGCAAGGT	60.0
162aF1	CGCAGAAGTCGATTAGGGCA	60.2
162aR1	GCTTCACCACTGATACATCGC	59.4
162aF2 [b]	TGATTGACGGCGCGTAAGAT	60.2
162aR2 [b]	AGGTCGAATTCTCACCAGAAACA	59.9
162bF1	ACGGTGAGTCATCAGATTCCT	58.5
162bR1	CCGATCTACTAATCTGCAAAGCTG	59.6

Note: If possible two forward and reverse primers were designed for each gene (referred to as F1, F2 and R1, R2 respectively).

[a] Tm = melting temperature of primer.

[b] Primer that worked in *P. cheesemanii* and/or *P. fastigiatum* (see Table 3).

[c] Primer was designed manually.

Supplementary Table 4 (continued)

162bF2	ACGGTGAGTCATCAGATTCCTC	59.6
162bR2	ACTTCACAAGATTCCCTCACCT	59.0
164cF1 [b]	ACGTGAATGAGCAAGCAGAAG	59.2
164cR1 [b]	CCCCCGCTTTCTAAGGCAA	60.0
164cF2	GATTCGGCGGCAGGAAGTAG	60.9
165aF1 [b]	AAAGCCCATCTTCGTCTCCG	60.1
165aR1 [bc]	AATCTACTCTTAAGAAGYCATGCA	60.3
165aF2	CGCCACTCATCATTCCCTCA	59.8
165aR2	AGCAAAAAGCATARACGGATCC	60.1
166aF1	GGACCAGAGACACCCATGTT	59.3
166aR1	AAATCCGATGAGCCATGGAA	57.3
166aF2 [b]	CGTGTGGGACCAGAGACAC	60.0
166aR2 [b]	CCATGCAAACAATCAATAACGCA	59.1
169gF1 [b]	CCTAAGAAACATACGACACGTGG	59.4
169gR1 [b]	TCAAGCCTTCCACGCATGAT	60.0
169gF2	CCTTCACAATGTGATGAGTCTCC	59.1
169hF1 [b]	CCAAGGCCATGAACACATGT	58.7
169hR1 [b]	TCTTGGCCACAGTGAAGCAA	60.1
169hF2	ATCAATTTCAGACCGCCCGT	60.0
172F1 [b]	CTCTCTCTCTCTCTCATCTGTGT	58.2
172R1 [b]	AGGTCATCCAAGGCATTGCT	59.7
172F2	TGAAGGTACGAGTTTCTAGTGTCT	58.9
391F1 [b]	AGTGACAATCCAGGAGCTGT	58.6
391R1 [b]	AGGCTCAAACCAAGGTCCTC	59.6
394aF1 [b]	ATCACCACCGTCCTTCTCTC	58.8
394aR1 [b]	ACCCTAGATCGAGGCTCTTCA	59.8
394aF2	GGGAGTGATGAGTGCGACAT	59.8
395dF1 [c]	AGAATGTCACCCATYCTATCTTC	60.9
395dR1 [c]	CAATATWGTTGGTAGTAGAGAGAG	60.3
396bF1	AGCAGGTTCACATATCTATCACTCA	59.4
396bR1 [b]	GCCTGAGGTGGAGTGGTTAG	59.7
396bF2 [b]	GGCATAGTATTTGGACTTGATGGA	58.6
398bF1	ATCGTGTGAACCGATGGTCC	60.1
398bR1	CCAGGAAGTTCACTCGTCCC	60.0
398bF2 [b]	TGCTAGTACAAAACTCGCCGT	60.0
398bR2 [b]	TCCGCCTACTCCGTGAAGTA	60.0
408F1 [b]	GGTGGGGACTTGTGATGCTT	60.3
408R1 [b]	ACACAACAAACAGCGCAGTC	59.9
408R2	ATGAACAGATTGGCAGCGGA	60.0
472F1 [b]	CGTGGAAGATCATGTCTGAGC	58.8
472R1 [b]	CCTCTGCACAAAACACGCAT	59.7
825F1	ATGCGTCACCAGGAGAAACA	59.6
825R1	ACAAGTTCACAGAGTACTTCCTGA	59.4

Supplementary Table 4 (continued)

825F2 [b]	GAAGCCAGTGGGGAGGTAAC	60.0
825R2 [b]	TCACAGAGTACTTCCTGAATTGGA	59.2
848F1 [b]	GGCACTCAGGCTTGAAACCT	60.5
848R1 [b]	TTCCCGTTGATTGAGGCCAA	59.9
852F1	TGCTTTCACCGCTCTCTCTG	60.0
852R1	TCACGCTTTGGTAGAGCTGG	60.0
852F2 [b]	GGCCGGAGAATATGGGAGTC	59.7
852R2 [b]	TGCCATGACGGACACCTAAC	60.0

Supplementary Table 5: Eighty-one genes up-regulated in *P. fastigiatum* (absolute logFC >2)

name	logFC	p.adjust	description
AT1G09310	-6.45416	1.57E-186	Protein of unknown function, DUF538
AT2G29150	-4.54474	1.12E-84	NAD(P)-binding Rossmann-fold superfamily protein
AT1G62510	-5.75526	2.09E-69	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
AT2G02120	-4.38801	4.33E-56	Predicted to encode a PR (pathogenesis-related) protein
AT2G44130	-3.78247	2.11E-54	Galactose oxidase/kelch repeat superfamily protein
AT2G38430	-2.96501	9.35E-54	unknown protein
AT1G21550	-4.81534	1.67E-47	Calcium-binding EF-hand family protein
AT2G19310	-3.29727	1.00E-44	HSP20-like chaperones superfamily protein
AT2G47710	-2.7091	2.51E-42	Adenine nucleotide alpha hydrolases-like superfamily protein
AT1G35720	-2.33137	3.46E-37	Encodes a member of the annexin gene family, a diverse, multigene family of calcium-dependent, membrane-binding proteins
AT1G23390	-3.89415	1.16E-36	Kelch repeat-containing F-box family protein
AT4G22100	-2.07845	2.89E-36	beta glucosidase 2 (BGLU3)
AT4G38690	-2.6734	2.26E-35	PLC-like phosphodiesterases superfamily protein
AT4G15440	-2.45383	1.73E-29	Encodes a hydroperoxide lyase. Also a member of the CYP74B cytochrome p450 family.
AT2G16360	-3.83096	3.62E-29	Ribosomal protein S25 family protein
AT3G61870	-2.56772	7.04E-28	unknown protein

AT2G42540	-3.0607	1.30E-27	A cold-regulated gene whose product is targeted to the chloroplast
AT2G36830	-2.0242	1.34E-26	Encodes a tonoplast intrinsic protein, which functions as water channel.
AT3G14210	-2.13758	5.36E-26	A semidominant QTL which has an epistatic effect on the Epithiospecifier gene. Represses nitrile formation and favors isothiocyanate production during glucosinolate hydrolysis
AT5G16010	-3.1576	1.82E-24	3-oxo-5-alpha-steroid 4-dehydrogenase family protein
AT3G19710	-2.27723	1.21E-22	Belongs to the branched-chain amino acid aminotransferase gene family. Involved in the methionine chain elongation pathway that leads to the ultimate biosynthesis of methionine-derived glucosinolates.
AT1G19000	-2.64692	2.31E-22	Homeodomain-like superfamily protein
AT1G49660	-2.12547	1.48E-21	Encodes a protein with carboxylesterase whose activity was tested using pNA.
AT1G28230	-2.1653	2.11E-21	Encodes a transporter that transports purines, cytokinins and other adenine derivatives
AT3G04880	-2.44593	5.09E-21	Encodes a novel protein involved in DNA repair from UV damage.
AT4G37310	-2.31136	1.28E-20	member of CYP81H
AT2G40610	-2.74288	2.39E-20	member of Alpha-Expansin Gene Family
AT1G20450	-2.09048	8.86E-20	Encodes a gene induced by low temperature and dehydration. Inhibits e.coli growth while overexpressed.
AT2G43820	-2.70625	1.88E-18	Induced by Salicylic acid, virus, fungus and bacteria. Involved in the tryptophan synthesis pathway
AT5G21105	-2.47876	2.20E-18	Plant L-ascorbate oxidase
AT4G22240	-2.41065	3.58E-18	Plastid-lipid associated protein PAP / fibrillin family protein
AT4G26530	-2.11843	8.78E-18	Aldolase superfamily protein
AT2G36320	-2.01237	1.02E-17	A20/AN1-like zinc finger family protein
AT1G63160	-2.22084	1.62E-17	replication factor C 2 (RFC2)
AT2G23120	-2.10335	1.99E-17	Late embryogenesis abundant protein, group 6
AT4G20830	-2.07925	1.99E-17	FAD-binding Berberine family protein
AT5G24780	-3.71435	2.01E-17	Encodes an acid phosphatase similar to soybean vegetative storage proteins. Gene expression is induced by wounding and jasmonic acid.
AT3G01790	-2.12645	9.05E-17	Ribosomal protein L13 family protein

AT1G18250	-2.54066	2.36E-16	encodes a thaumatin-like protein
AT1G52870	-2.89726	2.71E-16	Peroxisomal membrane 22 kDa (Mpv17/PMP22) family protein
AT3G21870	-3.0775	2.75E-16	cyclin p2
AT2G20560	-2.36966	6.58E-15	DNAJ heat shock family protein
AT3G32980	-2.9291	2.36E-14	Peroxidase superfamily protein
AT1G32470	-2.19576	3.02E-14	Single hybrid motif superfamily protein
AT1G12780	-2.15056	3.68E-14	Encodes a UDP-glucose epimerase that catalyzes the interconversion of the sugar nucleotides UDP-glucose UDP-galactose via a UDP-4-keto-hexose intermediate. Responsive to stress.
AT5G11190	-3.84262	6.72E-14	encodes a member of the ERF (ethylene response factor) subfamily B-6 of ERF/AP2 transcription factor family
AT4G38370	-2.01098	1.81E-13	Phosphoglycerate mutase family protein
AT4G31290	-2.21338	2.06E-13	ChaC-like family protein
AT4G04830	-2.04284	2.06E-13	methionine sulfoxide reductase B5 (MSRB5)
AT5G18650	-2.00444	2.36E-13	CHY-type/CTCHY-type/RING-type Zinc finger protein
AT5G04070	-3.62038	3.71E-13	NAD(P)-binding Rossmann-fold superfamily protein
AT5G20230	-3.02658	5.58E-13	Al-stress-induced gene
AT4G28400	-2.06192	1.12E-12	Protein phosphatase 2C family protein
AT3G13750	-2.32311	1.88E-12	beta-galactosidase, glycosyl hydrolase family 35
AT2G36830	-2.06247	9.60E-12	Encodes a tonoplast intrinsic protein, which functions as water channel
AT3G28220	-2.75998	2.32E-11	TRAF-like family protein
AT1G76650	-4.34765	7.45E-11	calmodulin-like 38 (CML38)
AT3G03900	-2.34499	1.99E-10	Provides activated sulfate for the sulfation of secondary metabolites, including the glucosinolates. Redundant with APK4.
AT2G15680	-2.13691	3.32E-10	Encodes a calmodulin-like protein.
AT1G10522	-2.01415	4.44E-10	Encodes PRIN2 (plastid redox insensitive 2).
AT1G26800	-2.71895	6.51E-10	RING/U-box superfamily protein
AT5G02090	-3.61218	1.03E-09	unknown protein
AT2G43550	-2.01011	1.12E-09	Encodes a defensin-like (DEFL) family protein.

AT1G59860	-2.74732	1.26E-09	HSP20-like chaperones superfamily protein
AT3G28900	-2.02165	1.56E-08	Ribosomal protein L34e superfamily protein
AT3G10020	-2.13072	2.28E-08	unknown protein
AT5G02450	-2.25198	2.56E-08	Ribosomal protein L36e family protein
AT3G24500	-2.07563	5.53E-08	One of three genes in *A. thaliana* encoding multiprotein bridging factor 1, a highly conserved transcriptional coactivator.
AT3G05570	-2.43783	6.33E-08	unknown protein
AT3G12580	-2.13084	9.95E-08	heat shock protein 70 (HSP70)
AT1G27290	-2.02289	1.00E-07	unknown protein
AT2G35520	-2.13844	1.07E-06	DEFENDER AGAINST CELL DEATH 2 (DAD2)
AT4G22490	-2.16861	1.27E-06	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
AT5G02830	-2.61641	2.41E-06	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G13470	-2.05869	9.64E-06	Protein of unknown function (DUF1262)
AT2G26150	-2.64226	1.71E-05	member of Heat Stress Transcription Factor (Hsf) family.
AT4G12480	-3.21616	2.12E-05	a putative lipid transfer protein, vernalization-responsive and cold-induced
AT2G29500	-2.41329	2.66E-05	HSP20-like chaperones superfamily protein
AT2G41760	-2.24733	5.99E-05	unknown protein
AT3G55240	-2.05763	0.000166	Overexpression leads to PEL (Pseudo-Etiolation in Light) phenotype.
AT2G38870	-2.9965	0.022744	Predicted to encode a PR (pathogenesis-related) peptide that belongs to the PR-6 proteinase inhibitor family

Supplementary Table 6: Genes up in *P. cheesemanii* (absolute logFC >2)

name	log.fc	p.value	description
AT4G11640	5.316915	1.76E-144	Serine racemase, which is a bifunctional PLP-dependent enzyme catalyzing racemization of serine and dehydration of serine to pyruvate in the same way as mammalian serine racemases. similar to mammalian serine racemases.
AT2G31955	4.012451	1.20E-78	COFACTOR OF NITRATE REDUCTASE AND XANTHINE DEHYDROGENASE 2.

AT3G44310	3.036303	6.26E-73	NIT1 catalyzes the terminal activation step in indole-acetic acid biosynthesis. Predominantly expressed isoform of nitrilase isoenzyme family.
AT1G45201	3.999271	2.01E-60	Target of AtGRP7 regulation.
AT3G23000	4.259074	4.36E-54	Encodes a serine/threonine protein kinase with similarities to CBL-interacting protein kinases, SNF1 and SOS2.
AT1G53430	3.612843	9.28E-54	Leucine-rich repeat transmembrane protein kinase
AT5G25180	3.025286	1.92E-50	putative cytochrome P450
AT3G56910	2.490051	1.33E-39	plastid-specific 50S ribosomal protein 5 (PSRP5)
AT3G09680	3.156958	6.56E-38	Ribosomal protein S12/S23 family protein
AT3G44990	4.178211	1.42E-37	xyloglucan endo-transglycosylase
AT1G73600	3.487197	4.41E-37	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT1G50110	2.464382	1.60E-31	D-aminoacid aminotransferase-like PLP-dependent enzymes superfamily protein
AT5G02240	2.12835	1.84E-29	Protein is tyrosine-phosphorylated and its phosphorylation state is modulated in response to ABA in *Arabidopsis* thaliana seeds.
AT4G16146	2.399652	3.11E-29	cAMP-regulated phosphoprotein 19-related protein
AT4G28390	2.202534	9.91E-29	Encodes a mitochondrial ADP/ATP carrier protein.
AT2G03240	2.5177	1.30E-28	EXS (ERD1/XPR1/SYG1) family protein
AT4G16590	2.358633	6.53E-28	encodes a gene similar to cellulose synthase
AT2G39920	2.43624	1.76E-27	HAD superfamily, subfamily IIIB acid phosphatase
AT4G25740	3.082253	4.15E-27	RNA binding Plectin/S10 domain-containing protein
AT5G04950	2.024402	4.57E-25	Encodes a nicotianamide synthase.
AT1G12080	2.949885	7.05E-25	Vacuolar calcium-binding protein-related
AT2G46940	2.908207	6.29E-23	unknown protein
AT2G01120	2.418289	8.69E-23	Origin Recognition Complex subunit 4. Involved in the initiation of DNA replication.
AT4G28590	2.383393	2.83E-22	FUNCTIONS IN: molecular_function unknown
AT5G02540	2.360638	5.49E-22	NAD(P)-binding Rossmann-fold superfamily protein
AT1G68600	2.0264	7.03E-22	Aluminium activated malate transporter family protein
AT2G01890	3.539962	1.61E-21	Encodes a purple acid phosphatase (PAP) belonging to the low molecular weight plant PAP group.
AT2G22310	3.203801	1.96E-21	Encodes a ubiquitin-specific protease.

AT1G78460	4.655361	6.90E-21	SOUL heme-binding family protein
AT3G14620	7.946483	8.59E-21	putative cytochrome P450
AT5G28020	2.714775	6.08E-20	Encodes cysteine synthase CysD2.
AT1G70780	3.34464	4.79E-19	unknown protein
AT1G12090	2.796123	1.10E-18	extensin-like protein (ELP)
AT5G07580	2.677803	5.69E-18	encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family.
AT1G17190	2.812665	8.38E-18	Encodes glutathione transferase belonging to the tau class of GSTs.
AT3G23050	2.710176	2.89E-17	Transcription regulator acting as repressor of auxin-inducible gene expression.
AT1G59700	2.218247	3.94E-17	Encodes glutathione transferase belonging to the tau class of GSTs. Naming convention according to Wagner et al. (2002).
AT3G10520	2.97501	1.18E-16	Encodes a class 2 non-symbiotic hemoglobin.
AT3G49120	2.180735	5.81E-16	Class III peroxidase Perx34. Expression activated by light. May play a role in generating H2O2 during defense response.
AT1G30760	2.597258	1.56E-15	FAD-binding Berberine family protein
AT3G14280	2.587157	3.57E-15	unknown protein
AT1G74940	2.048304	3.71E-15	Protein of unknown function (DUF581)
AT3G26840	3.507839	3.76E-15	Esterase/lipase/thioesterase family protein
AT1G05680	2.534944	2.38E-14	Encodes a UDP-glucosyltransferase, UGT74E2, that acts on IBA (indole-3-butyric acid) and affects auxin homeostasis.
AT5G06760	3.08553	2.61E-14	Encodes LEA4-5, a member of the Late Embryogenesis Abundant (LEA) proteins which typically accumulate in response to low water availability conditions imposed during development or by the environment
AT2G32170	2.258967	3.15E-14	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT1G54250	2.758065	4.98E-14	One of two highly similar proteins that can serve as non-catalytic subunits of Nuclear RNA polymerases II and V
AT2G41940	2.436052	7.09E-14	Encodes a zinc finger protein containing only a single zinc finger.
AT1G62780	2.410815	3.91E-13	unknown protein
AT2G28190	2.14102	4.31E-13	Encodes a chloroplastic copper/zinc superoxide dismutase CSD2 that can detoxify superoxide radicals. Its expression is affected by miR398-directed mRNA cleavage. Activation depends totally on CCS.
AT2G01850	2.136501	1.05E-12	EXGT-A3 has homology to xyloglucan

			endotransglucosylases/hydrolases (XTHs).
AT4G36640	2.809138	3.39E-12	Sec14p-like phosphatidylinositol transfer family protein
AT5G23250	2.595259	3.91E-12	Succinyl-CoA ligase, alpha subunit
AT2G31750	2.097837	4.48E-12	UDP-glucosyl transferase 74D1 (UGT74D1)
AT5G05300	7.38649	4.72E-12	unknown protein
AT4G02980	2.625989	1.30E-11	Auxin binding protein involved in cell elongation and cell division.
AT5G23750	2.040821	5.17E-11	Remorin family protein
AT3G03500	2.135186	7.23E-11	TatD related DNase
AT1G26210	4.143925	8.95E-11	AtSOFL1 acts redundantly with AtSOFL2 as positive regulator of cytokinin levels.
AT2G31880	2.14658	1.13E-10	Encodes a putative leucine rich repeat transmembrane protein that is expressed in response to Pseudomonas syringae.
AT2G30740	2.035091	1.93E-10	Protein kinase superfamily protein
AT4G11880	6.821523	3.62E-10	AGL12, AGL14, and AGL17 are all preferentially expressed in root tissues and therefore represent the only characterized MADS box genes expressed in roots.
AT3G07470	2.235512	4.08E-10	Protein of unknown function, DUF538
AT4G39190	2.945222	4.27E-10	unknown protein
AT1G11700	2.442467	6.32E-10	Protein of unknown function, DUF584
AT3G50440	2.051727	6.34E-10	Encodes a protein shown to have methyl jasmonate esterase activity in vitro.
AT1G47960	2.137507	7.89E-10	Plant cell wall (CWI) and vacuolar invertases (VI) play important roles in carbohydrate metabolism, stress responses and sugar signaling. This protein may inhibit their activity.
AT3G46320	2.875316	8.85E-10	Histone superfamily protein
AT2G40880	2.105228	1.01E-09	Encodes a protein with cysteine proteinase inhibitor activity. Overexpression increases tolerance to abiotic stressors (i.e.salt,osmitic, cold stress).
AT1G66180	2.366958	1.80E-09	The gene encodes a putative aspartyl protease (ASP). Its expression is induced in response to light and ascorbate.
AT3G02990	2.123644	3.07E-09	member of Heat Stress Transcription Factor (Hsf) family
AT3G53980	3.42487	3.29E-09	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein

AT4G14010	2.56687	6.32E-09	Member of a diversely expressed predicted peptide family showing sequence similarity to tobacco Rapid Alkalinization Factor (RALF), and is believed to play an essential role in the physiology of *Arabidopsis*.
AT1G58420	3.889357	1.05E-08	Uncharacterised conserved protein UCP031279
AT3G60530	2.0481	1.20E-08	Encodes a member of the GATA factor family of zinc finger transcription factors.
AT1G23120	2.016108	3.00E-08	Polyketide cyclase/dehydrase and lipid transport superfamily protein
AT1G15400	2.16981	4.01E-08	unknown protein
AT3G49120	2.606641	4.55E-08	Class III peroxidase Perx34. Expressed in roots, leaves and stems. Located in the cell wall. Involved in cell elongation. Expression activated by light. May play a role in generating H2O2 during defense response.
AT2G43580	6.826784	5.53E-08	Chitinase family protein
AT3G25882	3.220444	8.53E-08	encodes a kinase that physically interacts with NPR1/NIM1
AT1G73540	3.417729	9.28E-08	nudix hydrolase homolog 21 (NUDT21)
AT1G27300	2.512399	1.37E-07	unknown protein
AT1G35210	4.128684	1.90E-07	unknown protein
AT2G40316	2.275558	3.22E-07	FUNCTIONS IN: molecular_function unknown
AT1G50590	2.594044	4.28E-07	RmlC-like cupins superfamily protein
AT1G23410	2.077805	4.63E-07	Ribosomal protein S27a / Ubiquitin family protein
AT5G16160	3.288009	7.33E-07	unknown protein
AT4G38540	2.659722	1.16E-06	FAD/NAD(P)-binding oxidoreductase family protein
AT1G23730	2.269346	2.43E-06	beta carbonic anhydrase 3 (BCA3)
AT1G72170	2.148314	7.51E-06	Domain of unknown function (DUF543)
AT5G07010	3.175477	8.46E-06	Encodes a sulfotransferase that acts specifically on 11- and 12-hydroxyjasmonic acid.
AT1G26210	2.045304	1.26E-05	AtSOFL1 acts redundantly with AtSOFL2 as positive regulator of cytokinin levels.
AT2G27500	2.243015	1.89E-05	Glycosyl hydrolase superfamily protein
AT4G21470	2.12613	2.12E-05	Bifunctional enzyme that catalyzes hydrolysis of FMN to riboflavin, and phosphorylation of riboflavin to FMN.
AT1G25422	2.440415	2.75E-05	unknown protein
AT1G26820	2.337796	4.26E-05	Encodes ribonuclease RNS3.
AT3G26500	2.085414	5.18E-05	Encodes PIRL2, a member of the Plant Intracellular Ras-

			group-related LRRs (Leucine rich repeat proteins).
AT4G28940	2.292212	8.04E-05	Phosphorylase superfamily protein
AT3G23170	2.842559	8.38E-05	unknown protein
AT4G11280	2.642235	0.000126	encodes a member of the 1-aminocyclopropane-1-carboxylate (ACC) synthase (S-adenosyl-L-methionine methylthioadenosine-lyase, EC 4.4.1.14) gene family
AT2G41200	2.223025	0.000134	unknown protein
AT1G63390	5.55651	0.000157	FAD/NAD(P)-binding oxidoreductase family protein
AT1G62880	2.861482	0.000395	Cornichon family protein
AT1G60080	2.239952	0.000569	3'-5'-exoribonuclease family protein
AT4G25000	3.213529	0.001012	Predicted to be secreted protein based on signalP prediction. Involved in starch mobilization.
AT1G74458	3.214139	0.001021	unknown protein
AT1G32928	2.101805	0.001692	unknown protein
AT3G23440	2.35153	0.004011	EMBRYO SAC DEVELOPMENT ARREST 6 (EDA6)
AT1G22990	2.116599	0.004315	Heavy metal transport/detoxification superfamily protein
AT1G47820	3.445778	0.008538	unknown protein

Supplementary File 1: Script for identification of differentially expressed miRNAs

```
library(edgeR)

data <- read.delim("v0R_2.txt", header=TRUE, row.names=1)

dim(data)
data <- as.matrix(data)
lib.sizes <- colSums(data)
lib.sizes

group <- factor(c(1,2))
d <- DGEList(counts = data, group = group, lib.size = lib.sizes)
d

d <- estimateCommonDisp(d)
# No replicates, so no common dispersion. Common dispersion value instead taken #from
mRNA analysis
d$common.dispersion <- 0.02407237
d

########Cheesemanii vs Fasti############################
cn.common <- exactTest(d, pair = c("1", "2"))
```

```
dim(cn.common$table)
sum(cn.common$table$PValue<0.001)
cnres <- cn.common$table[cn.common$table$PValue <0.001,]
dim(cnres)
cnres$alogFC <- abs(cnres$logFC)
dim(cnres)
cnres2 <- cnres[cnres$alogFC > log2(2),]
dim(cnres2)

#UP IN SPECIES 1
cnres3 <- cnres2[cnres2$logFC < log2(2),]
dim(cnres3)
names <- row.names(cnres3)
length(names)
data.cn <- data[which(row.names(data)%in%names),c(1,2)]
dim(data.cn)
##names and original counts
data.cn
#print up in species 1
cnres3

###UP IN SPECIES 2
cnres4 <- cnres2[cnres2$logFC > log2(2),]
dim(cnres4)
names <- row.names(cnres4)
length(names)
data.cn <- data[which(row.names(data)%in%names),c(1,2)]
dim(data.cn)
 ##names and original counts
data.cn
#print up in species 2
cnres4
```

Supplementary File 2: Genbank accession numbers for *Pachycladon* pre-miRNA sequences

Pch-miR159b	KJ641580
Pch-miR160a	KJ641579
Pch-miR162a	KJ641565
Pch-miR164c	KJ641568
Pch-miR165a	KJ641567
Pch-miR166a	KJ641576
Pch-miR169g	KJ641563
Pch-miR169h	KJ641562
Pch-miR391	KJ641561
Pch-miR394a	KJ641558
Pch-miR396b	KJ641560
Pch-miR408	KJ641574
Pch-miR472	KJ641556
Pch-miR825	KJ641569
Pch-miR852-homeolog1	KJ641581
Pch-miR852-homeolog2	KJ641564
Pfa-miR157a	KJ641570
Pfa-miR157c	KJ641566
Pfa-miR159b	KJ641582
Pfa-miR162a	KJ641578
Pfa-miR166a	KJ641554
Pfa-miR169g	KJ641583
Pfa-miR169h	KJ641577
Pfa-miR391	KJ641571
Pfa-miR394a	KJ641559
Pfa-miR396b	KJ641557
Pfa-miR398b	KJ641573
Pfa-miR472	KJ641572
Pfa-miR852-homeolog1	KJ641575
Pfa-miR852-homeolog2	KJ641555

Supplementary File 3: Pre-miRNA sequences of *P. cheesemanii* and *P. fastigiatum* in FASTA format

>Pch-miR159b
AGAACAAAGGAAGAATTAGGAAGAGCTCTTTGAAGTTCAATGAAGGGTTTAGCAGGGTGAAGTAA
AGCTGCTAAGCTATGGATCCCATAAGCCTTATCAAATTCAAAATAATTGATGATAAGTTTTTTTATG
GATACCATATCTCAGGAGCTTTCACTTACCCCTTTAATGGCTTCACTCTTCTTTGGATTGAAGGGAGC
TCTTCATCTCTCCCTCCCTCT
>Pch-miR160a
GGTGTATTATATATGTATGCCTGGCTCCCTGTATGCCATACGCTTAGCCCATCGAGTATCGATGACC
TCCGTGGATGGCGTATGAGGAGCCATGCATATCCTCATACATATATATACATATTTCTCTAAT

>Pch-miR162a
CGCTGGAGGCAGCGGTTCATCGATCTCTTCCTGGGAAACAAAATAAAGAAAAAAAAAACATGAAT
AGATCGATAAACCTCTGCATCCAGTG
>Pch-miR164c
GTAATGGGTGAGTAACACTTGCTGGAGAAGCAGGGCACGTGCGAACACAAATGAGATCGGTCGGT
ACGTGTTGATCATACTTTCGCACGTGTTCTACTCCTCCAACACGTGTCTCTCCCCCTAC
>Pch-miR165a
GTTGAGGGGAATGTTGTCTGGATCGAGGATATTTATATATACGGACACATATATACATGTATGTTG
ATACAAGTGAGCATATATATGTATAGAGAGTATCCTCGGACCAGGCTTCATCCCCCCCAAC
>Pch-miR169g
GATGATGATGATGATGAGAGTCTCTAGTTTGTACCCAGAGAAGTCTTGCATGGAAAGAATAGAGA
ATGAGGTTGAGCCAAGGATGACTTGCCGATTTTACCAACGAATCTGAACUGATTTGGTGTCCGGCA
AGTTGACCTTGGCTCTATTTCCTTCTTCTTCTTTTCGATGTTAGACTTCTGGATATCTATTTCATCATA
GTCGTGAATC
>Pch-miR391
ATTTTGAAACTGCGAACAAAGATTTGCTTCGCAGGAGAGATAGCGCCCTCGCCTAAGTTTAACCGG
TGGTGACGGTATCTCTCCTACGTAGCAATCCTTATATATGCATCTTTATGCAGAGAGATGCATCTCG
AAGTTTAAAGGT
>Pch-miR394a
TCATGAGGGTTTGACAAAGAGTTTCTTACCGACTTCTTTGGCATTCTGTCCACCTCCTTCTATACATA
TATGCATGTGTGTGTATATAAGTGTATGTATTTGCGTTGTGTGTGGAAGAAGGAGGTGGATATACT
GCCAATAGAGATCTGTTAGGGTTTCTTCGTAAAACCCTCTTGA
>Pch-miR396b
TTCAGAAGAAGGAGATGATGAAGATCCTGGTCATATTTTTCCACAGCTTTCTTGAACTTTCTTTTTTA
TTTCCTTTTTTTAACAAATCAATATAGCTAAAAATCTAATTAGCACTTTGGAAACAAAGAAAAAGCTC
AAGAAAGGTGTGGGAAAACATGACAATACAGGGTTTCTCCATTGATTCAATTGTGCCATCAGATTC
TTC
>Pch-miR408
AGAAGTAGACAAAGTGGTGATGAGATAGACAGGGAACAAGCAGAGCATGGATTGAGTTTACTAA
AACATTAAACGACTGTGTTTTGTCTCTACCCATGCACTGCCTCTTCCCTGGCTCCCTCTTTCTCTCTAT
TCTTTCTCTCCTTTT
>Pch-miR472
TGGAGTCATATTCTCATCAAAGATGGATTCGGGCTTACGCCTCTGTATGTATGGGCGAAGAAGGCA
AAATCTCACCTTTCACGCAGATCAACAATAAATTTTGTTGAATAGATGCTGGATTTGTAAGGTTGTG
ATCTGGTTTTTGTGTTCGAAAGTGTAGAGATCCGCAAGATTCAAGTAAGATTTTGCCTACTCCGCCC
ATACCATACATACCCATAATCCTGAATCCATCTCTGGTGAGCAAGTACTGAAGTCCA
>Pch-miR825
CATCAACTCGTTCAAGCACCAGCTCGAAGAAGCGTAGCTAATTTATTTTAGAAAATCATGGATCTGA
AAGAGCTACGCTTCTCGAGAAAGTGCATGAACAAGTTGATG
>Pch-miR852
TCAGAACTAAGGCGCTTATCTTCTTTGATATTGCATGGAACATGCTTCTACTTCTCTGGGAGATGCAT
TTTATGGATATATCAGAGAAGATAAGCGCCTTAGTTCTGAA
>Pfa-miR157a
GGTTTGAGAGGCATCGATCGTGTTGACAGAAGATAGAGAGCACAGATGATGAGATACAATTTGGA
GCAATTTCTTTGCATCTTACTCCTTTGTGCTCTCTAGCCTTCTGTCATCACTTTTTATTTTCTGAATCC
>Pfa-miR157c
TAGGTTTGAGAGTGATGTTGGTTGTTGACAGAAGATAGAGAGCACTAAGGATGACATGCAAGTAC
ATACATATATATCATCACACCGCATGTGGATGTTAAAATATGTATAACAAATTCAAGAAAGAAATTA
GAGAGAGAGAGGGAGAGAAAGAGAAAGAGCCTGCATCTCTAATCTTTTGTGCTCTCTATACTTCTA
TCACCACCTTTATTTCTTCTTCTCTCTTACCTA

>Pfa-miR159b
AGAACAAAGGAAGAATTAGGAAGAGCTCTTTGAAGTTCAATGGAGGGTTTAGCAGGGTCAAGTAA
AGCTGCTAAGCTATGGATCCCATAAGCCTTATCAAATTCAAAATAATTGATGATAAGGTTTTTTATG
GATACCATATCTCAGGAGCTTTCACTTACCCCTTTAATGGCTTCACTCTTCTTTGGATTGAAGGGAGC
TCTTCATGATCTCTCCCTCCCT
>Pfa-miR162a
CGCTGGAGGCACCGGTTCATCGATCTCTTCCTGGGAAACAAAATTTAAAAAAAAAACAAAACATGA
ATAGATCGATAAACCTCTGCATCCAGTG
>Pfa-miR166a
GGGGAATGTTGTCTGGCTCGAGGACTCTGGCTCCCTCTATTCATGTTGGATCTTCTTCGATCTAATA
TTTGAATTGAACCTCAAGATTTCAGATCTGATTAGGGTTTTAGCGTCGTCGGACCAGGCTTCATTCC
CCCCAATTGTTGCTCC
>Pfa-miR169g
AAGATCACAAAATAACAAGAGAGGTAGAGAAAAAATGATGATGATGATGATGAGAGTCTCTAGTT
TGTACCAGAGAGTCTTGCATGGAAGAGTAGAGAATGAGGTTGAGCCAAGGATGACTTGCCGATTT
TACCAACGAATCTGAACTGATTTTGGTGTCCGGCAAGTTGACCTTGGCTCTGTTTCCTTCTCTTCTTT
TCGATGTTAGACTCTGGATATCTATTTCATCATAGTCGTGAATCGTGATCAAACTTTGCAATTTCATT
GAAATGTGACTCTT
>Pfa-miR169h
AAAGAAAATGGTGACATGAAGAATGAGAACTTGTGTGGTAGCCAAGGATGACTTGCCTGCGTTTT
AAACCATATCTATCAAAGACTCGATCGATAGTCATAAGGTTGGTTAGTCGTCAGGCAGTCTCCTCGG
CTATTAATTCAGACAATTCTCGTTCTCTTCATTTCACATTTCTCTTT
>Pfa-miR391
AAACTGCGAACAAAGATTTGCTTCGCAGGAGAGATAGCGCCATCGCCTAAGTTTAACCGGTGGTGA
CGGTATCTCTCCTACGTAGCAATCCTTATATATGCATCTTTATGCAGAGAGATGCAGCTCGAAGTTT
>Pfa-miR394a
TCATGAGGGTTTGACAAAGAGTTTCTTACATACTTCTTTGGCATTCTGTCCACCTCCTTCTATACATAT
ATATATGCATGTGTGTATATATATATAAGTGTATGTATTTGCGTTGTGTGTGGAAGAAGGAGGTGG
ATATACTGCCAATAGAGATCTGTTAGGGTTTCTTCGTAAAACCCTCTTGA
>Pfa-miR396b
GAAGGAGATGATGAAGATCCTGGTCATATTTTTCCACAGCTTTCTTGAACTTTCTTTTTTATTTCCTTT
TTTTACCAAATCAATATAGCTAAAAATCTCTAATTAGCACTTTGGAAACAAAGAAAAAGCTCAAGAA
AGCTGTGGGAAAACATGACAATTCAGGGTTTCTCCATTGATTCAATTGTGCCATCAGATTCTTC
>Pfa-miR398b
GATATTTTGAAGGTAGTGGATCTCGACAGGGTTGATATGAGAACACATGTGCAATCAACGGCTGTA
ATGATGCCATGTAATTGTTACATCTCTCGTTTTCATGTGTTCTCAGGTCACCCCTGCTGAGCTCTTTCT
CTACCGTCCATCATTATC
>Pfa-miR472
TTATGCGCCTAGTGGAGTCAAATTCTTATCAAAGATGGATTCGGGCTTACGCCTCTGTATGTATGGG
CAAAGAAGGCAAAATCTCACCTTTCACGCAGATCAACAATAAATTTTGTTGAATAGATGCTGGATTT
GTAAGGTTGTGATCTGGTTTTTGTGTTTGAAAGTGTAGAGATCCGCAATTCAAGTAAGATTTTGCCT
ACTCCGCCCATACCATACATACCCATAATCCTGAATCCATCTCTGGTGAGCAAGTACTGAAGTCCAA
AGCT
>Pfa-miR852
GCTCAGAACTAAGGCGCTTATCTTCTTTGATATTGCATGGAACATGCTTCTACTTCTCTGGGAGATG
CATTTTATGGATATATCAGAGAAGATAAGCGCCTTATCCATATAATGC

Supplementary File 4: Script for identification of differentially expressed mRNAs.

```
library(edgeR)

data <- read.delim("all_counts_60.txt", header=TRUE, row.names=1)
dim(data)
data <- as.matrix(data)

lib.sizes <- colSums(data)
lib.sizes
d <- DGEList(counts = data, group = rep(1:2, each = 3), lib.size = lib.sizes)
d$samples

d2 <- calcNormFactors(d)
d2$samples
d3 <- estimateCommonDisp(d2)
d4 <- estimateTagwiseDisp(d3, trend="none")

et <- exactTest(d4)
dim(et$table)
et
et$table$p.adjust <- p.adjust(et$table$PValue, method="BH")
dim(et$table)
et

res <- et$table[et$table$p.adjust <0.05,]
dim(res)
res$alogFC <- abs(res$logFC)
dim(res)
res2 <- res[res$alogFC >= log2(2),]
dim(res2)
res2F <- res2[res2$logFC <= -1,]
dim(res2F)
res2C <- res2[res2$logFC >= 1,]
dim(res2C)

summary(de <- decideTestsDGE(et, p=0.05, adjust="BH"))

anno <- read.delim("descriptions.txt", header=TRUE)
dim(anno)
head(anno)

select <- rownames(res2)
length(select)

anno2 <- anno[ which(anno[,1]%in%select), ]
dim(anno2)
anno3 <- anno2[order(anno2[,1]),]
res3 <- res2[order(rownames(res2)),]

anno.res <- cbind(res3, anno3)
```

```
anno.res[1:5,]
anno.res.2 <- anno.res[order(anno.res[,1]),]
anno.res.2[1:5,]

write.table(anno.res.2, "all_counts_60_edgeR results.txt", col.names=NA, sep="\t")

###draw volcano plot
names(et$table)
dim(et$table)
et$table[1:10,]
et$table$log.p.adjust <- abs(log10(et$table$p.adjust))
dim(et$table)
et$table[1:10,]
vp <- et$table
vp1 <- vp[vp$p.adjust < 0.05,]
dim(vp1)
vp1[1:10,]
vp2 <- vp1[ vp1$logFC >=1,]
dim(vp2)
vp3 <- vp1[ vp1$logFC <=-1,]
dim(vp3)
vp4 <- vp[vp$p.adjust >0.05,]
dim(vp4)
vp5 <- vp1[ (vp1$logFC < 1) &(vp1$logFC > -1) ,]
dim(vp5)

###volcano plot 1
jpeg("pc vs pf.jpg",width=800, height=800)
plot((vp[,1]),(vp[,5]), type="n", xlab="logFC", ylab="absolute log p.adjust",
ylim=c(0,10),cex.axis=1.6, cex.lab=1.6)
points((vp2[,1]),(vp2[,5]), pch=2, col="red")
points((vp3[,1]),(vp3[,5]), pch=1, col="green")
points((vp4[,1]),(vp4[,5]), pch=8, col="black")
points((vp5[,1]),(vp5[,5]), pch=8, col="black")
abline(h=0, lty=1)
abline(v=0, lty=1)
abline(v=log2(2), lty=3)
abline(v=-log2(2), lty=3)
abline(h= abs(log10(0.05)), lty=3)
dev.off()
```

Supplementary File 5: Script for identification of enriched GO terms

```
load("at_gomat.RData")# the go matrix and the gene list
load ("find.RData") #the annotations for gene ontology terms

c1 <- read.csv("averagedduplicates.csv", header=T)
dim(c1)
ind = which(!duplicated(c1$ID))
length(ind)
c2 <- c1[ind,c("ID", "logFC")]
```

```
dim(c2)
c3 <- c2[which(c2$logFC!="NA"),]
dim(c3)
rownames(c3) <- c3$ID
c4 <- c3[,-1, drop=FALSE]
dim(c4)

#up date gene list
gene.list <- intersect(row.names(c4), rownames(go.mat) )
length(gene.list)
#up date c4
c5 <- c4[ which(row.names(c4) %in% gene.list), ,drop=FALSE]
dim(c5)

mu <- mean(c5$logFC)
std <- sqrt(var(c5))
vec <- rep(NA, ncol(go.mat) )
names(vec)  <- colnames(go.mat)
z <- p <- vec

for (i in 1:ncol(go.mat) )
{
hit <- rownames(go.mat)[ go.mat[,i]==1 ]

d <- c5 [ which(row.names(c5) %in% hit ), ,drop=FALSE]
sm <- mean (d$logFC)
n <- length(hit)

z[i] <- (sm-mu)/ (std /sqrt(n) )
ifelse( z[i] <=0, p[i] <- -pnorm(z[i] ), p[i] <- pnorm(z[i], lower.tail=F)  )
}

adj.pval <- p.adjust (p, method="BH")
head(adj.pval)
c <- cbind(z, adj.pval)
anno <- find[ match( rownames(c), names(find))]
c <- data.frame(c, anno)
c <- c[ abs(c[,2])< 5e-2,]
c <- c[ order( c[,2] ),]

write.csv(c, "edgeR_averaged_duplicates.csv")
}
```