Non-coding RNA networks regulating leaf vegetative desiccation tolerance in the resurrection plant *Xerophyta humilis*.

Evan Milborrow

Thesis presented for the degree

Master of Science

In the Department of Molecular and Cell Biology

University of Cape Town

February 2018

Abstract

Common to orthodox seeds, desiccation tolerance (DT) is exceedingly rare in the vegetative tissues of modern angiosperms, being limited to a small number of "resurrection plants". While the molecular mechanisms of DT, as well as the transcription factors regulating the seed and vegetative DT programmes, have been identified, very little is known with regards to the role of regulatory non-coding RNAs (ncRNAs). To investigate the presence and roles of possible ncRNA players, RNA-Seq was performed on desiccating *Xerophyta humilis* leaves and a bioinformatic pipeline assembled to identify the potential decoy lncRNAs and miRNAs present. Interaction mapping was performed, identifying a number of small regulatory networks each regulating a small subset of the desiccation transcriptome. Predicted networks were screened for function related to DT and expression patterns consistent with functional regulatory interactions. Of the predicted networks, two appear highly promising as potential regulators of key DT response genes. The results indicate that differentially expressed (DE) desiccation response ncRNAs are present in the vegetative tissues of *X. humilis* and may play a key role in the regulation of DT. This suggests that ncRNAs could play a more important role in DT than previously thought, and may have facilitated the evolution of vegetative DT through reprogramming of seed DT programs in vegetative tissues.

Acknowledgements

I would like to thank my supervisors Professors Nicola Illing and Robert Ingle, for their support, guidance and input over the duration of this project, as well as for funding all sequencing work.

I would like to thank all members of lab 426 (the EvoDevo Lab), both past and present, who provided assistance, troubleshooting, advice and encouragement over the course of this project, often putting their own work aside to do so. Without their help much of this work would have been much more challenging and the lab much less enjoyable. Notably I'd like to thank Rafe Lyall and Stephen Schlebusch for their technical assistance with the bioinformatics, Rafe for assembling and providing access to the *X. humilis* transcriptome, and Stephen for access to the *X. humilis* genome.

I'd like to thank the National Research Foundation (NRF) for funding my tuition, as well as my parents for their ongoing financial support.

Lastly I'd like to thank all my friends (including the lab mates already mentioned), housemate Duncan, and particularly my family for their emotional support and encouragement throughout this process (Sarah your help was invaluable). Thank you everyone for the countless offers to proofread drafts. Thank you for putting up with me and for all the (often one-sided) effort you put into our friendships and my emotional wellbeing.

Table of Contents

Chapter 1: Vegetative desiccation tolerance and the possible role of regulatory non-coding RNAs.

Chapter 2: Construction of a bioinformatics pipeline for prediction of candidate regulatory lncRNAs in desiccating leaves of the resurrection plant Xerophyta humilis.

Chapter 1: Vegetative desiccation tolerance and the possible role of regulatory non-coding RNAs.

1. Introduction

Plants are estimated to have first transitioned from a purely aquatic to a terrestrial environment over 400 million years ago (Gensel and Andrews. 1984; Kenrick and Crane. 1997). While life on land presented a new ecological niche and many advantages, it is fraught with numerous challenges and stresses. Some of these stresses such as disease and herbivory (primarily biotic) are universal, while other stresses (primarily abiotic) are specific to a terrestrial existence. As sessile organisms, plants are unable to relocate to more favourable conditions when faced by such challenges, and instead have evolved numerous mechanisms to avoid, minimise and tolerate stress. One of the primary challenges facing terrestrial plants is desiccation (Farrant and Moore. 2011). Desiccation tolerance (DT) is one strategy for dealing with water loss, and is defined as the ability of an organism to survive the loss of almost all (> 95%) cellular water for extended periods of time, with no permanent damage upon rehydration (Alpert. 2005; Gaff and Oliver. 2013; Illing et al. 2005; Maia et al. 2011; Terrasson et al. 2013; Oliver et al. 2000). While common in bryophytes, vegetative desiccation tolerance (VDT) is extremely rare in modern angiosperms. Only 135 taxonomically diverse angiosperm species, collectively known as resurrection plants, possess vegetative DT (VDT), including *Xerophyta humilis* (Velloziaceae) (Gaff and Oliver. 2013).

1.1. The relevance of vegetative desiccation tolerance.

Drought is an unpredictable natural phenomenon in many regions of the world, capable of inflicting severe damage to agricultural crops, national economies and the livelihood of farming communities. Unlike *X. humilis* and other resurrection plants, the traditional agricultural crops on which we are completely dependent for our food, are overwhelmingly desiccation sensitive and dependent on a consistent water supply (Gaff and Oliver. 2013). This sensitivity of orthodox crops to water deficit, and the resulting loss of crops, is particularly relevant within the context of Sub-Saharan Africa, and a problem that is only likely to heighten and become more globally prevalent as a result of global climate change and the increasing demand for plant crops (Bruinsma, 2009). By gaining an understanding into the genetic basis and functional mechanics of vegetative desiccation tolerance hopefully strategies can be developed to activate similar mechanisms, or enhance resistance to a water deficit, within key crops.

1.2. Desiccation induced stress and molecular mechanisms of DT

Water loss due to desiccation results in various cellular stresses and stressors including 1) mechanical strain due to a loss of turgor, 2) reactive oxygen species and free radicals resulting from metabolic activity and 3) the destabilisation and denaturation of cellular constituents due to reduced intracellular hydration (Illing et al. 2005). Vegetative desiccation tolerance is achieved largely by induction of protection mechanisms, similar to those observed in orthodox seeds and during early germination (Illing et al. 2005; Oliver et al. 2004). The synthesis of protective reducing sugars (such as sucrose), which act as water replacement molecules, allows for the formation of a sugar-glass lattice that stabilises cellular components, provides structural support to the cell, and prevents the cell membranes from shearing away from the cell walls. The synthesis of protective late embryonic abundance (LEA) proteins and heat shock proteins (HSP) stabilize cellular structures and components (Buitink et al. 2006; Illing et al. 2005; Terrasson et al. 2013). Antioxidant compound synthesis and metabolic repression prevent damage by reactive oxygen species (Terrasson et al. 2013; Maia et al. 2011). This includes the disassembling of the photosynthetic apparatus and associated loss of chlorophyll in some species (poikilochlorophylly). Lastly, mRNAs and proteins essential for successful recovery from desiccation are produced and stored (Alpert. 2006).

1.3. Regulation of DT pathways.

The regulatory mechanisms governing desiccation tolerance are well characterised during late seed maturation. Signalling by the phytohormone Abscisic acid (ABA) controls seed maturation via a cascade of master regulatory transcription factors (TFs). In *A. thaliana* the most prominent TFs are *LEAFY COTYLEDON (LEC1)*, *ABISCISIC ACID INSENSITIVE 3 and 5* (*ABI3 and ABI5*), *FUSCA 3* (*FUS3*) and a second *LEAFY COTYLEDON (LEC2)*, which form part of the LEC1-*ABI3-FUS3-LEC2* (LAFL) gene network, and each of which controls a subset of seed maturation genes (Brocard-Gifford et al. 2003; Santos-Mendoza et al. 2008; Jia et al. 2014; Maia et al. 2014). As these TFs are essential for seed DT, and as ABA signalling also plays an essential role in abiotic stress response, it has been proposed that vegetative DT may have evolved through the co-option and possible modification to one of these prior existing desiccation response programmes. If this were the case, VDT would be expected to share many of the same TFs and regulatory pathways (Bewley and Oliver. 1992; Oliver et al. 2004; Illing et al. 2005). Recent work by Costa et al. (2017) has shown that not only are transcripts typically associated with seed DT induced in the drying vegetative tissues of the resurrection plant *Xerophyta*

viscosa, but orthologues of ABI3 and ABI5 were found to be present, as well as the majority of the ABI3 regulon. While ABI3 expression was not found to be significantly altered, the majority of the ABI3 regulon showed increased expression over the course of desiccation, exemplifying the seed-like character of vegetative desiccation tolerance and supporting the hypothesis that vegetative DT may have arisen through co-opting the regulatory networks present in DT seeds (Costa et al. 2017). In *X. humilis,* vegetative desiccation is also associated with an induction of a number of well-characterised LAFL target genes. *ABI3* and *LEC1* orthologues have been identified, but both with very low expression levels, and with only *ABI3* showing slight but significant differential expression (DE), making it unlikely that either is acting as a master regulator of VDT (Lyall. 2016). It therefore seems possible, and even likely, that VDT arose through co-option of the downstream components of the seed DT program. It does not appear however that this occurred through direct re-activation of the central gene networks that control seed development.

While the transcriptional response of a number of resurrection plants to dehydration has been analysed and the search for transcription factors that activate the desiccation/seed maturation genes is ongoing, little work has been carried out on ncRNAs which could also play important regulatory roles. It is possible that ncRNAs may play an important role in activating the desiccation response, both in vegetative tissue and during seed maturation. They seem prime candidates given their ability to implement rapid and extreme regulation shifts, as occurs during desiccation.

1.4. The emergence of regulatory RNAs.

1.4.1. RNA expression.

RNA is traditionally thought of as an intermediary in gene expression between DNA and proteins, as part of the DNA-RNA-Protein "central dogma" of molecular biology. Until fairly recently, the role of RNAs as anything other than an intermediate of protein synthesis has been largely overlooked. Any RNAs with roles in protein synthesis, gene regulation and nucleic acid processing, were considered outliers in a protein-centric understanding of cell function (Mercer et al. 2009).

With the advent of tiling arrays, oligonucleotide microarrays and then finally high throughput next-generation sequencing (NGS) technology, providing the ability to sequence whole transcriptomes (RNA-Seq), as well as bioinformatics tools to assemble and interpret the resulting data, a complete

high-resolution view of total gene expression became possible. NGS also allowed for cheaper and faster whole genome sequencing (DNA-Seq), to which the transcriptomic data could be compared.

While the transcriptomes of prokaryotes do consist primarily of protein coding (PC) genes, this does not hold true for eukaryotic organisms (Taft et al. 2007), with much of the non-coding eukaryotic genome being transcribed, indicating much more pervasive transcription than originally thought (Okazaki et al. 2002, Carninci et al. 2005, Mortazavi et al. 2008 Wilhelm et al 2008, Nagalakshmi et al. 2008). In mammals less than two percent of the total genome encodes for proteins (ENCODE Project Consortium. 2007) and in the relatively small and compact *Arabidopsis thaliana* (Arabodopsis) genome less than 50% of the genome is protein coding (Yamada et al. 2003). Despite this, transcriptional analyses have shown that pervasive transcription takes place at many non-coding regions with up to 90% of the eukaryotic genome being transcribed into RNA products (Mehler and Mattick. 2007; ENCODE Project Consortium. 2007; Mattick. 2011; Djebali et al. 2012). The vast majority of the genome therefore encodes for nc transcripts, with recent estimates of approximately 15000 long non-coding RNA (>200nt) transcripts in humans (Derrien et al 2012), compared to approximately 19000 protein coding (PC) genes (Ezkurdia et al. 2014). This challenged existing perceptions on the limitations of RNAs as functional effectors within biological systems (Atkinson et al. 2012, Liu et al. 2012a, Wang et al. 2014).

1.4.2. The coding vs the non-coding genome

The majority of the expressed, previously-underappreciated non-coding (nc) transcripts are derived from intergenic regions previously thought of as non-functional or "junk DNA" (Comings et al. 1972, Ohno. 1972), and are distinct from transposons and housekeeping RNAs, such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and splicing small nuclear RNAs (snRNAs) (Lander et al. 2001). The new transcripts exist in a vast range of sizes, from about 20 nucleotides to thousands of nucleotides in length. While distinct from mRNA, they share features with mRNAs; notably capped 5' ends, with many possessing spliced introns and polyadenylation (Guttman et al. 2009, Liu et al. 2012a). However, in comparison to mRNAs they possess no evolutionary conserved open reading frames and their expression levels are not always but often very low. While most of the RNA in a cell is rRNA and tRNA, and mRNA transcripts account for 3-7% of RNA, long non-coding RNAs only account for 0.03%–0.20% by mass (Palazzo and Lee. 2015), and have been shown to be between ten and seven fold less abundant than mRNAs (Kornienko et al. 2016). This often low expression, coupled with a higher

natural expression variation than protein-coding genes, lead to the existence of many ncRNAs initially being misinterpreted as transcriptional noise (Rymarqui et al. 2008).

1.4.3. Functionality.

1.4.3.1. Lack of conservation.

The identification of nc transcripts across unicellular and multicellular plants and animals raises the question of whether they are functionally important, or simply spurious transcriptional noise as a result of low RNA polymerase fidelity and/or transcriptional run-on ("leaky expression") (Struhl. 2007). Traditionally, primary sequence conservation across species has been used as a hallmark to identify functional important sequences, as a result of the high evolutionary conservation of open reading frames and amino acid sequences in proteins. This does not hold for ncRNAs however, with very low sequence conservation found between many ncRNAs in closely related organisms. Approximately 2-5.5% of long non-coding RNAs, in both plants and animals, show primary sequence conservation. While some short conserved elements do exist in some lncRNAs, it is believed that this lack of overall conservation has allowed for rapid evolution (Marques and Ponting. 2009; Ulitsky et al. 2011; Liu et al. 2012a; Zhang et al. 2014; Ponjavic et al. 2017). While the primary sequence of the small non-coding RNAs (sRNAs) is essential for function, lncRNAs have been shown to fold into complex secondary and tertiary structures, and it is through these structures/conformations that they are able to interact with RNA-binding proteins (Wang et al. 2014). Therefore, it is reasonable that so far as protein mediate functions are concerned, no primary sequence conservation is required for the longer ncRNAs.

1.4.3.2. Specificity of non-coding RNA expression.

Despite the lack of observed sequence conservation, individual ncRNAs show specific and tightly regulated expression during specific developmental stages/contexts (Amaral et al. 2008, Guttman et al. 2010), in specific cell types (Dinger et al. 2008), under specific biotic and abiotic conditions (Guttman et al. 2010), and within specific subcellular localisations (Mercer et al. 2009). Identification of transcription factors (TFs) binding to ncRNA loci and evidence of selection for these TF binding sites, all indicated that ncRNA expression is tightly and explicitly regulated in a spatio-temporal manner (Cawley et al 2004, Carninci et al 2005, Ponjavic et al 2007). This would only be the case if these transcripts were in fact functional and playing important roles within their specific contexts.

1.4.3.3. NC transcript abundance correlates to organismal complexity.

Further evidence of functionality is uncovered when nc transcript diversity is compared to organismal complexity. Genome size (DNA amount) correlates poorly with organismal complexity, a phenomenon known as the "C-value paradox" (Eddy. 2012). When the number of transcripts is compared across organisms with highly different levels of developmental complexity, the protein coding (PC) genes remain largely conserved in number and function (Mattick. 2011). This holds true for the vast majority of higher eukaryotic organisms, suggesting that regulatory differences rather than the number of encoded proteins must account for the range in organismal complexity, even when accounting for alternative splicing and post-translational protein modifications (Taft et al. 2007, Mattick. 2011). Instead, organismal complexity better correlates with the proportion of the genome encoding ncRNAs (Taft et al. 2007). The number of non-coding RNA transcripts, specifically long (>200nt) non-coding RNAs, does increase with complexity. As the number and diversities of ncRNAs expanded and evolved, they may have provided additional regulatory potential (Mattick et al. 2004, Mercer et al. 2009). This in turn allowed for more diverse and complex regulatory pathways to evolve, which may account for, or at least contribute towards, the observed phenotypic differences (Reviewed in Mattick. 2011).

1.4.4. Non-coding RNAs form a diverse group.

While it has become increasingly clear that nc transcripts are likely to play functional roles, in order to fully examine these roles it is important to understand that there is not a single type of ncRNA transcript. Non-coding RNAs can be categorised into several diverse groups on the basis of their size, known function, and origins. Housekeeping ncRNAs, such as ribosomal RNAs (rRNA), small nuclear RNAs (snRNA), small nucleolar RNAs (snoRNA), and many other regulatory RNAs, are involved in general cellular function and maintenance (Cech and Steitz. 2014). The remaining regulatory RNAs can be broken down into two main groups: small RNAs (sRNAs) and long non-coding RNAs (lncRNAs). This dissertation will be focussing on the potential role that these miRNAs and lncRNAs may be playing during desiccation tolerance in *X. humilis.*

1.5.1. Long non-coding RNAs

Currently lacking satisfactory functional classification long non-coding RNAs (lncRNAs) are lumped into a single diverse heterologous group. LncRNAs are defined transcripts that are longer than 200 nucleotides, expressed in a time, tissue, cell and/or condition specific manner and that possess little

to no coding potential (non-coding). They therefore exert their functions as RNAs and are not translated into any protein products. The 200nt size cut-off has no physiological relevance and was chosen purely on the basis of it being a convenient and practical cut-off in RNA purification protocols to exclude small RNAs (Kapranov et al. 2007). While it was originally accepted that only polyadenylated lncRNAs transcribed by RNA Polymerase II (RNA pol II) were stable "Typical lncRNA", this is no longer the case (Ulitsky and Bartel. 2013). Transcripts may be polyadentylated or non-polyadentylated, dependent on the polymerase by which they are transcribed (Reviewed in Liu et al. 2015a). LncRNAs are incredibly diverse (abundant). *A. thaliana* has approximately 40 000 putative lncRNA transcripts (Jin et al. 2013; Liu et al. 2012a; Wang et al. 2014), 6480 of which have been characterised as long intergenic non-coding RNAs (lincRNAs) (Liu et al. 2012a). In both cotton (*Gossypium hirsutum L.*) and humans, approximately ten thousand lncRNAs have been identified (Derrien et al. 2012; Lu et al. 2016).

1.5.2. LncRNA classification by transcriptional origin and molecular function.

Due to the high diversity of transcripts currently referred to as lncRNAs, multiple approaches have been used to classify transcripts into related groups, namely based on how they map to the genome (transcriptional origin), or alternatively by their mode of function.

1.5.2.1. Transcriptional origin relative to protein coding genes.

Genomic origin, relative to PC genes, is a good initial way to classify lncRNAs on the basis of transcript mapping and in the absence of any functional information. LncRNAs can be transcribed from genomic regions between PC genes, within the promoters of PC genes, within the introns of PC genes, or from the antisense strand of the PC region of PC genes, as shown in Figure 1.1. Long intergenic ncRNAs (lincRNAs) are transcriptionally independent lncRNAs transcribed from the regions between PC genes, separated from PC genes by at least 1kb (Bonasio et al. 2014). Intronic lncRNAs (incRNAs) initiate within an intron of a PC gene and may be transcribed in either direction, without overlapping any exons. At times, lincRNAs and incRNAs may overlap slightly with PC regions further complicating identification and effective classification, and incRNAs specifically are difficult to distinguish from the primary (pre-splicing) transcript of related PC genes (Liu et al. 2015b). Long non-coding natural antisense transcripts (lncNATs or NATs) are lncRNAs that are transcribed from the antisense strand of a PC gene. They may initiate within or 3' to the PC and overlap with at least one exon. Promoter lncRNAs are transcribed from the promoters of PC genes (Rinn and Chang.2012).

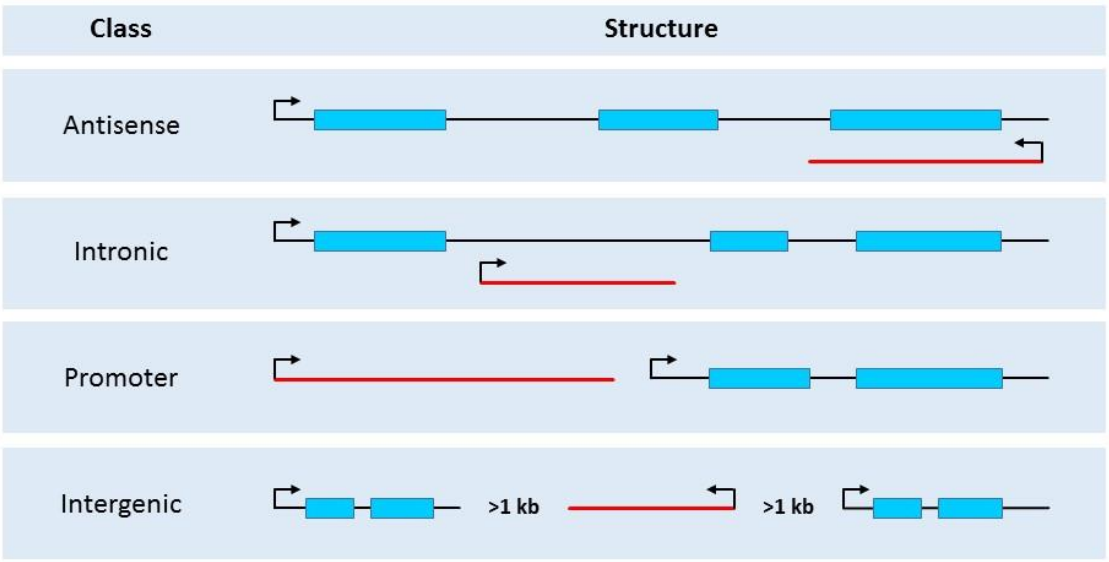

Figure 1.1: Classification of long non-coding RNAs by genomic location relative to protein coding (PC) genes. Natural antisense transcripts (lncNATs) are anti-sense to and overlap at least one coding exon. Intronic (in)RNAs initiate within an exon, are transcribed in either direction and terminate without overlapping an exon. Promoter lncRNAs are transcribed from the promoter region of a PC gene. Long intergenic (linc)RNAs are independently transcribed lncRNAs located at least 1kb from any PC genes. Protein coding genes are represented by blue exons and black introns. Red lines represent NC genes. Arrows indicate the transcriptional start sites and direction of transcription.

1.5.2.2. Modes of long non-coding RNA action

A second further method by which lncRNAs are classified, given additional information, is by their molecular modes of action. These are determined by the exact physical/direct interactions that can occur between lncRNAs and possible molecular effectors.

1.5.2.2.1. Mechanisms of lncRNA targeting.

In order for lncRNAs to perform many of their functional biological roles, they must possess the ability to recognise their relevant effectors or targets. These interactions fall into two broad categories: sequence and structural mediated interactions. The nucleotide nature of lncRNAs gives the ability to accurately bind complementary DNA or RNA target sequences, with RNA-RNA interactions being common for antisense transcripts. In mammals as many as 70% of coding mRNA transcripts have corresponding antisense ncRNAs (Samani et al. 2007). RNA-DNA interactions may occur as duplexes or triplexes with either single or double stranded DNA (Martianov et al. 2007). LncRNAs also have a

propensity for intramolecular base-pairing resulting in the formation of loops, particularly hairpins, and complex secondary and tertiary structures. These higher order structures, beyond the primary sequence level, allow for the formation of complex protein binding domains, not evident at the nucleotide level.

By forming similar tertiary structures, unrelated or non-conserved lncRNA can be of vastly differing nucleotide sequences and have no evident homology, but still be able to perform identical roles within or between organisms (Torarinsson et al. 2006). As such predicting function from lncRNA sequences alone is currently not possible. Disruption of the tertiary structure, however, results in a complete loss of function (Kino et al. 2010).

1.5.2.2.2. Molecular modes of action.

The proposed modes or "Archetypes" of molecular function are: Decoys, Guides and Scaffolds, as shown in Figure 1.2. These modes are not mutually exclusive, and form a hierarchy of lncRNA functional complexity that may represent the process of lncRNA evolution (Wang and Chang. 2011). Although not a mode of lncRNA transcript function, it should be noted that the act of lncRNA transcription itself may function to enhance transcription of surrounding genes by adopting an open chromatin state.

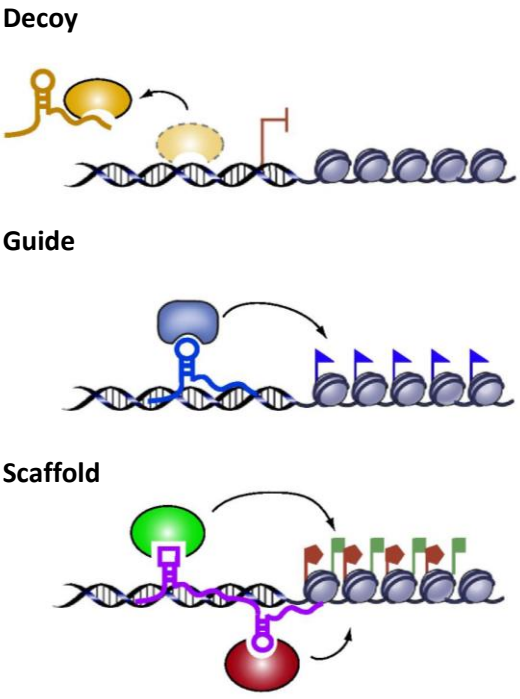

Figure 1.2: Illustrative diagram showing the three primary modes of lncRNA molecular action. Decoy lncRNAs titrate miRNAs, transcription factors and other regulatory proteins by sequestering them away from their intended chromatin (shown) or miRNA targets. Guide lncRNAs recruit chromatin-modifying enzymes (shown) or other transcriptional effectors to their specific target sites, either in *cis* or *trans*. As scaffolds, lncRNAs recruit and enable or stabilize the structural assembly of multisubunit ribonucleoprotein complexes. This may alter or enhance complex function. Images adapted from Wang and Chang. 2011.

Decoys are lncRNAs with miRNA or protein binding sites and function as competitive inhibitors, titrating key regulatory effectors away from their intended targets and thereby inhibiting effector function. Titration allows for rapid and efficient up and down adjustment (binding and release) of available effectors, without the need for lengthy degradation and

translation steps. The large size of lncRNAs allows a single transcript to contain multiple effector binding sites, and in the case of miRNA effectors, mismatches in the miRNA binding motif prevent miRNA induced cleavage. This ability to act as a non-degradable sink or "molecular sponge" for miRNAs and RNA-binding proteins (RBPs) is known as "target mimicry" (Franco-Zorrilla et al. 2007). Non-degradable target mimicry is a common mechanism of lncRNA mediated regulation of miRNA activity in plants (Ivashuta et al. 2011, Meng et al. 2012, Wu et al. 2013). Putative target mimics for at least 20 miRNAs have been identified in Arabidopsis and rice, with functionality successfully demonstrated by transgenic experiments (Wu et al. 2013). The discovery of target mimicry in plants has since been followed by the discovery of a similar process in animals known as competitive endogenous lncRNA (ceRNAs) (Kartha and Subramanian. 2014).

Other effectors regulated by decoy lncRNAs are TFs and chromatin modifiers. Depending on the specific lncRNA and the associated effector involved, decoys may play a central role in either positive or negative regulation of transcription (Reviewed in Wang and Chang. 2011 & Morriss and Cooper. 2017). Overall decoy lncRNAs possess the simplest structural requirements for lncRNA function, requiring only nucleotide motifs that mimic effector target sites, and may represent the most primitive mode of lncRNA function. Decoy lncRNAs are often transcribed from promoter and enhancer regions, which are rich in TF binding sites (Guenther et al. 2007).

The dual ability to be part of both nucleotide and structurally mediated interactions allows **Guide lncRNAs** to directly bind proteins, and as a ribonucleoprotein complex direct their localisation to specific target sites. This explains how many RNA binding proteins (RBPs) lack any sequence specificity of their own, yet are still able to act at specific genomic loci. Alternatively, guides may first bind to targets sites and then recruit regulatory proteins directly to these sites. The ability of guide lncRNAs to both bind to effector proteins (like decoys) and localise to specific genomic loci or DNA regions, through the presence of additional transcript sequence complimentary to the genomic target site, suggests they likely evolved more recently from existing decoy lncRNAs.

Guides may function in *cis* or in *trans*. Guides which function in *cis* guide regulatory proteins to neighbouring genes. As they act at or near their site of transcription, they need not be expressed at high levels, and may even act as "tethers" acting directly at their site of synthesis, while still bound to the RNA polymerase complex (Lee et al. 2009). Guides that function in *trans* act on distantly located genes often on another chromosome, and must be expressed at much higher levels to compensate for diffusion (Lee et al. 2012, Kung et al. 2013, Yang et al. 2014b). In order for *trans* acting guide

lncRNAs to function, they must be able to properly localise to their target sites. This occurs through the formation of DNA:RNA heteroduplexes or RNA:DNA:DNA triplexes, or through RNA recognition and binding to specific chromatin features (Bonasio et al. 2010). In both cases a knockdown would result in altered or lost localisation of effector molecules, as well as a resulting loss of effector function.

Guide lncRNAs may perform an activating or repressive function, as determined by the specific effectors involved. These effectors may be either single activating proteins, such as the Trithorax group proteins (TxG), repressive proteins, such as the Polycomb group proteins (PcG), or TF proteins, such as transcription factor II B (TFIIB). Alternatively, the effectors may be multisubunit complexes such as the polycomb repressive complexes (PRC) or activating mixed-lineage leukemia (MLL) complexes (Reviewed in Wang and Chang. 2011).

The third mode of lncRNA function is as **Scaffolds** – serving as a central structural platform upon which the subunits of a regulatory complex may assemble (Spitale et al. 2011). The ability of lncRNAs to be comprised of multiple domains and to form complex secondary and tertiary structures, allows lncRNAs to interact with the multiple protein subunits which traditionally make up regulatory complexes (Good et al. 2011). Scaffolds may (1) be required for protein complex assembly, (2) enhance complex function by thermodynamic stabilization and/or allosteric activation, and/or (3) bring together distinct transcriptional regulators (without forming a true protein complex) to independently exert their respective functions but in a coordinated manner (spatio-temporally). This is distinct from guides which take an assembled complex and direct it to its target, in that the scaffold is required for complex assembly itself, allowing careful control over when and where regulatory complexes are able to assemble and function. Scaffold lncRNAs are believed to have evolved from guide lncRNAs by addition or multiplication of effector protein binding sites (Wang and Chang. 2011)

1.5.2.2.3. Examples of plant lncRNAs functioning as decoys, guides and scaffolds.

By and large, many of the roles and molecular mechanisms of lncRNAs have been uncovered in animal systems (Ulitsky and Bartel. 2013, Cech and Steitz. 2014). These same mechanisms however have been shown to be present in plant systems.

Phosphate starvation induced lncRNA ISP1 functions as a decoy for miR399.

As discussed, decoy lncRNAs function as competitive miRNA inhibitors, titrating miRNAs away from their original target, and thereby repressing or fine tuning their activity. One of the best understood examples of decoy lncRNA activity in plants is in the regulation of phosphate homeostasis by the miRNA miR399 and the lncRNA INDUCED BY PHOSPHATE STARVATION1 (IPS1). MiR399 is a key regulator of plant phosphate homeostasis. Induced under condition of phosphate starvation, miR399 binds and induces transcript cleavage of its target mRNA *PHOSPHATE 2* (*PHO2*). *PHO2* encodes a ubiquitin-protein ligase that mediates degradation of the phosphate transporters PHOSPHATE 1 (PHO1) and PHOSPHATE TRANSPORTER 1s (PHT1s) at the endomembrane (Aung et al. 2006; Liu et al. 2012b). MiRNA induced PHO2 repression thereby increases phosphate uptake. The lncRNA INDUCED BY PHOSPHATE STARVATION1 (IPS1) possesses a 23nt motif conserved across all plants. This motif is complimentary to miR399, other than a 3nt mismatched loop, where the miRNA would normally induce cleavage in a mRNA target. This mismatch allows ISP1 to effectively act as a target mimic or decoy binding miR399, without miRNA induced cleavage, to effectively sequester miR399 transcripts and attenuate miR399-mediated repression of *PHO2* (Franco-Zorrilla et al. 2007).

LncRNAs COOLAIR and COLDAIR guide Chromatin remodelling.

One of the key roles of lncRNAs in plant systems is as regulators of transcription through mediating epigenetic state and chromatin remodelling. It is well known that chromatin state plays an important role in regulating gene expression. Histone methylation specifically plays an important role in regulating heterochromatin state, with trimethylation of Histone 3 at Lysine 27 (H3K2me3) leading to and being a well-recognised mark of compact heterochromatin, and repression of transcription.

FLOWERING LOCUS C (*FLC*) is a MADS-box TF and transcriptional repressor in plants that functions to repress genes required for the developmental shift from vegetative growth to flowering (Sheldon et al. 2000). Repression is highly regulated in response to seasonal information and environmental cues to ensure flowering occurs at the correct time of year. In cold climates many plants undergo a process known as vernalisation, in which exposure to cold leads to an induction of rapid flowering in preparation for spring or germination. This occurs through the epigenetic silencing of *FLC* and is mediated by two cold induced lncRNAs, COOLAIR and COLDAIR (Figure 1.3).

COOLAIR is a lncNAT to the whole FLC locus and is expressed in response to cold, during vernalisation, leading to cold-dependent transcriptional silencing of *FLC* (Liu et al. 2010). Expressed during the first two weeks of cold, COOLAIR expression leads to a switch from sense to antisense transcription, repressing *FLC* expression by competing for the transcriptional machinery present at the *FLC* locus (Helliwell et al. 2011; Swiezewski et al. 2009).

COLDAIR is a sense lncRNA encoded by the first intron of FLC (Heo and Sung. 2011, Ietswaart et al. 2012). Transcribed by RNA PolII, COLDAIR contains a cold-response cis-acting element in its promoter region and is transiently induced by cold and the onset of vernalisation (Heo and Sung. 2011). COLDAIR plays a key role in vernalisation by directly associating with and recruiting CURLY LEAF (CLF), a component of the PRC2, to the *FLC* locus (Swiezewski et al. 2009). COLDAIR therefore functions as a guide lncRNA. PRC2 is a repressive complex that leads to epigenetic silencing by repressive histone modification H3K27me3, leading to a stable silenced state (Buzas et al. 2011; Heo and Sung. 2011, Csorba et al. 2014). The ability of ncRNAs to interact and recruit the PRC2 is an evolutionary conserved mode of gene repression, being found in both plants and animals.

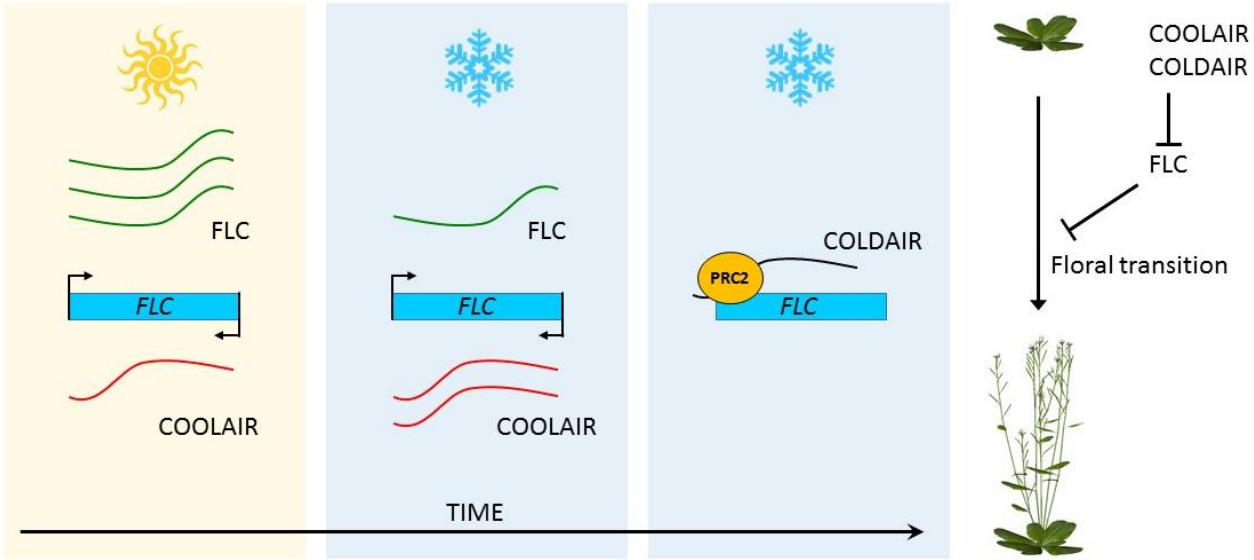

Figure 1.3: The lncRNAs COOLAIR and COLDAIR silence *FLC* expression during vernalisation. Vernalisation is the acceleration of flowering in response to prolonged cold. During warm conditions *FLC* (green) is expressed at high levels, inhibiting a transition from vegetative growth to flowering, and COOLAIR (red) is expressed at low levels. Exposure to cold induces an upregulation in COOLAIR expression from the antisense strand of the *FLC* locus, and a decrease in *FLC* expression. During prolonged exposure to cold, COLDAIR (black) is expressed, leading to repressive histone modification H3K27me3 and associated epigenetic silencing of *FLC* by Polycomb Repressive Complex 2 (PRC2).

1.6.1. Plant sRNAs

Plants express large numbers of 18-25nt small RNAs (sRNAs) which function to mediate gene silencing (GS) at either the transcriptional (TGS) or post transcriptionally (PTGS) level (Voinnet. 2009). SRNAs are divided into several major classes, by function, mode of action, size/composition and biosynthetic origin (Fei et al. 2013). These classes include small interfering RNA (siRNA), microRNA (miRNA), Piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNAs), tRNA-derived small RNA (tsRNA), small rDNA-derived RNA (srRNA), and small nuclear RNA (snRNAs/U-RNAs). While often difficult to distinguish from sequence alone, the advent of high throughput next generation sequencing (NGS) and bioinformatics provided the ability to combine information on nucleotide sequence, genomic origin, post-transcriptional processing (biosynthetic intermediates) and putative target sites to readily distinguish between these classes and their fundamental roles in transcriptional regulation. This resulted in a rapid expansion in our understanding of sRNA diversity, ubiquity and importance. MicroRNAs (miRNAs) are the second most abundant of these sRNA classes, after siRNA, and are recognised as a key post-transcriptional repressor in all eukaryotes (Voinnet. 2009).

1.6.2. Defining miRNAs and distinguishing miRNAs from other sRNAs

MiRNAs are ubiquitous, being found in both plants (Park et al. 2002; Reinhart et al. 2002) and Metazoa (animals) (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros, 2001.). While all miRNAs function as post-transcriptional repressors, they have diverse functional roles. In plants and animals miRNAs have different methods of biosynthesis, differences in target binding characteristics and differences in their preferred mechanism of action. It is thus difficult to develop a single comprehensive and unambiguous definition for all current eukaryotic miRNAs. It is plausible that the current understanding of miRNAs may in fact encompass classes of similar sRNAs that will in fact be subdivided in the future. Despite this, similar features do exist, and a general definition can be given.

MiRNAs are small, silencing, endogenous, non-translated RNAs transcribed as an inverted repeat precursor transcript, that undergoes cleavage by Rnase III enzyme of the Dicer and/or Drosha family of proteins to yield transcripts approximately 20-24nt in length (Brodersen et al. 2008; Voinnet. 2009). In plants most miRNAs are ~21-22nt in length. These transcripts function as *trans*-acting sequence specific guides, leading Argonaute (AGO) protein to complementary RNA targets. As sRNAs all function in similar ways, miRNAs are differentiated primarily by their genomic origin and unique mode of biosynthesis (Axtell. 2013; Fei et al. 2013).

1.6.3. Biogenesis of plant miRNAs

The biosynthesis of plant miRNAs is best described by the canonical biosynthetic pathway in *A. thaliana*, though a plethora of slight variations exist. A general overview of the pathway is as follows: DNA – primary miRNA – precursor miRNA – miRNA/miRNA* duplex – mature single stranded (ss) miRNA, as illustrated in Figure 1.4.

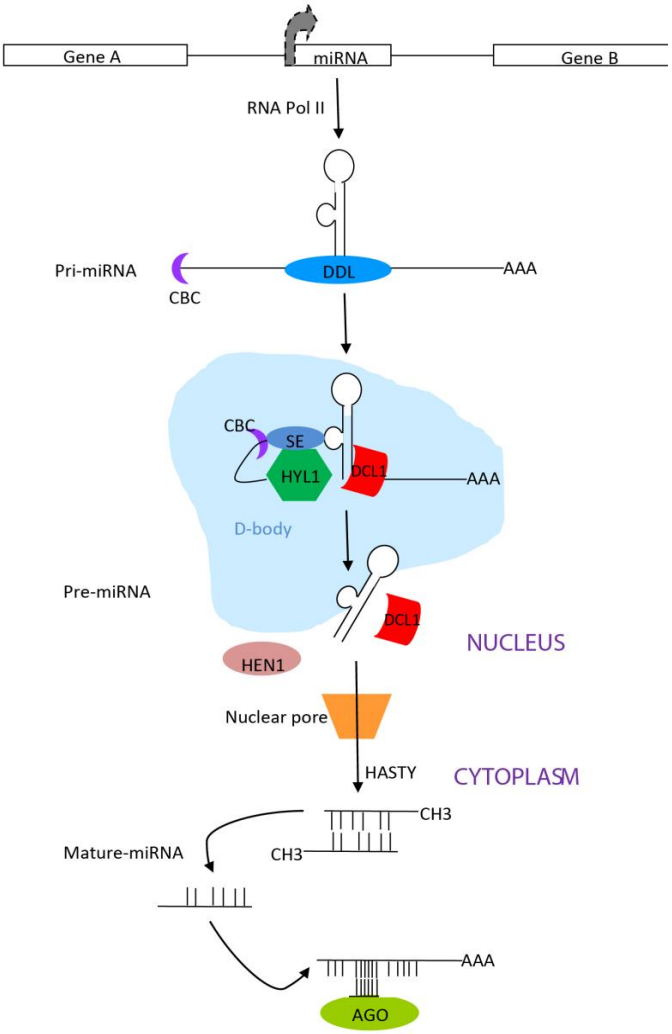

Figure 1.4. Overview of the canonical miRNA biosynthesis pathway in plants. The primary miRNA transcript (pri-miRNA) is transcribed from the miRNA gene by RNA Poltmerase II (RNA PolII), before folding back on itself to form a stem-loop structure. Dicer-like 1 (DCL1) excise the stem loop to yield precursor miRNA (pre-miRNA), and then further excises the miRNA/miRNA* duplex from the stem of the pre-miRNA. The duplex is transported to the cytoplasm for loading of the mature miRNA strand into the RNA-induced silencing complex (RISC). The enzymes and complexes involved in miRNA biosynthesis are Dawdle protein (DDL), binding complex (CDC), Serrate protein (SE), Hyponastic leaves 1 (HYL1), Dicer-like 1 (DCL1), exportin-5 ortholog (HASTY), methylation protein (HEN1), and Argonaute (AGO).
Figure from Christopher et al. 2016

The vast majority of characterised miRNA genes exist as independent transcriptional units not associated with any protein coding (PC) genes (Reinhart et al. 2002). Transcription by RNA Polymerase II (PolII) yields a single long primary transcript (pri-miRNA). The pri-miRNA undergoes 5' capping and polyadenylation, before folding back on itself to form a hairpin stem-loop structure. While fairly uniform in animals, the length and stem-loop structure of plant pri-miRNAs is highly variable, with pri-miRNAs able to exceed 1kb in length (Lee et al 2004; Voinnet. 2009).

Processing of the pri-miRNA by a RNase III endonuclease Dicer-like (DCL) protein, excises the stem-loop to yield precursor miRNA (pre-miRNA). Most pre-miRNAs are excised by Dicer-like (DCL) 1, 1 of 4 DCL proteins in Arabidopsis (Park et al. 2002; Reinhart et al. 2002; Rajagopalan et al. 2006). Binding and processing pri-miRNA to pre-RNA by DCL1 requires dsRNA-

binding protein (dsRBP), HYPONASTIC LEAVES1 (HYL1), and C2H2-Zinc finger protein SERRATE (SE). This occurs in nuclear processing centres, the D-bodies or SmD3/SmB-bodies (Kurihara et al. 2006; Fang and Spector. 2007).

The pre-miRNA transcript appears to be very short lived in plants. DCL1 further excises the miRNA/miRNA* duplex from the stem of the pre-miRNA (Kim. 2005). In plants which lack a Drosha-like protein present in Metazoa, a single DCL protein therefore performs all miRNA processing in a single concerted step. This differentiates plant miRNA biogenesis from the two-enzyme/two-step compartmentalised process present in Metazoa.

The miRNA/miRNA* duplex is exported to the cytoplasm by the protein HASTY (HST), an Exportin 5 homolog (Bollman et al. 2003; Park et al. 2005). It is not clear however, if miRNA/miRNA* is excised from the pri-miRNA transcripts before, during, or after transport.

The resulting miRNA/miRNA* duplex is methylated on the 2' hydroxyl groups of the 3' terminal nucleotides by the methyltransferase HUA ENHANCER1 (HEN1). This protects the RNA transcripts from uridylation and subsequent degradation (Li et al. 2005; Yu et al. 2005).

The miRNA strand of the miRNA/miRNA* duplex has weaker 5' base pairing. This energetic asymmetry marks the miRNA strand for preferential loading into a RNA-protein complex, containing the mature miRNA strand as well as a protein of the Argonaute family, referred to as a RNA-induced silencing complex (RISC) (Khvorova et al. 2003; Schwarz et al. 2003). *A. thaliana* possesses 10 paralog Argonaute (AGO) proteins. AGO1 and AGO10 play roles in miRNA mediated mRNA repression (Brodersen et al. 2008). The majority of characterised miRNAs form a RISC with AGO1 (Reviewed in Vaucheret. 2008). Protected by the RISC, miRNA strands accumulate to higher levels than their associated star strands (Reinhartet al. 2002; Lim et al. 2003), which being preferentially excluded, are subject to degradation (Tomari et al. 2004). AGO itself may degrade the star strand during miRNA loading, as observed for cleavage assisted loading of siRNAs (Matranga et al. 2005). The PAZ domain of AGO is a RNA binding domain and directly binds the mature miRNA (Lingel et al. 2003; Song et al. 2003; Yan et al. 2003). All perceived activity is catalysed by the AGO protein of the RISC. The miRNAs therefore act in *trans* as sequence specific guides for target recognition (Lai. 2002; Rhoades et al. 2002; Axtell et al. 2008).

1.6.4. Target recognition.

MiRNAs recognize target RNA transcripts through nucleotide Watson-Crick complementarity. They are able to bind and mediate transcript levels of both coding and non-coding RNAs. Target binding to protein coding transcripts may occur in any region of the mRNA - within the 5' or 3' UTRs, or the ORF.

In animals, initial target recognition requires near perfect binding of an eight nucleotide seed sequence, followed by relaxed, imperfect complementarity to multiple target sites in the 3' UTRs of RNA targets. This imperfect pairing allows a single miRNA to bind and regulate a large number of direct targets (Voinnet. 2009). The strict perfect (Llave et al. 2002) or near perfect (Rhoades et al. 2002) target complementarity of plant miRNAs distinguishes them from animal miRNAs (Rhoades et al. 2002; Jones-Rhoades and Bartel. 2004). A result of strict target requirements in plants is that, unlike animal miRNAs, most plant miRNAs appear to directly regulate a single transcript, or few related transcripts. (Bartel. 2009). The differing complementarity requirements and targeting characteristics of animal and plant miRNAs also directly influences their predominant mode of action.

1.6.5. Molecular mode of action of plant miRNAs.

MiRNAs repress gene expression post-transcriptionally through two primary mechanisms: translational inhibition (Brodersen et al. 2008) and mRNA transcript degradation by guided site-specific cleavage, referred to as target “slicing” (Bartel. 2004; Baumberger and Baulcombe. 2005, Brodersen et al. 2008; Eulalio et al. 2008). The mode of action is dictated by the extent and pattern of target complementarity (Valencia-Sanchez et al. 2006, Voinnet. 2009).

1.6.5.1. Plant mRNA slicing; miRNA directed cleavage.

In plants, slicing appears to be the predominant mode of miRNA action. Almost every miRNA transcript has associated cleavage targets – as expected given the high level of target high complementarity in plant miRNAs (Jones-Rhoades et al. 2006).

While miRNAs act as complimentary guides, they do not themselves possess any catalytic ability. RISC-catalysed cleavage of target mRNA sequences is carried out by the Piwi domain of AGO, which resembles RNase H in structure and possesses RNA endonucleolytic activity (Liu et al. 2004; Song et al. 2004; Baumberger and Baulcombe. 2005). Cleavage is achieved through hydrolysis of a single

phosophodiester bond, opposite the 10th and 11th miRNA nucleotide positions, within the backbone of the complementary RNA transcript (Llave et al. 2002). High miRNA-target complementarity favours cleavage, insuring the correct target and cleavage position, as well as access to the cleavage site by the Piwi domain. The cleaved fragments are released, and the free RISC is able to bind and cleave subsequent target RNAs. The presence of predictable cleavage products allows for experimental validation of miRNA cleavage activity.

The vast majority of RNA sequences targeted and cleaved are mRNAs. Cleavage can mediate transcript levels, but also enables rapid "mRNA clearance" when drastic and rapid transcriptional shifts are required, such as during key developmental stages or times of stress. Plant mRNA slicing is therefore of particular interest with regards to vegetative desiccation tolerance where rapid metabolic shutdown is required, and major transcriptional shifts occur towards protective programmes (Voinnet. 2009).

1.6.5.2. Translational repression.

AGO also functions to represses gene expression and mediate protein levels through inhibiting translation of mRNAs into functional protein products (Axtell. 2008; Brodersen et al. 2008). This mode of action is favoured in animal systems, where high miRNA-target complementarity is not required for target occupancy and subsequent translational repression. Transcriptional repression does occur in plants but is much less common.

1.6.6. Conservation of plant miRNAs

Unlike in metazoans, where miRNAs are generally highly conserved, plants possess many novel species-specific miRNAs. As such, plant miRNAs can be classified into two categories by sequence conservation/diversity: Ancient miRNAs, which are highly conserved across multiple plant lineages, and young miRNAs, which are not (Axtell and Bowman. 2008; Tang. 2010; Cuperus et al. 2011)

1.6.6.1. Conserved plant miRNAs families

MiRNA families are miRNA genes grouped on the basis of shared mature miRNA sequence and/or shared pre-miRNA structure – suggesting shared function (Kaczkowski et al. 2009; Ding et al. 2011; Kozomara and Griffiths-Jones. 2011). A small number of plant miRNA families, and their respective

targets, are present and highly conserved across phylogenetically distant land-plant lineages (Zhang et al. 2006; Cuperus et al. 2011; Chávez Montes et al. 2014). Examples of this are miR156, miR160, miR165/166, miR167, miR319, miR390, miR395, and miR408, which appear to be universally present in all Embryophyta (Voinnet. 2009; Taylor et al., 2014; You et al. 2017). This suggests that these miRNA families, and plant miRNAs as a whole, evolved very early in plant evolutionary history, and predate not only the divergence of eudicotyledons (dicots) and monocotyledons (monocots) approximately 100 million years ago (MYA), but also that of the gymnosperms and angiosperms (305 MYA) as well as the tracheophytes and bryophytes (490MYA) (Axtell. 2008; Cuperus et al. 2011; Taylor et al., 2014). The emergence of these conserved plant miRNAs may well have coincided with adaptation to a terrestrial existence – at roughly the same point which desiccation tolerance is believed to have originally evolved.

While the number of conserved plant miRNAs is relatively low, these conserved miRNAs have multiple gene copies, arising through genome duplication events. These are all highly expressed, accounting for the majority of miRNA abundance (Vazquez et al. 2008; Chávez Montes et al. 2014). These miRNAs predominantly regulate ancestral transcription factors or physiological enzymes which are involved in key biological processes, basic plant development, or stress responses (Garcia. 2008; Todesco et al. 2010; Yan et al. 2012; Qin et al. 2014). For example, the *A. thaliana* miR395 and miR399 miRNAs are induced by sulphur and phosphate starvation respectively (Fujii et al. 2005; Chiou et al. 2006; Kawashima et al. 2009).

1.6.6.2. Non-conserved plant miRNAs

In Arabidopsis, non-conserved miRNAs far outnumber conserved miRNAs (Rajagopalan et al. 2006; Zhang et al. 2006; Fahlgren et al. 2007). This is true for all land plants (Voinnet. 2009). Having evolved recently, these relatively young miRNAs correspond to a single or low number of genomic loci and family members (Voinnet. 2009). They tend to be weakly expressed, at levels significantly lower than conserved miRNAs, and are often processed imprecisely (Fahlgren et al. 2007; Chávez Montes et al. 2014). Some non-conserved miRNAs are expressed abundantly but only in specific tissues or after being induced under very specific conditions, suggesting that non-conserved miRNAs may play a role in environmental adaptation (Cuperus et al. 2011; Taylor et al. 2014; Qin et al. 2014). Evolution of new miRNAs has been rapid with each plant species appearing to have a unique set of species-specific miRNAs. Other miRNAs are only present in a few closely related species (Qin et al. 2014). Of ~100

miRNA families present in *A. thaliana*, at least 29 are not present in the *Arabidopsis lyrata* genome – a closely related species with only 5 million years divergence (de Felippes et al. 2008).

It is possible that due to this rapid evolution, some non-conserved miRNAs appear to lack functional mRNA targets and have been considered 'energy wasters' (Axtell. 2008; Axtell. 2013; Qin et al. 2014). The low abundance of many non-conserved miRNAs has also thrown doubt on their biological significance. The low abundance miR1916, miR1917, miR1918 and miR1919 families, described in S*olanum lycopersicum*, however have been shown to play an active role in fruit ripening – suggesting that other low abundance miRNAs may also have biologically relevant functions (Moxon et al. 2008; Chávez Montes et al. 2014). Between these functional low abundance miRNAs, and the abundant but highly condition specific miRNAs, non-conserved plant miRNAs are implicated in a vast array of biological functions. While there appears to be few processes they aren't involved in, some of their roles include: regulation of plant defence and immune response processes, as well as adaptation to both biotic and abiotic stresses (Sunkar et al. 2007; Pedersen and David. 2008; Voinnet. 2008). Co-ordinating and resetting stress response gene expression appears to be a major emerging function of many plant miRNAs (Sunkar et al. 2007). This involvement in stress response programmes, by both conserved and non-conserved miRNA families, is particularly pertinent in the context of desiccation tolerance.

1.7. Aims of the current study

This study was built on the hypothesis that non-coding miRNAs may play key roles in regulating the VDT programmes of *X. humilis* and other resurrection plants. The aim of this project was to identify the lncRNAs and miRNAs present during desiccation, as well as to predict possible regulatory lncRNA-miRNA and miRNA-mRNA interactions. In Chapter 2, I describe the total RNA sequencing of the *X. humilis* desiccation transcriptome, as well the construction of a bioinformatics pipeline to screen the de novo assembled transcripts for a core set of putative desiccation response lncRNAs that may play a role as miRNA decoys. Chapter 3 describes the sequencing of the small RNA complement present in the desiccating *X. humilis* leaves. The sRNA-Seq data is then analysed to predict a set of high-confidence desiccation response miRNAs, which are subjected to expression analysis and categorised by homology to known miRNA families. In Chapter 4, I predict target interactions between the putative lncRNAs, predicted miRNAs and the *X. humilis* leaf desiccation transcriptome. The predicted interactions are used to assemble a number of small regulatory networks, each of which is analysed with regards to composition, expression consistent with regulatory interactions, and annotated

transcript function. The identified networks give insight into possible key ncRNA regulators as well as their immediate regulatory interactions, which may play a role governing the vegetative desiccation response in *X. humilis*.

Chapter 2: Construction of a bioinformatics pipeline for prediction of candidate regulatory lncRNAs in desiccating leaves of the resurrection plant *Xerophyta humilis*.

2.1. Introduction

Resurrection plants are of great interest due to their extreme phenotype and its potential to give insight as an ideal model for improving water stress tolerance. While the origin, underlying genetics, and molecular mechanisms of DT are well characterized in pollen and orthodox seeds, the study of DT in the vegetative tissues of resurrection plants has been limited by the costs and difficulty of working with non-model organisms, particularly in plants which possess large scale genomic variability and complexity (Dinakar and Bartels, 2013).

With increasing accessibility and affordability of NGS technology, as well as computational tools and computing power, it is now possible to examine the desiccation transcriptome of resurrection plants in greater depth, as well as to effectively meet the bioinformatics challenges posed by working with these non-model organisms. The transcriptomes of three resurrection plants (*Craterostigma plantagineum, Haberlea rhodopensis* and *Xerophyta viscosa*) and the genomes of two resurrection plants (*Boea hygrometrica and Xerophyta viscosa*) have already been published, with the *X. humilis* transcriptome and genome complete, but pending publishing by our research group (Rodriguez et al. 2010; Gechev et al. 2013; Xiao et al. 2015; Costa et al. 2017; Lyall. 2016; Illing & Schlebusch, personal communication). Almost all the studies to date, have however, been focused on coding transcripts, with studies into the regulatory mechanics focusing on the key transcription factors regulating desiccation tolerance. In this study, I have focussed on the role of non-coding RNAs, including long non-coding RNAs in regulating the onset of desiccation tolerance.

2.1.1. Long non-coding RNAs are involved in plant stress.

Long non-coding RNAs are abundant in plants. A recent study by Wang et al. identified 30,550 lincRNAs and 4,718 lncNATs present during cotton fibre development (Wang et al. 2015). In addition to being numerous, lncRNAs have been shown to play key roles in eukaryotic gene regulation, with regulatory lncRNAs appearing to be involved in every aspect of plant biological function, including growth, development, reproduction, and response to both biotic (disease) and abiotic stresses, as discussed in Chapter 1 (Rymarquis et al. 2008; Guttman et al. 2009; Ponting et al. 2009; Liu et al. 2012; Rinn and Chang. 2012; Zhu et al. 2014; Wang et al. 2015; Joshi et al. 2016). A major theme emerging in plant

lncRNA function is their involvement in regulating stress response programs of gene expression (Deng et al 2018). This includes responses to nutrient deficiency and temperature changes (*IPS1*, *COOLAIR* and *COLDAIR*; Chapter 1), as well as plant stress response programs during times of water deficit.

An example is *Drought Induced RNA* (*DRIR*), a novel regulator of the water stress response in *Arabidopsis thaliana*. While low levels of *DRIR* expression occur under non-stress condition, exposure to osmotic stress, through water deficit or high salt levels, as well as exposure to the plant stress hormone abscisic acid (ABA) result in *DRIR* being highly upregulated. Transgenic *A. thaliana* plants engineered to overexpress *DRIR,* exhibit significantly increased stress tolerance and survival under water deficit conditions. Overexpression was found to accelerate stomatal closure under stress conditions and to have impacted a number of downstream ABA signalling, water transport, and other stress relief genes (Qin. 2017). Other studies have identified a number of plant stress lncRNAs. These include the identification of 125 plant stress lncRNAs in wheat (Xin et al. 2011) and 504 drought responsive lncRNAs in Populus trichocarpa (Shuai et al. 2014).

2.1.2. Requirements for identifying lncRNAs.

While the approach and methods used to identify lncRNAs are relatively simple and straight forward, the actual ability to confidently identify and annotate lncRNA transcripts is more challenging. Unlike mRNAs which have a clear PC function, lncRNAs are defined by their lack of PC ability – a more difficult characteristic to definitively prove. The exact process used to identify lncRNAs from deep-sequencing data varies between studies, but always seeks to satisfy the two main defining requirements of all lncRNAs: 1) existence of a reliably expressed transcript sequence ≥200 nucleotides in length, and 2) the absence of any protein coding potential and/or translated protein product (Mattick and Rinn. 2015; Wang et al. 2017).

2.1.2.1. Identifying biologically relevant long RNA transcripts.

The first step in any lncRNA discovery pipeline is the identification of an initial set of candidate transcripts. This can be achieved through RNA-Seq and de novo transcript assembly, or through a genome wide survey in which RNA-Seq reads are mapped to an existing genome to identify and extract putative transcriptional units (Li et al. 2014; Li et al. 2017). Once a set of all possible candidate transcripts have been identified, a size cut-off is applied to ensure all transcripts meet the ≥200 bp criteria of all lncRNAs (Deng et al. 2018). This step is often achieved indirectly as a result of RNA-Seq

library preparation. Once a set of appropriately sized candidate transcripts have been identified, the transcripts can be screened for active and consistent expression under the conditions of interest. This ensures that transcripts are actively transcribed, likely to be biologically relevant, and not simply a result of leaky expression, incorrect transcript assembly, or an artefact of the sequencing process. This screening can be performed through application of 1) a minimum coverage cut-off (number of mapped reads, times coverage or FPKM) and 2) conserved expression across multiple samples or replicates (Li et al. 2016; Deng et al. 2018). From previous studies, a minimum coverage of ≥3 (Li et al. 2016; Deng et al. 2018) appears to be an acceptable threshold, and/or an FPKM cut-off of 0.5 (Li et al. 2017). Once a set of appropriately sized transcripts showing dependable expression have been identified, the nucleotide sequences must be assessed for the absence of any PC potential.

2.1.2.2. Assessing candidate long RNA transcripts for Coding Potential.

Coding potential is a measure of the probability that a RNA nucleotide sequence possesses one or more valid open reading frames (ORFs) and has the ability to be transcribed into a functional protein. Multiple complementary bioinformatics approaches exist to assess PC potential, and assess transcripts on either 1) Similarity to known protein coding sequence, or 2) Coding statistics (Mattick and Rinn. 2015; Wang et al. 2017). Both methods should be applied, and can be performed in any order.

A general first step in screening for non-coding transcripts is removal of all transcripts belonging to known PC genes. If the goal is to identify novel lncRNAs, then known NC transcripts can also be removed (Deng et al. 2018). Removal of known sequences can be performed using any sequence alignment tool, such as BLASTn. An annotated transcriptome or list of protein sequences may not be available and it is very likely that some remaining candidate transcripts may code for proteins not present in the list of known coding transcripts. Furthermore, transcripts coding for small proteins or peptides can fall under the bioinformatics radar of initial genomic and transcriptomic annotation. Alignment of transcripts (BLASTX) to an annotated and comprehensive protein database allows effective removal of any transcripts with shared homology to other known proteins (Li et al. 2017; Deng et al. 2018). NCBI, Swiss-Prot or a central repository such as UniProt are commonly used protein databases. Furthermore, remaining transcripts can be screened for the presence of recognisable protein coding domains / motifs. Pfamscan can be used to compare the transcript sequences to the Pfam databases of all known coding domains (Punta et al. 2012; Li et al. 2016; Li et al. 2017; Deng et al. 2018). Retrotransposon derived elements common in lncRNAs, can however lead to false positives for protein coding potential (Mattick and Rinn. 2015).

Once all transcripts with homology to any known PC genes have been removed, coding statistics can be used to examine and assess the nucleotide sequences directly. This is done on the basis of open reading frame (ORF) presence, size and coverage, Fickett statistics, and Hexamer nucleotide usage (Fickett. 1982; Wang et al. 2013; Li et al. 2016). A number of tools have been developed and can be used to score and assess coding potential. Commonly used tools include: Phylogenetic codon substitution frequency (PhyloCSF) (Lin et al. 2011.), Coding Potential Calculator (CPC) (Kong et al. 2007), Coding-Non-Coding Index (CNCI) (Sun et al. 2013a), and Coding Potential Assessment Tool (CPAT) (Wang et al. 2013). PhyloCSF uses codon substitution frequency (CSF) to differentiate between PC and NC transcripts on the basis of synonymous mutation patterns. PC genes show a strong evolutionary pressure to conserve amino acid content, which is not observed in NC transcripts (Cabili et al. 2011; Lin et al. 2011; Mattick and Rinn. 2015). This approach is however dependent on having a large number of comparison species, the evolutionary history of the plant, and an ORF size – meaning it is often not feasible (Gascoigne et al. 2012). CPC and CNCI are based on support vector machines (SVM) and CPAT uses a logistic regression model with sequence features. As these tools rely on different approaches to assess coding potential, they can be used in combination for added stringency.

The exact threshold for permissible PC potential appears rather arbitrary and varies between studies, with some using a protein coding score cut-off of less than zero, indicating the complete absence of PC capacity (Deng et al. 2018), and others applying a requirement that no ORF >100 (Li et al. 2014) or >120 amino acids (Li et al. 2017) be present. The exact cut-off varies with the level of stringency required as the coding potential score does not directly indicate the presence or absence of a protein product.

More recently, standalone tools have been developed for the direct analysis and identification of lncRNAs from RNA-Seq data. These include iSeeRNA (Sun et al. 2013b) and DeepLNC (Tripathi et al. 2016), but will not be used for this study.

2.1.3. Aims.

The aim of the work presented in this chapter is the construction of a bioinformatics pipeline for the identification of lncRNAs from RNA-Seq data. This pipeline is then used to identify a set of candidate lncRNAs that are present and which may be involved in regulation of vegetative desiccation tolerance in the resurrection plant *Xerophyta humilis.*

2.2. Materials and methods

2.2.1. Plant material

Mature *X. humilis* plants were collected as desiccated mats from Borakalalo National park (North West Province, South Africa) and transported to the University of Cape Town (North West Provincial Government Permit 062 NW-12; Cape Nature Permit AAA007-01733). The plants were transferred into three growth trays containing soil from the collection site, and were rehydrated in the UCT Botany greenhouse. Three weeks prior to desiccation, the trays were transferred into a Conviron Adaptis A350 climate-controlled growth chamber for acclimatisation. The chambers were maintained at a constant 25°C with a 16 hour daylight period and an average luminosity of 250 $\mu mol.m^{-2}.s^{-1}$. Plants were top-watered until media saturation twice a week.

2.2.2. Plant desiccation, Leaf collection and RWC calculation

Plant desiccation was performed under aforementioned growth conditions, by cessation of watering. Leaf harvesting was performed at the same time daily (11:30am) over the course of approximately two weeks, until all plants were fully desiccated. Leaves were harvested by pulling single undamaged leaves from multiple rosettes in each of the three trays. Each individual leaf was immediately torn in half, down the mid-vein. One half was flash frozen in liquid nitrogen and stored at -80°C until RNA extraction could be performed, approximately 2 weeks following the completion of leaf collection. This preserved RNA integrity, from natural degradation and RNases released during tearing, as well as freezing gene expression at the point of harvesting, preventing any post-harvesting or damage-induced expression changes. The second half was weighed (wet weight, W_{wet}) and placed in a drying oven at 60°C. Drying-leaves were weighed multiple times following harvesting, until fully desiccated – at which time their weights remained constant (dry weight, W_{dry}).

To determine the relative water content (RWC) of leaves, before drying, the following formula was used:

$$\mathrm{RWC}\ (\%) = \left(\frac{\mathrm{W_{wet}} - \mathrm{W_{dry}}}{\mathrm{W_{dry}}}\right) - \mathrm{AWC_{avg}} \times 100\%$$

Where AWC_{avg} is average absolute water content (AWC) of the leaves: an estimate of the maximum turgor of a leaf based on its dry mass (Gechev et al. 2013).

AWC_{avg} was estimated by collecting 15 hydrated *X. humilis* leaves, 5 from each tray, before the plants were desiccated. The leaves were immersed in distilled H_2O for 16 hours at 4°C, at 100% humidity in an enclosed Petri dish. The weight of the fully turgid leaves (max weight, W_{max}), wiped dry, was then measured and the leaves were dried as described above (W_{dry}). The individual AWC (AWC_i) for each leaf was calculated using the following formula:

$$\mathrm{AWC_i} = \left(\frac{\mathrm{W_{max}} - \mathrm{W_{dry}}}{\mathrm{W_{dry}}}\right) \quad \mathrm{gH_2O.\,g^{-1}} \qquad \text{(Gechev et al. 2013).}$$

The AWC_i values of all 15 leaves were averaged to give the AWC_{avg} for determination of leaf RWCs.

2.2.3. Experimental design, total RNA extraction and quality assessment of RNA extracts.

Total RNA sequencing was performed to enable both assembly and of the *X. humilis* desiccation transcriptome, including lncRNA transcripts, as well as downstream identification of these lncRNAs. Five stages of desiccation were selected in order to assay for transcriptional changes occurring over the entire course of the vegetative desiccation response. The leaf RWCs selected for RNA-Seq were 100% (Fully hydrated), 80%, 60%, 40% and 5% (fully desiccated). RNA was extracted from leaves that had been calculated to have a RWC closest to and within ±6% of these values.

Total RNA was extracted from individual leaves using QIAzol Lysis Reagent (QIAGEN) and a protocol modified from the manufacturer's instructions. Selected leaves were removed from storage at -80°C and transferred to individual 2ml Eppendorf tubes containing 400µl of QIAzol Lysis Reagent and three stainless steel ball bearings. Ball bearings were pre-cleaned in isopropanol, rinsed in chloroform and air dried in sterile environment before transfer into each Eppendorf tube. Leaf tissue was disrupted by grinding in a Retsch MM400 Oscillating mill at 30Hz for 15 minutes, or until homogenized. 200µl of chloroform and a further 600µl of QIAzol Lysis Reagent was added to each tube, and the tubes vortexed for 30 seconds. The samples were then incubated on ice for 15 minutes and centrifuged at 12000 x g and 4°C for 15 minutes, to pellet any debris and to separate the aqueous and organic phases. The upper aqueous phase (supernatant) was transferred to a new 1.5ml eppendorf tube and an equal volume of 70% ethanol made up in DEPC-treated sterile H_2O was added to each sample, mixed by gentle pipetting. An on-column RNA clean-up was performed using the RNeasy Mini Kit (QAIGEN) with an on-column DNase I digestion using the RNase-Free DNase Set (QAIGEN), as per the manufacturer's instructions. The silica membrane captures RNA longer than 200bp. The purified RNA was eluted in 25µl of RNase-free water. The extracted RNA samples were placed in storage at -80°C, except for a small volume of each retained for quantification and quality assessment.

The individual RNA extractions were analysed, and their concentrations were measured by running 1µl of each sample on a Nanodrop ND-1000 spectrophotometer. The purity and RNA integrity (degradation) was visually assessed using denaturing agarose gel electrophoresis. 1µl of each RNA sample was diluted in 2x volume of RNA sample loading buffer, and denatured at 60°C for 5 minutes, before being electrophoresed on an 1.2% denaturing formaldehyde agarose gel (1.2g agarose, 43 ml H2O, 6ml 10X MOPS and 11ml Formaldehyde) in 1X MOPS at 100V for 30 minutes. The individual total RNA extractions determined to be of sufficiently high quality, were selected for subsequent pooling.

The selected samples were evenly divided and pooled to form three independent biological samples for each of the five RWCs. Where possible, 5 leaf extractions were selected per pool. In reality, pools consisted of between 1 and 5 RNA extracts. Each pool was comprised of equal amounts (µg) of RNA from its constituent RNA extraction samples, with no RNA being shared between pools, to make up a total of between 20 and 30 µg RNA per sample. Following pooling, the concentration and quality of each pool was reassessed by both Nanodrop and gel electrophoresis.

2.2.4. Stabilization of RNA samples with RNAstable

In order to stably transport the RNA for sequencing, single aliquots containing 5µg of total RNA were prepared for each pooled RNA sample and treated with 20µl of RNAstable LD (Biomatrica) – as per the manufacturer's instructions. Samples were gently mixed and dried without heat using a Speedvac Plus SC2 10A (SAVANT) vacuum concentrator. The dried samples tubes were wrapped in Parafilm and sealed in a protective heat-sealed moisture-resistant bag with a separate sachet of silica-based desiccant, ready for transport to Beijing Genomics Institute (BGI) for analysis on a Bioanalyzer (Agilent), library construction and RNA sequencing.

In order to independently assess and verify the integrity of RNA post-transport, replicate samples were also treated with RNAstable LD, desiccated, wrapped in parafilm and stored in an air-tight container with silica desiccant for 7 days. The samples were then resolubilized in sterile DEPC-treated water, frozen and sent for analysis on a Bioanalyzer (Agilent) at the Centre for Proteomic and Genomic Research (CPGR) in Cape Town, South Africa.

2.2.5. RNA Sequencing

The 15 RNAstable-protected samples (5 RWCs, 3 replicates) were sent to the Beijing Genomics Institute (BGI), Shenzhen, China, for sequencing. TruSeq sequencing libraries were constructed using the TruSeq Stranded Total RNA with Ribo-Zero Plant kit (Illumina). Sequencing was performed on an Illumina HiSeq2000 sequencing instrument using a 90bp paired-end amplification protocol to a depth of $40x10^6$ reads per sample. The resulting raw reads were pre-processed by BGI, to remove sequencing adapters and poor-quality sequence, and the clean read data downloaded from the BGI FTP server.

2.2.6. Read quality checking, pre-processing and de novo transcriptome assembly and annotation.

All quality checking, read pre-processing and the assembly of the *X. humilis* desiccation transcriptome was performed by Rafe Lyall, a PhD candidate in our lab at the time (Lyall. 2016). A brief overview of the approach is given in order to give clarity on the origin of the 'Dropset' data from which the lncRNA sequences were subsequently obtained.

Multiple de novo transcriptome assemblies were performed using multiple sets of parameters, Kmer sizes, and assembly tools: Trinity v2014-04-12 and Trinity v2.0.6 (Grabherr et al. 2011), TransABYSS v1.5.2 (Robertson et al. 2010) and Bridger v2014-12-1 (Chang et al. 2015). Of these, eight assemblies were selected (Trinity k=25, Bridger k=21, 25, 31, Transabyss k=25, 41, 61, 81) and merged. From these eight assemblies the best coding representative for each of the assembled transcripts was selected using the Evidential Gene (Evigene) pipeline (Gilbert. 2013; Lyall. 2016). Evigene outputs a primary transcript set (primary coding sequence for each assembled transcript), a secondary transcript set (Isoforms of the primary coding sequences) and a dropset. The resulting *X. humilis* transcriptome, primary, and secondary datasets, were evaluated and annotated by protein coding identity as well as GO terms.

2.2.7. Bioinformatic prediction of lncRNA sequences.

In order to identify a set of putative decoy lncRNAs that may be part of a lncRNA-miRNA network (Chapter 4), a series of bioinformatic filtering and selection steps were applied. An overview of the entire bioinformatic filtering pipeline is given in Figure 2.1 below.

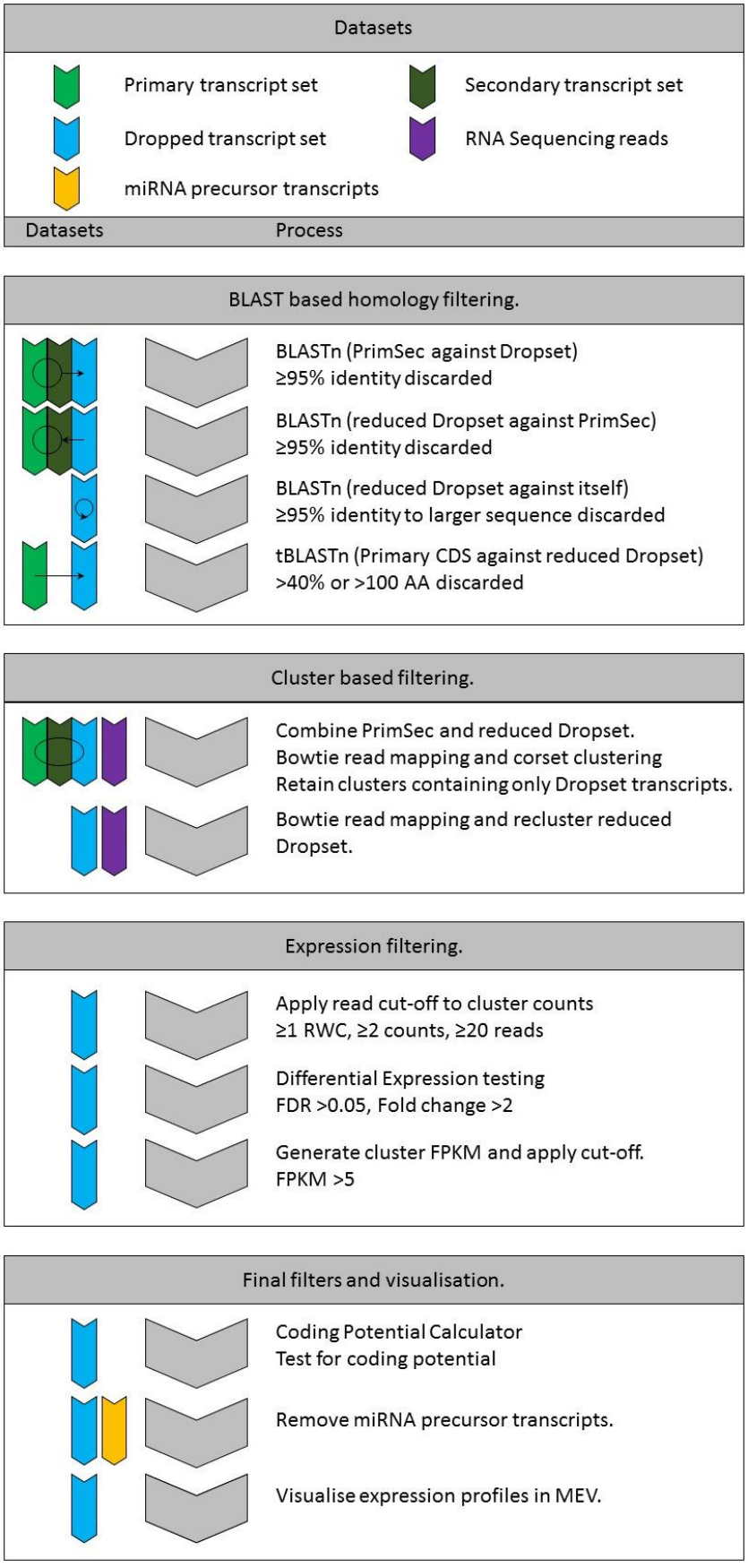

Figure 2.1: Overview of the lncRNA filtering and prediction pipeline.

2.2.7.1. Removing redundancy to dropset and coding PrimSec sequences.

The transcriptome dropset contains all sequences not in the primary or secondary datasets. It includes duplicates and fragments of coding sequences, misassemblies, intronic contamination, chimeric transcripts as well as all non-coding (<50% CDS) transcripts. To remove all transcripts similar to coding transcripts present in the primary + secondary (PrimSec) transcriptome assembly, all dropset nucleotide sequences were compared to the PrimSec dataset using BLASTn (evalue >$1x10^{-5}$, strand = plus). This was performed in two rounds, as shown in Figure 2.2. First all PrimSec sequences were blasted against the dropset sequences, removing all dropset sequences with query coverage ≥95%. This removed all dropset sequences containing a match to an almost full length PrimSec sequence. Secondly, the remaining dropset sequences were blasted against the PrimSec database, and all dropset sequences with query coverage ≥95% were discarded. This removed all almost full length dropset sequences that mapped within a PrimSec sequence.

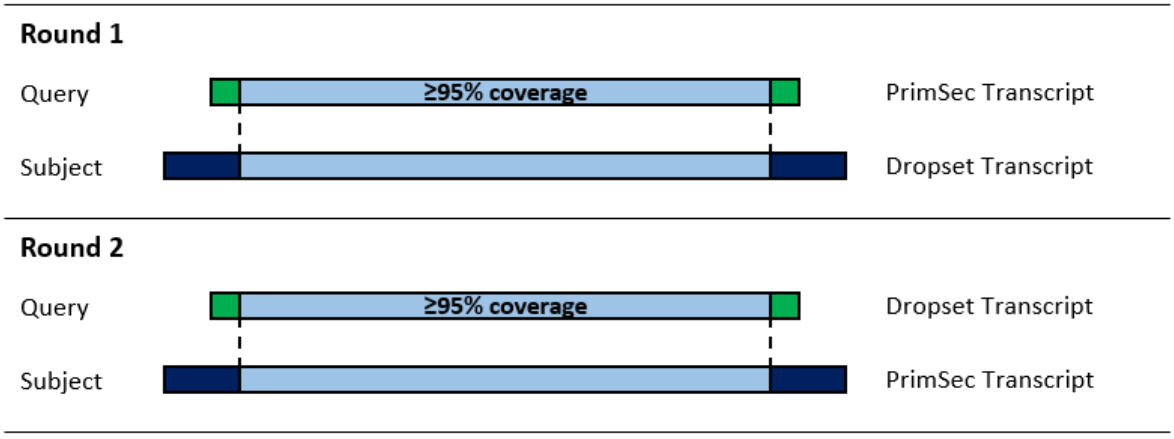

Figure 2.2: BLASTn strategy for removing dropset sequences with partial or full homology to PrimSec sequences. Two rounds of BLASTs were performed: 1) PrimSec sequences were blasted against the dropset sequences and 2) the remaining dropset sequences were blasted back against the Primsec sequences. All dropset sequences with ≥95% query coverage were discarded in each round. All BLASTS were performed with an Expect (E) value cut-off of

The remaining dropset sequences were blasted back against themselves. All transcripts with ≥95% sequence identity to a larger sequence in the database were removed, resulting in a non-redundant set of dropset specific transcript sequences.

The 95% similarity cut-offs however, may not remove chimeric sequences or coding transcripts with wildly differing or incorrectly assembled 5' and 3' terminal regions / untranslated regions (UTRs) relative to the PrimSec transcripts. In order to remove these chimeric sequences, the predicted coding sequences (CDS) of the assembled Primary transcripts were blasted (tBLASTn) against the remaining

non-redundant dropset sequences (Figure 2.1). Only the Primary transcripts CDSs were used, as any open reading frames (ORFs) found in the secondary transcript set should be truncated variants of those found in the primary transcript set. All dropset sequences with hits to a predicted ORF, in the positive frame, with >40% sequence identity, or a match of >100 amino acids, were removed. All BLAST runs were performed with an Expect (E) value cut-off of $1x10^{-5}$ and a maximum of 20 reported hits.

2.2.7.2. Bowtie 2 Mapping and Corset Clustering

In order to determine the expression levels of the putative lncRNAs, the pre-processed and error-corrected paired-end sequencing reads for each of the 15 samples (5 RWCs, 3 biological repeats) were mapped to a combined index of the remaining dropset and all PrimSec sequences using Bowtie2 (v2.2.4; Langmead and Salzberg. 2012). The parameters "--fr --nofw --no-mixed --no-discordant" were used in order to account for the paired-end reads as well as the stranded RNA-Seq libraries. The parameters "--end-to-end" and "--all" were also used, the latter to report all alignments for each read. The BAM alignment files for all 15 samples were then used in conjunction with Corset (v1.03; Davidson and Oshlack. 2014), to simultaneously cluster the combined set of transcripts on the basis of both shared sequence identity, as well as expression (single clustering event). These clusters of similar transcripts were each given an identifier, and for every cluster, total read counts generated corresponding to each of the 15 RNA-Seq samples. Default Corset settings were used.

The resulting clusters fell into one of three categories: 1) dropset-specific clusters containing only dropset transcript sequences, 2) PrimSec-specific clusters containing only PrimSeq transcript sequences, and 3) shared clusters containing both PrimSec and dropset sequences. Dropset transcripts clustering with transcriptomic transcripts are less likely to be of interest and were removed.

The sequencing reads were then remapped to the remaining dropset transcripts, as done previously. The dropset transcripts were reclustered and cluster counts were generated (15 raw read counts for each cluster).

2.2.7.3. Raw Read Count Cut-off.

One of the properties of lncRNAs, that has been identified, is that they exhibit higher natural expression variation than PC genes, with some expressed at low levels (relative to mRNAs)

originally raising questions about their biological significance (Kornienko. 2016). I reasoned that the very lowly expressed lncRNAs were unlikely to play a major role in VDT, and the dataset was simplified by applying a read count filter to remove these sequences. A total read count was not used as it fails to take into account the distribution of contributing reads between samples. Instead, all clusters were removed that did not possess at least one RWC with at least 2 replicates of at least 20 reads each.

2.2.7.4. DESeq2 Analysis

2.2.7.4.1. Principle component analysis

RNA-Seq read counts were used to perform a PCA analysis on the entire set of 15 RNA-Seq datasets in order to assess the robustness of the datasets. These datasets include all sequenced reads, including both reads assembled into primary, secondary or dropset transcripts, as well as all unassembled reads. Read counts were regularised-log transformed (rlog, blind=TRUE) and a two axis Principle Component Analysis (PCA) performed using DESeq2 (v1.10.1; Love et al. 2014). The regularised-log transformed (rlog, blind=TRUE) cluster count data for the remaining dropset-only clusters was also compared using a two axis Principle Component Analysis (PCA), in order to compare the remaining data subsets across the 5 RWCs and between replicates. Both PCAs were performed using default DESeq2 settings.

2.2.7.4.2. Differential Expression testing and fold change cut-off.

Differential expression analysis was performed for all remaining dropset clusters, comparing expression between the 5 RWCs, using DESeq2 (v1.10.1; Love et al. 2014). The raw read counts for each cluster were used as DESeq2 performs its own read count normalisation and size correction. Differential expression analysis was performed using a log-likelihood ratio test (LRT), and default DESeq2 parameters. Sequences were designated as differentially expressed if the false discovery rate (FDR) was less than 0.05. The DESeq2 normalised cluster read counts for the DE clusters were obtained. DESeq2 was run in Rstudio Desktop (v0.99.892), utilizing the x64 Windows version of R (v3.2.3; R Development Core Team 2015) and the latest version of DESeq2 (v1.10.1; Love et al. 2014), with all required dependencies.

For each DE dropset-cluster, the average normalised read count was calculated for each of the 5 RWCs and the fold change (FC) between the RWCs with the highest and lowest count determined. All clusters with a FC greater than 2 were retained.

2.2.7.5. Filtering by FPKM.

The selection of differentially expressed transcripts with high read counts allows transcripts likely to be of little or no biological significance to be discarded. Due to the long length and length variability of the remaining lncRNA transcripts, and lncRNAs in general, it was decided that Fragments Per Kilobase of transcript per Million mapped reads (FPKM) would provide a more uniform and unbiased strategy for exclusion of low expression transcripts, than raw read counts alone. As such the FPKM values for each cluster were calculated. In order to determine FPKM, a transcript length, and hence a representative sequence, was required for each cluster. For clusters of two or more transcripts, the longest transcript was identified. Concurrently, the sequence assembly program CAP3 was used to generate a consensus sequence for every cluster (Huang and Madan. 1999). The longer of these two sequences was selected as the representative sequence for that cluster, and its length was used to calculate the cluster FPKM values. The FPKM values for each cluster for each sample were calculated using the following formula:

$$\mathrm{FPKM} = (10^9 \times \mathrm{C}) \div (\mathrm{N} \times \mathrm{L})$$

Where C is the number of paired-ends mapped to the gene (read count / 2), N is the library size (total mappable reads) and L is the number of base pairs in the gene (transcript length). All clusters with their highest individual FPKM ≤ 5 were discarded, so as to reduce the size of the dataset.

2.2.7.6. Coding Potential Calculator

All individual transcripts from the filtered dropset clusters were assessed with regards to their own protein coding potential using Coding Potential Calculator (v0.9) – which assesses the protein-coding potential of a transcript based on six biologically meaningful sequence features (Kong et al. 2007). CPC was run locally using default settings and UniREF90 as the reference protein database (Suzek et al. 2014; www.uniprot.org). UniREF90 was selected as all entries have been manually reviewed and the smaller size of the collapsed database allows for faster analysis, with minimal loss of the power of homology detection. All sequences designated as coding sequences were discarded.

2.2.7.7. Removing putative miRNA precursors.

The combined set of all predicted miRDeep-P and ShortStack precursor sequences (Chapter 3) were blasted (BLASTn) against the indexed set of remaining transcript sequences. Transcript hits covering at least 90% of both the query (precursor) and subject (nc transcript) with >99.5% sequence identity were taken to be precursor miRNA sequences and were discarded from the remaining set of candidate lncRNA sequences.

2.2.7.8. Gene expression clustering

The expression vectors for the remaining DE candidate lncRNA genes were predicted using Multi-Experiment Viewer (MeV v4.9.0; www.tm4.org). The normalised read count data, obtained from DESeq2 and averaged for each set of 3 RWC replicates, were used as input for MeV. The expression levels for each transcript were mean-centred and normalised to facilitate visualisation. The expression profiles were then clustered using K-Means clustering (Pearson correlation, 5 clusters, 50 iterations). To further compare the overall relation between transcript expression over the 5 RWCs, hierarchical clustering was used to predict a sample tree (Pearson correlation, average linkage clustering) for the 5 RWCs.

2.3. Results

2.3.1. Plant material.

In order to calculate the RWCs of individual leaves for RNA extraction and pooling, the global AWC was first determined. The AWCs of the 15 individual *X. humilis* leaves was found to be highly variable, with the average of the 15 samples being 3.4 $gH_2O.g^{-1}$ of leaf tissue (± 0.66 SD) (Table 2.1.). While minor variability is expected for biological systems, this high level of variance between samples may unavoidably impede a precise and accurate determination of true leaf RWCs, possibly leading to the less precise pooling of leaves than desired.

Table 2.1: Absolute water contents (AWC) of individual *X. humilis* leaves from each of the three collection trays. The 15 individual AWCs used to determine the global AWC (± SD) for calculating leaf RWCs.

	Absolute Water Content ($gH_2O.g^{-1}$)					
	Leaf 1	**Leaf 2**	**Leaf 3**	**Leaf 4**	**Leaf 5**	**Average**
Tray A	3,64	3,36	3,55	2,24	2,29	3,02
Tray B	4,26	4,13	3,72	3,40	3,11	3,73
Tray C	3,68	4,18	3,81	3,24	2,37	3,45
					AWC	**3.4 ± 0.66**

Table 2.2: Leaves collected from dehydrating *X. humilis* plants.

Leaves collected	
Tray A	103
Tray B	89
Tray C	93
Total	285

Additional Leaves were collected from dehydrating *X. humilis* plants, once daily at 11:30am for a period of 11 days. In total, 285 individual leaves were collected (Table 2.2), the desiccation curve for which is given in Figure 2.3. Following leaf collection, leaves were selected and grouped into five bins corresponding to 100%, 80%, 60%, 40% and 5% RWC (±6% RWC), for subsequent RNA extractions. Total RNA was extracted from individual leaves, and quality assessed via Nanodrop and gel electrophoresis. Suitable samples showing clear RNA bands, little to no degradation and low contamination, were pooled to form three independent biologicals, for each of the five selected RWCs (Table 2.3, Figure 2.4). While 5 individual leaf samples were desired per pool, some pools were comprised of as little as 1 sample. More leaves were not collected to ensure all samples originated from a single desiccation event. The quality of RNA for the pooled samples were assessed on a Bioanalyzer.

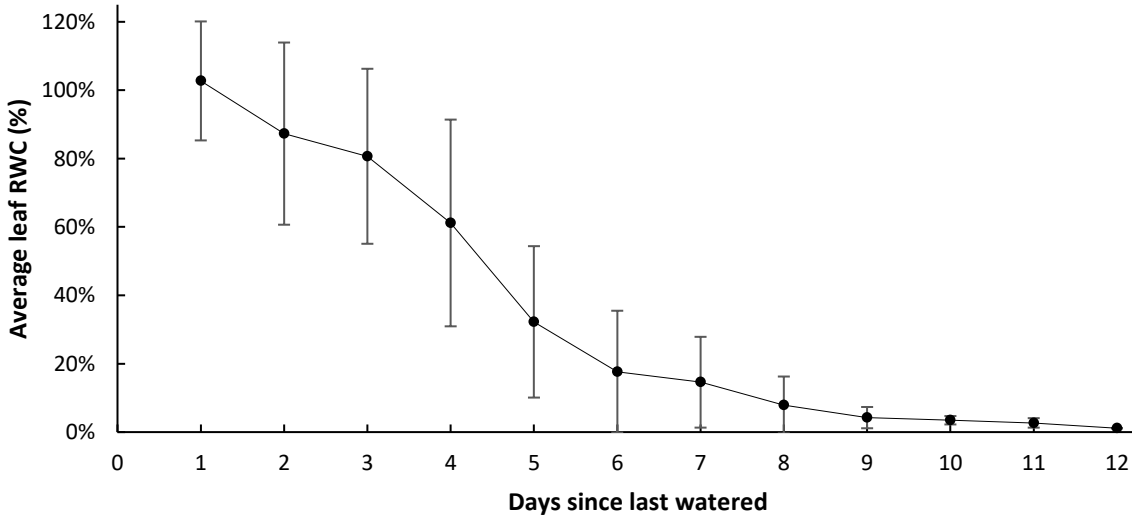

Figure 2.3: Desiccation curve for *X. humilis* leaves, showing the average RWC for all *X. humilis* leaves collected on each day following the cessation of watering. Error bars show the standard deviation from the mean.

Table 2.3: Leaf samples for RNA Pooling. RNA extractions corresponding to the indicated leaves, letter in brackets denoting the source tray, contributed equally (ng) to each pool. The RNA integrity number as determined by BGI is given for each pooled sample.

Pool	100% RWC			80% RWC			60% RWC			40% RWC			5% RWC		
	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl
A	(A)003	107,5%	352,6	(B)034	82,0%	212,0	(A)010	60,5%	324,2	(A)027	40,4%	94,9	(A)091	5,1%	435,3
	(A)009	101,3%	165,4	(B)035	79,0%	291,3	(A)005	55,8%	229,9	(C)039	40,5%	793,5	(A)102	4,9%	119,5
	(A)002	100,7%	293,3							(B)013	37,3%	98,4	(A)045	4,8%	346,1
	(A)008	97,5%	398,9							(B)053	36,8%	108,5	(A)099	4,6%	193,9
													(A)005	4,5%	291,5
	RIN: 6,0			RIN: 4,6			RIN: 5,5			RIN: 5,6			RIN: 6,1		
B	(B)019	112,6%	160,9	(B)032	77,3%	264,9	(C)003	64,3%	328,0	(A)031	40,0%	422,9	(B)066	6,7%	437,2
	(B)002	105,0%	429,2	(C)001	85,2%	391,8				(C)032	37,3%	951,3	(B)067	5,7%	434,8
	(B)012	102,4%	287,7							(B)052	40,9%	250,0	(B)070	5,2%	644,6
	(B)001	99,5%	463,5							(B)057	38,2%	648,0	(B)083	5,1%	553,4
	(B)018	97,5%	283,5										(B)092	4,4%	362,7
	RIN: 5,7			RIN: 4,7			RIN: 5,2			RIN: 5,6			RIN: 6,4		
C	(C)009	120,7%	804,2	(B)030	75,0%	575,1	(A)006	59,4%	594,5	(A)020	36,0%	84,6	(C)097	5,4%	628,8
	(C)002	108,7%	707,3	(C)011	82,6%	734,4	(C)018	63,7%	791,5	(C)033	36,2%	210,5	(C)074	5,1%	785,4
	(C)007	108,4%	418,4							(B)061	43,7%	326,5	(C)066	5,0%	444,5
	(C)005	100,8%	567,0							(A)033	35,1%	184,7	(C)049	4,6%	522,5
	(C)012	100,7%	629,6										(C)047	4,6%	665,8
	RIN: 5,1			RIN: 3,5			RIN: 5,1			RIN: 5,9			RIN: 6,0		

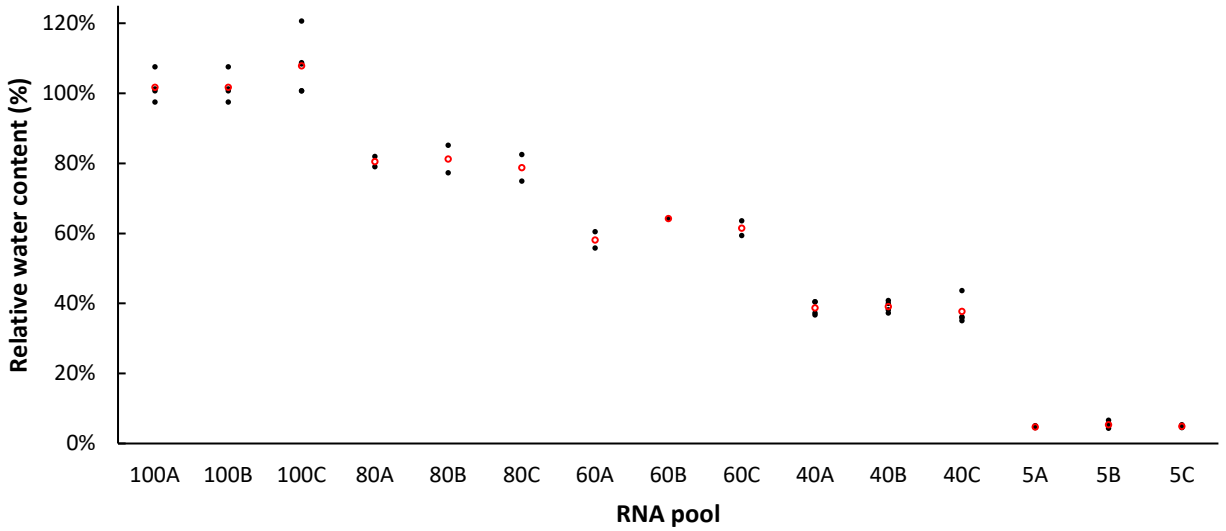

Figure 2.4: Leaf sample RWCs for RNA pooling. The RWC of individual contributing leaves are shown in black, with the average RWC for each RNA pool shown in red.

The RNA Integrity Number (RIN) values obtained from BGI were low, relative to the desired value being >6.5 (BGI), especially the 80% pooled samples. BGI uses a mammalian RNA standard for their Bioanalyzer (Agilent) and calculation of RIN scores. This was expected to skew obtained RIN values. In order to reassess the quality of samples, as well as to compare Bioanalyzer results using plant and animal standards, selected samples were reanalysed on a Bioanalyzer (Agilent) at the CPGR, using a grape vine leaf RNA standard.

The six individual component RNA extractions for the three pools with the lowest BGI RIN scores (80% RWC pools), the pool (60% RWC, Pool C) with the highest BGI RIN score, and a pool (5% RWC, Pool B) with an intermediate BGI RIN score, were removed from storage at -80°C and the frozen samples were directly delivered to CPGR for analysis. The BGI and CPGR results are compared in Figure 2.5.

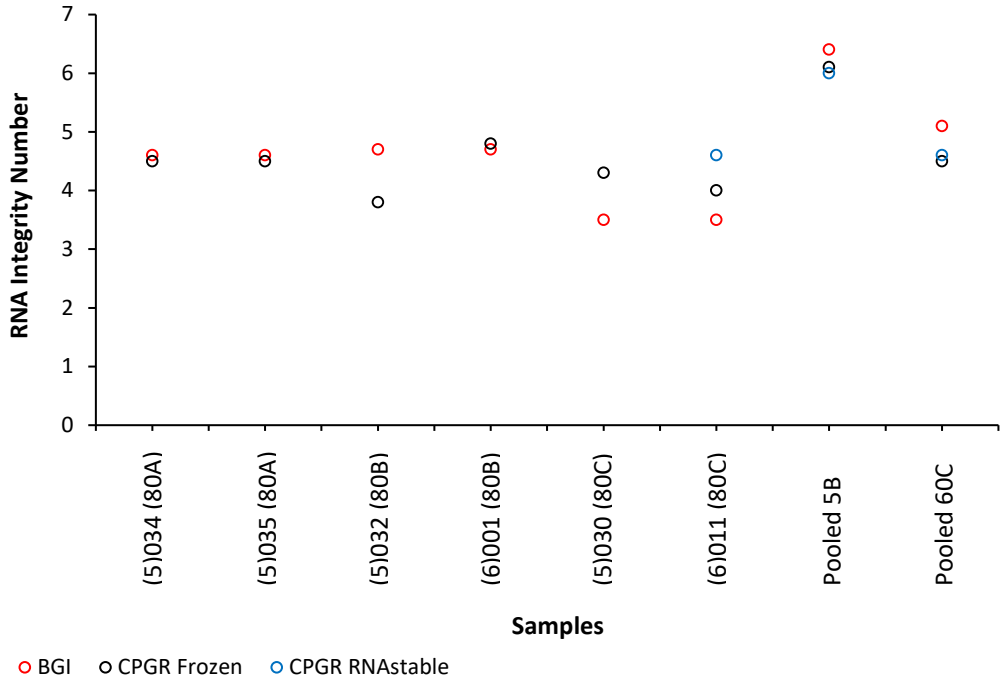

Figure 2.5: BGI and CPGR RIN scores for comparison of plant and animal standards, as well as assessment of RNAstable effectiveness. Pooled RNA, as well as both Pooled RNA and individual RNA extracts were sent to BGI (Red) and CPGR (Black) respectively for RNA integrity analysis and RIN determination. BGI uses an animal standard for RIN determination, while CPGR used a grape leaf RNA extract as standard. To test the effectiveness of RNAstable, RNA samples stabilized with RNAstable, stored and then resolubilized were also sent to CPGR for assessment (Blue). The BGI RIN scores for pooled samples are used as proxy for the RIN scores of contributing extracts.

The RIN scores obtained from CPGR do not differ largely from the BGI RIN scores, nor do they all shift in a specific direction. This suggests that the low RIN values obtained are not purely a result of the type of standard used. The RIN calculation makes use of the ratios of the 28S and 18S ribosomal RNA (rRNA) peaks. In plants however, three types of ribosomal RNA are present; mitochondrial, chloroplastic (23S, 16S) and nuclear (28S, 18S). RIN analysis of green plant samples on a Bioanalyzer usually results in low RIN values due to the chloroplastic rRNA, and there is no sure assay to find accurate RIN numbers for green plants. Furthermore, variation in RIN values can differ between tissues, developmental stages, species and RNA extraction techniques, and many not reflect the true integrity or degradation of the RNA samples (Johnson et al. 2012). This is likely the reason for the poor RIN values. Treatment and storage with RNAstable was also not found to negatively impact RNA integrity. The BGI and CPGR Bioanalyzer electropherograms showed RNA clear peaks and Bioanalyzer electrophoresis images, as well gel electrophoresis of the pooled samples showed multiple clean RNA

bands with minimal degradation – indicating high RNA integrity. On account of the gel visualisation and Nanodrop results, the RNA was deemed suitable for subsequent processing and sequencing.

2.3.2. RNA sequencing and de novo transcriptome assembly.

Following pre-processing, to remove adapter sequence and low-quality reads, over 350 million paired-end sequences, spread across the 15 pooled samples (5 RWCs with 3 biological repeats), were downloaded from the BGI servers. These datasets were handed over to Rafe Lyall, then a PhD candidate in our lab, for subsequent processing and analysis. All details of how the data was handled and the de novo transcriptome assembly performed can be found in his PhD thesis (Lyall. 2016). The final combined transcriptome assembly consisted of 2,611,123 transcripts divided into three datasets, as designated by EviGene. A "primary" set of 72 893 (2.8%) putative transcripts, containing the most complete predicted ORF-containing sequences, a "secondary" set of 93 478 (3.6%) sequences with >50% coding sequence similarity (putative isoforms or transcript variants) to the "primary" set, and all remaining 2 444 752 (93.6%) contigs in a "dropped" set, as shown in Figure 2.6.

2.3.3. lncRNA filtering

2.3.3.1. Remove partial PrimSec sequences and dropset duplicates.

In order to remove all duplicates and fragments of PrimSec coding sequences from the dropset, the PrimSec transcript sequences were blasted against the dropset sequences, removing all dropset sequences with query coverage ≥95%. The remaining 869 982 (35.6%) dropset sequences were blasted back against the PrimSec sequences, with 562 674 (23.0% of dropset) remaining after again removing transcripts with ≥95% coverage. The large reduction in the number of transcripts indicates that the majority of the dataset was partially assembled fragments of transcripts existing in the PrimSec datasets. To remove redundancy within the remaining transcript set, the transcripts were blasted against themselves, and all transcripts with ≥95% sequence identity to a larger sequence in the dataset were removed, leaving 200 845 (8.2% of dropset) transcripts. Remaining chimeric sequences, or sequences with wildly differing or incorrectly assembled terminal/untranslated regions (UTRs), relative to the PrimSec transcripts, were removed by blasting the Primary transcript predicted coding sequences against the dropset sequences (tblastn), removing positive frame hits with >40% sequence identity or a match of >100 amino acids. This left 157 154 (6.4%) remaining dropset sequences. The contribution of each filtering process is shown in Figure 2.7 below.

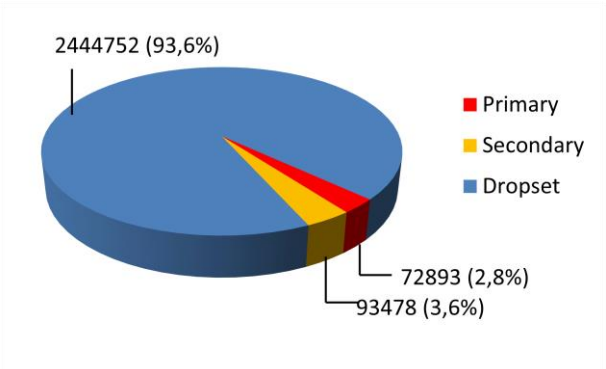

Figure 2.6: Categorisation of all assembled transcripts from the merged *X. humilis* de novo transcriptome assemblies. The Primary, Secondary and Dropped transcript sets consist the most complete predicted ORF-containing sequences, all sequences with >50% coding sequence similarity (putative isoforms or transcript variants) to the "primary" set, and all remaining assembled transcripts respectively. Any lncRNA sequences will be found in the dropset.

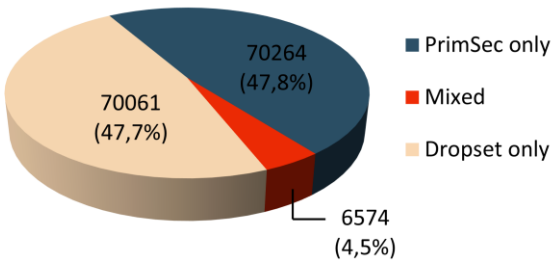

Figure 2.8: Corset clustering of PrimSec and dropset transcripts. Clusters contain either only PrimSec sequences, only dropset sequences, or a combination of the two. "Dropset only" sequences were retained.

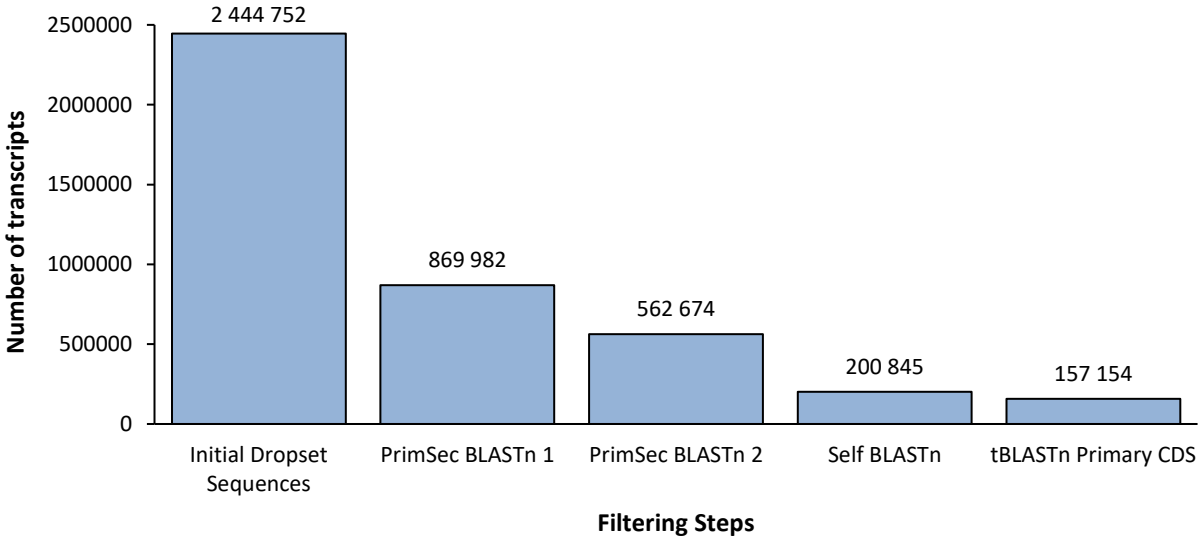

Figure 2.7: Dropset filtering pipeline for PrimSec complementarity and transcript redundancy. The number of initial dropset transcripts as well as the number of transcripts remaining following each filtering step are shown. PrimSec transcripts were blasted (BLASTn) against dropset sequences, retaining dropset transcripts with <95% query complementarity (PrimSec BLASTn 1). Remaining dropset sequences were blasted back against all PrimSec transcripts, again retaining dropset transcripts with <95% query complementarity (PrimSec BLASTn 2). The dropset was blasted against itself removing transcripts with ≥95% sequence identity to a larger sequence in the dataset (Self BLASTn). tBLASTn was used to blast the predicted coding sequences of the primary transcripts against the dropset transcripts, removing positive frame hits with >40% sequence identity or a match of >100 amino acids (tBLASTn Primary CDS).

It is clear that the dropset resulting from the transcriptome assembly pipeline contained a diverse variety of redundant sequences of multiple possible types and origins: misassemblies, intronic contamination, chimeric transcripts, duplicates and fragments of larger coding sequences. This is an expected result of the assembly strategy in which the final Transcript set resulted from merging 8 assemblies, by separate 3 assembly tools using a number of different assembly parameters (Kmer sizes), which resulted in a high number of similar and duplicate transcripts being assembled (Lyall. 2016). Less than 6.5% of the dropset represented new, non-redundant sequences, not found in the PrimSec set of PC trancripts. By removing all sequences with homology to the PC transcripts, either by nucleotide or translated sequences, the vast majority of PC transcripts were expected to have been removed.

2.3.3.2. Filtering by transcript clustering

Sequencing reads were aligned to the combined set of PrimSec and remaining dropset transcripts, and Corset was used to cluster the transcripts on the basis of shared sequence identity and expression. The sequencing reads from the 15 RNA-Seq samples aligned with an average overall alignment rate of 85.22% (± 1.18% SD), with 157 154 transcripts clustering into 146 899 separate clusters. These consisted of 70 264 (47.8%) PrimSec clusters, 6 574 (4.5%) clusters comprised of both PrimSec and dropset transcripts, and 70 061 (47.7%) clusters containing only dropset transcripts, as shown in Figure 2.8. Reads were realigned to the 130 418 sequences from the clusters containing only dropset sequences, with an average overall alignment rate of 16.98% (± 1.83% SD), and the sequences re-clustered into 70 242 “dropset-only” clusters. The removal of PrimSec trancripts and the re-allocation of ambiguous reads, results in the seperation of transcripts into 181 new clusters.

2.3.3.3. Filtering of dropset clusters by read count.

The total raw read counts from each of the 15 sRNA-Seq datasets (5 RWCS, 3 replicates) were generated for all “dropset-only” clusters by summing the individual contributions of all transcripts within a cluster. Clusters with low and inconsistent expression were excluded by removing all clusters that did not possess at least 1 RWC stage with at least 2 samples of ≥20 reads. This excluded 54.5% of clusters, leaving 31 983 clusters containing 81 483 putative lncRNA sequences.

2.3.3.4. Principle Component Analysis

In order to test the quality of the initial sampling and datasets, as well as the robustness of the replicate samples for each RWC, the regularised-log transformed (rlog) raw read data for the RNA-Seq datasets was compared using Principal Components Analysis (PCA). The two principle components most able to explain the data are plotted in Figure 2.9A. Generally, biological replicates for each of the 5 RWCS are expected to cluster together. Likewise, RWCs with similar patterns of transcript expression are expected to group closer together. As seen, the vast majority (81%) of the variance between the initial 15 RNA-Seq datasets is explained by Principle component 1 (PC1), with all RWCs appearing to be arranged along PC1 (left to right) by increasing water content. Low RWCs (5%, 40%) cluster together to the far left, with the higher RWCs (60%, 80%, 100%) clustering to the far right. The separating of samples between the 40% and 60% RWCS indicates that major transcriptional changes may occur between these stages, which is consistent with previous findings indicating that around 60% RWC (Lyall. 2016) is the point at which the major transcriptional changes dictated by the desiccation tolerance programme come into effect. The ordered separation of samples by RWC also suggesting a fairly linear progression of gene expression changes. Each RWC most closely resembles (fewest changes in transcripts and transcript abundance) 'neighbouring' stages.

All RWCs, bar the 80% samples, cluster together and away from the other RWCs indicating that the replicates show robust behaviour with very similar read identity and expression within each RWC. This indicates good datasets for lncRNA prediction. The 80% samples do not cluster together, with 80B falling along PC1 closest to the 100% replicates, 80C falling closer to the 60% replicates and 80A falling between the 40% and 60% replicates. The source of the variation between these samples is unclear, and while the 80% replicates do have a low number of pooled leaves (2 each), possibly indicating separation is a biological phenomenon not masked due to the low number of leaves per sample, the 60% samples, which also have few contributing leaf extracts per sample, show good grouping. It is possible 80% RWC may also be a point of transcriptional changes, with some samples falling on either side of a "transcriptional watershed". The datasets proved to be of sufficient quality for transcriptome assembly and analysis (Lyall. 2016) and were deemed of suitable quality for lncRNA prediction and analysis.

In order to assess the subsets of the original read data retained in the filtered putative lncRNA clusters, a second PCA was performed on cluster read counts (Figure 2.9B). The same clustering patterns are observed, with similar separation across almost identical PCAs (percentage contribution to variance)

and with a decrease in variation between replicates. This indicates the filtered lncRNA dataset is still representative of the original 5 RWCs.

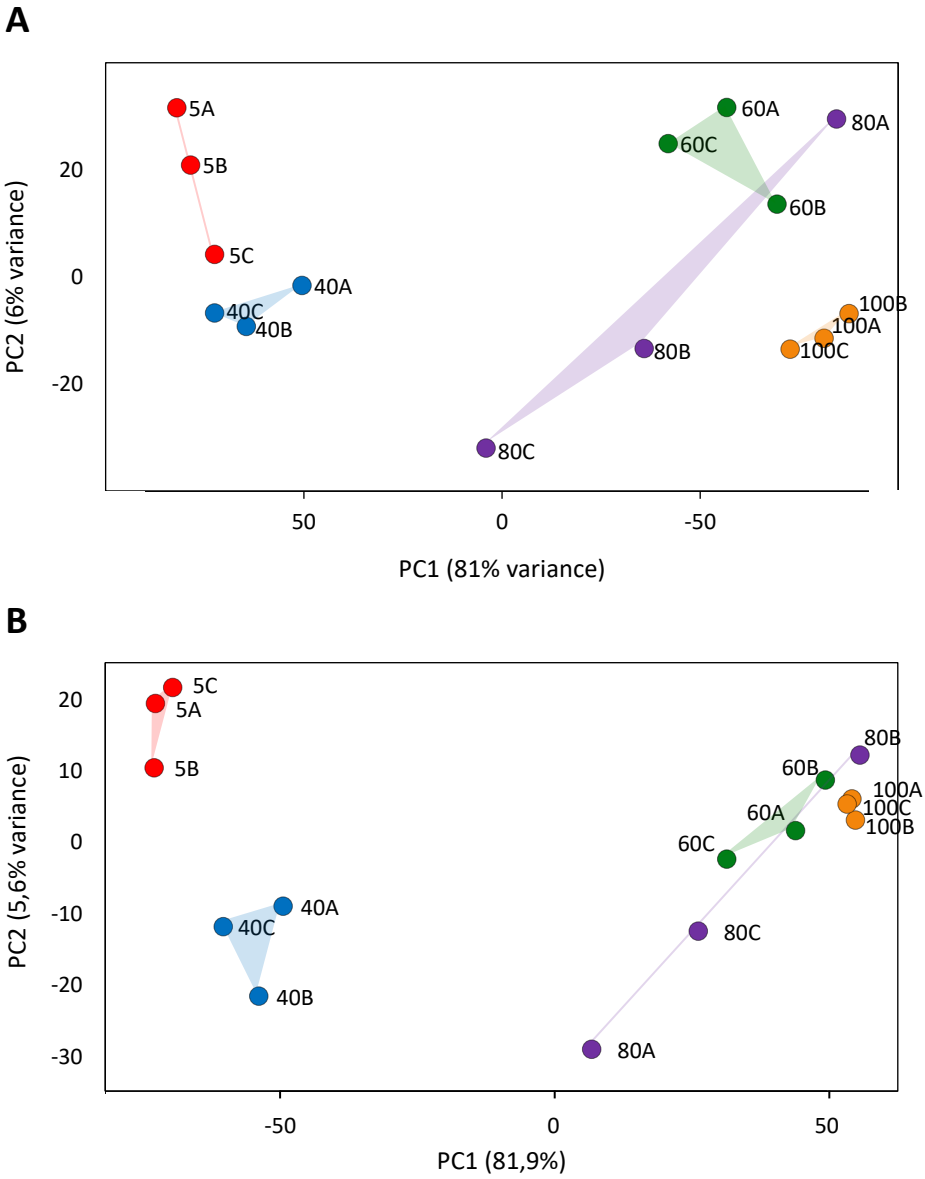

Figure 2.9: Principle component analysis of the regularised log read counts for the original 15 RNA-Seq libraries (A) and the filtered putative lncRNA cluster datasets (B). For the filtered lncRNA datasets, raw read counts were generated by Bowtie2 for all lncRNA transcripts. Transcript read counts were summed to obtain the cluster read counts for each of the 15 RNA-Seq samples. Clusters without at least 1 RWC with 2 or more replicates of at least 10 reads each were removed. For A and B, counts were converted to regularized log (rlog) values using DESeq2 (blind=TRUE), and the two primary principle components used to create the diagnostic PCA plot. The PCA plot for the 15 RNA-Seq libraries (A) has been reversed across PC1 to facilitate visual comparison to the PCA for the filtered putative lncRNA cluster set (B).

2.3.3.5. Differential expression testing and Fold-change cut-off.

Differential expression is a key indicator of possible lncRNA functionality and may be indicative of either miRNA induced degradation or time-specific functional expression. The remaining dropset clusters and their respective raw read counts were assessed for differential expression using the R DESeq2 package from Bioconductor. Differential expression was tested for, using a log-likelihood ratio test (LTR), testing for lncRNA gene clusters that are differentially expressed between any of the five stages of desiccation. A significance cut-off was selected at a false discovery rate (FDR) less than 0.05. Of the 31 983 clusters remaining after the read count filter, 97 (0.3%) were discarded by DESeq2 as outliers (FDR cannot be determined) and 0 were discarded for having read counts too low to reliably determine significance (mean count <3) as any such clusters were removed during read count filtering. 21 885 (68.6%) of the remaining clusters were found to be differentially expressed over the course of desiccation, at an FDR <= 0.05. This corresponds to 56 958 individual transcripts. 11580 (52.9%) of these clusters showed up regulation (LFC>0) over the course of desiccation, while 10305 (47.1%) were found to be down regulated (LFC<0). 18476 of these DE clusters have an FDR <= 0.05. The high number of transcripts showing differential expression indicates a potential key role played by lncRNAs as overall modulators and/or regulators of the *X. humilis* desiccation programme of gene expression, but also indicates further filtering may be required.

DESeq2 was also used to normalise the raw cluster read counts. For each cluster, the normalised read counts were averaged within each RWC and the fold change (FC) between RWCs with the highest and lowest average read counts determined. 10809 (49.4%) of the DE clusters were found to have a FC>2 and were retained for further filtering and analysis.

All remaining clusters therefore showed both statistically significant differential expression and an absolute fold change greater than 2.

2.3.3.6. Applying the FPKM cut-off

In order to apply a more stringent cut-off, unbiased by transcript size, the 15 raw read counts for each cluster were converted to FPKM values. Clusters without a single FPKM value greater than 5 were discarded as being of low biological significance. The remaining 8011 (37.8%) clusters contain 18165 putative lncRNA sequences. All sequences within these clusters were taken as being differentially expressed as it is almost impossible to separate out the expression of some of these transcripts, given

the high extent of shared sequence identity. Each individual transcript was assigned their respective normalised cluster counts as a proxy for their own individual read counts.

2.3.3.7. Coding Potential Calculator

The lack of protein coding potential is a defining characteristic of all lncRNAs. While the majority of PC sequences are expected to have been removed during the initial PrimSec transcript filtering step, the high number of differentially expressed transcripts suggests that there may still be transcripts, other than lncRNAs, present in the dataset. To ensure any remaining transcripts with coding potential are removed, and to ensure all putative lncRNA transcripts meet the required non-coding criteria, all 18165 remaining putative lncRNA transcripts were independently assessed for intrinsic protein coding potential using Coding Potential Calculator (CPC), and Uniref90 as the reference database. 15192 (83.6%) of sequences were found to be non-coding (NC), using default parameters. This corresponds to 2.7% of original transcriptome dropped set sequences. We can be very confident that all PC transcripts have been effectively removed.

2.3.3.8. Remove miRNA precursor (filter 15 miRDeep precursors)

All putative precursor miRNA (pre-miRNA) transcripts predicted by both miRDeep-P and ShortStack (Chapter 3) were blasted against the NC transcripts, removing transcripts with hits covering at least 90% of both the query (precursor) and subject (nc transcript) with >99.5% sequence identity. Very few possible pre-miRNAs were found with only 7 transcripts, from 7 clusters, being removed.

The results for all filtering steps – from cluster filtering to the removal of pre-miRNAs – are given in Figure 2.10 below.

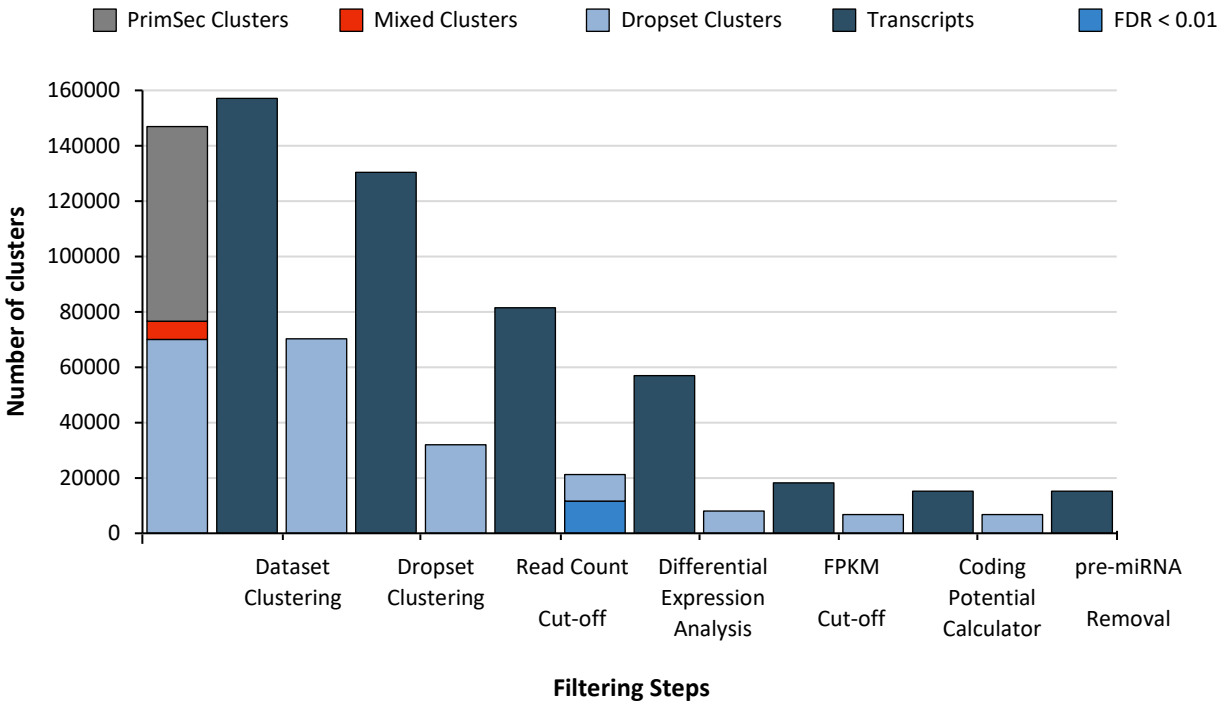

Figure 2.10: lncRNA prediction following removal of redundant sequences and PrimSec homologs. The number of sequences (Dark blue) and clusters (other colours) remaining after each filtering step are shown. All PrimSec and remaining dropset sequences were clustered by sequence identity and expression. Clusters containing only dropset sequences were re-clustered. All clusters without ≥1RWC, with ≥2 samples of ≥20 raw read were discarded. DESeq2 was used to assess all clusters for differential expression (FDR <0.05) between the 5 RWCs. Normalised read counts were converted to FPKM values and a FPKM cut-off of >5 was applied. Remaining sequences were individually assessed for coding potential (CPC), retaining only non-coding sequences. Any sequences matching predicted miRNA precursors were then removed.

2.3.3.9. Visualisation of expression profiles in Multiple Experiment Viewer.

The remaining DE lncRNA genes were clustered into 5 expression profiles using the K-means algorithm in MeV (Fig 2.11). The expression levels for each transcript were mean-centred and normalised to facilitate visualisation of expression changes (rather than absolute expression).

Clusters A and C both show consecutive waves of expression in an apparent biphasic positive-expression pattern. Cluster A, the largest cluster, shows expression when fully hydrated (100% RWC), mid-desiccation (60%) as well as transcripts being present during complete desiccation (5% RWC). As leaves at 5% RWC are in a state of anhydrobiosis and expression isn't taking place these may represent transcripts expressed late during desiccation, needed and stored in preparation for the rehydration process. Expression by many of these transcripts at 60% RWC suggests they may play a key role regulating desiccation tolerance pathways. Cluster C shows an induction of transcript expression

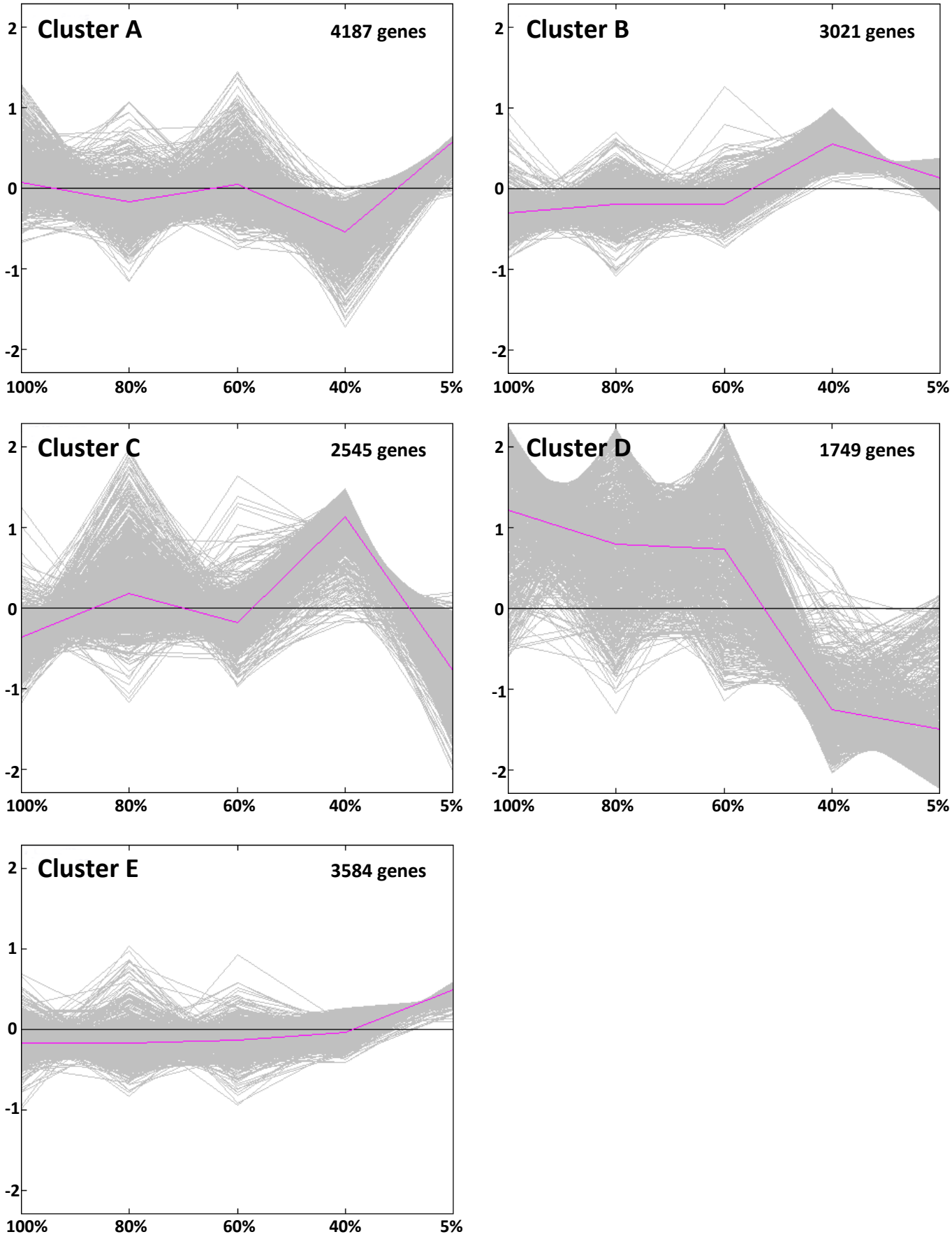

Figure 2.11: Expression profiles for the final lncRNAs of the *X. humilis* desiccation transcriptome. The final set of lncRNA transcripts was clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 5 clusters, 50 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number of lncRNA genes in each cluster. RWCs are given below each time point.

during early dehydration and late dehydration (80% and 40% RWC). It seems likely some key lncRNAs regulating desiccation tolerance pathways would be found within this set. Clusters B, D and E are generally comprised of transcripts with a single point at which expression is permanently induced or repressed. Clusters B and E both show an induction of expression during late dehydration, 40% and 5% respectively. Cluster D shows expression during early dehydration (100% - 60% RWC) with all transcripts showing clear repression from 40% RWC onward. Cluster D also has the largest change in expression levels relative to the other cluster, while clusters B and E show low relative expression level changes. Cluster D may however be primarily lncRNAs that are involved in normal cell function and that are repressed during dehydration. These would be of little interest. Further functional/target analysis is required to differentiate between the growth phase and desiccation response lncRNAs.

Logically, clusters showing large relative increased or decreased expression in response to water loss are most likely to contain interesting regulatory lncRNAs. As such, clusters B and C (Up regulation at low RWCs) and cluster D (drastic down regulation) appear most interesting. As the clustering of all transcripts into 5 expression profiles limits visualisation of all possible expression patterns, the 5 clusters were all retained for further interaction mapping and in-depth expression analysis/comparison (Chapter 4).

2.4. General Discussion

LncRNAs are known to be key regulators of gene expression, both at a transcriptional and post transcriptional level, and play an integral role as key regulatory molecules in plant stress response (Valadkhan and Valencia-Hipólito. 2017; Wang. 2017). I set out to determine whether lncRNAs are present as part of the vegetative desiccation tolerance regulatory program, and to identify a core set of putative lncRNAs for subsequent identification of decoy / competitive endogenous lncRNAs (ceRNA) via miRNA interaction mapping.

This chapter had two main outcomes, namely: 1) construction of a bioinformatics pipeline to identify differentially expressed lncRNAs and 2) the actual identification and expression clustering of DE lncRNAs for use in chapter 4. The stages and significance of findings are discussed below:

2.4.1. Leaf collection

In order to analyse the miRNAs present and active during vegetative desiccation tolerance in *X. humilis*, leaves were collected from plants during a single desiccation event. Leaves corresponding to five stages of desiccation (100%, 80%, 60%, 40% and 5% RWC) were pooled into three replicates per stage. By pooling multiple leaves per replicate, I aimed to minimise the relative contribution of natural leaf variation to the sample as a whole – thereby effectively amplifying the features specific to the desiccation response itself, at each specific RWC. This is important, not only due to natural leaf variation, but also as not all leaves resurrect (possibly due to natural senescence) and often the tips of resurrected leaves may die. Which leaves will or will not resurrect cannot be predicted beforehand. By including more leaves per replicate sample, the relative contribution of each individual leaf is minimised, thereby minimising the effect of any dying leaves or leaf tissue. Unfortunately, due to the rapid rate of desiccation at certain stages, relative to other stages, equal numbers of leaves were not obtained for each desiccation stage. Further desiccation events were not performed to obtain more leaves as it was desired that sRNA data be generated from the same desiccation event as the miRNA dataset. As such the number of leaves per pooled sample, as well as the relative contribution of each collection tray to each sample, differed between RWCs.

2.4.2. A high-quality RNA-Seq library.

Despite variation between the numbers of leaves pooled in each RNA-Seq sample, PCA indicate that a robust high-quality RNA-Seq dataset was obtained. 15 RNA-Seq datasets all show clear RWC specific behaviour, separating by high and low RWC along PC1 (Figure 2.9). Separation by RWC, as well as the close clustering of RWC replicates, indicates that a robust group of sequencing datasets was obtained, that well reflects the expression changes occurring over the course of desiccation, and from which reliable analysis and predictions can be made.

This also indicates that the variation between the numbers of leaves included in each pooled RNA sample should not have negatively affected analysis. The same RNA-Seq dataset was used to effectively assemble and analyse the full *X. humilis* desiccation transcriptome (Lyall. 2016), validating its quality.

2.4.3. Bioinformatic filtering pipeline effectively reduced dataset size.

The primary goal of this chapter was to assemble a pipeline to identify differentially expressed putative lncRNA that may interact with miRNAs as competitive endogenous lncRNAs (decoys). The 2,611,123 initial dropset transcripts were subjected to a pipeline of filtering steps to obtain a final set of putative lncRNA transcripts. Transcripts were filtered on the basis of similarity to coding sequences, read count and FPKM, differential expression, coding potential and similarity to miRNA precursor sequences (Figure 2.1). Size based filtering (>200bp) was achieved by RNA-Seq library preparation so no specific size cut-off was applied. This filtering pipeline was able to effectively reduce the number of dropset transcripts to 15,185 putative lncRNA sequences. This is an elimination of 99.4% of the original dropset sequences (172 fold reduction), that did not meet the lncRNA classification criteria.

The presence of large numbers of lncRNA transcripts showing DE and high levels of expression during key stages of the desiccation response pathway of gene expression supports the notion that at least some of these lncRNAs may be playing key regulatory roles: modulating miRNA activity in order to either regulate metabolic shutdown processes or activate desiccation response pathways of gene expression. It is not possible however to speculate on the roles of any specific lncRNAs at this stage as too many lncRNA sequences remain for in depth analysis.

2.4.4. High-confidence putative lncRNAs were identified.

The 15,185 putative lncRNA sequences identified is still still represent a large number of sequences to work with. It is unlikely that these all represent individual and functional lncRNAs. The final set of 15,185 predicted lncRNA transcripts includes multiple transcripts that were clustered together by Corset on the basis of shared reads and sequence identity. These often represent paralog lncRNA genes that may act redundantly to fulfil similar roles or may have evolved to fulfil new independent regulatory roles. NC RNAs have a much higher rate of evolution, with function deriving from structure, and only small nucleotide regions required for complementary nucleotide interactions. It is therefore not unexpected to find highly similar lncRNA paralogous sequences. All such transcripts were retained for downstream analysis.

In order to assess the success of the filtering pipeline, to identify a biologically reasonable number of lncRNAs, it is useful to compare the number of putative lncRNAs identified to the numbers found in similar plants studies. These comparisons highlight a number of key findings. Firstly, our set of putative

lncRNAs has more than double the number of known *A. thaliana* lncRNAs. 6480 lncRNAs have been identified in *Arabidopsis thaliana* through combined transcriptomic studies (Liu et al. 2012) and 6,584 potential lncRNA have been identified in trifoliate orange (Poncirus trifoliata L. Raf.) through genome-wide screening for lncRNAs (Wang et al. 2017). This suggest that too many putative lncRNAs remain and that further filtering is required to identify the true lncRNA transcripts.

15 thousand putative lncRNAs appears, however, to be a reasonable number of identified putative lncRNAs when compared to other transcriptomic studies. 70% of *A. thaliana* annotated mRNAs have been found to have associated antisense transcripts, with a total of 37,238 long non-coding natural antisense transcripts (lncNATs) being identified (Wang et al. 2014). In maize (*Zea mays*), a total of 20,163 putative lncRNAs have been identified (Li et al. 2014). We must however take into account that not all identified putative lncRNAs are likely to function as lncRNA, and many are likely sRNA precursors. Of the 20,163 putative maize lncRNAs, 1704 were deemed high-confidence lncRNAs, while comparison to the full set of all known maize sRNAs showed that 90% were in fact long non-coding sRNA precursors. Despite screening for precursor miRNA sequences, my set of predicted miRNA precursors is not comprehensive, and miRNAs are only 1 of many types of sRNAs present in the cell. Furthermore, the transcripts which do function as lncRNAs, will be a diverse mixed set of long non-coding natural antisense transcripts (lncNAT), intronic lncRNAs, promoter lncRNAs and long intergenic RNAs (lincRNAs), each of which may function as competitive endogenous RNAs (ceRNAs/decoys), facilitators of an open chromatin state, molecular scaffolds or guides.

2.4.5. Large dataset is permissible for future prediction of ceRNA interactions.

If I was interested in further analysis of the lncRNAs as a whole, too many putative lncRNAs remain. Further refinement and reduction of the dataset would be required, as well as further classification of the lncRNAs. This could be achieved by increasing stringency or adding additional filtering steps, such as increasing the required fold change cut-off in order to select only the lncRNAs showing very extreme up or down regulation in response to water loss. Mapping of the lncRNA transcript sequences to the *X. humilis* genome to identifying their mapped positions relative to PC genes, would be a good first step to lncRNA classification. In this study however, I focus solely on identifying ceRNAs interacting with miRNAs (Chapter 3) as part of regulatory RNA networks (Chapter 4). As such, the interaction mapping and network expression analysis performed in chapter 4 will function as an additional selection/filtering step and I can afford to be more permissive with the initial identification of putative

lncRNA transcripts, presented in this chapter, as well as the number and specific transcripts allow through for further analysis.

2.5 Conclusion

In summary, this subsection of the non-coding RNA study set out to perform RNA sequencing on desiccating *X. humilis* leaves, and from the resulting RNA-Seq data to predict a set of putative lncRNA transcripts that may contain lncRNAs playing key regulatory roles as part of the desiccation tolerance program of gene expression. RNA sequencing allowed for successful de novo assembly of the *X. humilis* dessication transcriptome and identification of key transcriptional changes occurring as desiccation proceeds (Lyall. 2016), as well as assembly of a large and diverse set of "dropped" sequences. When subjected to a vigorous filtering pipeline this dropset yielded 15,185 predicted lncRNA transcripts. These transcripts are all distinct from any Primary transcripts, are all non-coding, larger than 200bp, show significant differential expression (FDR < 0.05, FC >2) across desiccation stages, match no known *X. humilis* pre-miRNA sequences (Chapter 3), and are present at copy numbers indicating possible biological significance. While a relatively large number of putative lncRNA sequences remain, it is likely that only a small subset are playing important roles during desiccation, regulating the desiccation response network of gene expression. The expression profiles of these transcripts have been visualised and at least two clusters show expression vectors indicating RWC-specific induction highlighting them as high probability candidates as key regulators. By mapping the predicted lncRNAs to the miRNAs identified in Chapter 3, as well as to the coding transcripts targeted by these miRNAs, the key lncRNAs will be identified and explored further in Chapter 4. The full set of 15,185 putative lncRNA transcripts will likely contain many other interesting lncRNAs, functioning as molecular scaffolds and guides for other transcription effectors. This dataset therefore also provides a platform for future work into these other classes of regulatory lncRNAs.

CHAPTER 3: Bioinformatic prediction of putative regulatory miRNAs in desiccating leaves of the resurrection plant *Xerophyta humilis*.

3.1. Introduction

MiRNAs are small non-coding RNAs which function as highly specific post-transcriptional repressors through targeted mRNA degradation or translational inhibition (Brodersen et al. 2008). Able to facilitate rapid and precise shifts in the transcription landscape, such as the dramatic metabolic shutdown and stress response that occurs during the VDT response (Lyall. 2016), miRNAs are of interest as key regulators of both target mRNA transcript levels as well as lncRNAs abundance. MiRNAs are closely linked to lncRNA function, not only regulating target lncRNA levels, but also themselves being regulated by decoy lncRNAs (ceRNAs) through competitive endogenous inhibition. As a result of this close regulatory interplay, the lncRNAs involved in VDT (Chapter 2) cannot be considered in isolation. The role of the miRNAs regulating VDT must also be examined.

With the increasing availability and accessibility of next generation sequencing (NGS), miRNAs have emerged as master regulators of plant growth, development and the maintenance of genome integrity. Studies into the role of miRNAs in plant stress response programmes have shown that the expression profiles of most miRNAs involved in plant growth and development are significantly altered under stress conditions, indicating stress responsive miRNA expression (Zhao et al. 2007; Trindade et al. 2010; Kulcheski et al. 2011). For example, under drought conditions the Arabidopsis miR168 is down-regulated leading to an increase in its mRNA target, nuclear transcription factor Y subunit A-5 (NFYA5). Arabidopsis lines overexpressing nuclear transcription factor Y subunit A-5 (NFYA5) show increased drought tolerance, while miR168 overexpression increases drought sensitivity, indicating a functional role by miR168 in mediating the drought stress response (Li et al. 2008). Similar studies have shown miRNAs to be involved in a number of other stress response programmes including: osmotic, cold, heat and other abiotic stress responses, biotic (bacterial pathogensis) stress responses, phosphate/sulfate/copper/nitrogen nutrient deficiency and response to mechanical damage (Reviewed in Guleria et al. 2011 & Sunkar et al. 2012). Many plant stress response genes have been found to be targets of miRNA activity, such as miR938 which targets two Cu/Zn superoxide dismutases (CSD1 and CSD2) (Sunkar & Zhu. 2004). Links between phytohormone signalling and sRNA activity, such as dehydration-related ABA-inducible *Craterostigma* desiccation tolerant (*CDT-1*), indicate sRNAs may function as a bridge between exogenous and endogenous cell signalling pathways. It has been suggested that generation of these novel miRNA and other sRNAs may have allowed for or driven

the evolutionary adaptation to allow survival under such extreme stress conditions (Phillips et al. 2007).

The presence of miRNAs in these stress response programmes, which often require rapid transcriptional shifts and transcript silencing, is not surprising given that these traits are hallmarks of sRNAs function. VDT requires rapid shutoff of metabolic pathways and activation of key protective pathways during desiccation, and is associated with extensive transcriptional re-programming (Lyall. 2016). In turn rehydration is often extremely rapid with desiccated tissues being pre-primed; the necessary transcripts being transcribed but not translated and instead stably stored in desiccated tissues until water becomes available again (Dace et al. 1998; Alpert. 2006). While it is clear that miRNAs and other sRNAs play key roles in desiccation tolerance, little is known with regards to their functional roles in the desiccating leaves of *X. humilis*. Furthermore, the mechanisms by which miRNAs confer dehydration stress tolerance at different stages of desiccation, and compared to other resurrection plant species, is unclear and often confusing. For example, drought stress leads to a down-regulation of miR169 in *A. thaliana* and *Medicago truncatula*, but an increase in expression in rice (*Oryza sativa*). These conflicting results indicate the role or mechanism of miRNA action may vary between species (Zhao et al. 2007; Li et al. 2008; Trindade et al. 2010).

3.1.1. Bioinformatic prediction of miRNAs

The traditional experimental approach to miRNA discovery and identification has relied on isolating sRNAs from high resolution gels, cloning and direct Sanger sequencing, followed by experimental verification of miRNA activity (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros, 2001; Aravin and Tuschl, 2005; Wark et al. 2008). Time and resource intensive, these limited throughput protocols severely hampered the rate of miRNA discovery, with very few sRNAs examined actually representing mature miRNA sequences. Low sensitivity and the sheer volume of high copy miRNAs also obscured lowly expressed miRNAs from discovery.

Early computational approaches attempted to screen genomes for sequences, that if transcribed would form hairpin structures resembling miRNA precursors. Structural predictions were then screened by scoring, applying various set rules, or applying parameters derived from a training set of known miRNAs (Kang and Friedländer, 2015). The human genome has at least 11 million putative hairpins (Bentwich et al. 2005), with only a few thousand actually coding for mature miRNAs (Kozomara and Griffiths-Jones, 2011; Friedländer et al. 2014). The high number of genome sequences

able to form hairpins, were they transcribed, means such an approach has an extremely high false positive rate (Friedländer et al. 2008).

The advent of Next generation deep sequencing allowed high throughput surveying of the complete sRNA pool present, at an unprecedented level of sensitivity. Both lowly expressed miRNAs as well as rare degradation products were now detectable. Small RNAs corresponding to millions of genomic loci could be screened, as opposed to the few thousand previously possible with Sanger sequencing (Kang and Friedländer, 2015). The vast quantities of sequence data generated however, posed new computational challenges, the central problem being how to differentiate miRNAs from other sRNAs and degradation products (Berezikov et al. 2006; Ruby et al. 2006). This is not trivial, with billions of possible sRNAs, many of which are simply products of RNA degradation (Kang and Friedländer, 2015). Furthermore, specialized skills were required to adequately analyse and interpret data (Williamson et al. 2013).

While homology to known miRNAs can be used to identify conserved miRNA families (Kang and Friedländer, 2015), in order to avoid bias to identify novel miRNAs, including the non-conserved species-specific miRNAs abundant in plants (Voinnet. 2009), de novo prediction of miRNAs from sRNA-Seq data is needed. Many tools now exist for de novo prediction of miRNAs form sRNA-Seq data (Williamson et al. 2013; Reviewed in Kang and Friedländer, 2015). Most of these tools rely on the variations of the same basic approach for prediction.

3.1.2. Probabilistic scoring of putative pre-miRNA sequences for adherence to a classical biogenesis model.

While miRNA sequences themselves are not particularly distinctive, their very unique mode of biogenesis (chapter 1) is key to the discovery and annotation of novel miRNAs (Kang and Friedländer, 2015). This key feature can be used to distinguish miRNAs from other small RNAs by implementing a probabilistic model, which compares the predicted pre-miRNAs and sequenced biosynthetic by-products to the known model of pre-miRNA processing. The position and frequency of mapped RNA-Seq reads allows candidates incompatible with miRNA biogenesis to be discarded and compatible candidates to be assigned statistical probability scores, a measure of likelihood that they represent true miRNA transcripts. This forms the core module of the in silico prediction pipeline, indicated in Figure 3.1.

In order to identify putative miRNA precursors, sRNA-Seq read sequences are first aligned to a reference genome. The mapped loci are expanded to include the flanking genome sequence bracketing each alignment, and the expanded genomic regions are excised. As such, the majority of computational tools used for novel miRNA prediction require a reference genome. A full, high quality genome assembly is an important resource, as gaps in the genome would limit comprehensive miRNA discovery. In the absence of an available genome, the genome of a closely related species can be used as proxy, with the disadvantage of excluding all species-specific miRNAs (Williamson et al. 2013; Kang

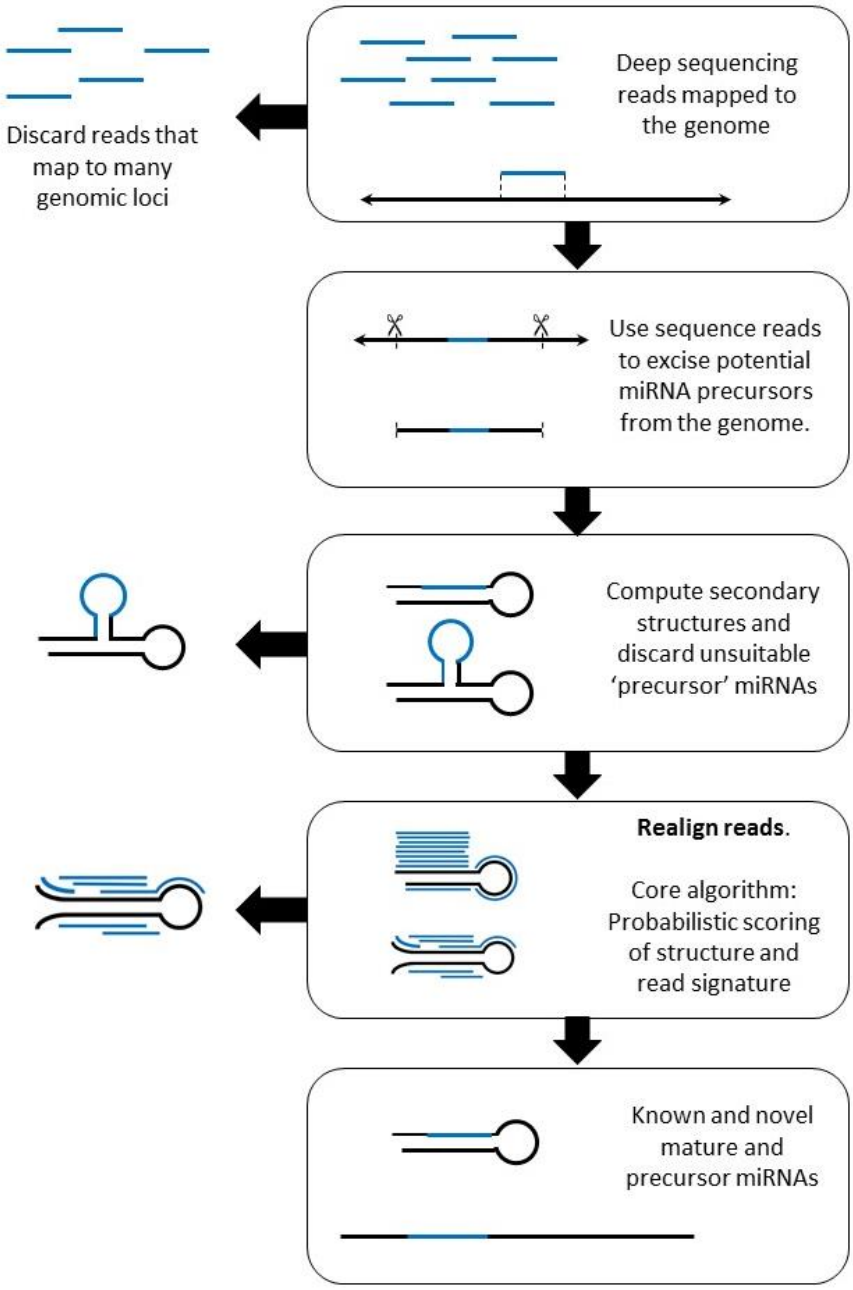

Figure 3.1: Flowchart showing the steps of in silico miRNA prediction from sRNA-Seq data.

and Friedländer, 2015). The secondary structures of the expanded reference sequences are then computed. Sequences predicted to form appropriate stem-loop structures, a feature of all miRNA precursors, are retained and sRNA sequencing reads are realigned to the remaining putative precursor structures (Figure 3.2).

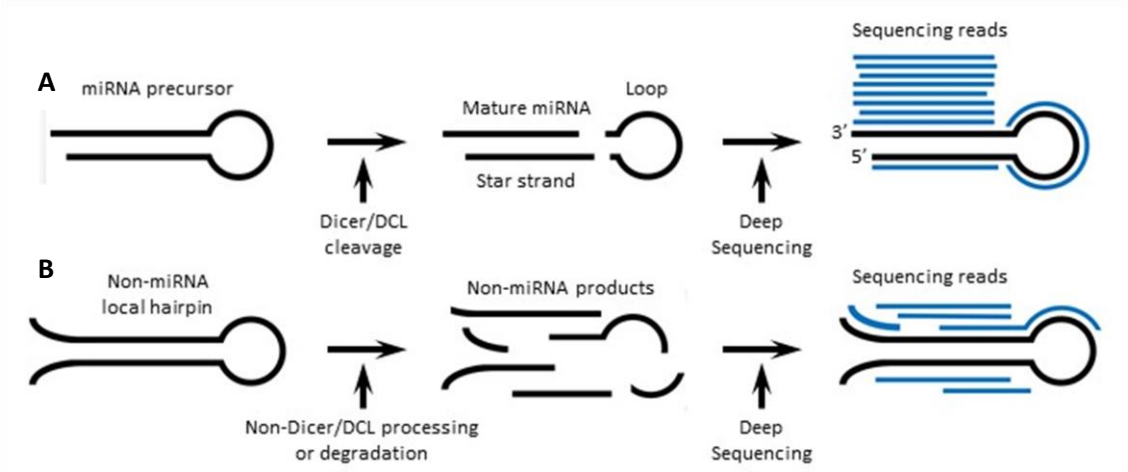

Figure 3.2: Compatibility of sRNA sequencing reads to miRNA biogenesis allows prediction of true miRNA sequences. (**A**) Cleavage of a stable miRNA precursor by Dicer or a DCL protein results in three products - the mature miRNA sequence (most abundant), the loop sequence and the star strand – all of which can be sequenced. The sRNA sequencing reads, when mapped to the predicted miRNA precursor, correspond to the positions of these 3 products. The statistics of the read positions and frequencies (the read "signature") is highly unique to, characteristic and indicative of miRNAs. (**B**) While many other hairpin structures are also transcribed, and produce sRNAs through non-Dicer processing or degradation, the resulting sRNAs and sequencing reads are inconsistent with miRNA biogenesis.

Adapted from Friedländer et al. 2008, Figure 1.

Dicer/DCL cleavage produces consistent predictable products of known length and position from within the miRNA precursor (Friedlander 2008). Reads related to miRNAs are therefore expected to correspond to one of three products of miRNA precursors processing: the mature miRNA strand, the complimentary miRNA star strand and the loop sequence (Fig 3.2) (Friedländer et al. 2008; Williamson et al. 2013). Very few other reads, inconsistent with these expected products, should map to the putative precursor (Friedländer et al. 2008). Furthermore, reads must demonstrate characteristics resembling those of known miRNAs. This includes short (2nt) 3'overhangs on both the miRNA and miRNA* strands, characteristic of Drosha/ Dicer/DCL processing. The 5' ends should be clearly defined and align precisely (Ruby et al. 2006). Mature miRNA reads are expected to be more abundant than other miRNA-related reads, which are less stable and degraded upon precursor cleavage and miRNA loading. These rules apply to both animal and plant sRNA-Seq reads. The read alignments patterns (position and abundance) to the putative precursor are referred to as the "Signature" of read alignment (Friedländer et al. 2008; Williamson et al. 2013).

Evaluation of expanded sequence on the basis of secondary structure (Fig 3.1), negative free energy (energetic stability) and the "signature" features of read alignment are used to discard reference sequences inconsistent with miRNA biogenesis, and to statically determine a probabilistic likelihood score (usually a logs odd score) for the reference structures being real miRNA precursors (Friedländer et al. 2008). Energetic stability is a hallmark of miRNA precursors, with the stability of precursor hairpins known to exceed that of non-precursor hairpins (Bonnet et al. 2004). Higher abundance of miRNA and miRNA* reads, as well as the presence of the miRNA* strand and the relative and absolute stability of the precursor hairpin all increase the strength of the prediction. The power of miRNA discovery is thus proportional to the depth of sequencing (Friedländer et al. 2008).

3.1.3. Assessing prediction Quality: Sensitivity vs Specificity

Many tools have been developed for the de novo miRNA prediction from sRNA-Seq data. While many use the same basic approach, the software used can dramatically affect the number and quality of predictions made (Williamson et al. 2013). In order to compare tools, understand their specific priorities, and select the best tool for a specific application, it is important to be aware of their individual levels of sensitivity and specificity (BOX 3.1, Table 3.1.). Sensitivity is the ability of an algorithm to correctly identify as many of the true miRNAs present in a dataset as possible. Specificity refers to the algorithm's ability to correctly identify and discard non-miRNAs. While high sensitivity and specificity are both desirable, there is a trade-off between the two (Kang and Friedländer, 2015). In the case of miRNA prediction from sRNA-Seq, the number of reads from true miRNAs is far surpassed by non-miRNA sequences. This trade-off is therefore skewed. While a slight reduction in sensitivity, increasing false negatives, is tolerable, an equivalent percentage decrease in specificity will result in a much larger number of false positives. These can quickly drown out the true positives. Most miRNA prediction tools therefore aim to maximise specificity, while sacrificing some sensitivity (Kang and Friedländer, 2015).

BOX 3.1: Key definitions pertaining to miRNA prediction algorithms.

Sensitivity: The fraction of the known/true distinct miRNAs in the sRNA dataset recovered by the algorithm.

Specificity: The fraction of (assumed) non-miRNA sequences correctly discarded by the algorithm

False positive rate: The fraction of non-miRNAs incorrectly reported as true miRNAs, or 1 - Specificity

Accuracy: The fraction of distinct sequences correctly classified by the algorithm, summing over all sRNAs (miRNAs and non-miRNAs).

Table 3.1: Sensitivity vs Specificity.

		miRNA state	
		Genuine miRNA	**Not genuine miRNA**
miRNA prediction	**Positive**	**True positive (TP)**	**False positive (FP)**
	Negative	**False negative (FN)**	**True negative (TN)**
Formulas	**Sensitivity (True positive rate)**		**TP / (TP + FN)**
	Specificity (True negative rate)		**TN / (FN + TN)**
	Accuracy		**(TP + TN) / (TP + FP + TN + FN)**

***** From Kang and Friedländer, 2015, Table 3.

3.1.4. Selection of Bioinformatic tools

3.1.4.1. MiRDeep-P

MiRDeep, one of the first tools for de novo miRNA prediction, is still the most popular programme (by citations) for miRNA discovery in use today (Williamson et al. 2013). MiRDeep pioneered the use of a Bayesian probabilistic model of miRNA biogenesis for miRNA prediction (Friedländer et al. 2008). The use of Bayesian statistics framed on an explicit biogenesis model produces very robust results, circumventing the need for sRNA read filtering, and without the need for species specific prediction parameters. This makes miRDeep ideal for use on emerging model systems or non-model organisms (Kang and Friedländer, 2015). Performance of this tool has been extensively benchmarked and validated both experimentally and in comparison to other computational tools (Friedländer et al. 2008; Friedländer et al. 2009; Metpally et al. 2013; Kang and Friedländer. 2015). A comparison of three miRNA prediction tools, miRDeep, miRDeep2, and miRanalyzer, found that miRDeep possessed the best overall balance between sensitivity and specificity, for the assessment methods used (Area under the curve) (Williamson et al. 2013). Many other programmes (miRTool, miReNA etc.) make use of the core miRDeep algorithm as part of their pipeline or toolkit (Zhu et al. 2010).

Most of the bioinformatics tools for predicting miRNAs for have been developed for use animal sequence datasets, and few bioinformatics tools exist for predicting plant miRNAs. However, although miRDeep is designed for use on animal systems, the robustness of the miRDeep algorithm means it is applicable to, and been successfully used, on plant systems (Yang and Li. 2011). Despite this, two

significant changes are required to account for unique features of plant systems: 1) Plant pre-miRNAs are much longer and variable in length than their relatively uniform animal counterparts, and 2) more plant miRNAs belong to paralogous families, possessing multiple identical or near identical members (Yang and Li. 2011). Independent groups have therefore modified and expanded the core miRDeep algorithm (Yang and Li. 2011, Yang and Qu. 2012, Wu et al. 2013). MiRDeep-P is one such modified version, making use of the core miRDeep-P algorithm and probabilistic biogenesis model already discussed (Section 3.1.2), but with plant specific optimisation. Following Bowtie alignment of sRNA-Seq reads to the reference genome, miRDeep-P utilizes a larger optimized window size for precursor excision. In *Arabidopsis thaliana* the optimal window size has been shown to be 250 bp, a window size that has also been shown to perform well in various monocots and dicots (Yang and Li. 2011). Alternatively, the window size can be either manually set or empirically determine, if a set of known miRNAs is available (Yang and Li. 2011). Further changes include implementation of a unique plant specific scoring system and a set of post-prediction plant-specific miRNA filtering criteria (Yang and Li. 2011). This final filtering step screens for known characteristics of plant miRNA genes resulting from precise excision of a ~21-nucleotide miRNA/miRNA* duplex from the stem of a single-stranded, stem-loop precursor. These include 1) the presence of duplex forming miRNA and miRNA* strands that possesses two nucleotide 3'overhang, 2) four or less mismatches in the nucleotide binding between the miRNA and the complementary arm of the precursor hairpin and 3) asymmetric bulges are minimal in size (one or two nucleotides) and frequency (typically one or less), especially within the miRNA/miRNA* duplex (Meyers et al. 2008). MiRDeep-P has been tested in a broad number of plant species, including *A. thaliana*, *Oryza sativa* and *Carica papaya* (Yang and Li. 2011).

The major strengths of miRDeep and miRDeep-P algorithm, other than the latter's optimisation for use in a plant system, are its ability to predict novel miRNA genes de novo; in species without detailed annotation, and without the needing to use homology to known miRNAs. Provided robust genome assembly exists and sufficient sequence coverage is obtained, even lowly expressed miRNAs are detectable. This is important for enabling the detection of species-specific miRNAs so prevalent in plants (Fahlgren et al. 2007). miRDeep-P allows both mature and precursor miRNA sequences to be identified, as well their genomic locations thereby distinguishing between paralogous genes coding for identical or near identical miRNAs, as well as enabling the expression status of paralogous miRNAs to be uncoupled from one another (Yang and Li. 2011).

A limitation to consider, is that MiRDeep-P is computationally demanding and requires access to high performance computing. MiRDeep-P is comprised of 9 independent yet sequentially executed Perl

scripts, and is run in a command line environment. Utilizing the core miRDeep algorithm (Friedlander. 2008), short read alignments are performed using Bowtie (Langmead et al. 2009), and ViennaRNA package (RNAfold) predicts RNA secondary structure (Lorenz et al. 2011).

3.1.4.2. ShortStack

ShortStack, like miRDeep, makes use of a probabilistic model of miRNA biosynthesis to annotate reference aligned sRNA-Seq data by strict structural and expression-based criteria, without the need for prior annotation or homology. Unlike miRDeep however, ShortStack is designed for comprehensive annotation and quantification of all small RNA in a single command, while at the same time requiring only moderate computational resources and computing time (Axtell. 2013).

ShortStack is by design, highly specific, prioritizing specificity over sensitivity. To illustrate this, the default plant settings used by ShortStack, exclude 12% of all known Viridiplantae miRNAs found in the miRBase database (Axtell. 2013). One of the ways ShortStack achieves this is through a stringent requirement that all miRNAs have detectable star strands. This is distinct from miRDeep which instead applies a heavy scoring penalty in the absence of the star strand. Designed for use on both animal and plant (monocot and dicot) sRNA-Seq data, users are able to specify the “miRTYPES” (plant or animal). This further increases specificity of miRNA prediction and removes putative miRNAs with unusual structural characteristics, an option unavailable in miRDeep. The default maximum window size for sRNA precursors is 300 nt. Despite the high default stringency most parameters are user-adjustable for greater flexibility, in case less specificity and greater sensitivity are required.

Another advantage to ShortStack is its comprehensive output format. The tabular output for all successfully annotated sRNAs includes: de novo defined sRNA clusters, annotation of miRNAs and other hairpin-associated sRNA loci and any detectable sRNA phasing (repeating arrangement of aligned sRNAs). Furthermore, information on sRNA size composition, strandedness and repetitiveness is given for all sRNA genes/loci. ShortStack also generates read counts for all sRNA loci, mapping reads to the full precursor sequences and using the distribution of surrounding reads to inform the partitioning and assignment of ambiguous reads between paralogous miRNAs (Axtell. 2013; Johnson et al. 2016). ShortStack therefore offers a second, highly specific, detailed and more recent tool designed for use on plant systems, able to annotate the miRNAs present while accounting for the other sRNA classes.

3.1.5. Aims.

To investigate the possible role of the lncRNAs (Chapter 2) as competitive endogenous inhibitors (ceRNAs) of miRNAs regulating desiccation tolerance, I require a set of high confidence miRNAs found in the desiccating leaves of *X. humilis*. The work presented in this chapter therefore seeks to fulfil 3 main objectives: 1) Identify miRNAs present in the desiccating leaves of *X. humilis*, 2) Identify which of these miRNAs may be involved in regulating VDT, and 3) To select a set of high confidence miRNAs that can be used to investigate the interplay between the regulatory ceRNAs and miRNAs during VDT. In order to achieve these objectives, I set out to sequence the small RNA complement of the desiccating X. humilis leaf transcriptome and use a bioinformatic approach to predict and select the core miRNA set on the basis of expression and homology to known miRNA families.

3.2. Materials and methods

3.2.1. Plant material, sRNA extraction and quality assessment.

Leaves used for sRNA sequencing and miRNA prediction originate from the same *X. humilis* plants, the same desiccation event and the same original leaf collection, as those used for total RNA sequencing, assembly of the *X. humilis* desiccation transcriptome (Lyall. 2016), and lncRNA prediction (Chapter 2).

On account of the 100%, 80%, 60%, 40%, and 5% RWC leaves having already been used during the first RNA-Seq experiment, alternative RWC stages had to be selected. The RWCs selected for sRNA-Seq were 90%, 70%, 50%, 30% and 10%, to as closely reflect the stages used for total RNA extraction as possible. RNA was extracted from leaves with RWCS calculated to be within ±5% of these values.

Total RNA, including all sRNAs, was extracted from each individual leaf using QIAzol Lysis Reagent (QIAGEN) following a protocol modified from the manufacturer's instructions. The initial extraction procedure was performed as detailed in chapter 2, up until and including centrifugation for 15 minutes at 12 000 x g and 4°C to pellet any debris and separate the organic and aqueous layers. Extractions were not performed on columns. Instead the upper aqueous layer from each sample was transferred to a new 1.5ml tube, containing 250 µl isopropanol and 250 µl High Salt Precipitation Buffer[1], mixed by inversion and incubated at room temperature for 5 minutes. Each sample was centrifuged for 8

1 **High Salt Precipitation Buffer:** 0.8 M sodium citrate, 1.2 M sodium chloride made up in sterile DEPC-treated H_2O.

minutes at 12 000 x g and 4°C to pellet the RNA. The remaining supernatant was removed, 1 ml of 75% (v/v) ethanol made up in DEPC-treated sterile H_2O added to each tube and the tubes centrifuged for 5 minutes at 12 000 x g and 4°C, to clean the pellet of any remaining contaminants. All supernatant was removed and the pellet air dried for approximately 5 minutes. Pellets were re-suspended in 30 µl of RNase-free water at 55°C for 10 minutes and transferred to a new 1.5ml tube. The samples were placed in storage at -80°C, other than a small volume of each retained for quantification and quality assessment. A DNase I digestion was not performed on the total RNA extracted for sRNA sequencing, as instructed by Beijing Genomics institute (BGI)

The individual total RNA extractions were analysed for contamination and their concentrations measured by running 1 µl of each sample on a Nanodrop ND-1000 spectrophotometer. The purity and integrity of the extracted RNA was visually examined on a denaturing agarose gel. One µl of each RNA sample was diluted in 2x volume of RNA sample loading buffer[1] and denatured at 60°C for 5 minutes before being electrophoresed on a 1.2% denaturing formaldehyde agarose gel[2] at 100V for 30 minutes.

All samples determined to be of sufficient quality and quantity, taking into account their absorption curves, A260/280 and A260/230 ratios, as well as their gel images, were selected for subsequent pooling. The selected samples were pooled to form three independent biological samples for each of the five RWCs. While RNA from five independent leaves was the desired pool, there were insufficient samples at each RWC to allow this, and so each pool comprised equal amounts of RNA between one and six contributing leaf extracts, with no RNA being shared between pools. The concentration and quality of each pool was then reassessed by both Nanodrop and gel electrophoresis. In order to stably transport the RNA for sequencing, single aliquots containing 20µg of total RNA were prepared for each pooled RNA sample and treated with 20µl of RNAstable LD (Biomatrica) as per the manufacturer's instructions. Samples were gently mixed and dried without heat using a Speedvac Plus SC2 10A (SAVANT) vacuum concentrator. The tubes were wrapped in Parafilm and sealed in a protective heat-sealed moisture-resistant bag with a separate sachet of silica-based desiccant, ready for transport.

1 **RNA sample loading buffer**: 105µl Formaldehyde, 300µl Formamide, 60µl 10x MOPS (0.2M MOPS, 0.05M sodium acetate, 0.001M EDTA made up in DEPC-treated water), 6µl Ethidium bromide)

2 **1.2% denaturing formaldehyde agarose gel**: 1.2g agarose, 43 ml H_2O, 6 ml 10x MOPS and 11 ml 37% (v/v) formaldehyde) in 1x MOPS

3.2.2. Small RNA selection, library preparation and Sequencing.

The 15 total RNA samples were sent to the BGI for RIN analysis, small RNA fraction purification, strand specific library construction and sRNA sequencing. Fifty bp single-end sequencing was performed on a HiSeq4000 sequencing instrument, to produce at least 10 Mb of reads per sample, after removing adapter sequence and low-quality reads. The resulting pre-processed, clean read data was downloaded from the BGI FTP server. All reads less than 18 base pairs in length were removed.

Small reads were discarded as recommended by Yang and Li. 2011, in order to facilitate faster computational analysis and prevent flooding of the mapping output. These reads are unlikely to represent any full-length miRNA transcripts (~21 bp) and partial miRNAs that are discarded should have full length transcripts remaining in the dataset.

3.2.3. Bioinformatics prediction of putative miRNA sequences.

In order to best predict and obtain high confidence putative miRNAs, two bioinformatics tools for miRNA prediction were selected – miRDeep-P and ShortStack.

3.2.3.1. Predicting putative miRNAs using miRDEEP-P

miRDeep-P, v1.3, was used to perform independent de novo miRNA predictions for each of the 15 sequencing samples. The miRDeep-P pipeline consists of 6 sequentially executed Perl scripts, run from a command line environment. Scripts were run with default settings, as per the miRDeep-P manual. Any modifications or specifics are given below.

The sRNA-Seq reads were aligned to a draft version of the *X. humilis* genome (Schlebusch & Illing, unpublished data) using Bowtie (Langmead et al. 2009), selecting only perfect full length, 100% identity, end-to-end alignments. Bowtie was used instead of the more recent Bowtie2, as Bowtie performs better for short reads up to 50bp and is able to restrict gaps in alignment (Langmead & Salzberg. 2012). The remaining sRNA reads were filtered to remove reads mapping more than 50 times to the genome. Although a cut-off of >15 times is recommended for Arabidopsis, this is based on the fact that its largest known miRNA family (miR169) has 14 known members (Yang and Li. 2011). Other species of plants require different cut-offs, based off known family sizes or other empirical considerations, such as genome size. As nothing is known about *X. humilis* miRNAs, and the *X. humilis*

genome is at least 4 times as large as the *A. thaliana* genome (See Box 3.2), it was decided that a cut-off (-c) of 15 was possibly too conservative, and a cut-off of 50 was selected to match the ShortStack default ("--bowtie_m").

BOX 3.2: Xerophyta humilis genome size.

The published genome size for *X. humilis* is 532mbp (Hanson et al. 2001). A flow cytometry experiment performed by our laboratory to confirm this published value, estimated the genome size to be 944Mbps (Milborrow, unpublished data), almost double (1.77X) the published value. Furthermore, this new value is consistent with the size of the *X. humilis* draft genome assembly (Schlebusch & Illing, unpublished data) used for this experiment. While the exact size of the *X. humilis* genome has not been finalized, what is certain is that it is significantly larger than the ~135Mbps *Arabidopsis thaliana* genome (TAIR10).

Following read mapping, precursor sequences were excised from the *X. humilis* genome and RNAfold (Lorenz et al. 2011) was used to predict potential precursor secondary structures. A precursor length of 300 bp was selected to coincide with the length of ShortStack precursor lengths.

For mapping reads, back to excised precursor sequences, Bowtie was again used, requiring perfect alignment. In order to remove redundant predicted miRNAs and filter by plant criteria, chromosome lengths are required. As this information is not available in *X. humilis*, and is used by miRDeep-P to compare distances between miRNA loci, scaffold lengths were used in their place.

3.2.3.2. Predicting putative miRNAs using ShortStack.

ShortStack v3.4, was used to independently predict putative miRNAs from each of the 15 sRNA-Seq samples. Default parameters were used (50 hit cut-off and 300 bp size), for read mapping and to excising possible miRNA precursor sequences from the *X. humilis* genome. The comprehensive ShortStack output was processed to pull out all results corresponding to predicted putative miRNAs, as well as the results corresponding to "N15", predicted to be miRNAs but lacking the presence of the key miRNA star strand.

3.2.4. Generation of read counts and Principle Component Analysis.

Raw read counts were generated for all unique miRNA and N15 sequences by mapping the original sRNA-Seq data (reads of 18 or more nucleotides) for each of the 15 samples (five RWC with three biological repeats) to the corresponding pre-miRNA sequence set, using Bowtie (Langmead et al.

2009). The relevant counts were then extracted. The parameters -a and -v 0 were used, to ensure full end to end hits with zero mismatches. The resulting counts were filtered, retaining only sequences with at least 10 reads in at least two independent pools for at least 1 RWC. This was performed to reduce any noise from false predictions, sequences not uniformly found within a set of replicates, or candidates of few counts and likely low biological relevance.

In order to check the robustness of the sample data used for miRNA prediction, and for variability between independent replicates, which should cluster together by RWC, the regularised-log transformed (rlog, blind=TRUE) sample data were compared using a two axis Principal Components Analysis (PCA). The R DESeq2 package from Bioconductor was used to perform the PCA and the two principle components most able to explain the data were plotted.

In order to also assess the full original sRNA-Seq dataset, all sRNA-Seq reads (≥ 18bp) were mapped to the *X. humilis* genome using Bowtie (-v 0, by default only reporting the 1 best alignment for each read). All contiguous genomic sequences (contigs) to which sRNA-seq reads mapped, as well as their raw read counts, were compiled into a new genomic read count dataset. Contigs that did not possess at least one RWC with at least 2 replicates possessing at least 10 reads were discarded, and the remaining genomic read count dataset underwent the same rlog transformation and PCA.

3.2.5. Differential expression Analysis

Differential expression analysis was performed on the filtered set of all unique 'miRNA' sequences at each RWC, using DESeq2. DESeq2 performs its own read count normalisation and size correction during analysis (Love et al. 2014). Differential expression analysis was performed using a log-likelihood ratio test (LRT), and default DESeq2 parameters. Sequences were classified as differentially expressed if the false discovery rate (FDR) was less than 0.01. The alpha value used for DESeq2 independent filtering was set to match the FDR value of 0.01. The DESeq2 analyses were all run in RStudio Desktop (v0.99.892), utilizing the x64 Windows version of R (v3.2.3; R Development Core Team 2015) and DESEq2 (v1.10.1; Love et al. 2014), with all required dependencies.

3.2.6. Selecting high confidence putative miRNAs for further study.

The resulting mature sequences from both prediction tools were sorted by complete sequence homology into five groups: 1) mature miRNA sequences (with star strand) predicted by ShortStack

only, 2) mature 'N15' sequences (without star strand) predicted by Shortstack only, 3) mature miRNA sequences (with or without star strand) predicted by miRDeep-P only, 4) mature sequences present in both the miRDeep-P and ShortStack miRNA sets, and 5) mature sequences present in both the miRDeep-P and ShortStack 'N15' sets. These are illustrated in Figure 3.3. The miRNAs predicted by both ShortStack (with star strands) and miRDeep-P (Group 4) were selected as a high potential, high confidence list for further analysis.

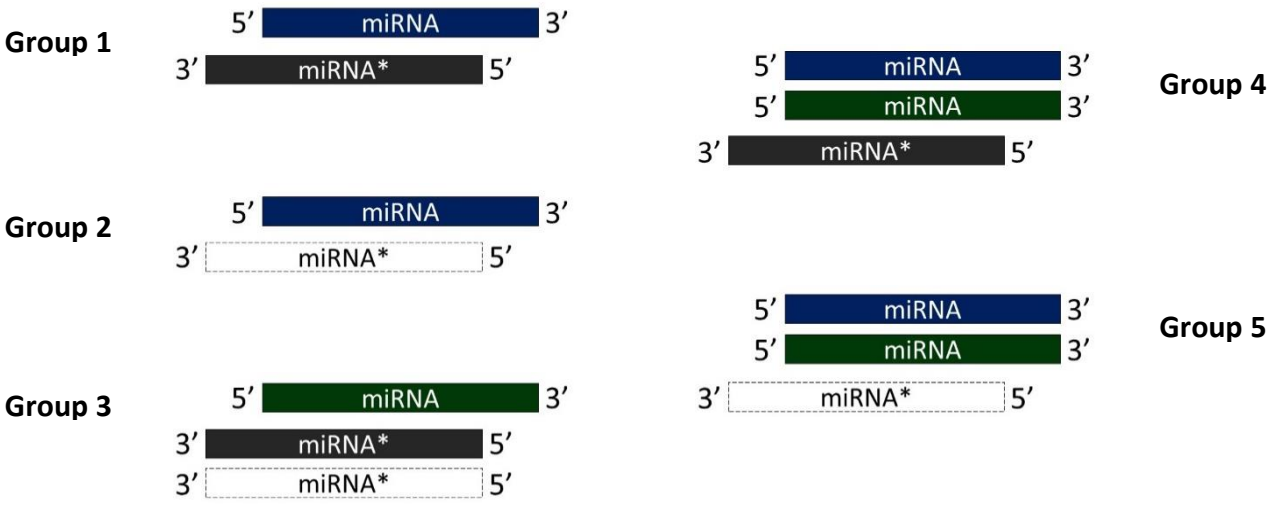

Figure 3.3: The combined putative miRDeep-P and ShortStack 'miRNAs' form 5 distinct groups. Mature miRNA strands predicted by each tool are shown in blue (ShortStack) and green (miRDeep-P). Where both colours are shown, the identical miRNA sequence was predicted by both tools. The presence or absence of a detectable star strand is indicated in grey or white respectively. Where both miRNA* options are shown, either may be the case.

3.2.7. Searching for the predicted miRNAs in miRBase.

Each of the selected high confidence mature miRNAs was individually blasted (BLASTN) against miRBase (v21), the primary online repository for miRNA sequence data and annotation (Griffiths-Jones et al. 2006), in order to identify homologous miRNAs and possibly the miRNA family to which they belong. An E-value cut-off of 10 was used to search against all mature miRNA sequences belonging to members of the Viridiplantae.

3.3. Results

3.3.1. Plant material.

X. humilis leaves were previously collected, with all leaves used for sRNA sequencing (miRNA prediction) and total RNA sequencing (transcriptome assembly and lncRNA prediction) being collected during a single desiccation cycle (Chapter 2). As such leaves corresponding to the stages used for the transcriptome study and lncRNA prediction could not be used. Therefore, following leaf collection and total RNA sequencing, the remaining leaves were selected and grouped into five bins corresponding to relative water contents of 90%, 70%, 50%, 30% and 10% RWC (±5% RWC). Total RNA was extracted from individual leaves, without a column in order to retain all sRNAs, and quality assessed via Nanodrop and gel electrophoresis. Samples deemed to be of sufficient quality, showing clear RNA bands, little to no degradation and low contamination were pooled to form three independent biologicals for each of the five selected RWCs (Table 3.2, Figure 3.4). While no pooled samples or extracts fall within the range of the previous total RNA sequencing pools (Chapter 2), it is apparent that the RWC of some new extracts are closer to RWC of extracts in the previous RWC pools, than to some of the samples in their own (Figure 3.5). Rather than representing clear distinct RWCS, the two sets of 5 RWCs chosen for the two sequencing experiments appear to form more of a continuous spectrum. Pooled samples were assessed on a Bioanalyzer. The RNA Integrity Number (RIN) values obtained were lower than the desired value of at least 6.5 (BGI). This was previously seen for the total RNA sequencing experiment, and the reasons for the low observed RIN values are discussed Chapter 2. The obtained RIN values were nevertheless compared to those obtained for the previous total RNA pools, which despite their low RIN values were of high integrity and suitable for RNA-Seq. The RNA pools for sRNA-Seq showed more consistent quality, within the bounds of what was previously obtained (Figure 3.6). On account of this, and the gel electrolysis images showing clear RNA bands with minimal degradation, the RNA was deemed suitable and subsequent processing and sRNA sequencing performed

Table 3.2: Leaf samples for RNA Pooling and sRNA sequencing. RNA extractions corresponding to the indicated leaves, letter in brackets denoting the source tray, contributed equally (ng) to each pool. The RNA integrity number as determined by BGI is given for each pooled sample.

Pool	90% RWC			70% RWC			50% RWC			30% RWC			10% RWC		
	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl	Leaf	RWC	ng/µl
A	(A)001	91,2%	263,4	(A)012	67,4%	756,2	(A)013	53,9%	505,3	(A)071	24,1%	251,8	(A)053	7,4%	1048,2
	(B)007	89,4%	258,6				(B)043	48,1%	554,1	(A)036	31,9%	261,8	(A)070	7,7%	415,3
							(C)016	51,1%	629,9	(C)037	27,4%	815,4	(C)054	12,3%	1114,4
													(C)076	7,5%	1180,7
													(B)048	10,9%	392,0
	RIN: 4,5			RIN: 5,2			RIN: 4,6			RIN: 5,0			RIN: 5,3		
B	(C)010	96,6%	561,8	(C)020	66,2%	430,1	(A)028	52,9%	327,0	(A)034	31,4%	218,3	(A)046	6,8%	734,2
	(B)020	91,8%	297,1	(C)017	73,5%	330,6	(B)041	48,8%	742,0	(A)044	32,5%	690,1	(C)050	12,1%	828,9
							(B)056	45,8%	542,1	(C)042	29,4%	695,2	(C)082	8,3%	1015,0
							(C)019	48,3%	775,2				(B)071	9,8%	673,6
													(B)063	10,3%	645,4
	RIN: 4,9			RIN: 5,0			RIN: 4,4			RIN: 4,8			RIN: 5,1		
C	(B)009	95,3%	629,2	(B)042	68,8%	572,1	(A)029	54,4%	378,6	(A)032	32,4%	487,2	(A)059	12,6%	304,9
	(B)010	89,0%	307,8				(A)025	48,1%	605,2	(A)022	25,0%	691,2	(A)052	9,8%	373,9
							(B)014	45,6%	230,4	(B)059	30,5%	379,7	(C)060	9,7%	907,2
							(C)006	54,6%	246,1				(C)038	9,1%	692,9
													(B)076	9,4%	587,4
													(B)072	10,9%	494,1
	RIN: 5,1			RIN: 5,1			RIN: 4,6			RIN: 4,5			RIN: 5,5		

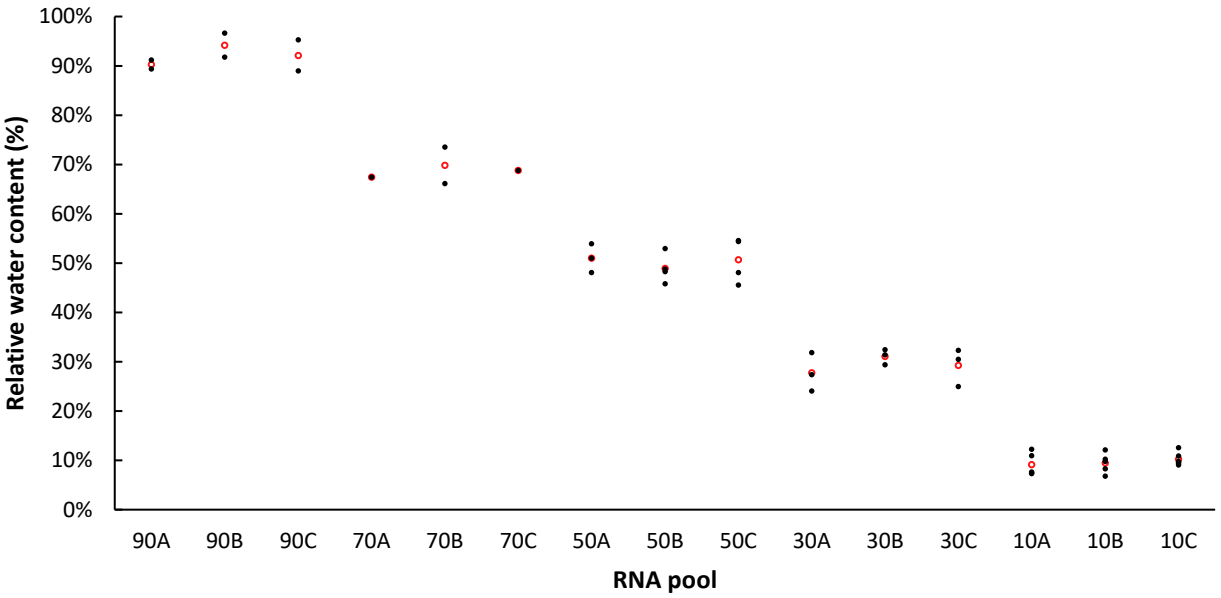

Figure 3.4: Leaf sample RWCs for RNA pooling and sRNA sequencing. The RWC of individual contributing leaves are shown in black, with the average RWC for each RNA pool shown in red.

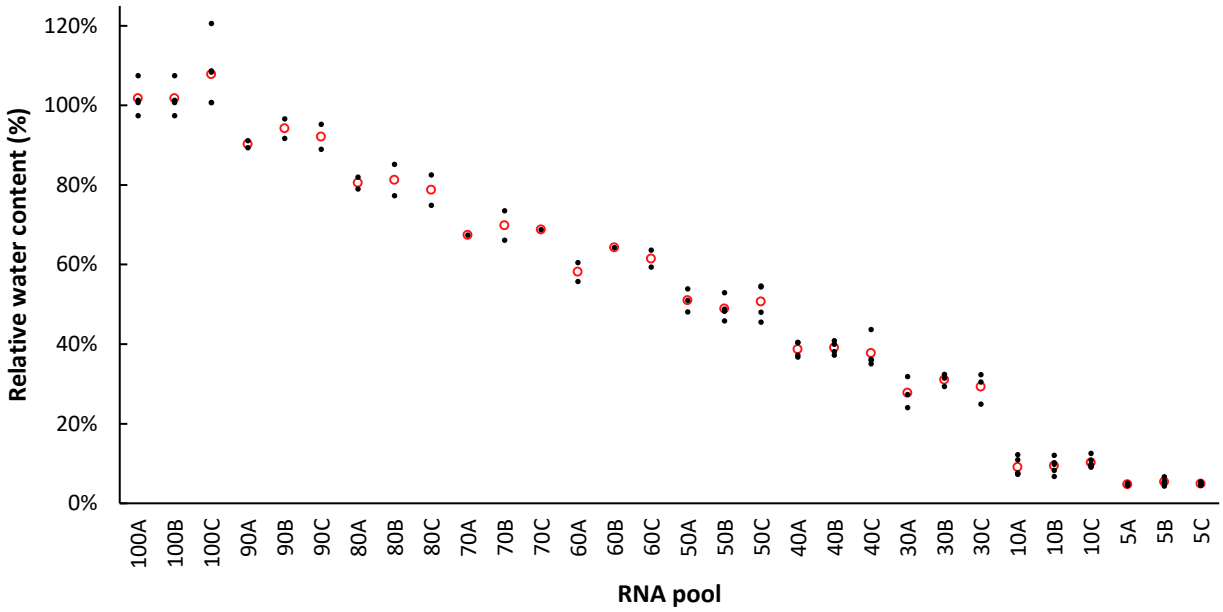

Figure 3.5: All leaf sample RWCs for both RNA pooling events. The RWC of individual contributing leaves are shown in black, with the average RWC for each RNA pool shown in red.

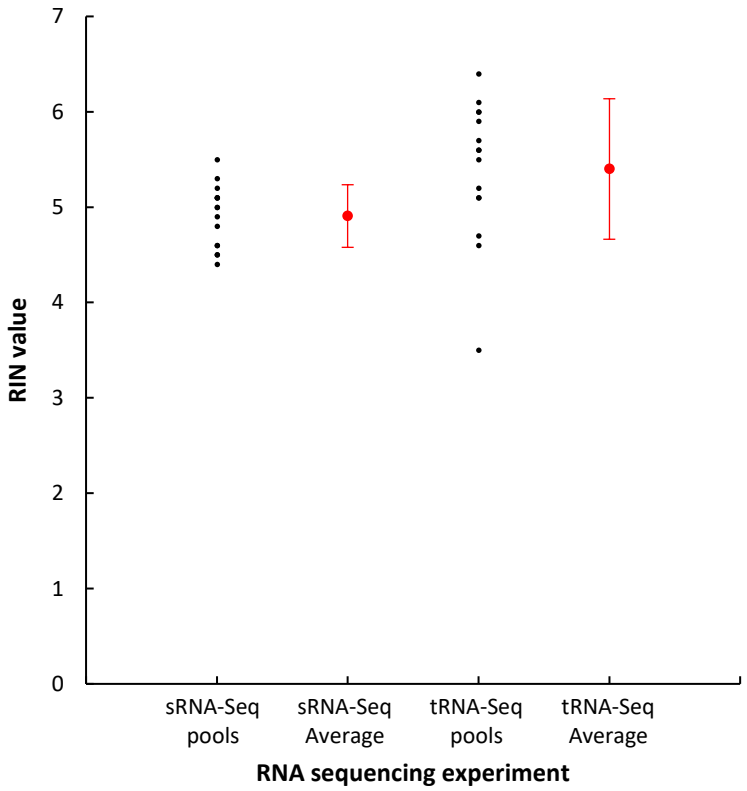

Figure 3.6: 15 RNA pools RIN values for both RNA pooling events. The RWC of individual contributing leaves are shown in black, with the average RWC for each RNA pool shown in red.

3.3.2. sRNA sequencing data

Following pre-processing, to remove adapter sequence and low-quality reads, the clean read data consisted of 182 million single-end sRNA sequences, spread across the 15 experimental samples with at least 10 Mb of reads per sample. Each sample had approximately 12.1 million reads (± 0.25 million SD), and approximately 14.65 % (± 1.55% SD) of these were unique reads (Figure 3.7). These correspond to 16.3 million unique reads overall. The sRNA sequences were analysed for size distribution (Figure 3.8). As seen, the overall number of reads and the individual read distributions are relatively uniform for all 15 samples. In the libraries of all 15 samples, peaks in read abundance are observed at 21 (25.9%) and 24 (25.1%) nucleotides in length. These have previously been observed as the major size classes for plant sRNAs both during (Thiebaut et al. 2014) and independent of desiccation (Wan et al. 2012), and correspond to typical length of plant repeat associated small interfering RNAs (*rasiRNAs*) and miRNAs.

Due to the importance of low abundance, possibly in single copy, products of miRNA biosynthesis to the miRNA prediction algorithms e.g. star strands, as well as the low abundance of some miRNA species, further pre-processing, trimming and error correcting was not performed. Any artefacts resulting from sequencing error should not map to the genomic sequence and therefore not impact on subsequent prediction. The clean reads were therefore considered of sufficient quality for subsequent miRNA prediction. As all plant miRNAs are expected to be 20-22bp in length, all reads less than 18bp were discarded to facilitate faster miRNA prediction, as indicated in Figure 3.8. The effect of removing reads less 18bp was minimal on overall read count, resulting in a 2.2% reduction, as seen in Figure 3.7.

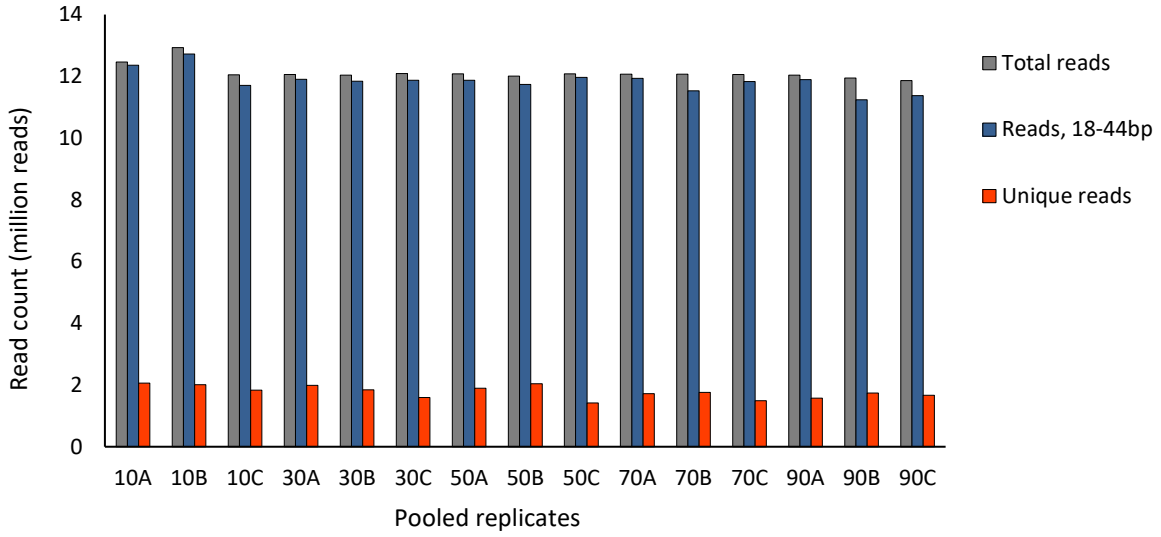

Figure 3.7: Read counts for sRNA sequencing data. The overall clean read count, number of reads remaining following exclusion of reads less than 18bp and the number of unique reads in each of the 15 samples are shown.

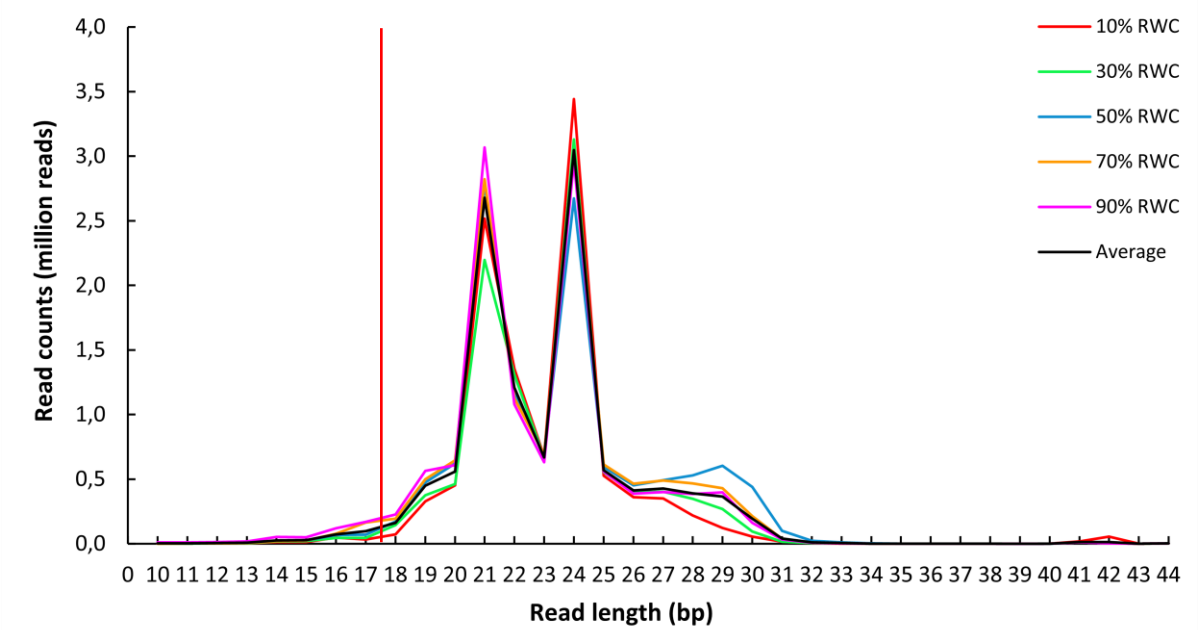

Figure 3.8: Small RNA reads sizes distribution. Graph depicts the length distribution of redundant small RNA sequences averaged for each of the 5 relative water contents. Peaks in read abundance are observed at 21 (25.9%) and 24 (25.1%) nucleotides in length, corresponding to the typical length of plant *rasiRNAs* and miRNAs respectively. The red line indicates the 18bp minimum read length cut-off point.

3.3.3. Bioinformatic prediction of putative miRNA sequences.

ShortStack and MiRDeep-P were used to independently predict putative miRNAs from each of the 15 sequencing samples, using as close to equivalent settings as possible. ShortStack is computationally light and fast to run, and outputs putative miRNAs with detectable star strands (miRNAs) and putative

miRNAs without detectable star strands (N15s) separately. Overall ShortStack predicted 71 miRNAs and 6262 N15s, while miRDeep-P predicted 952 putative miRNAs. These correspond to a combined set of 6964 unique putative miRNA sequences.

3.3.4. Principle Component Analysis

In order assess the initial, full sRNA-Seq datasets for variation and separation by RWC, as well as to compare the robustness of the pooled biological replicates for each RWC, a Principal Components Analysis (PCA) was performed. All sRNA-Seq reads were mapped to the *X. humilis* genome, generating read counts for all contigs with mapped reads. A PCA was then performed on the regularised-log (rlog) transformed raw counts for the full complement of mapped sRNA-Seq reads present in each sample (5 RWCs, 3 replicates). The PCA was performed on the regularised-log transformed (rlog) read count data for the 1080 remaining 'miRNAs', using the R DESeq2 package from Bioconductor. The two principle components most able to explain the data, are plotted in Figure 3.9A.

Generally, biological replicates for each of the 5 RWCS are expected to cluster together. Likewise, RWCs with similar patterns of expression are expected to group closer together. In Figure 3.9A the principle components each account for a relatively small fraction of the total sample variance, 9.9% (PC1) and 9.6% (PC2) respectively, especially in comparison to PC1 (81%) for the lncRNA analysis (Chapter 2). The small PCAs suggest that the data is noisier with more general data variability. Despite this, we can see that the 15 RNA-Seq samples cluster by replicates and separate by RWC along principle component 1 (PC1). Only the 90% and 50% RWC samples show any overlap. While the PCs are smaller, this separation appears as good, if not cleaner, than that observed for the lncRNA data. This indicates RWC specific expression of sRNA transcripts is detectable in the sequencing data and that the RNA-Seq datasets are of good quality for lncRNA prediction and analysis. PC2 does not appear to explain any sample variance on account of RWC.

In order to also compare the predicted miRNA complements for the 15 sequences samples, another Principal Components Analysis (PCA) was performed. Raw read counts were generated for the combined set of 6964 unique putative miRNA sequences predicted by ShortStack and miRDeep-P, for each of the 15 sRNA-Seq datasets (5 RWCS, 3 replicates). The combined set of read counts for each miRNA gene were filtered to remove all miRNAs not possessing at least 1 RWC with at least 2 replicates

of at least 10 reads each. The PCA was performed on the regularised-log transformed (rlog) read count data for the 1080 remaining 'miRNAs'. The two principle components most able to explain the data, are plotted in Figure 3.9B.

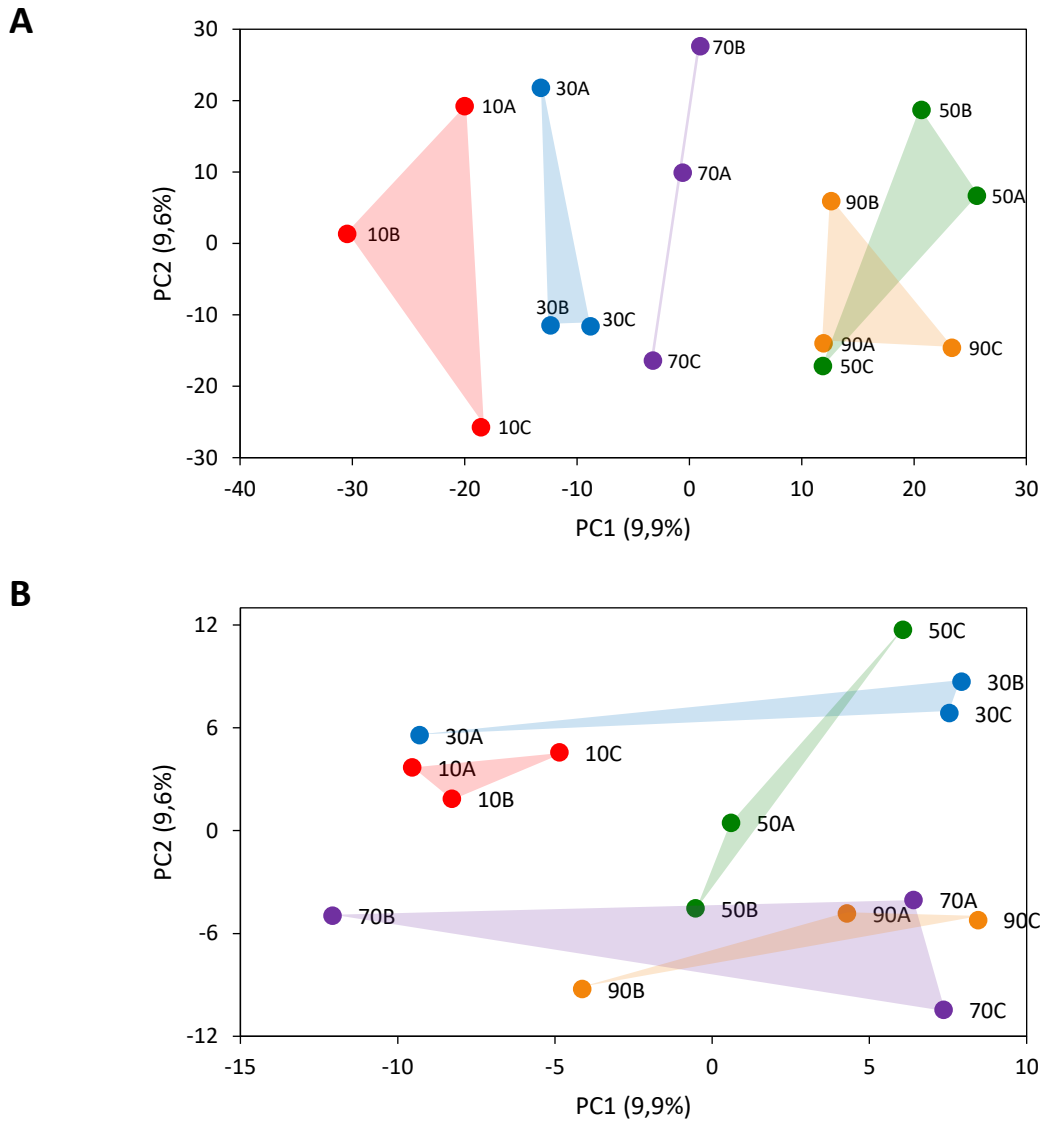

Figure 3.9: Principle component analysis of the regularised log read counts for the original 15 sRNA-Seq libraries (A) and the predicted set of all putative miRNA genes (B). To obtain read counts for only the full length transcribed genes present in both the *X. humilis* genome and the initial sRNA-Seq data, raw scaffolds counts were generated by Bowtie mapping the 15 sRNA-Seq libraries against the *X. humilis* genome. miRNA read counts were generated by Bowtie mapping the 15 sRNA-Seq libraries against the combined set of 6964 putative miRNAs/'N15s' predicted by miRDeep-P and ShortStack. In both cases, the obtained counts were filtered to remove scaffolds/miRNAs not possessing at least 1 RWC with at least 2 replicates of at least 10 reads each. The remaining raw counts were converted to regularized log (rlog) values using DESeq2 (blind=TRUE), and the two primary principle components used to create the diagnostic PCA plot. By default, only the 500 scaffolds/miRNAs showing the highest read count variance between samples are used to create the plot. The regions between lines directly connecting RWC replicates has been coloured, to give an indication of when replicates from two or more RWCs share space on the component plane, indicating shared RWC characteristics.

While the PCA for the putative miRNAs does not directly resemble the full sRNA-Seq PCA, it does share many features. The relative contribution of the principle components (%) remain very similar, and the 15 samples still separate by RWC, although now less neatly and along PC2. All low RWCs (10%, 30%, 50%) appear to group high along PC2, while the high RWCs (70%, 90%) group low down on the PCA. Only the 50% RWC, the most intermediate water content, appears to not group along PC2 and has replicates found on both sides of the RWC divide. It is unclear why the PCA is less neat, with less clear separation, but the fact that the RWC samples do tend to group by shared RWC and separate by high or low RWC indicates RWC specific miRNA activity should be detectable and analysis can proceed.

3.3.5. Selecting a high confidence set of miRNAs.

In order to select high confidence candidate miRNAs for interaction mapping (Chapter 4), the overlap between the predicted ShortStack miRNAs (with star strands) and the miRDeep-P miRNA predictions was identified, as shown in Figure 3.10. Overall 41 miRNAs, with star strands, predicted by both bioinformatics tools, make it through to the final list of selected miRNAs. While these are regarded as the highest confidence dataset, downstream analysis will focus on them alone.

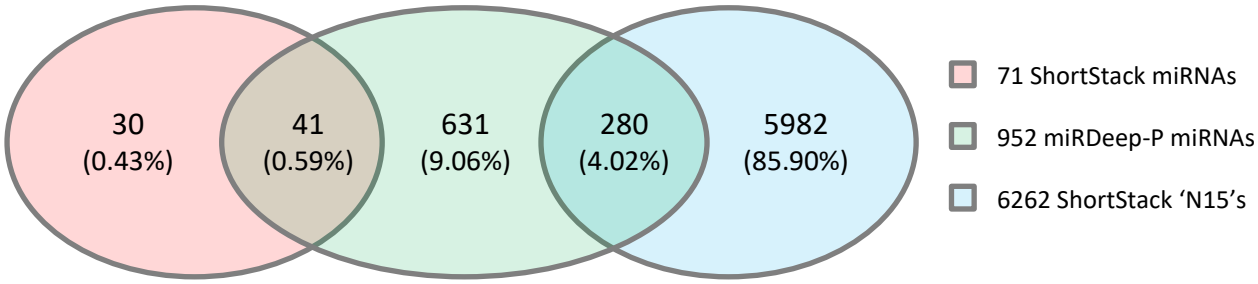

Figure 3.10: In silico predicted putative miRNAs. Two tools, miRDeep-P and ShortStack, bioinformatically predicted 6964 unique putative miRNA sequences. The 'N15' denotes ShortStack sequences that meet the criteria of miRNAs, but without a detectable star strands. The 41 putative miRNAs predicted by both tools were selected as high confidence candidates for subsequent analysis.

3.3.6. Differential expression analysis

While differential expression is not required for miRNAs to be of interest, in that their effective levels of availability can be modulated through competitive endogenous lncRNA levels and the presence of lncRNA 'sponges', miRNAs that are found to be DE are of particular interest. Raw read counts for the

full set of 1080 count-filtered putative miRNAs were assessed for differential expression using the R DESeq2 package from Bioconductor.

Differential expression was tested for using a log-likelihood ratio test (LTR), testing for genes that are differentially expressed between any of the five RWCs. A significance cut-off was selected at a false discovery rate (FDR) less than 0.01. Of the 1080 putative miRNAs, 0 were discarded for having read counts too low to reliably determine significance, likely as a result of the read count filtering. 2 (0.19%) were discarded as outliers (FDR cannot be determined), and 21 (1.98%) were found to be differentially expressed over the course of desiccation, at an FDR <= 0.01. 18 (1.7%) of these showed up regulation (LFC>0) over the course of desiccation, while 3 (0.28%) were found to be downregulated. This indicates that some miRNAs are themselves differentially expressed during dehydration, but does not yet confirm any regulatory role. Of these 21 DE miRNAs only one, designated 'Xh_shared_miRNA41', belongs to the core set of high confidence 41 miRNAs predicted by both miRDeep-P and ShortStack. An additional 11 candidate miRNAs were found to be differentially expressed at an FDR <= 0.05. Of these, only 1 additional core miRNA was found, 'Xh_shared_miRNA14'. These two core set miRNAs both show high levels of expression (high read counts) as shown in Table 3.3, and are both upregulated over the course of desiccation, as indicated in Figure 3.11.

Table 3.3: Core set miRNAs showing differential expression.

miRNA	Mean replicate reads	Adjusted p-value
Xh_shared_miRNA14	4713.9	0.0410
Xh_shared_miRNA41	8493.5	0.0058

While only 2 of the 41 high confidence miRNAs were found to have statistically significant differential expression, this may not be a result of constant expression, but rather may simply reflect the high variability observed between RWC replicates, as indicated by the large error bars in their expression profiles (Figure 3.11) and as suggested by the PCA of predicted miRNAs. Furthermore, many of the miRNAs had relatively low counts making it difficult to differentiate biological variation from simple variation in the number of reads sequenced between samples.

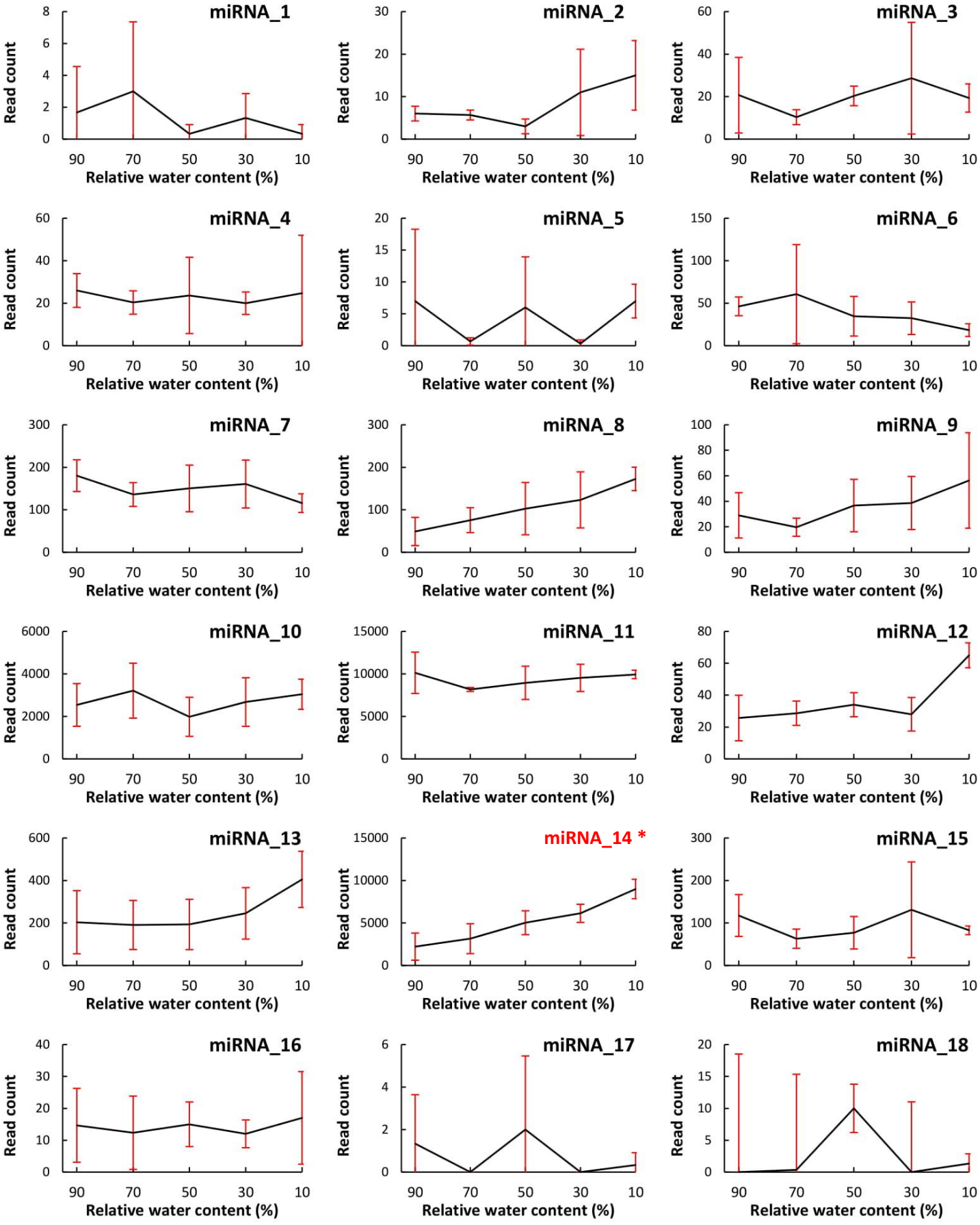

Figure 3.11 part 1: Expression profiles for the 41 high confidence miRNAs, selected for interaction mapping. For each miRNA, normalised read counts were averaged at each RWC and plotted along with the standard deviation at each RWC (red).The names of differentially expressed miRNAs are indicated in red, with an asterisk.

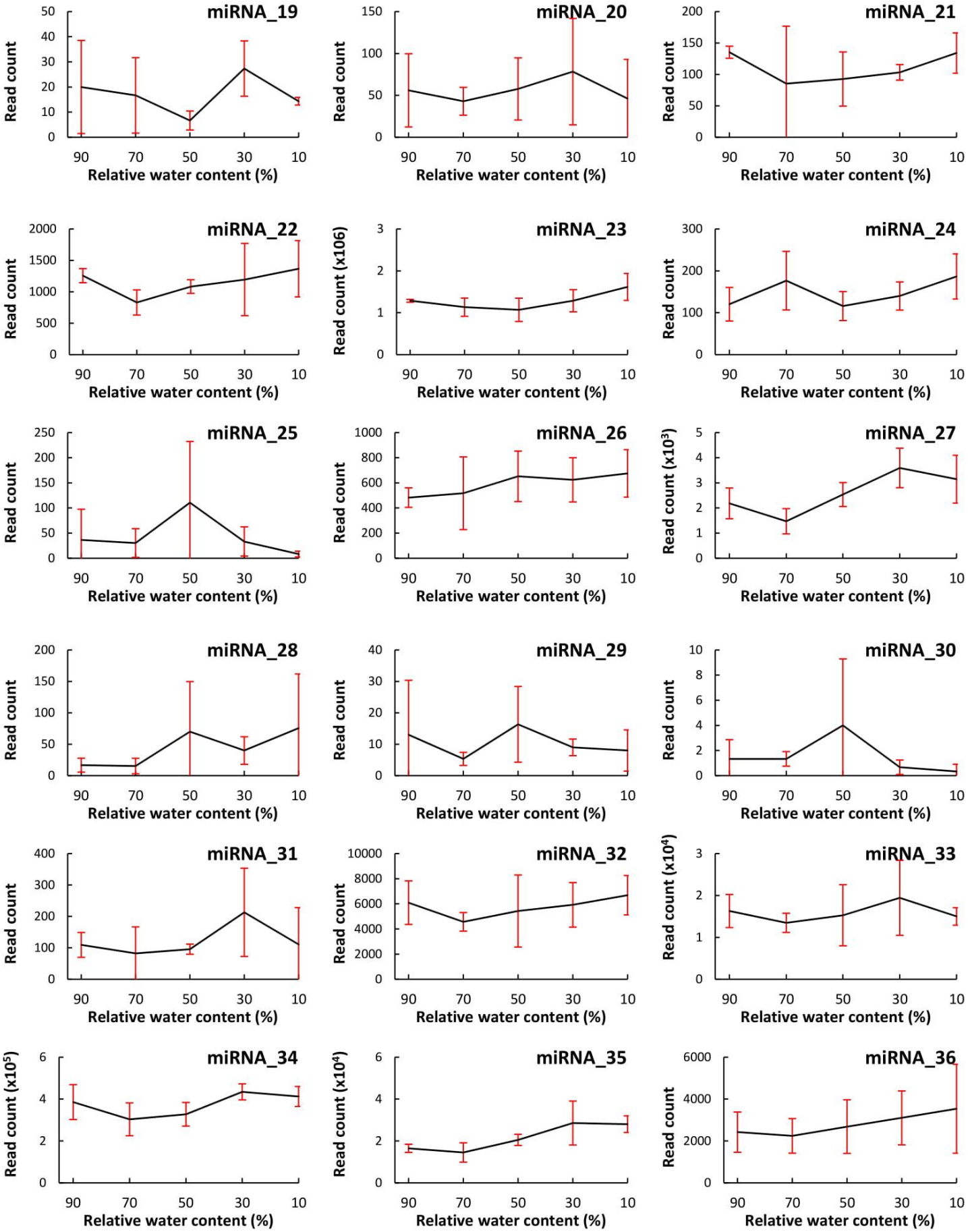

Figure 3.11 part 2: Expression profiles for the 41 high confidence miRNAs, selected for interaction mapping. For each miRNA, normalised read counts were averaged at each RWC and plotted along with the standard deviation at each RWC (red). The names of differentially expressed miRNAs are indicated in red, with an asterisk.

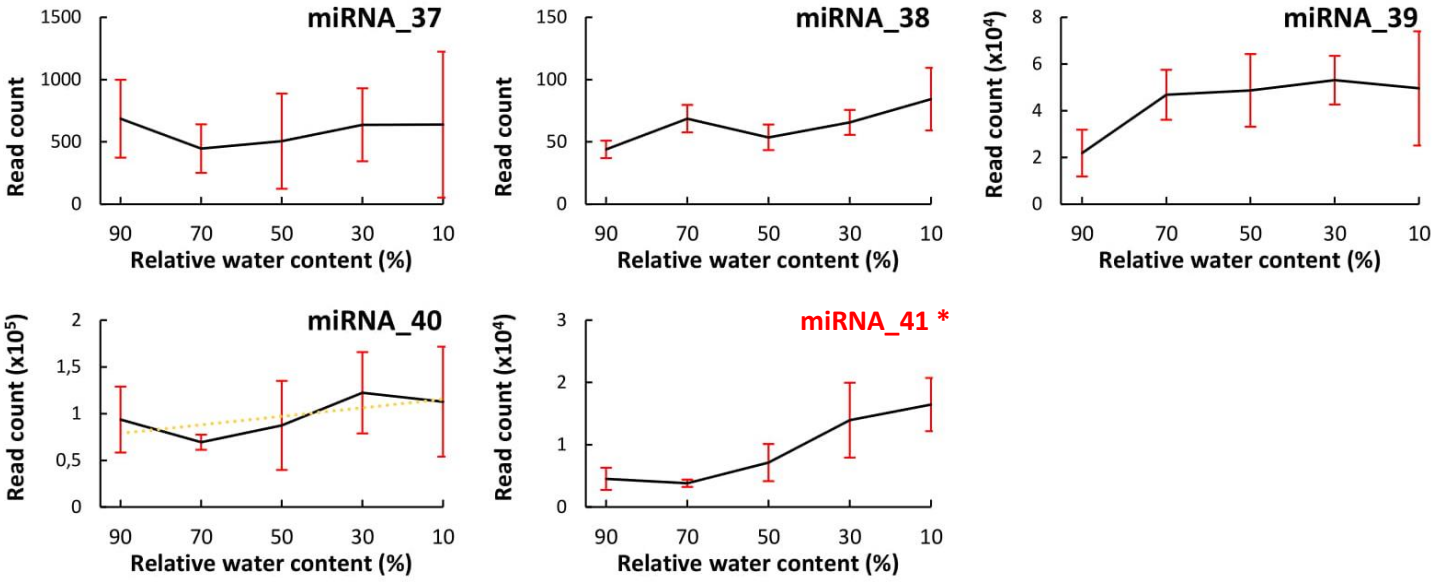

Figure 3.11 part 3: Expression profiles for the 41 high confidence miRNAs, selected for interaction mapping. For each miRNA, normalised read counts were averaged at each RWC and plotted along with the standard deviation at each RWC (red).The names of differentially expressed miRNAs are indicated in red, with an asterisk.

3.3.7. miRBase Annotation

The mature sequences of the 41 highest confidence miRNAs were blasted against all mature plant (Viridiplantae) miRNA sequences in the miRBase online repository (Table 3.4). Twenty-six of the 41 selected miRNAs showed significant hits to known plant miRNAs. In each case, the selected miRNA hit multiple separate miRNA sequences, from multiple plants and studies, all belonging to the same miRNA family. The high to complete sequence identity of these matches indicates the selected miRNA is likely to be a homologue of the respective conserved miRNA family. 14 such definite families were identified, with 7 families having multiple representatives present in the 41 selected miRNAs (Figure 3.12).

Table 3.4: miRBase BLASTN results for the 41 selected miRNAs. The mature miRNA sequence for each of the selected miRNAs was blasted against the miRBAse Viridiplantae mature miRNA sequences. An E-value cut-off of 10 was used, and the number of hits limited to 100. DE miRNA genes (FDR <= 0.05) are shown in red.

miRNA	ID	Mature sequence	Length	miRBase Hits	Top hit	Alignment length	Evalue	Identity of miRNA family
1	Xh_shared_miRNA_1	AAAGGGGTTCGAGCTGTTGGAAGA	24	0	-	-	-	-
2	Xh_shared_miRNA_2	ACATTCCTCATCTTCGGCAAA	21	0	-	-	-	-
3	Xh_shared_miRNA_3	AGAAGAGAGAGAGTACAGCTT	21	98	osa-miR529b	21	0,002	miR529
4	Xh_shared_miRNA_4	AGAATCTTGATGATGCTGCAT	21	100	ath-miR172a	21	0,002	miR172
5	Xh_shared_miRNA_5	ATCATGCTGTCCCTTTGGATC	21	75	ptc-miR393a-3p	20	0,022	miR39*
6	Xh_shared_miRNA_6	ATTGGCTGCAGCGCACGGGGTCTG	24	0	-	-	-	-
7	Xh_shared_miRNA_7	ATTGGCTGCAGTGCACGGGGTCTG	24	0	-	-	-	-
8	Xh_shared_miRNA_8	CCATTAAAGACCTCGATTGCT	21	0	-	-	-	-
9	Xh_shared_miRNA_9	CGACAGAAGAGAGTGAGCAC	20	100	ath-miR156g	20	0,004	miR156
10	Xh_shared_miRNA_10	CGAGCCGAACCAATGTCACTC	21	94	mdm-miR171i	20	0,022	miR171
11	Xh_shared_miRNA_11	CGGATCCCGCCTTGCATCAAC	21	19	aau-miR168	19	0,01	miR168
12	Xh_shared_miRNA_12	CGTTGGCATGGTACTCCTACC	21	0	-	-	-	-
13	Xh_shared_miRNA_13	CTCAGGTCGGATTTCAATCGTC	22	0	-	-	-	-
14	Xh_shared_miRNA_14	CTTCGGCATTGTACTCCTACA	21	0	-	-	-	-
15	Xh_shared_miRNA_15	GAGGATGCTAAATAGGACGATAAG	24	0	-	-	-	-
16	Xh_shared_miRNA_16	GCTGTACCCTCTCTCTTCTTC	21	43	zma-miR529-3p	21	0,002	miR529*
17	Xh_shared_miRNA_17	GGGCAATTCTCCTCTGGCAGT	21	100	ppe-miR399b	20	0,022	miR399
18	Xh_shared_miRNA_18	GGGCTACTCTACTTTGGCAGG	21	100	cme-miR399g	21	0,27	miR399
19	Xh_shared_miRNA_19	TACCCCGTATGCTGTAGTCAACTT	24	0	-	-	-	-
20	Xh_shared_miRNA_20	TAGCATCTAGGAGTATGTTTT	21	0	-	-	-	-
21	Xh_shared_miRNA_21	TATCCGTCGACTGATACCCGAGAT	24	0	-	-	-	-
22	Xh_shared_miRNA_22	TCGCTTGGTGCAGGTCGGGAA	21	55	ath-miR168a-5p	21	0,002	miR168
23	Xh_shared_miRNA_23	TCGGACCAGGCTTCATTCCCC	21	100	pvu-miR166a	21	0,002	miR166
24	Xh_shared_miRNA_24	TCGGCAAGCTGTCCTTGGCTAC	22	100	zma-miR169r-3p	20	0,024	miR169
25	Xh_shared_miRNA_25	TGAAGTGTTTGGGGGAACTC	20	100	rco-miR395d	20	0,004	miR395
26	Xh_shared_miRNA_26	TGACAATGAGAGAGAGCACTC	21	53	csi-miR535	21	0,008	miR535*
27	Xh_shared_miRNA_27	TGACAGAAGAGAGTGAGCAC	20	100	csi-miR156	20	0,004	miR156
28	Xh_shared_miRNA_28	TGCACTGCCTCTTCCCTGGCTC	22	60	ppt-miR408b	21	0,002	miR408
29	Xh_shared_miRNA_29	TGCCAAAGGAGAATTGCCCTG	21	100	osa-miR399a	21	0,002	miR399
30	Xh_shared_miRNA_30	TGCCAAAGGAGAGTTGCCCTA	21	100	osa-miR399j	21	0,002	miR399
31	Xh_shared_miRNA_31	TTAGCGTCAAGGAGGACATTT	21	0	-	-	-	-
32	Xh_shared_miRNA_32	TTCCACAGCTTTCTTGAACTG	21	100	ath-miR396a-5p	21	0,002	miR396
33	Xh_shared_miRNA_33	TTCCACAGCTTTCTTGAACTT	21	100	ath-miR396b-5p	21	0,002	miR396
34	Xh_shared_miRNA_34	TTCCACGGCTTTCTTGAACTA	21	100	ptc-miR396f	20	0,004	miR396
35	Xh_shared_miRNA_35	TTGACAGAAGAGAGTGAGCAC	21	100	gma-miR156k	21	0,002	miR156
36	Xh_shared_miRNA_36	TTGACAGAAGATAGAGAGCAC	21	100	ath-miR157a-5p	21	0,002	miR156 / miR157
37	Xh_shared_miRNA_37	TTGAGCCGCGTCAATATCTCC	21	100	ctr-miR171	21	0,002	miR171
38	Xh_shared_miRNA_38	TTGGCAAGCTGTCCTTGGCTAC	22	99	zma-miR169r-3p	20	0,024	miR169
39	Xh_shared_miRNA_39	TTTCCAAAGTTCCTCCGGGCA	21	2	stu-miR7984a	12	8.4	- *
40	Xh_shared_miRNA_40	TTTGGATTGAAGGGAGCTCTA	21	100	ath-miR159a	21	0,002	miR159*
41	Xh_shared_miRNA_41	TTTTTTCTGATGCTGCCCGAAC	22	0	-	-	-	-

* Indicates the predicted Xh_miRNA hit to additional miRNA families. These are low identity or partial hits and are therefore likely hits to miRNAs from distinct but related miRNA families.

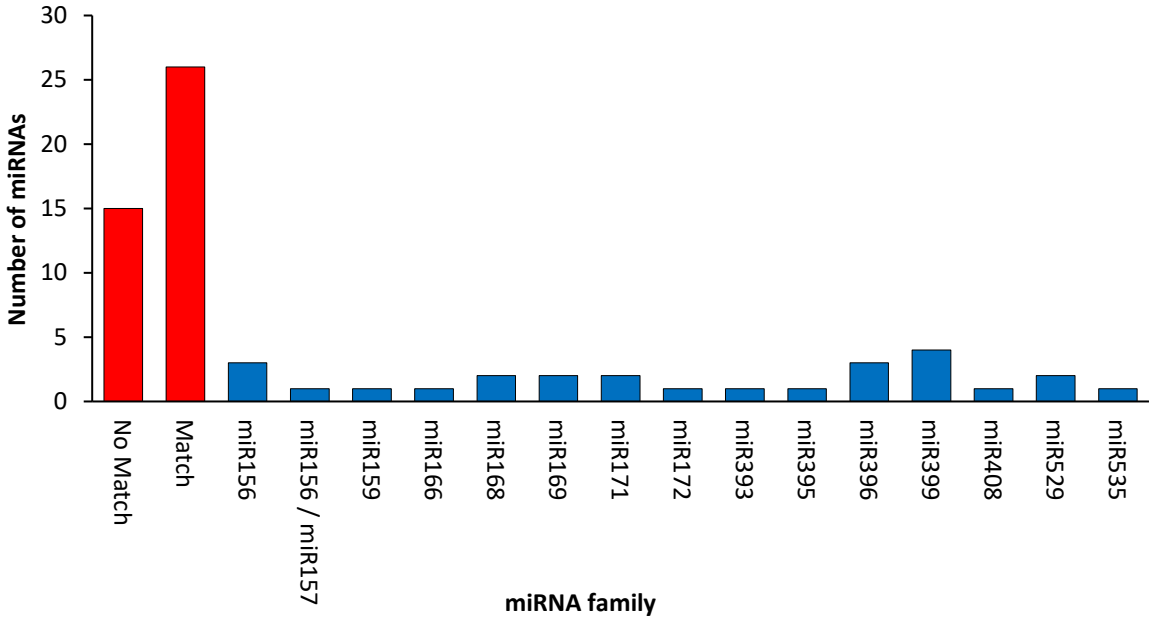

Figure 3.12: Number of the 41 selected miRNAs with sequence homology to each of the identified miRNA families. Homology was identified by blasting the mature miRNA sequences against the Viridiplantae (green plant) mature miRNA sequences deposited in miRBase. The number of selected miRNAs with and without homology to known miRNA families are shown by the bars in red. The number of selected miRNAs with homology to known miRNAs in each respective miRNA family present are shown by the bars in blue.

3.4. Discussion

3.4.1. sRNA-Seq and miRNA prediction.

Following sRNA-Seq, miRNA prediction was performed for each of the 15 samples using two miRNA prediction tools. A set of 6964 unique possible miRNA sequences were predicted, including the N15 sequences predicted by ShortStack to meet all criteria of miRNAs bar the presence of detectable star strands. This set is highly likely to contain a number of false positives particularly within the subset lacking star strands. These may instead be members of other classes of similar sRNAs. Of these possible sequences, a core set of 41 high confidence miRNA sequences, predicted by both bioinformatics tools used, were selected for downstream interaction mapping (Chapter 4). The selection of only 41 candidate miRNAs for further study excludes many predicted miRNAs that are likely to play important roles in vegetative desiccation tolerance. It was decided to focus on a small core set of miRNAs, however, in order to allow for more comprehensive analysis of each candidate, within the limited time frame of the current project. Once a pipeline has been created for downstream

analysis, miRNA target prediction and network analysis, further candidates can be analysed at a later date.

The vast majority of these 41 selected sequences showed no significant differential expression over the course of desiccation. Of the 41 selected miRNA genes, only two showed DE for a FDR <=0.05. The low number of differentially expressed miRNAs, appears surprising given our hypothesis that miRNAs play key roles activating and regulating the vegetative desiccation response. We know however, that there are two mechanisms by which miRNA transcript abundance may change to affect the desiccation phenotype: 1) Transcript abundance may change or 2) Transcript abundance may not change, but availability may change through sequestration by lncRNAs. The lack of DE indicated that the latter may be the main means of miRNA regulation during VDT. This will be explored further in Chapter 4. The two DE miRNAs both show high levels of expression (high read counts) and are both upregulated over the course of desiccation, consistent with a possible functional role. Furthermore, the lack of statistical significance may simply be a result of the high variation observed between replicates, evident in the PCAs and expression profiles. Expression profiles of all 41 miRNAs (Figure 3.11) were plotted to possibly identify interesting trends, even without statistically significance, but the high sample variation and resulting error bars make this difficult to achieve. Larger numbers of leaves in each pooled RNA sample and more biological repeats are needed to confidently tell whether or not more of the 41 core miRNAs are differentially expressed. Alternatively, RT-qPCR experiments should be done on fresh biological samples to validate the predicted mean expression profiles and to check for significance.

Although not examined in this study, the remaining 20 DE putative miRNAs (not part of the core analysis set) may play interesting roles and hold potential as good candidates for further study.

3.4.2. miRBase annotation.

In order to better annotate the selected miRNA sequences, the mature sequences were blasted against all mature miRNA sequences from green plants (Viridiplantae) within miRBase. 26 of the 41 selected miRNAs were found to be homologous to known members of 14 miRNA families. Neither of the two DE miRNAs were found to be homologues of any known families. Function is known for many of these miRNA families.

3.4.2.1. The miR156, miR529, miR172 and miR159 families.

At least 7 of the core miRNAs appear to be members of the miR156, miR529, miR172 and miR159 families. These four miRNA families are closely related through either function or regulatory interactions. miR156 and miR529 are evolutionary related miRNAs which target and repress members of the *SQUAMOSA promoter-binding protein like (SPL)* family of plant specific TFs. *SPLs* are generally activators of transcription with roles in leaf development, vegetative phase changes, control of flowering, vegetative morphology and hormone (Gibberilic acid) signalling (Morea et al. 2016). Expression of these miRNAs represses vegetative development, as would be required during drought induced dormancy/anhydrobiosis. miR159 acts via repression of the MYB33 TF, a key TF required for miR156 expression, thereby repressing miR156 levels and allowing vegetative development to proceed (Guo et al. 2017). miR172 acts downstream and is repressed by miR156 via the transcriptional regulator SQUAMOSA PROMOTER BINDING PROTEIN LIKE 9 or 10 (*SPL9/10)* (Wu et al. 2009). miR172 family miRNAs target mRNA transcripts coding for the ethylene-responsive element binding protein (EREBP) / APETALA2-like family of TFs (Riechmann & Meyerowitz. 1998). The AP2/ERF TF superfamily, however, contains 147 members in Arabidopsis, which regulate almost aspect of plant development, floral repression and stress responses between them (Nakano et al. 2006). As such, without knowing more about specific binding interactions, and the specific TFs involved, the exact role played by the identified miR172-family miRNA cannot be determined. miR156, miR529, miR172 and *SPL* have all been previously identified during studies into plant abiotic stress responses, both to drought stress (Bertolini et al. 2013; Maoa et al. 2016) and cold (Zhang et al. 2009). This suggests that our detected miRNAs are consistent with the miRNA families detected in similar studies, and that these three miRNAs appear to be promising candidates for interaction mapping.

3.4.2.2. The remaining identified miRNA families

The miR171 family of miRNAs target mRNA transcripts coding for GRAS domain or SCARECROW-like proteins, playing an important role in signalling by the phytohormones gibberellin and auxin, light signalling and meristem maintenance (Rhoades et al. 2002, Huang et al. 2017). miR171-targeted scarecrow-like proteins (SCL6/22/27) repress chlorophyll biosynthesis (Ma et al. 2014), a process that would seem necessary during dismantling of the photosynthetic machinery and loss of chlorophyll observed in poikilochlorophyllous desiccation-tolerant plants, such as *X. humilis*. The miR168 family of miRNAs facilitates AGO1-catalysed cleavage of AGO1 transcripts, maintaining AGO1 homeostasis in Arabidopsis (Vaucheret et al. 2006). This explains their general presence but does not suggest any

specific role during desiccation. miRNAs of the miR166 family regulate diverse aspects of plant development; including meristem formation, leaf patterning and floral development (Jung and Park. 2007). In *Glycine ma*x miR169 has been shown to target a gene encoding the Nuclear transcription factor Y subunit alpha (NF-YA) TF, implicated in enhancing drought stress tolerance (Ni et al. 2013). Although drought tolerance is distinct from desiccation tolerance, both forms of tolerance overlap in the expression of protective LEAs and heat-shock proteins. miR169 family miRNAs also play a key role in stress-induced flowering in plants, as previously mentioned a phenomenon consistent with the rapid post-rehydration flowing seen in *X. humilis* (Xu et al. 2013). miR399 miRNAs and their target gene PHOSPHATE 2 (PHO2) regulate inorganic phosphate homeostasis (Lin et al. 2008). Similarly, miR395 play crucial roles in sulfate homeostasis in *A. thaliana*: uptake, transport and assimilation (Liang et al. 2010). miR408 is a highly conserved family of plant miRNAs targeting and repressing genes encoding copper-containing proteins. Copper is essential for photosynthesis with the two most abundant copper proteins, plastocyanin and copper/zinc superoxide dismutase (SOD) being located in the chloroplast (Abdel-Ghany & Pilon. 2008). While it would seem that miR408 would therefore act to repress the photosynthetic machinery, during anhydrobiosis and poikilochlorophylly, SODs are in fact upregulated during desiccation. Elevation of SOD activity in resurrection plants, including *X. humilis* from below 70% RWC, acts as a free radical scavenging system protecting tissues from excess light and ionising radiation (Illing et al. 2005; Sherwin and Farrant. 1998; Farrant. 2000; Farrant et al. 2015.). The role of miR408 during desiccation is therefore unclear. miR396 is a highly conserved family of plant miRNAs which regulate conserved GROWTH-REGULATING FACTOR (GRF) family TFs, which are known to control cell proliferation in Arabidopsis leaves (Rodriguez et al. 2010b; Debernardi et al. 2012).

The 26 miRNAs identified in this study with homology to known miRNA families are known to play functional roles in key developmental and stress response pathways in other plants. It is thus likely that they may play important roles during the desiccation response in *X. humilis.* Although, further experimental testing is needed to validate their exact interactions, activity, and roles, as well as those of the 15 miRNAs with unclear relation to existing miRNA families. This finding suggests that these are good candidates for further target interaction mapping and network analysis. The fact that they are likely members of known miRNA families suggests that these miRNAs may have been co-opted into the desiccation tolerance pathway, or rather they may simply be expressed as downstream effectors of the metabolic shutdown observed during desiccation. Some may also play roles unrelated to VDT, which would be consistent with the lack of DE. They seem less likely to represent new key activators of the desiccation response. The remaining 15 miRNAs likely belong to new miRNA families, and

possibly represent unique species-specific miRNAs, common to plants. These unique miRNAs, not found in other plants, may represent good candidates for activators and regulators specifically evolved with the unique mode of desiccation tolerance present in the resurrection plants – the two differentially expressed miRNA transcripts fall into this category.

3.5. Conclusion.

In summary, this study set out to predict a set of putative miRNA sequences, expressed in the desiccating leaves of the resurrection plant *X. humilis*, to be used as candidate regulators for further investigation into regulation of the vegetative desiccation response. Numerous putative miRNAs are expressed, many differentially and at high levels in the leaves of *X. humilis* over the course of desiccation. Some of these may be involved in regulation of the desiccation response, but experimental evidence for this is required. From these, 41 high confidence miRNAs were selected for further study. The remaining putative miRNAs may also play key regulatory roles during desiccation but are not carried forward for network analysis (Chapter 4). These would be of interest for future study, possibly focusing on those showing differential expression. Of the 41 selected miRNAs, one showed differential expression for an FDR <=0.01, and two for an FDR <= 0.05. While most of the selected miRNA did not show DE, they may still represent key regulators, with titration by lncRNAs controlling their effective abundance. Twenty-nine of these 41 selected miRNAs appear to be members of known miRNA families. The interplay between the selected 41 miRNAs, the predicted desiccation lncRNAs (Chapter 2), and the desiccation transcriptome to form a regulatory network(s) will be examined in the subsequent chapter (Chapter 4).

Chapter 4: Prediction and analysis of the ceRNA-miRNA networks involved in vegetative desiccation.

4.1. Introduction

4.1.1. Introduction / background

LncRNAs act as 'miRNA decoys' also known as 'target-mimics' or 'miRNA-mimics' in plants (Franco-Zorrilla et al. 2007) and 'miRNA sponges' in animals (Ebert et al. 2007), interfering with and regulating the extent of miRNA binding and miRNA mediated regulation of mRNA targets. This ability to function as competitive endogenous RNAs (ceRNAs) allows for the rapid and precise fine-tuning of effective miRNA levels without the need for transcript degradation or miRNA biosynthesis (Rubio-Somoza and Weigel. 2011; Salmena et al. 2011). Modulation of miRNA levels through lncRNA expression and activity, is effectively able to create a differential in miRNA activity under constant miRNA expression levels (Wu et al. 2013). By sequestering and inactivating repressive miRNA transcripts, lncRNAs relieve miRNA mediated post transcriptional repression, acting as activators of post transcriptional activity. The classic example of this in plants is the lncRNA Induced by Phosphate Starvation 1 (*ISP1)* which is a decoy for ath-miR399. A 3nt bulge in the lncRNA between the 10th and 11th of the miRNA prevents miRNA mediated cleavage of the lncRNA. Inactivation of ath-miR399 by *ISP1* results in upregulated expression of the miR399 target *PHO2* (Franco-Zorrilla et al. 2007). Binding between miR160 and ath-eTM160-1, another miRNA-decoy pair, is illustrated in Figure 4.1 below. Some evidence to suggest that binding of miRNAs by target mimics can not only sequester miRNAs out of the active miRNA pool, but also induce miRNA degradation (Todesco et al., 2010; Ivashuta et al., 2011; Banks et al., 2012; Yan et al., 2012).

Figure 4.1: Predicted base-pairing interaction between miR160 and its decoy ath-eTM160-1, showing the nucleotide bulges (red) between the 9th and 11th positions of the miRNA which prevent miRNA directed cleavage of the lncRNA decoy. Adapted from Wu et al. 2013, Fig 3A.

4.1.2. Prediction of lncRNA-miRNA interactions.

The prediction of regulatory RNA-RNA interactions rests on two primary considerations: 1) Does the nucleotide composition of the ncRNA and target allow for effective binding, and 2) is binding and co-folding between the two RNAs thermodynamically favourable.

Plant miRNAs generally show near perfect sequence complementarity to targets, which facilitates easy and confident computational prediction (Jones-Rhoades & Bartel. 2004), with most plant miRNAs appearing to regulate a single or small group of related transcripts. Many tools exist for the direct mapping of miRNAs and systematic target prediction. Conservative direct mapping approaches, based purely on nucleotide composition and requiring near perfect sequence, have been performed using BLAST (Altschul et al. 1990), FASTA search (Pearson and Lipman. 1988) and GUUGle (Gertach and Giegerich. 2006), which functions like BLAST but allows the consideration of GU mismatches, to name a few (Tafer and Hofacker. 2008). Utilising sequence complementarity alone however has multiple shortfalls. These include the uncertainty of how many mismatches should be permitted, the loss of decoy lncRNAs as a result of mismatched loops, the absence of known authentic pairs in predicted interaction sets and the absence of any thermodynamic considerations (MFE of interaction) being taken into account. These shortfalls suggest more rigorous and refined approach is required (Rhoades et al. 2002; Jones-Rhoades and Bartel. 2004). The exclusion of thermodynamic analysis in particular leads to a lack of sensitivity when assessing interactions with large or complex targets, such as lncRNAs (Tafer and Hofacker. 2008).

4.1.2.1. Minimum free energy

Many algorithms exist to predict RNA binding and folding. Currently the most accurate and widely accepted algorithms function through free energy minimization (Turner et al. 1988; Zuker. 2000; Zuker and Stiegler, 1981). Gibb's free energy is the thermodynamic stability of binding between two

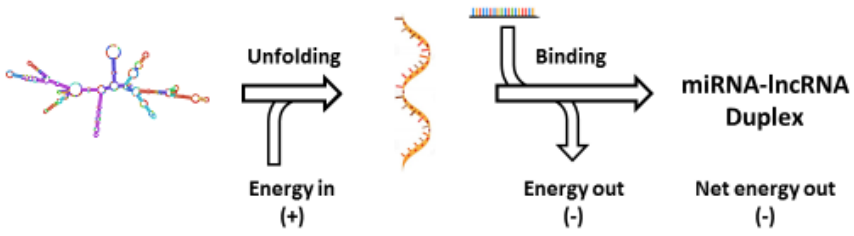

Figure 4.2: Diagram illustrating the primary energy considerations for miRNA-lncRNA binding.

complementary RNA transcripts. The energy required to unfold the lncRNA secondary structure and the miRNA-lncRNA binding energy contribute in an additive fashion (Fig 4.2) for a miRNA-lncRNA binding interaction (Matthews et al 1999; Matthews 2004).

4.1.2.2. RNA-folding algorithms

Multiple algorithms, including RNAhybrid, RNAduplex, RNAplex, RNAcofold, Pairfold, NUPAK and RNAup, have been developed to predict stable RNA-RNA interactions (Table 4.1).

Table 4.1: Summary of RNA-folding algorithms and their key features.

Tool	Advantages	Disadvantages	Key Features
RNAhybrid	Fast	Omits computation of secondary structures following binding	Simple. Relies on near perfect complementarity
RNAduplex			
RNAplex	Fastest	Less accurate energy prediction	Further simplifies the energy model
RNAcofold	Accounts for co-folding	Restricts types and positions of secondary structures. Excludes legitimate target sites.	Concatenates the two RNA to applying a modified RNA folding algorithm
Pairfold			
NUPACK			
RNAup	Allows for loop structures and binding to any unpaired region	Interactions are limited to a single binding site. Computationally demanding.	More complex and less restricted folding algorithm.

RNAhybrid (Rehmsmeier et al. 2004) and RNAduplex (Vienna RNA package, R. Lorenz et al. 2011) are the simplest and fastest algorithms for computing RNA binding and MFE. Both tools are able to quickly screen large datasets to predict multiple potential binding sites, representing the most energetically favourable hybridization, in large target RNAs. While the energy required to unfold the target RNA is calculated, these tools aim to exploit the near perfect complementarity (miRNA seed region and target) and completely omit computation of the secondary structures following binding. RNAplex (Tafer and Hofacker. 2008) is a modification to RNAduplex, with a simplified energy model. This allows RNAplex to run 10-27 times faster than RNAhybrid with a relative energy difference < 7%.

To improve accuracy and sensitivity by taking co-folding into account, RNAcofold (Hofacker et al 1994, Berhhart et al 2006), Pairfold (Andronescu et al 2005) and NUPACK (Dirks and Pierce. 2004) all function by concatenating the miRNA and target RNA into a single sequence and applying a modified RNA folding algorithm. The algorithms significantly simplify energy calculations by restricting the positions and types of secondary structures allowed (disallowing pseudoknots) as well as restricting the location

of possible binding sites. This unfortunately also excludes many legitimate target sites and interactions.

RNAup (Vienna RNA package) (Mückstein et al 2006) improves the specificity of this approach, by allowing loop structures such as pseudoknots and allowing binding to any unpaired region. However, the downside is that interactions are limited to a single binding site. Furthermore, the problem with including the pseudoknot structures is a significant increase in complexity and computational time required. According to Mückstein et al. (2008) it takes about 52 CPU days to compute all sRNA-mRNA pairs for a transcriptome on a computer with a 2.4GHz Intel Core Duo. A lack of experimental data on the energetics of these complex loop structures also means that, the optimal predicted structures may not correspond to reality (Tafer and Hofacker.2008).

The optimal strategy for assessing RNA-RNA binding and determining MFE is therefore likely a combination of using RNAplex for an initial screening for possible binding sites, followed up by a more accurate but CPU intensive screening method (for example RNAup) if further validation or a more accurate estimation of MFE is required.

4.1.2.3. Predicting the type of miRNA-lncRNA interaction and the role played by the lncRNA

While the prediction of interacting ncRNAs and target binding sites is essential for understanding lncRNA and miRNA function, the specific type of interaction also needs to be determined. Fortunately, miRNA-lncRNA interactions follow a specific and well defined set of binding characteristics, depending on the role played by the lncRNA – either as a miRNA decoy or a miRNA target (Fan et al. 2015). This allows us to differentiate between modes of lncRNA action on the basis of miRNA-lncRNA structural binding site characteristics alone.

Binding between the 9th to 12th position from the 5' end of the miRNA (middle of the miRNA binding site) represents the site at which miRNA targets are cleaved and is critical for effective target cleavage, thereby dictating the ultimate fate of the miRNA target (Jones-Rhoades et al. 2006). While up to one mismatch or gap between the miRNA and target will still allow for AGO to effectively access the cleavage site of the miRNA target RNA, further mismatches or gaps will obstruct access by the catalytic site and prevent target cleavage. Decoy lncRNAs therefore usually possess a 3nt bulge at this site (Ivashuta et al. 2011; Rubio-Somoza and Weigel. 2011). The regions outside of the miRNA binding site

appear to be highly varied without any apparent major constraints on sequence composition (Wu et al. 2013).

The presence or absence of this bulge in the miRNA cleavage site represents the major determining feature between cleavable target and decoy binding. Other rules however have been defined on the basis of known miRNA-lncRNA interactions. These rules for defining miRNA targets and decoys - which vary slightly between authors, studies and the level of desired stringency - are summarized in Table 4.2.

Table 4.2: Structural binding characteristics for miRNA binding to ncRNA cleavage targets and non-cleavable target mimics. In all cases the presence of a bulge, loop structure, or multiple mismatches in the target cleavage site acts to prevent access and cleavage of the target RNA. All positions indicated are from the 5' end of the miRNA strand.

miRNA target type	**Cleavage target**			**Target mimic**
Position	Restrictions and Requirements			
Between 9th and 12th position	1<mismatches/indels<6	3nt bulge required.	1-5 nt indel/bulge or mismatches	≤ 1 mismatch/indel
Mismatches in other regions	- miRNA 2nd-8th pos, perfect nucleotide pairing - ≤ 4 mismatches/ indels	- miRNA 2nd-8th pos, perfect nucleotide pairing - ≤ 3 mismatches and G/U pairs (other than bulge)	- No bulge allowed - Mismatch allowed at miRNA base 1 - ≤ 2 consecutive mismatches - ≤ 3 mismatches overall	- ≤ 4nt mismatches/bulges - No continuous mismatches
References	Fan et al. 2015 (Adapted from Wu. et al.)	Wu et al. 2013	Ivashuta et al. 2011	Fan et al. 2015

In order to apply such a set of rules, and to differentiate between lncRNAs acting as miRNA targets and miRNA decoys, two key pieces of information are required: 1) The identification of interacting miRNA and lncRNA species using MFE estimates and 2) The structure of the respective target site. Algorithms such Generic Small RNA-Transcriptome Aligner considers both these points.

4.1.2.4. Determining binding site structure: Generic Small RNA-Transcriptome Aligner (GSTAr)

Generic Small RNA-Transcriptome Aligner (GSTAr.pl) is a perl script for the flexible RNAplex-based alignment of small RNAs (miRNAs and siRNAs, 15-26 nts) to an established transcripome transcript set (M.J. Axtell. 2013). Following sequence-based alignment, the MFE of each query is calculated using RNAplex with default parameter settings. While GSTAr is explicitly not a miRNA target predictor, and alignments do not directly indicate interactions or the presence of miRNA cleavage sites, it does allow

for the identification of possible targets and putative splice sites. These can then be verified with further independent data, such as by degradome analysis.

One of the major strengths of GSTAr above and beyond its ability to identify miRNA-target alignments based on RNA-RNA hybridisation thermodynamic predictions, is the fact that it provides a structural output for all alignments. Following target mapping by GSTAr.pl, the structural output can be parsed, and interacting pairs either discarded, as aberrant binding, or classified, by application of the rules for miRNA decoy and target interactions. This is generally applied through use of a custom perl script (Ivashuta et al. 2011; Fan et al 2015). An example of the output structure is given in Figure 4.3 below.

Output structure: .(((((((.(((((((((-((&)).)))))))))).)))))))

miRNA Query 3'-5' .)))))))).)))))))))

Transcript 5'-3' .(((((((.(((((((((-((

Binding site structure

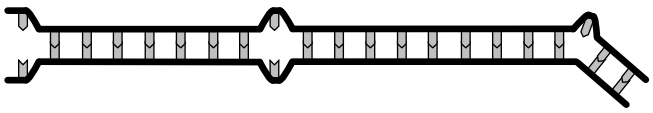

Figure 4.3: GSTAr outputs a structural representation of the RNA-RNA binding site. The region before the "&" represents the transcript 5'-3', while the region after the "&" represents the miRNA query, 5'-3'. "(" represents a transcript base that is paired, ")" represents a query based that is paired, "." represents an unpaired base, and "-" represents a gap inserted to facilitate alignment. This structural output allows for screening and differentiation of target and decoy lncRNAs by their unique binding site characteristics.

4.1.2.5. Assembling regulatory networks

The ability of both miRNAs and lncRNAs to bind other RNA species, allows for an important regulatory hierarchy: mRNAs are regulated by miRNAs, which in turn are regulated by and are able to regulate lncRNAs (Fig 4.4). It is theoretically possible that lncRNAs may interact directly with mRNA transcripts, but this is not a major recognised mode of action. Furthermore, it is difficult to ascertain any such affects as they may be mediated through specific lncRNA conformations and would likely be related to lncRNA-genome interactions, rather than mRNA binding. The recognised mRNA-miRNA and miRNA-lncRNA interactions form a number of small networks of interacting transcripts, each able to regulate

a subset of related developmental, metabolic or stress response processes. These small sub-networks may then interlink to form larger, more complex networks, especially if both decoy and target interactions are included. While the role of post transcriptional regulation and ncRNA crosstalk is largely overshadowed by more conventional modes of regulation, they can play major roles, especially when rapid transcriptional shifts and highly precise transcriptional fine tuning are required.

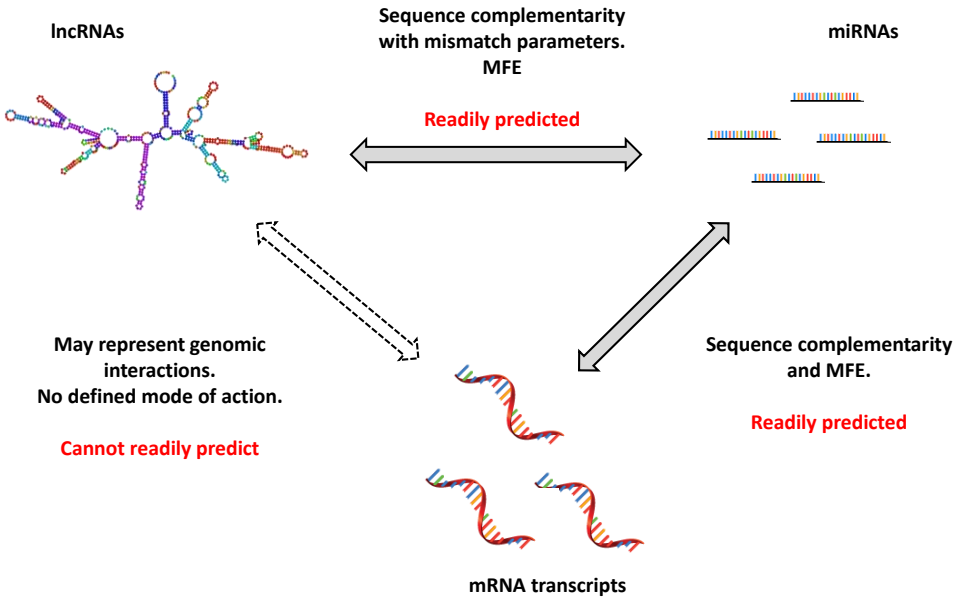

Figure 4.4. The prediction of RNA-RNA interactions between cross-regulatory ncRNAs and miRNA

Once lncRNA-miRNA and mRNA-miRNA interactions have been predicted and networks of interacting RNAs have been assembled, the predicted regulatory networks must be evaluated. Multiple criteria can be used. Firstly, as mentioned, each small network is expected to regulate mRNA transcripts relating to the same or related processes and functions. Thus, mRNAs that are targets of the same miRNA would be expected to be enriched in particular gene ontology (GO) terms. Secondly the transcript levels of interacting RNAs should make sense in light of the proposed regulatory interactions and modes of interaction. For example, if the abundance of a decoy lncRNA transcript increased, more miRNAs transcripts would be sequestered, resulting in an effective decrease their free transcript levels. This would in turn, decrease miRNA mediated degradation resulting in the accumulation of target mRNA transcripts. The consequences of changing lncRNA levels, depending on their mode of action (decoy or target), are given in Table 4.3 below. Correlating the expression patterns of RNAs in a putative network can be used to lend weight to these predicted interactions.

Ideally predicted interactions should also be experimentally validates to ensure that what is predicted in silico truly represents what takes place in vivo.

Table 4.3. The effect of regulatory noncoding RNA transcript changes on downstream target RNA transcript abundance. Arrows indicate the direction the applied effect.

LncRNA Mode of action	Δ lncRNA transcript levels	Δ effective miRNA transcript levels*	Δ mRNA transcript abundance
Decoy	Up →	Down →	Up
Decoy	Down →	Up →	Down
Target	Down ←	Up	
Target	Up ←	Down	

* The changes shown refer to effective miRNA levels available to bind miRNAs, not the total number of transcripts including those bound by lncRNAs.

4.1.3. Aims:

The aim of the work presented in this chapter is to construct regulatory networks between competitive endogenous lncRNA, miRNAs and mRNA transcripts. LncRNAs acting as miRNA decoys and targets were identified by mapping lncRNAs (identified in chapter 2) to target miRNA transcripts (identified in chapter 3) and classifying their modes of action on the basis of binding site structure. To explore the biological function of the identified decoy lncRNAs and interacting miRNAs, a network of mRNA interactions was constructed from the *X. humilis* desiccation transcriptome. The function of the predicted decoy lncRNAs and interacting miRNAs was then explored via co-expression analysis, mRNA transcript annotation and GO term enrichment analysis of the individual networks. Through this I aim to demonstrate that competing endogenous RNA may play key roles in regulating the desiccation phenotype of *X. humilis* leaves in response to water loss.

4.2. Method

4.2.1. RNA datasets

Sequence and count data for the 41 mature miRNA sequences were obtained as detailed in chapter 3. The 15 192 lncRNA sequences identified in chapter 2 were used and all treated as independent sequences (unclustered). The 14 614 differentially expressed mRNA transcripts (FPKM >5, FC >2) from the *X. humilis* leaf desiccation transcriptome (Chapter 3) were used for mapping miRNA targets. The annotation information for these mRNA transcripts was obtained from the *X. humilis* seed-leaf desiccation transcriptome (Lyall. 2016).

4.2.2. Mapping lncRNA-miRNA interactions and differentiating target from decoy lncRNAs.

All plausible interactions between the miRNA and lncRNA sequences were predicted using the GSTAr.pl script (V1.0, github.com/MikeAxtell/GSTAr). To identify high-quality interactions and distinguish lncRNAs, acting as miRNA decoys from those acting as miRNA targets, the binding site structures of all predicted interactions were parsed with a custom in-house Awk script. This script applied two sets of rules (Table 4.4) consolidated from those given in Table 4.2.

Table 4.4: Rules used to identify high confidence miRNA-lncRNA interactions, and to differentiate between lncRNAs acting as either miRNA targets or ceRNAs. All indicated positions are given from the 5' end of the interacting miRNA. The given rules are consolidated from those used by Ivashuta et al. 2011, Wu et al. 2013 and Fan et al. 2015.

	Restrictions and Requirements	
Position	**Decoy / ceRNAs**	**Target**
Between 9th and 12th position	- 1-5 mismatches/indels	- ≤1 mismatch/indel
Mismatches in other regions	- No bulges/Indels - 1st position, mismatch allowed - 2nd - 8th position, perfect nucleotide complementarity - ≤4 mismatches - ≤2 mismatches in a row	- ≤4 mismatches/indels - No continuous mismatches

Once distinguished, only competing endogenous (ceRNA) lncRNAs were retained for further analysis, thereby focussing attention on the ceRNAs directly regulating miRNA levels, and thereby mRNA levels.

4.2.3. Mapping miRNAs to the leaf transcriptome

The 41 high confidence miRNAs were independently mapped, with MFE considerations, against the entire *X. humilis* leaf desiccation transcriptome using two tools: A local instalment of TargetFinder (Bo & Wang. 2005) and the online TAPIR web server (Bonnet et al. 2010). Both tools were run using default settings, and for TAPIR the precise algorithm was selected. A MFE cut-off of -20kcal.mol^{-1} was applied. The outputs were combined to obtain a union of predicted interactions, an approach reported by Mishra et al 2015.

4.2.4. Selecting transcripts part of complete ceRNA-miRNA-mRNA networks

In order to assemble regulatory RNA networks comprised of all three RNA types (ceRNA, miRNA and mRNA) only miRNAs that were found to interact with both ceRNA and mRNA transcripts were selected. These were used to as central 'seeds' around which the ceRNA-miRNA-mRNAs networks were constructed.

4.2.5. Cytoscape mapping

The combined set of predicted pair-wise miRNA-ceRNA and miRNA-mRNA interactions were used to assemble and visualise all interacting RNAs into a number of interaction networks. This was performed using Cytoscape v3.4.0 (Shannon et al. 2003). RNAs were set as nodes, grouped by RNA class (ceRNA, miRNA or mRNA), with all predicted pairwise interactions set as edges to join these nodes.

4.2.6. Network analysis

Analysis of the putative networks of interacting RNAs took part in four stages: 1) clustering and analysis of intra-network transcript expression patterns, 2) mRNA transcript annotation, 3) analysis for functional enrichment and 4) functional prediction for both the decoy lncRNAs and networks as a whole.

4.2.6.1. Gene expression clustering and visual assessment for possible expression correlation.

For each network assembled by Cytoscape, the expression vectors for each RNA type – mRNA, miRNA and ceRNA – were constructed using Multi-Experiment Viewer (MeV v4.9.0; www.tm4.org). The

normalised read count data, obtained from DESeq2 (See previous chapters) and averaged for each set of 3 RWC replicates, were used as input for MeV. The expression values for each transcript were then re-normalised for each predicted network to facilitate visualisation. The expression profiles were then clustered using K-Means clustering (Pearson correlation, 10000 iterations), and plotted. The multiple expression plots for each subnetwork were examined in terms of the proposed modes of regulatory interaction, in order to correlate measured expression changes to proposed functionality and the expected relation between interacting transcript levels (Table 4.3).

4.2.6.2. Transcript annotation.

The protein identity and associated GO terms for all protein coding (PC) genes in the predicted Cytoscape networks were obtained from the updated *X. humilis* Seed-Leaf desiccation transcriptome (Lyall. 2016).

4.2.6.3. GO term enrichment analysis and functional prediction.

Statistical overrepresentation of GO categories was assessed and visualised for each of the PC expression clusters in each of the predicted Cytoscape networks using the BiNGO (v3.03, Maere et al. 2005), a Cytoscape plugin. GO enrichment was determined using a Hypergeometric test and a Benjamini & Hochberg False Discovery Rate (FDR) correction (P-value < 0.05). A custom annotation file containing the full complement of all Seed-Leaf GO terms assigned to each Leaf gene was used, with the full set of annotated differentially expressed *X. humilis* Leaf transcripts used as a reference set for enrichment.

4.3. Results

4.3.1. Mapping lncRNA-miRNA interactions

Interactions between miRNAs and lncRNAs were identified by mapping all possible miRNA-lncRNA sequence interactions. 41 predicted mature miRNA sequences were assessed, along with 15192 putative lncRNA sequences. Application of binding criteria allowed for the identified interactions to be classified as either ceRNA-miRNA or 'target lncRNA'-miRNA interactions (or discarded).

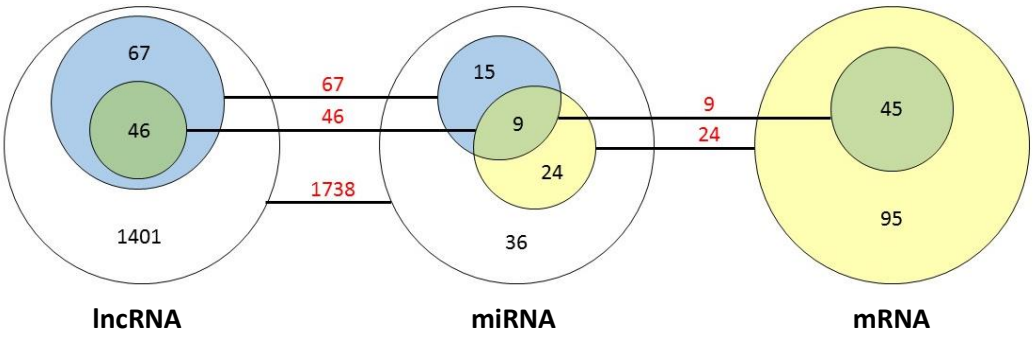

Figure 4.5: Diagram showing the results of the interaction predictions, and lncRNA classification. The values given are inclusive of all subdomains. The results for lncRNA mapping to miRNA sequences are shown in white (target) and blue (decoy/ceRNA) respectively. All predicted decoy lncRNAs and their pairwise interactions were a subset of the predicted target lncRNA interaction set. The results for miRNA|mRNA mapping are shown in yellow. The transcripts involved in linked lncRNA|miRNA|mRNA networks are indicated in green. The number of interactions is indicated in red above the lines joining the interacting subdomains within each RNA class.

Overall 1401 of the putative lncRNAs were found to map to 36 of the 41 predicted miRNAs, in a manner consistent with lncRNAs functioning as cleavable miRNA targets (Figure 4.5, White). This occurred through 1738 unique interactions (lncRNAs per miRNA: Average 48.28; SD 38.4; Max 198). The vast majority of target lncRNAs were found to be targeted by a single miRNA (Average 1.24; SD 0.61) as demonstrated in Figure 4.6.

67 of these 1401 'target lncRNAs' were also found to map to 15 of the 31 miRNAs in a manner consistent with ceRNA function (Figure 4.5, Blue). This suggests possible dual roles played by some lncRNAs, as ceRNAs themselves regulated by other miRNAs. The 15 miRNAs each bound multiple decoy lncRNAs (Average 4.47; SD 5.62; Median 2, Max 23). No decoy lncRNA was found to interact with more than one miRNA sequence.

4.3.2. Mapping miRNAs to the leaf transcriptome

In order to identify all mRNA transcripts targeted by regulatory miRNAs, the 41 miRNAs sequences were independently mapped against the *X. humilis* leaf desiccation transcriptome using two tools, TargetFinder and TAPIR. TargetFinder identified 102 unique interactions between 19 miRNAs and 76 mRNA transcripts. MiRNAs were found to bind 5.37 mRNA targets (SD 3.93) with the most mRNA targets for a single miRNA being 10. The mRNA targets had on average 1.34 associated miRNAs (SD 0.81) with a single mRNA being targeted with the maximum of 5 miRNAs. Tapir identified a greater number of miRNAs associated with a lower number of interactions and mRNA targets. 75 Unique interactions were identified between 22 miRNAs and 57 unique mRNA transcripts. The interacting miRNAs were associated on average with 3.41 mRNAs (SD 2.81) with a maximum of 17 mRNAs. The mRNA targets were on average targeted by 1.34 miRNAs (SD 0.79) with a single mRNA being targeted with the maximum of 5 miRNAs.

The outputs were then combined to obtain a comprehensive set of predicted interactions. The combined comprised 121 interactions between 24 miRNAs and 95 unique mRNA transcripts, indicating both tools predicted interactions not found by the other. Overall each miRNA was associated with an average of 5.04 mRNAs (SD 4.85), with the distribution given in Figure 4.7. Each mRNA transcript has on average 1.27 miRNAs (SD 0.64) miRNA interactions. The distribution of miRNAs per unique mRNA transcript sequence are given in Figure 4.8.

Nine miRNAs were found to both have putative mRNA targets, and be targets of ceRNAs. These 9 miRNAs form the central link between the ceRNA and mRNAs as part of ceRNA-miRNA-mRNA networks. Of the 95 targeted mRNAs, 45 were found to putatively interact with these 9 miRNAs, and of the predicted 67 ceRNAs, 46 were found to putatively interact with the 9 miRNAs (Figure 4.5, Green).

In summary target prediction identified 46 putative ceRNAs, 9 miRNAs and 45 mRNAs as part of one or more regulatory ceRNA-miRNA-mRNA networks.

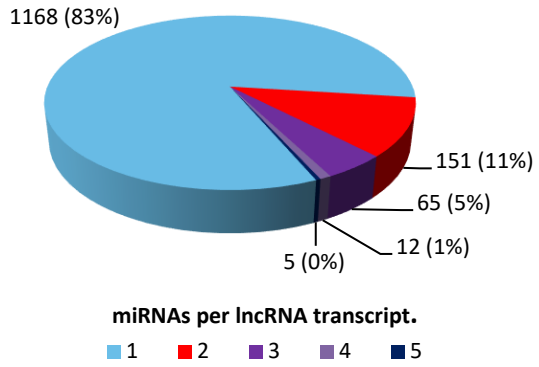

Figure 4.6: Distribution of the number of miRNAs targeting individual target lncRNAs.

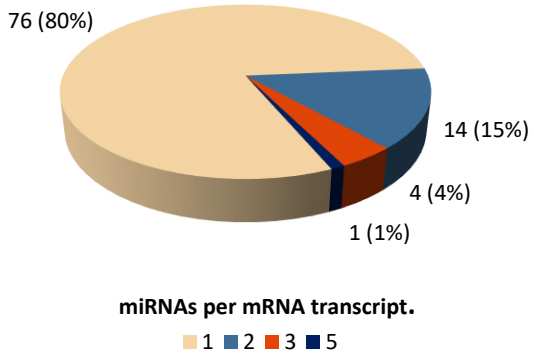

Figure 4.7: Distribution of the number of miRNA interactions with individual mRNA transcript sequences. The distribution is for the combined, non-redundant set of TargetFinder and TAPIR interactions.

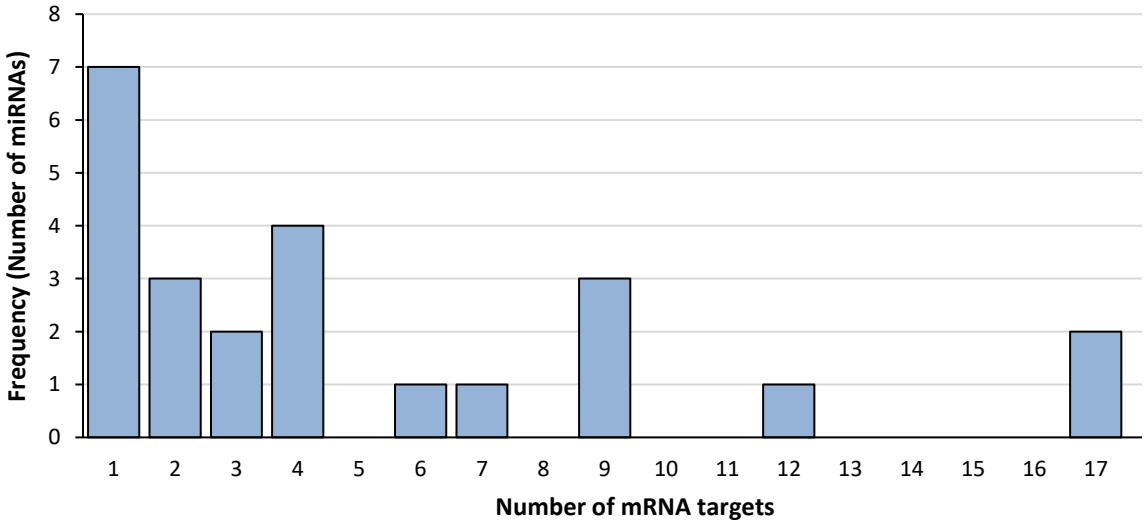

Figure 4.8: Distribution showing the number of mRNA sequences targeted by individual miRNA species. The distribution is for the combined, non-redundant set of TargetFinder and TAPIR interactions.

4.3.3. Construction of cross-regulatory RNA networks using Cytoscape.

All RNA transcripts (ceRNA, miRNA and mRNA) part of a regulatory network linking all three RNA types (Figure 4.5) were selected for network mapping and visualisation. Cytoscape was used to assemble all remaining pairwise interactions into 9 discrete regulatory RNA networks. Network maps

are given in Figures 4.9 and 4.10 part 1 & 2 (divided into parts for readability). These 9 networks may be independent or may interlink via other regulatory players, RNA or proteins, at either upstream or downstream level relative to the observed interactions. For ease of reference, the individual RNA networks will be referred to by citing the single unique miRNA linking the lncRNA and miRNA components.

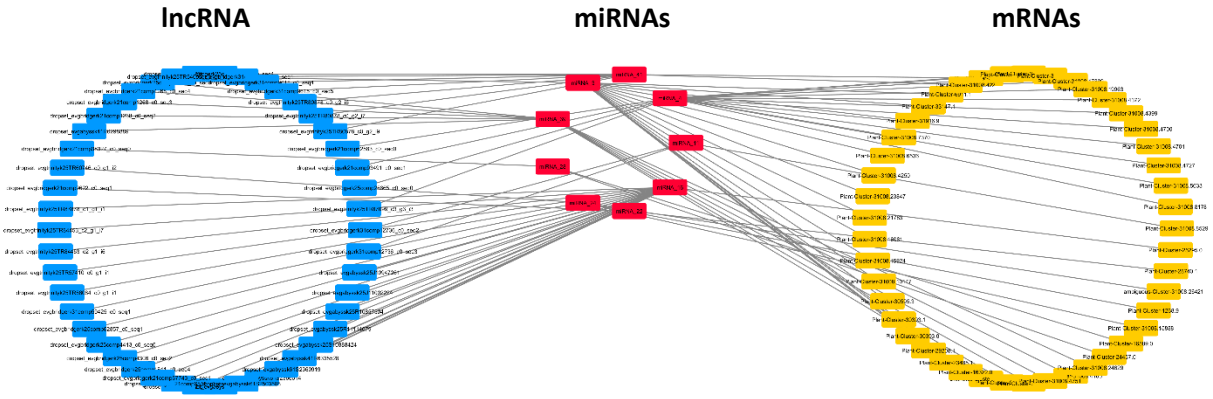

Figure 4.9: Network map showing the predicted regulatory interactions between RNA species of the 9 predicted lncRNA-miRNA-mRNA networks. Interactions between the between the 46 lncRNAs (Blue), 9 miRNAs (Red) and 45 mRNAs (Yellow), separated by type, are indicated by grey lines.

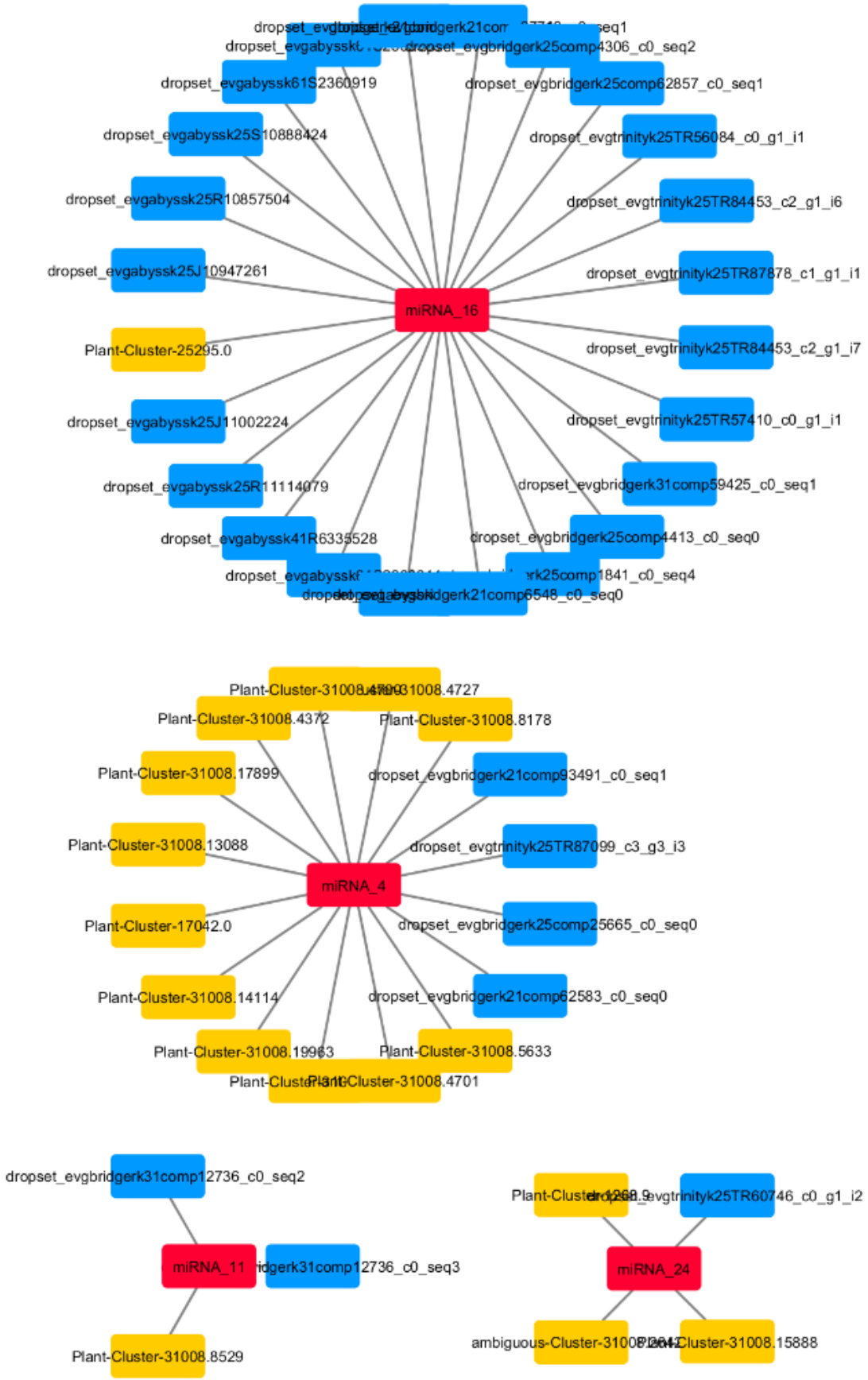

Figure 4.10 part1: Network maps for 4 of the 9 predicted lncRNA-miRNA-mRNA networks. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. miRNAs are centrally positioned, as they represent the key links in the networks, between the lncRNA and mRNAs.

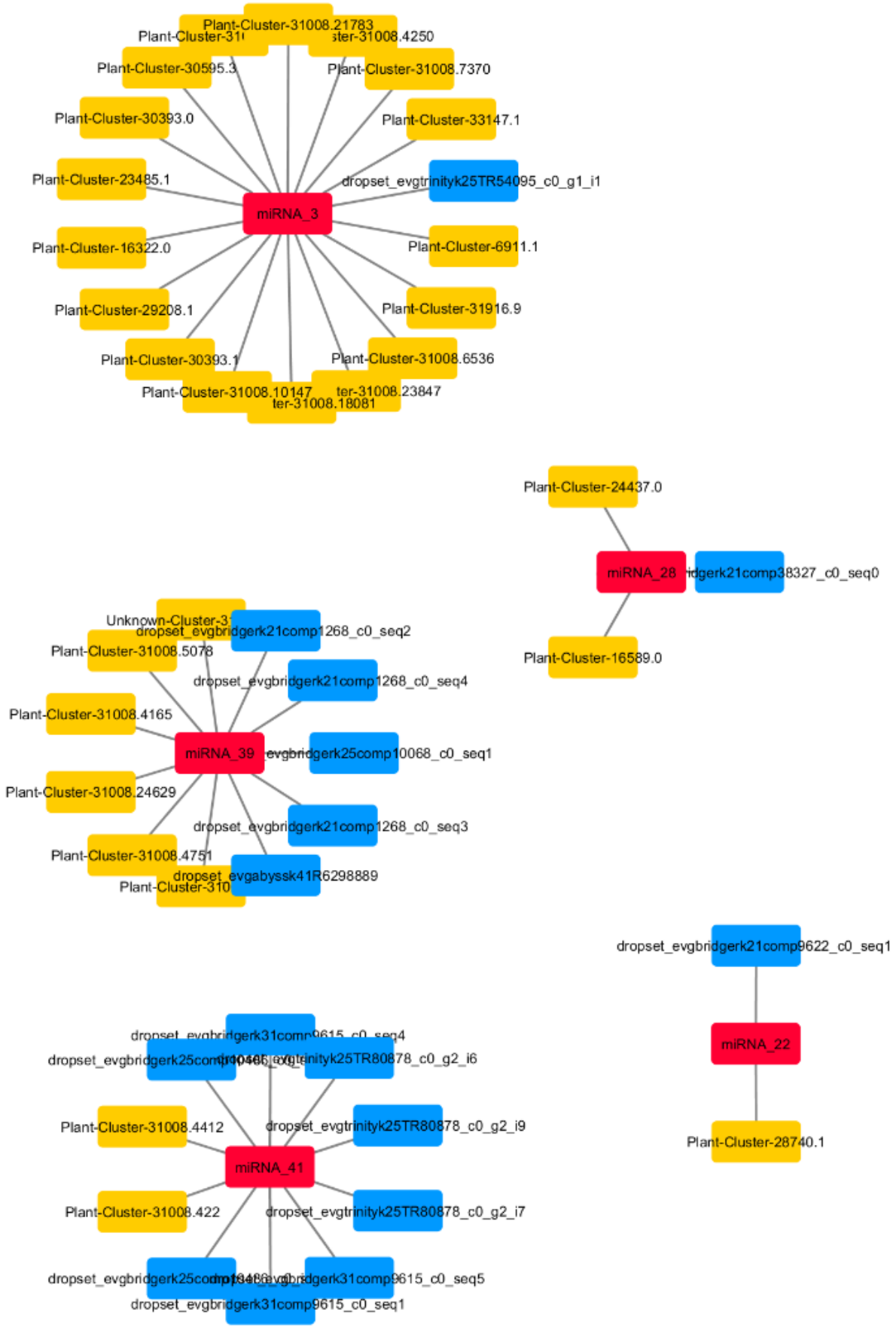

Figure 4.10 part 2: Network maps for 5 of the 9 predicted lncRNA-miRNA-mRNA networks. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. miRNAs are centrally positioned, as they represent the key links in the networks, between the lncRNA and mRNAs.

The 9 predicted RNA regulatory networks differ greatly in size and composition, as illustrated in Figure 4.11. The smallest RNA network is comprised of just three RNAs, with the largest network consisting of 23 unique RNAs. The ratio of lncRNAs to mRNAs in each network also differs greatly, with the minimum for each type being a single transcript, and the maximum being 23 and 17 unique transcripts respectively.

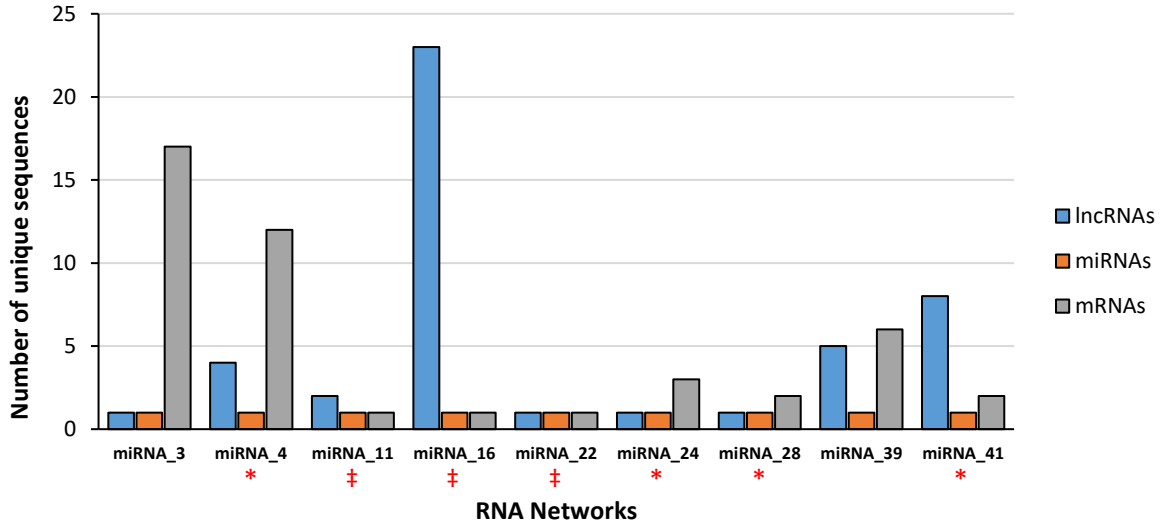

Figure 4.11: Composition of the 9 networks of interacting RNAs by RNA type and number. The number of unique transcripts of each type in the network of interacting RNAs is given. * indicates significant GO enrichment, no asterisk indicates no significant GO enrichment and ‡ denotes that only a single mRNA is present preventing meaningful enrichment analysis.

4.3.4. Evaluation of networks: correlation of expression profiles and Gene Onthology enrichment analysis.

In order to evaluate these networks, the expression vectors for all transcripts within each of the 9 RNA networks was plotted. Within each individual network, transcripts were separated by RNA type and then clustered using the K-means algorithm in MeV. Annotations for all mRNA transcripts including all associated gene ontology (GO) terms were used to perform a GO enrichment analysis (FDR < 0.05) on each mRNA expression cluster using BiNGO, a Cytoscape plugin. This was performed to identify the possible functional roles played by each network and their respective sub-clusters of mRNA expression. The individual results for each of the 9 RNA networks are presented below.

The 'miRNA_3' Network

The 17 mRNA sequences in the 'miRNA_3' network cluster into three groups by expression (Figures 4.12 & 4.13). Expression Profiles D and E both shown a rapid downregulation of expression between 70% and 30% RWC, which recovers at very low RWCs. Genes in these expression clusters serve a number of functions, such as signalling, protein synthesis and modification, and transcription factors (Table 4.5). Notable is the role of two genes (E) involved in chloroplast structure/function. Downregulation of these genes corresponds to the poikilochlorophylly associated loss of chlorophyll that occurs below 60% RWC, with upregulation possibly being in preparation for rehydration. Expression cluster-C shows upregulation below 80% RWC and then rapid downregulation at 5% RWC. These genes correspond to flavonoid biosynthesis, possibly produced as a UV protectant, a stress response transcription factor and an enzyme involved in ascorbic acid and cell-wall polymer synthesis. Of these three mRNA expression clusters, only mRNA cluster-C shows expression behaviour consistent with the expected function and expression profile (A) of the single decoy lncRNA. miRNA_3 is not significantly differentially expressed (chapter 3) so the true relationship between the miRNA and mRNAs in clusters D and E are unclear, as is the reason for their observed expression changes.

Enrichment analysis results presented in Figure 4.14 indicate that no GO terms were significantly enriched in the 3 cluster-C PC genes. The mRNA clusters D and E, despite not correlating with the lncRNA expression in any meaningful way, were found to have GO term enrichment, for a diverse set of plant processes.

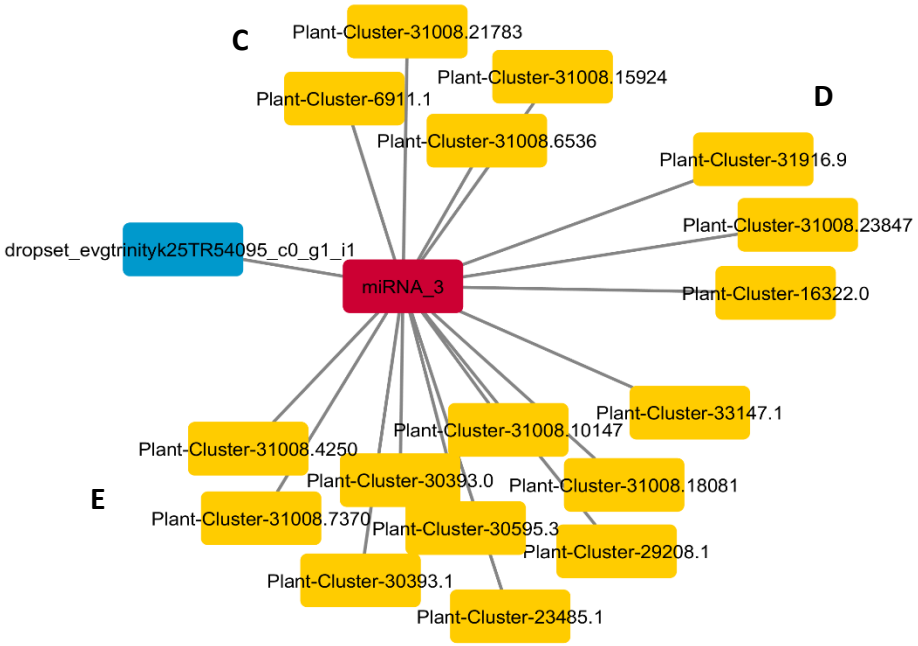

Figure 4.12: Network map of the 'miRNA_3' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters C, D and E correspond to mRNA expression clusters in Figure 4.13.

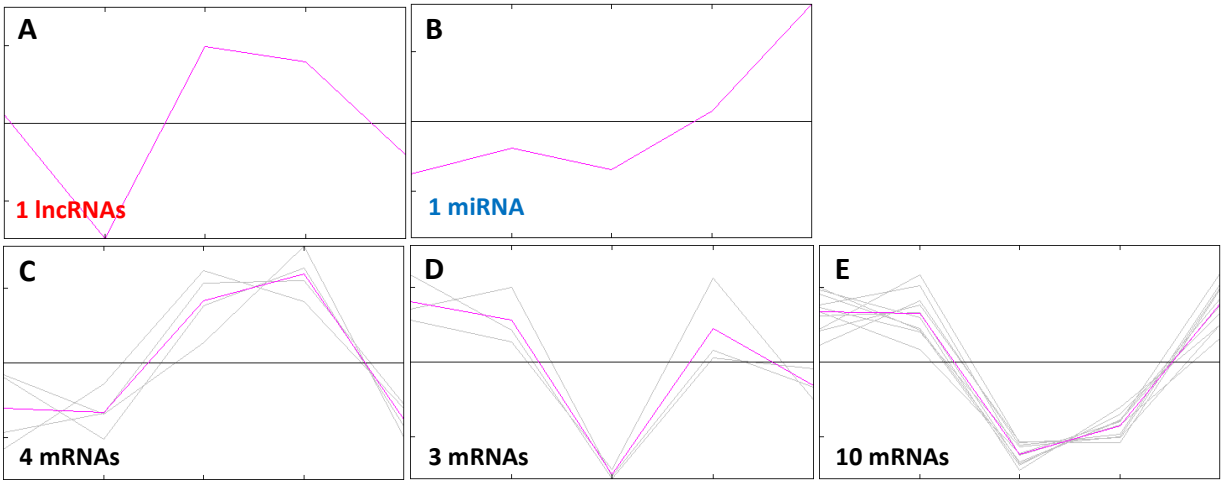

Figure 4.13: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_3' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & miRNA) and 90%, 70%, 50%, 30%, 10% (mRNAs). All RNAs are differentially expressed excluding the miRNA.

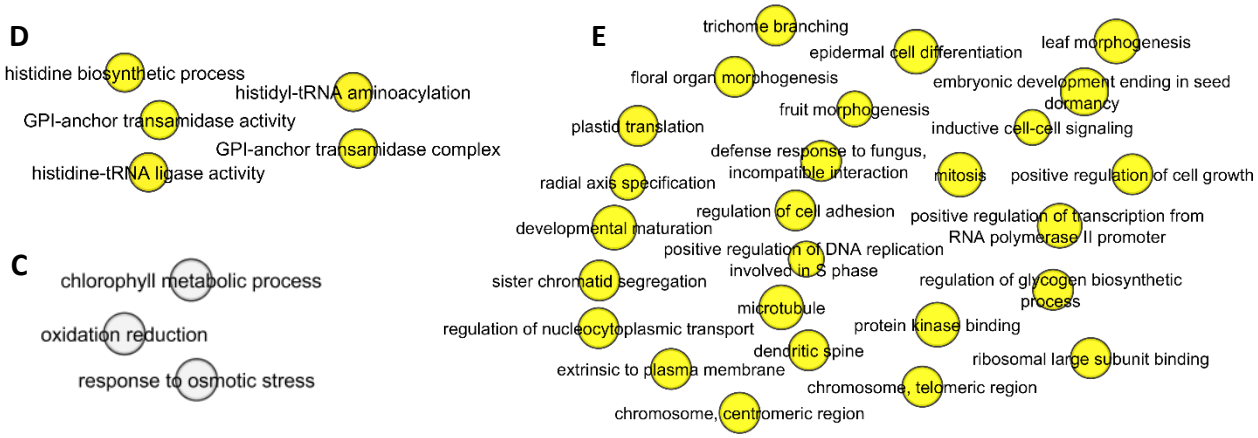

Figure 4.14: Enriched GO terms of genes present in the 'miRNA_3' RNA network. Terminal nodes -containing no outgoing edges - were isolated from the networks of significantly enriched GO terms (FDR < 0.05) determined using BINGO, a Cytoscape plugin. D and E correspond to mRNA expression clusters in Figure 4.13. As Cluster C showed no significant enrichment, the insignificant terminal nodes related to desiccation stress are given (white) to indicate possible transcript function.

The 'miRNA_4' Network

The 13 mRNA sequences in the 'miRNA_4' network cluster into three groups by expression (Figures 4.15 & 4.16). The 4 lncRNAs also separate into two clusters by expression. The mRNA clusters E and F show roughly the same expression pattern with significant downregulation at 60% and 40% RWC with an expression recovery at 5% RWC. Both genes in cluster-E code for transcription factors involved in stress response and delayed flowering (Table 4.5), both of which are expected during the *X. humilis* desiccation, which is followed by rapid flowering following rehydration. These two sequences may simply represent gene variants for a single gene, or may represent paralogs. 5 of the 7 sequences in cluster-F were annotated, 3 of which were transcription factors involved in floral induction and seed development (Table 4.5). mRNA cluster-D shows upregulation at 60% RWC, and repression at 5% RWC. 1 of 3 genes are annotated, as an enzyme involved in cobalamin/siroheme biosynthesis. Of the three mRNA expression clusters only E and F appear to correlate with lncRNA expression. lncRNA cluster B shows downregulation at 60% RWC which is mirrored by mRNA clusters E and F (60% RWC). A recovery in lncRNA expression at 40% RWC is mirrored by a delayed recovery in mRNA expression seen at 5% RWC. This suggests a possible functional regulatory relationship. lncRNA cluster A and mRNA cluster-D do not appear to show any correlation with other clusters. All three clusters were found to be enriched for GO terms, shown in Figure 4.17, related to the functions already discussed.

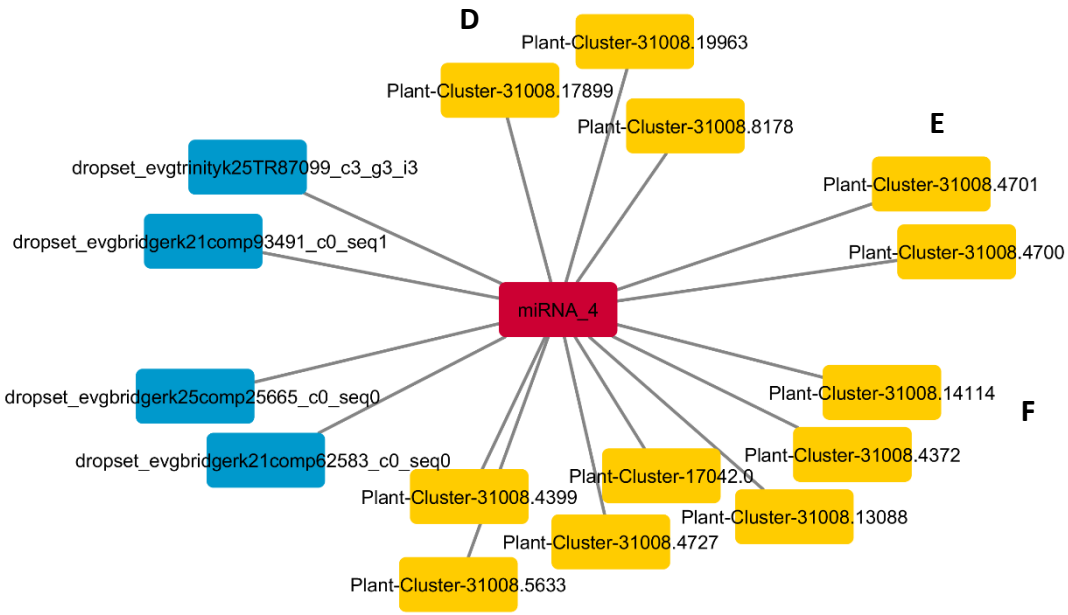

Figure 4.15: Network map of the 'miRNA_4' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters D, E and F correspond to mRNA expression clusters in Figure 4.16.

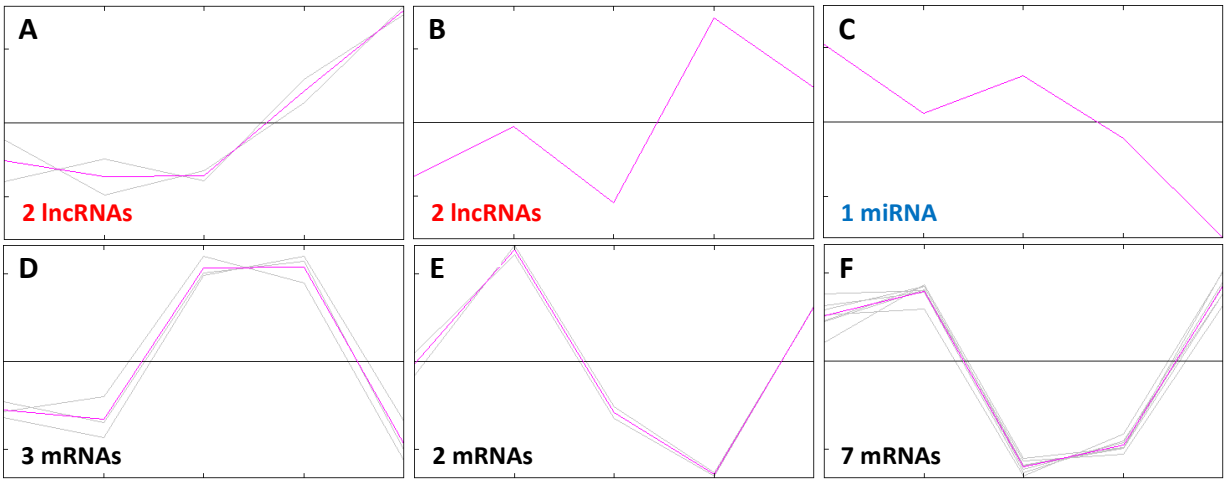

Figure 4.16: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_4' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & miRNA) and 90%, 70%, 50%, 30%, 10% (mRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_11' Network

The 'miRNA_11' Network is comprised of only a single mRNA and 2 lncRNAs, illustrated in Figure 4.18. This single mRNA gene encodes an enzyme involved in folate homeostasis (Table 4.5). Folate is required for methylation reactions and nucleic acid synthesis. Its possible role in the desiccation response is unclear, as all growth ceases. It is possible that it may play a role Histone tri-methylation (H3K27me3) by the Polycomb Repressive Complex (PRC) 2, thereby facilitating chromatin changes and epigenetic silencing of metabolic genes as cells enter anhydrobiosis. The expression changes observed in the 2 lncRNAs – gradually down regulated from 80% RWC to 40% RWC, followed by up regulation between 40% and 5% RWC – appears to be roughly mirrored by the mRNA levels which drop off rapidly from 80% RWC to 60% RWC, and then increase from 60% RWC and 50% RWC (Figure 4.19). This expression behaviour appears consistent with a functional regulatory interaction between the decoy lncRNAs, miRNA and mRNA transcripts. Due to the presence of only a single miRNA gene, meaningful functional enrichment of GO terms in the network could not be analysed. GO terms related to the transcript are however shown in Figure 4.20.

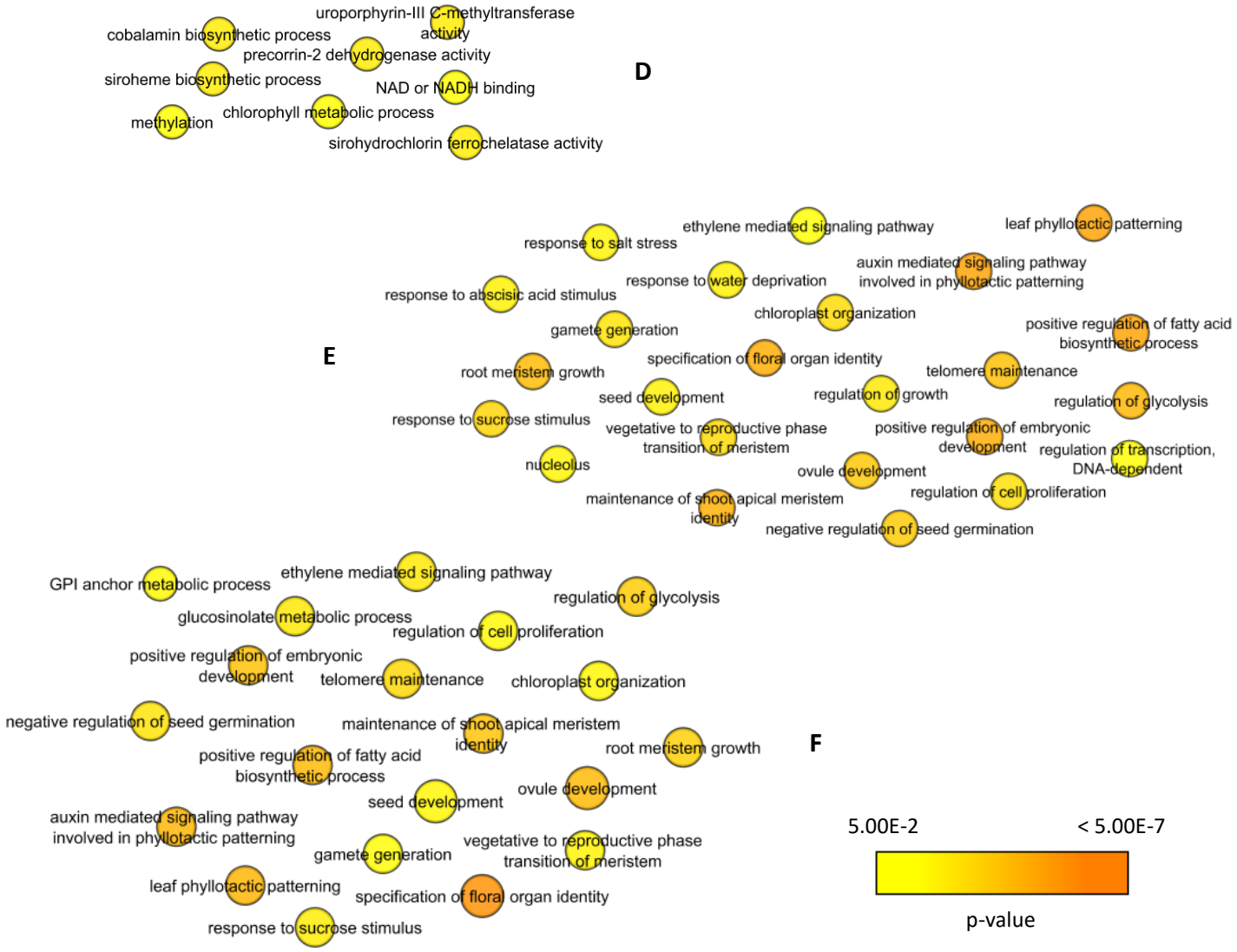

Figure 4.17: Enriched GO terms of genes present in the 'miRNA_4' RNA network. Terminal nodes -containing no outgoing edges - were isolated from the networks of significantly enriched GO terms (FDR < 0.05) determined using BINGO, a Cytoscape plugin. E, D and F correspond to mRNA expression clusters in Figure 4.16. The size of each node correlates with the number of input genes containing that GO term and node colour indicates significance (darker colour indicating greater significance).

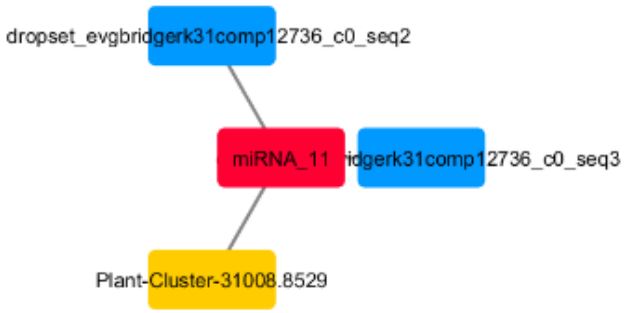

Figure 4.18: Network map of the 'miRNA_11' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines.

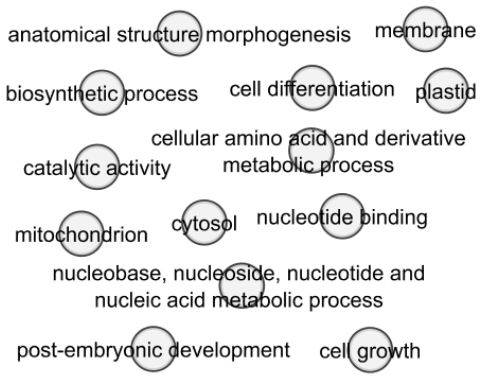

Figure 4.20: GO terms of the single mRNA gene present in the 'miRNA_11' RNA network. Terminal nodes -containing no outgoing edges - were isolated from the Plant GOSlim GO ontology. GO enrichment was not analysed due to the presence of only a single mRNA gene.

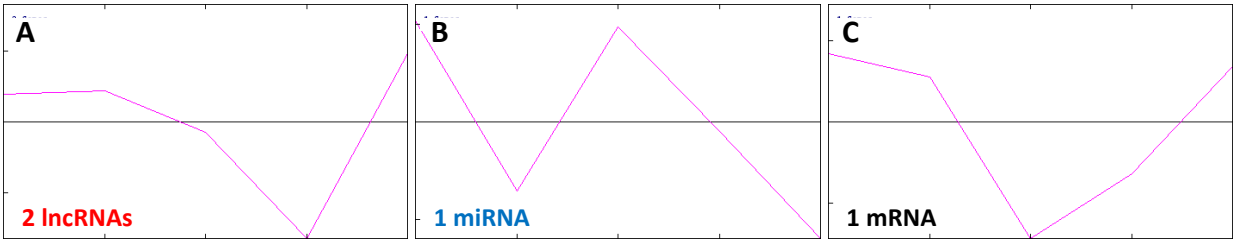

Figure 4.19: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_11' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_16' Network

The 'miRNA_16' network is comprised of a single mRNA, as well as 23 lncRNA genes divided into three clusters by expression (Figures 4.21 & 4.22). This protein coding gene encodes an enzyme required for pyridoxine biosynthesis (Table 4.5). Pyridoxine (vitamin B6) is an enzyme cofactor, though its function in plants has not been well elucidated. Due to the presence of only a single PC gene, meaningful functional enrichment of GO terms in the network could not be analysed. GO terms related to the transcript are however shown in figure 4.23. The expression profile of the PC gene, as shown in Figure 4.22, does not appear to in any way correlate with any of the three lncRNA expression profiles, drawing into question the legitimacy and functionality of the predicted interactions or roles played by these lncRNAs. Furthermore the presence of so many lncRNAs to a single target, and with such a diverse set of expression profiles, seems implausible for what is proposed as a set of key regulators.

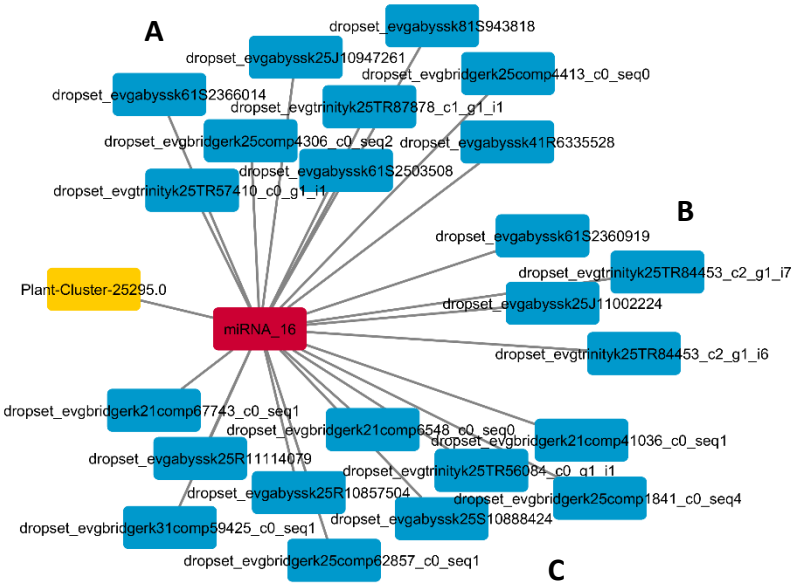

Figure 4.21: Network map of the 'miRNA_16' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters A, B and C correspond to lncRNA expression clusters in Figure 4.22.

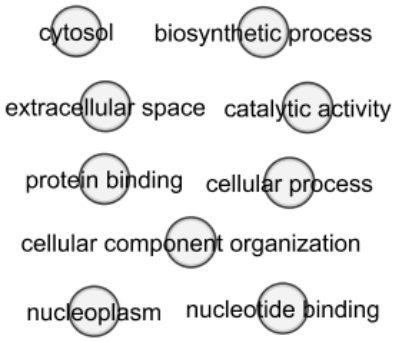

Figure 4.23: GO terms of the single mRNA gene present in the 'miRNA_16' RNA network. Terminal nodes were isolated from the Plant GOSlim GO ontology. GO enrichment was not analysed due to the presence of only a single mRNA gene.

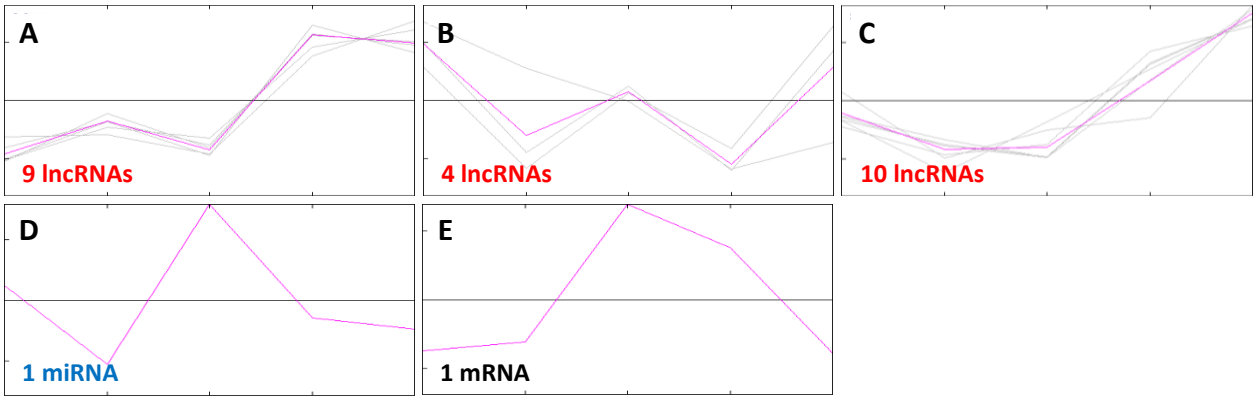

Figure 4.22: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_16' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_22' Network

The smallest of the 9 predicted RNA networks, the 'miRNA_22' network is comprised of only a single lncRNA, miRNA and mRNA gene (Figure 4.24). The single PC gene encodes the Ycf54 protein (Table 4.5). This is annotated as a bacterial protein, although a similar protein has been identified in plants hypothesised to be a plastid protein. The expression profile of this gene does not appear to correlate with the lncRNA expression, and is therefore inconsistent with the proposed lncRNA function (Figure 4.25). Overall, considering the annotation, as well as the expression pattern observed, it seems unlikely that the proposed network is of any further interest with regards to plant desiccation.

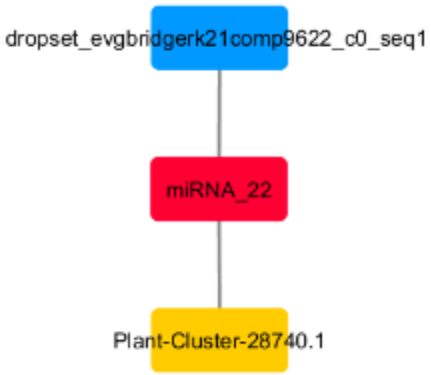

Figure 4.24: Network map of the 'miRNA_22' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines.

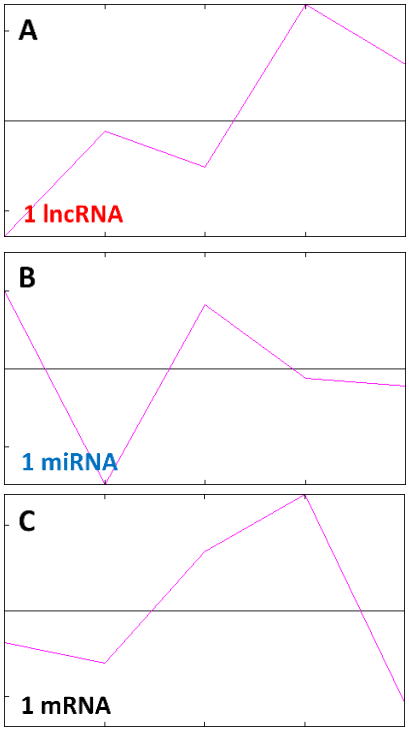

Figure 4.25: Network map of the 'miRNA_22' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_24' Network

The proposed 'miRNA_24' RNA network is comprised of three PC gene, regulated by a single miRNA and a single lncRNA (Figure 4.26). The three PC genes are clustered into two expression groups (Figure 4.27). The two genes in cluster-C are bacterial genes involved in conjugation and rhamnolipid synthesis (Table 4.5), and are of no interest to this study. They signify the presence of bacterial or fungal contamination. The annotated *X. humilis* desiccation transcriptome used (Lyall. R, PhD thesis, 2016) had known fungal transcripts removed during assembly. The presence of these genes in the predicted networks indicates that some transcripts of contamination origin evaded removal. These transcripts are of no interest. The single PC gene in cluster-D codes for a DNA helicase involved in DNA repair. Due to the presence of only a single PC gene in expression cluster-D, meaningful functional enrichment of GO terms could not be assessed. The full complement of GO terms related to the transcript was large with a diverse number of functions. A subset however indicate possible function related to stress response, as shown in Figure 4.28.

Expression profiles (Figure 4.27, B) indicate that the observed upregulation of the miRNA transcripts at 70% RWC coincides with a decrease in mRNA transcript levels (expression cluster-D) between 80% and 60% RWC, consistent with a regulatory interaction. It should be noted that miRNA_28 was not

found however to be significantly differentially expressed (Chapter 3). The gradual up regulation of the lncRNA with desiccation may contribute towards the recovery in observed cluster-D mRNA levels at RWCs less than 60%. This suggests a functional regulatory network may exist between the lncRNA, miRNA and the PC gene in expression cluster-D.

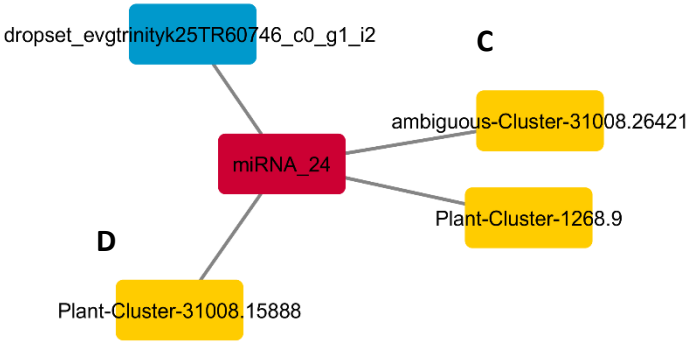

Figure 4.26: Network map of the 'miRNA_24' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters C and D correspond to mRNA expression clusters in Figure 4.27.

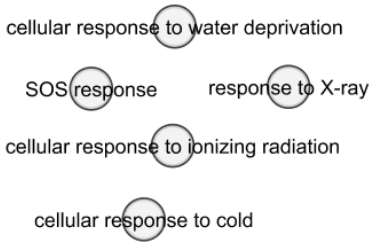

Figure 4.28: Stress response related GO terms for the single mRNA gene present in the 'miRNA_24' RNA network D Cluster. Terminal nodes corresponding to stress response GO terms were isolated from the Full GO ontology. GO enrichment was not analysed due to the presence of only a single mRNA gene in the cluster.

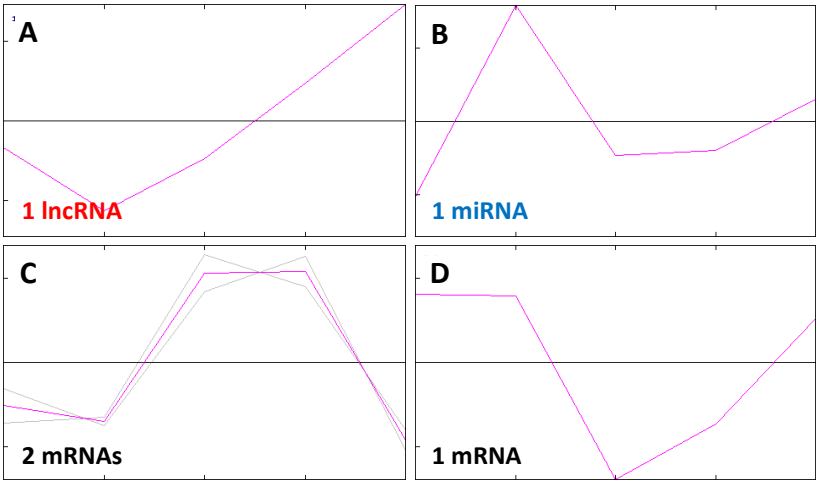

Figure 4.27: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_24' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_28' Network

The 'miRNA_28' Network' is comprised of single lncRNA and miRNA genes, as well as two PC genes that show shared expression (Figures 4.29 & 4.30). The 2 PC genes appear to be gene variants or paralogous genes, both of which code for the same blue copper protein (Table 4.5), involved in metal ion binding and electron transport, as confirmed by GO term enrichment analysis (Figure 4.31). Electron transport is essential to almost all biological processes. The exact role played by the identified protein is unclear. Expression of the PC genes shown in Figure 4.30 roughly mirrors the expression pattern observed for the proposed decoy lncRNA, which is expected for a function decoy lncRNA-miRNA-mRNA regulatory interaction. This supports the functionality of the network.

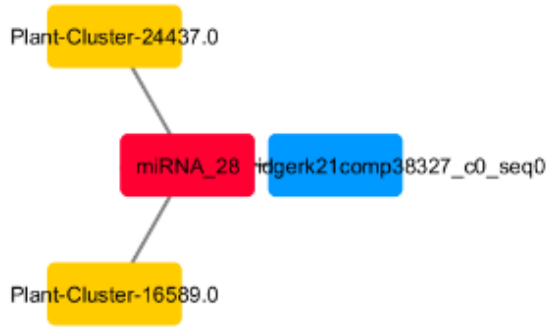

Figure 4.29: Network map of the 'miRNA_28' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines.

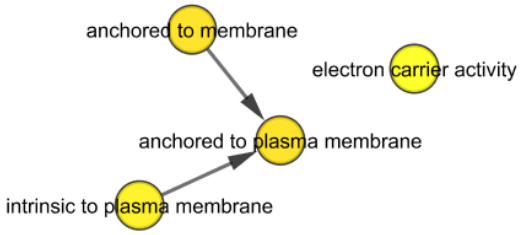

Figure 4.31: Enriched GO term network for genes present in the 'miRNA_28' RNA network. The networks of all significantly enriched GO terms (FDR < 0.05) determined using BINGO, a Cytoscape plugin, are shown. The size of each node correlates with the number of input genes containing that GO term and node colour indicates significance (darker colour indicating greater significance).

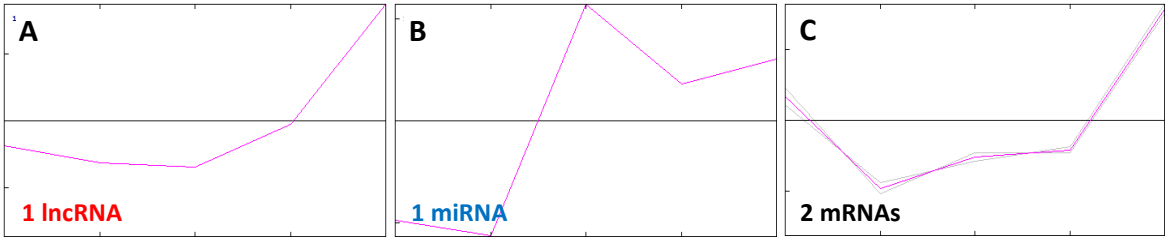

Figure 4.30: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_28' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The 'miRNA_39' Network

The 6 PC genes in the putative 'miRNA_39' RNA network form three clusters by expression (Figures 4.32 & 4.33). The single PC gene in cluster-C codes for a ubiquitin ligase (Table 4.5). Ubiquitination of proteins can mark them for degradation, or alter their activity, interactions, or cellular location. Protein degradation is logically required for the rapid metabolic shifts that occur during vegetative desiccation. Only one of the three PC genes in expression cluster-D was annotated in the *X. humilis* transcriptome. The gene codes for the receptor kinase PERK3 involved in cell surface receptor signalling. The two PC genes in expression cluster-E are an ADP_ATP antiporter and the histone protein H2B.2. None of the three expression clusters showed any GO term enrichment. Neither the protein identities, nor the associated GO terms appeared to be of much interest regarding vegetative desiccation, consisting of generalised and ubiquitous plant GO terms. The one exception is that the histone protein may play a role in repressive chromatin modifications by the PRC2, thereby repressing general metabolic processes as the leaves prepare for anhydrobiosis. None of the three PC gene expression clusters in Figure 4.33 appear to correlate in any way with the observed lncRNA or miRNA expression profiles. This brings the validity of the proposed functional interactions into question.

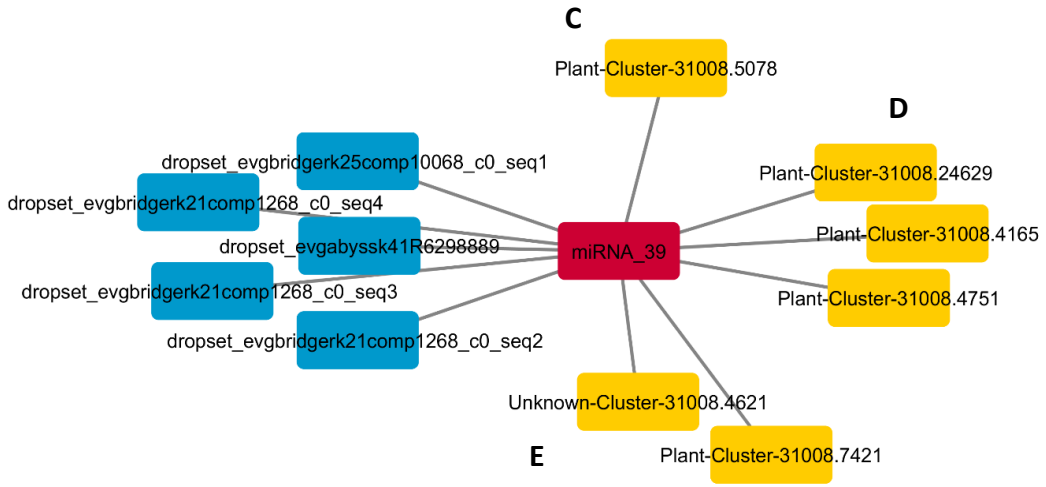

Figure 4.32: Network map of the 'miRNA_39' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters C, D and E correspond to mRNA expression clusters in Figure 4.33.

The 'miRNA_41' Network

While the 8 lncRNAs present in the proposed 'miRNA_41' RNA network all show the same pattern of expression, the 2 PC genes are clustered into two groups by expression (Figures 4.34 & 4.35). MiRNA_41 is also the only miRNA in 9 predicted networks that was found to be significantly differentially expressed over the course of desiccation (FDR <0.05, Chapter 3). At 70% RWC the miRNA is upregulated, plateauing at between 30% and 5% RWC. This up regulation coincides with a sharp decrease in mRNA transcript levels from 80% RWC. Although it appears that miRNA up regulation occurs after mRNA degradation and repression, this may be a result of the limited observation points. The correlation, if real, suggests a functional interaction between the miRNA and mRNA transcripts. MRNA transcript levels for expression clusters C and D appear to recover from 60% and 40% RWC respectively. While this cannot be explained by miRNA levels, a sharp increase in lncRNA transcript levels is observed at 60% RWC. This is consistent and supports the predicted decoy lncRNA interactions as being functional.

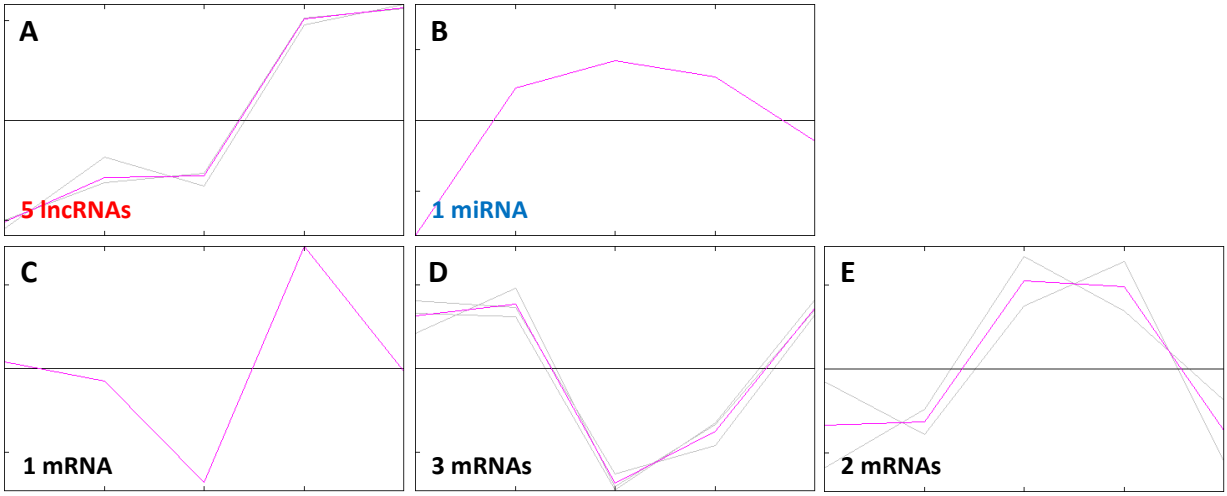

Figure 4.33: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_39' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs are differentially expressed excluding the miRNA.

The PC genes have been annotated as BEL1-like homeodomain protein 7, a transcription factor, (expression cluster-C) and a phosphatase involved in inositol deacylation of GPI-anchored proteins (expression cluster-D) (Table 4.5). Due to the presence of only a single PC gene in each expression cluster, meaningful functional enrichment of GO terms could not be assessed. The relevant GO terms associated with each gene are however given in Figure 4.36.

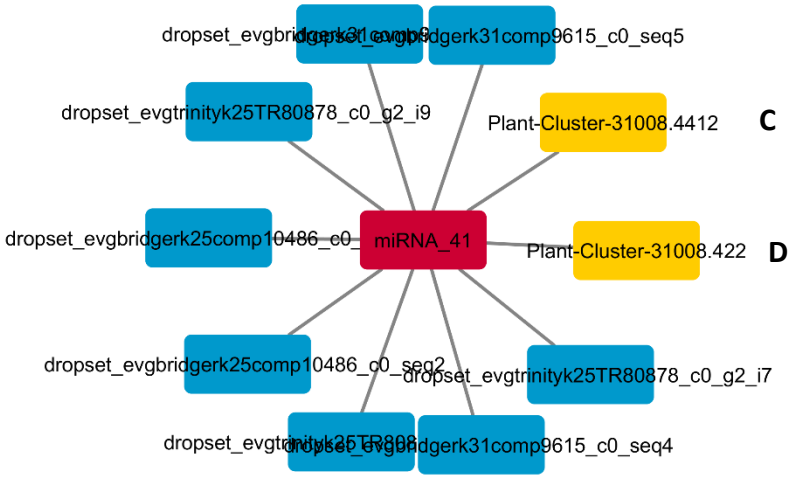

Figure 4.34: Network map of the 'miRNA_41' RNA network. Predicted functional binding interactions between lncRNAs (Blue), miRNAs (Red) and mRNAs (Yellow) are indicated by lines. The letters C and D correspond to mRNA expression clusters in Figure 4.35.

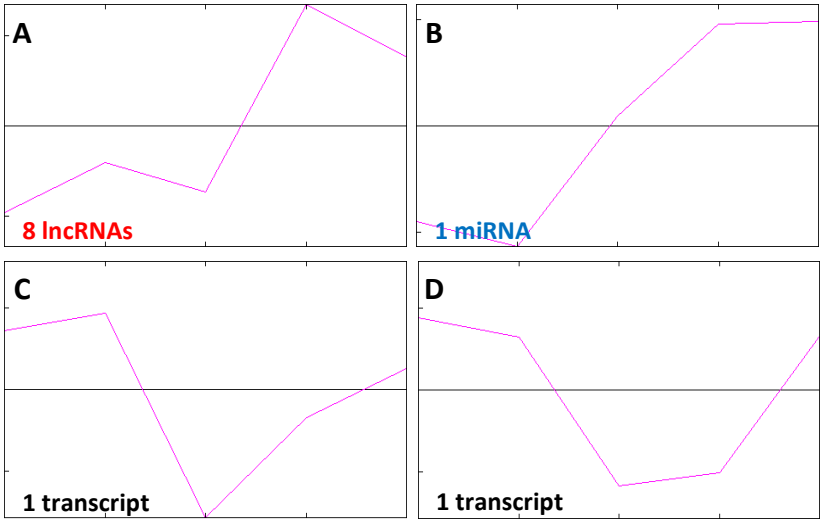

Figure 4.35: Expression profiles for the interacting lncRNA, miRNA and mRNA transcripts in the 'miRNA_41' RNA network. For each RNA type, normalised read counts were clustered into expression profiles using the K-Means algorithm in MeV (Pearson correlation, 10000 iterations). The expression profiles of both each individual transcript (grey) and the average expression for each cluster (pink) are shown, the as well as the number and type of transcripts in each cluster. The 5 RWCs represented on each plot are from left to right 100%, 80%, 60%, 40%, 5% (lncRNA & mRNA) and 90%, 70%, 50%, 30%, 10% (miRNAs). All RNAs shown are significantly differentially expressed (FDR <0.05).

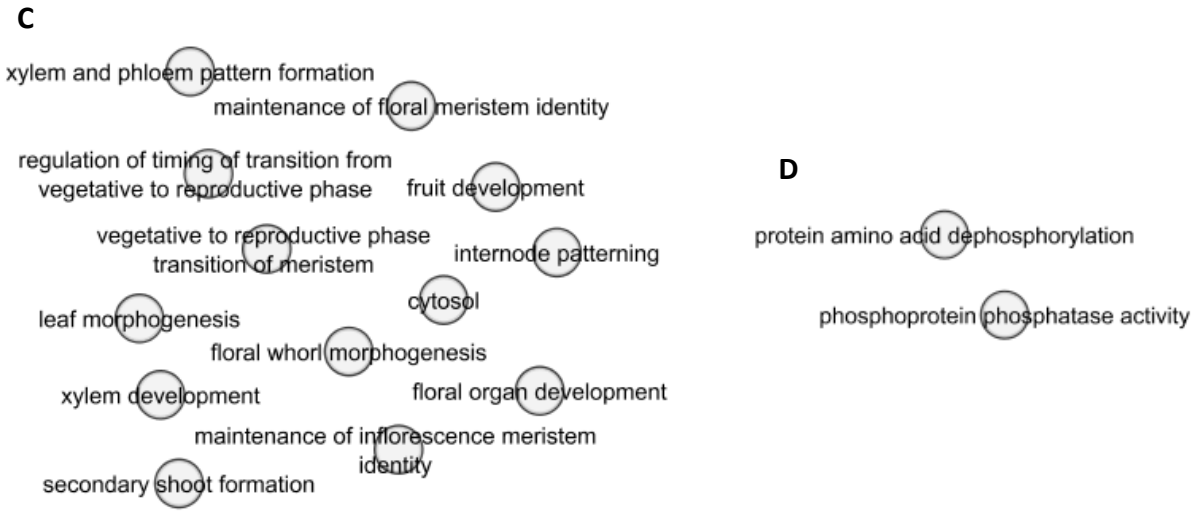

Figure 4.36: GO terms of genes present in the 'miRNA_41' RNA network. Terminal nodes -containing no outgoing edges - were isolated from the network of GO terms for each mRNA gene, plotted using BINGO, a Cytoscape plugin. C and D correspond to mRNA expression clusters in Figure 4.35. Selected GO terms corresponding to plant biological function are given for C. All GO terms are given for D. GO enrichment was not analysed due to the presence of only a single mRNA gene in each cluster.

Summary of Results

In order easily compare finding for each of the 9 predicted RNA networks, a summary of results is given in Table 4.6. Of the 9 networks, 4 were analysed for GO enrichment, with three being found to have enriched GO terms, all of which may play a role in the desiccation response. These three networks also showed gene expression patterns consistent with a functional regulatory network, and functional decoy lncRNA activity. An additional two networks, not tested for functional enrichment but with transcripts that could plausibly be involved in the desiccation response, were found to have expression patterns consistent with a functional RNA network.

Table 4.6: Summary of results for each of the 9 predicted RNA networks. Possible correlations are purely visual, no statistical testing has been performed.

Network (by miRNA)	GO Enrichment	Biological function(s)	Possible ceRNA-mRNA correlation
miRNA_3	Yes	Transcriptional regulation. Stress response. Chlorophyll metabolism. Floral development. Seed dormancy.	Yes.
miRNA_4	Yes	Transcriptional regulation. Plant Stress. Signalling. Meristem. Chloroplasts. Seed development. Phase transitions.	Yes.
miRNA_11	Untested	Assorted GO terms.	Yes.
miRNA_16	Untested	Assorted GO terms.	No.
miRNA_22	Untested	Bacterial contamination	NA.
miRNA_24	Untested	Assorted GO terms.	Possibly.
miRNA_28	Yes	Membrane, electron transport	Yes.
miRNA_39	No	Ubiquitination, Cell surface signalling, ADP transport.	No.
miRNA_41	Untested	Transcriptional regulation. Protein modification.	Yes.

Table 4.5: Summary table of the annotated UniProt idnetities and functions for all mRNA genes part of the 9 predicted RNA interaction networks. All annotations are from the unpublished *X. humilis* leaf-seed desiccation transcriptome assembled by Rafe Lyall, a past PhD student in our research group.

RNA Network (By miRNA)	Cluster	Sequence ID	UniProt Identifier	Protein	Function
3	C	Plant-Cluster-31008.15924	ZAT10_ARATH	Zinc finger ZAT10, Salt-tolerance zinc finger	Transcriptional repressor involved in abiotic (dehydration and cold) stress responses.
3	C	Plant-Cluster-31008.21783	MGDG1_ARATH	Monogalactosyldiacylglycerol synthase chloroplastic	Photosynthetic membrane synthesis.
3	C	Plant-Cluster-31008.6536	UFOG3_FRAAN	UDP-glucose flavonoid 3-O-glucosyltransferase 3, Flavonol 3-O-glucosyltransferase	Flavonoid biosynthesis.
3	C	Plant-Cluster-6911.1	MIOX_ORYSJ	Probable inositol oxygenase	UDP-glucuronic acid biosynthesis. Providing nucleotide sugars for cell-wall polymers. Plant ascorbate biosynthesis (possible role).
3	D	Plant-Cluster-16322.0	SYHM_ARATH	Histidine--tRNA chloroplastic mitochondrial, Histidyl-tRNA synthetase	Protein biosynthesis.
3	D	Plant-Cluster-31008.23847	ABHDD_DANRE	ABHD13, Alpha beta hydrolase domain-containing	Hydrolase.
3	D	Plant-Cluster-31916.9	-	-	-
3	E	Plant-Cluster-23485.1	RPK2_ARATH	LRR receptor-like serine threonine- kinase RPK2, TOADSTOOL 2, Receptor kinase	Regulation of floral and meristem activity, abiotic (dehydration and cold) stress response.
3	E	Plant-Cluster-29208.1	INO80_ARATH	DNA helicase INO80	Chromatin remodelling, DNA repair.
3	E	Plant-Cluster-30393.0	RRFC_SPIOL	Ribosome-recycling chloroplastic, Ribosome-releasing chloroplastic	Chloroplastic protein biosynthesis.
3	E	Plant-Cluster-30393.1	RRFC_SPIOL	Ribosome-recycling chloroplastic, Ribosome-releasing chloroplastic	Chloroplastic protein biosynthesis.
3	E	Plant-Cluster-30595.3	ROC2_ORYSJ	Homeobox-leucine zipper ROC2	Probable transcription factor.
3	E	Plant-Cluster-31008.10147	PP1_ORYSJ	Serine threonine- phosphatase	Protein phosphorylation, red light signaling.
3	E	Plant-Cluster-31008.18081	KSG9_ARATH	Shaggy-related kinase iota	Plant hormone mediated signaling pathway, hyperosmotic salinity response.
3	E	Plant-Cluster-31008.4250	CTBP_ARATH	C-terminal binding AN Short	Regulates leaf growth, prevents lipid peroxidation as a result of abiotic stress response.
3	E	Plant-Cluster-31008.7370	MD19A_ARATH	Mediator of RNA polymerase II transcription subunit	Positive transcriptional regulator.

3	E	Plant-Cluster-33147.1	SEH1_ARATH	SEH1	Export of mRNAs from the nucleus to the cytoplasm.
4	D	Plant-Cluster-31008.8178	-	-	-
4	D	Plant-Cluster-31008.17899	-	-	-
4	D	Plant-Cluster-31008.19963	CYSG2_CROS8	Siroheme synthase 2, Uroporphyrinogen-III C-methyltransferase	cobalamin/siroheme biosynthesis.
4	E	Plant-Cluster-31008.4700	RAP27_ARATH	Ethylene-responsive transcription factor RAP2-7	Transcriptional activator involved in stress response. Flowering time repression/delay.
4	E	Plant-Cluster-31008.4701	RAP27_ARATH	Ethylene-responsive transcription factor RAP2-7	Transcriptional activator involved in stress response. Flowering time repression/delay.
4	F	Plant-Cluster-17042.0	BST1_ASPOR	GPI inositol-deacylase	Involved in inositol deacylation of GPI-anchored proteins. ER-associated degradation of GPI-anchored proteins.
4	F	Plant-Cluster-31008.14114	-	-	-
4	F	Plant-Cluster-31008.5633	-	-	-
4	F	Plant-Cluster-31008.13088	UPI0004E57797	zinc finger NUTCRACKER-like	Metal ion/nucleic acid binding.
4	F	Plant-Cluster-31008.4372	AP2_ARATH	Floral homeotic APETALA 2	Transcriptional activator. Vegetative to floral growth transition. Floral and seed development.
4	F	Plant-Cluster-31008.4399	AP2_ARATH	Floral homeotic APETALA 2	Transcriptional activator. Vegetative to floral growth transition. Floral and seed development.
4	F	Plant-Cluster-31008.4727	AP2_ARATH	Floral homeotic APETALA 2	Transcriptional activator. Vegetative to floral growth transition. Floral and seed development.
11		Plant-Cluster-31008.8529	FPGS2_ARATH	Folylpolyglutamate synthase	Folate homeostasis.
16		Plant-Cluster-25295.0	PPOX1_ARATH	Pyridoxine pyridoxamine 5 -phosphate oxidase chloroplastic	Pyridoxine biosynthetic process.
22		Plant-Cluster-28740.1	YC54L_SYNY3	Ycf54	Bacterial protein, conserved hypothetical plastid protein.
24	C	Plant-Cluster-1268.9	PRM1_EMENI	Plasma membrane fusion prm1	Fungal/bacterial conjugation.
24	C	ambiguous-Cluster-31008.26421	RHLG_PSEAE	Rhamnolipids biosynthesis 3-oxoacyl-	Bacterial rhamnolipids synthesis.
24	D	Plant-Cluster-31008.15888	RQL3_ARATH	ATP-dependent DNA helicase Q-like 3	DNA helicase. DNA repair.
28		Plant-Cluster-16589.0	BCP_PEA	Blue copper protein	Metal ion binding. Electron transport.
28		Plant-Cluster-24437.0	BCP_PEA	Blue copper protein	Metal ion binding. Electron transport.

39	C	Plant-Cluster-31008.5078	PUB33_ARATH	U-box domain-containing 33, E3 ubiquitin ligase	Ubiquitination (Protein modification).
39	D	Plant-Cluster-31008.24629	PERK3_ARATH	Proline-rich receptor kinase PERK3, Proline-rich extensin-like receptor kinase	Cell surface receptor signaling. Protein phosphorylation.
39	D	Plant-Cluster-31008.4165	-	-	-
39	D	Plant-Cluster-31008.4751	-	-	-
39	E	Plant-Cluster-31008.7421	TLC2_ARATH	ADP,ATP carrier chloroplastic, ADP ATP translocase 2, Adenine nucleotide translocase	ATP:ADP antiporter.
39	E	Unknown-Cluster-31008.4621	H2B2_ARATH	Histone H2B.2	Core component of nucleosome.
41	C	Plant-Cluster-31008.4412	BLH7_ARATH	BEL1-like homeodomain 7	Transcription factor activity.
41	D	Plant-Cluster-31008.422	PPP7L_ARATH	Serine threonine- phosphatase	Involved in inositol deacylation of GPI-anchored proteins. ER-associated degradation of GPI-anchored proteins.

4.4. Discussion

4.4.1. Mapping lncRNA-miRNA interactions: RNA specificity and redundancy.

Results from predicting lncRNA-miRNA interactions are interesting with regards to the observed RNA specificity and abundance. More miRNA species were found to interact with target lncRNAs than decoy lncRNAs. This is not surprising given the high levels of target complementarity found in plant miRNAs, a trait required for target binding and degradation but not decoy interactions. While approximately 2 times as many miRNAs bind target lncRNAs than decoy lncRNAs, almost 23 times as many target lncRNAs were found than decoy lncRNAs. This suggests that a much greater proportion of lncRNAs function as miRNA targets than as decoys. It also suggests that single miRNAs can oversee the regulation of many unique target lncRNAs, whereas far fewer decoy lncRNAs regulate the abundance of a single miRNA. This is logical given the hierarchical nature of regulatory networks, with many target lncRNAs being downstream of (regulated by) the miRNA regulator, and the few decoy lncRNAs being upstream regulators of the target miRNA. The presence of more than one decoy lncRNA per target miRNA suggests a level of regulatory redundancy, in some but not all cases.

In all predicted miRNA-decoy lnRNA interactions, and the majority of miRNA-target lncRNA interactions, each lncRNA shows high levels of miRNA specificity only interacting with a single unique miRNA, within their specific interaction type (decoy lncRNAs may interact with another miRNA via a target interaction). mRNA transcripts were also all found to map to only a single miRNA each. This suggests no regulatory redundancy by miRNAs, and high lncRNA specificity. The greater lncRNA sequence flexibility (only a small percentage is required for binding) and the greater likelihood of miRNA sequence changes generating adverse effects may account for the lack of redundant miRNA sequences. This ability of a single miRNA to exert control over target mRNA and lncRNA levels may facilitate the tight regulatory control and rapid regulatory changes required and observed during desiccation. The high specificity of decoy lncRNAs means they cannot exert any direct overarching control over multiple miRNAs. They must therefore regulate either key upstream regulatory miRNAs that in turn govern multiple processes, or alternatively are focused towards regulation of specific key processes.

4.4.2. Mapping miRNAs to the leaf transcriptome

Mapping of miRNAs to the leaf transcriptome was performed using two tools. Both TargetFinder and TAPIR identified a similar number of possible interactions. While many pair-wise interactions were identified by both tools, unique interactions were identified by both tools and included in the final intersection set. As sRNA-Seq was performed to identify the miRNAs, it may be possible at a future date to extract and use the degradome data to validate the predicted miRNA targets (Fan et al. 2015). It should also be noted that although the prediction tools used do utilize MFE considerations when predicting miRNA-mRNA interactions, it would also be possible to independently and more rigorously compute the MFE of the predicted duplexes. This could also be performed for the predicted miRNA-lncRNA interactions. Due to MFE considerations already having been taken into account and that further screening - both with regards to expression and function - were still going to be performed, it was decided that this was not necessary. Both interactions predicted by both tools and the lower confidence interactions predicted by a single tool were included in the final interaction set, on the basis of limited interactions and further screening.

4.4.3. Selecting transcripts part of complete lncRNA-miRNA-mRNA networks

In order to be able to infer ceRNA function from mRNA annotations, only RNAs part of complete ceRNA-miRNA-mRNA networks were used for network mapping. Only ceRNAs were included for two reasons: 1) target lncRNA function cannot be inferred without knowledge of the target lncRNA targets, and 2) in order to facilitate interpretation of results simple small networks of interacting RNAs was desired. By including target RNAs, complex interlinking networks may form, making any interpretation of the data impossible and regulatory relationships difficult to unravel with multiple confounding interactions. By breaking the regulatory pathways present during desiccation into small units, they can be properly examined. The decoy lncRNAs acting as target lncRNAs as well can then later be identified in the key networks identified to be of interest, and additional understanding can be built on the foundation of the small networks I have identified. These small sub-networks are the 9 possible regulatory networks that were identified in Figure 4.7. While these are discrete, they may form part of larger networks, either linking up at an upstream or downstream level.

4.4.4. General discussion of Network analysis and Networks of interest.

When examining the predicted RNA networks a number of key properties need to be evaluated, in order to determine whether each individual network is both likely to represent a valid regulatory network, as well as to determine if the network is of interest with regards to vegetative desiccation tolerance. These can be broken down into four main considerations: 1) Network composition, 2) relative miRNA expression patterns, 3) Annotated gene functions, and 4) Functional enrichment.

The composition of the 9 predicted networks was observed to be highly variable (Fig. 4.8), in the numbers of lncRNAs and mRNAs present. While variation is expected in biological systems, networks such as those containing miRNA_16 and miRNA_41 with large numbers lncRNAs are suspect. It seems unlikely that available transcript levels of a single miRNA (miRNA_41) would be regulated by 23 separate decoy lncRNAs. This suggests that the validity of such networks, and all predicted interactions, needs to be closely scrutinized.

The expression patterns of RNAs within a network, relative to both the RWCs being examined as well as the other RNAs classes (lncRNA, miRNA, mRNA) are useful both in validating function as well as predicted interactions. To be of relevance as possible regulators of the desiccation response RNAs must be expressed at times of interest, such as during the desiccation process as opposed to general metabolic processes which are simply repressed upon the onset of desiccation begins. The expression patterns of RNAs regulated by another class of RNA, or interacting, must also correlate and make sense biologically. The expectation is that decoy lncRNAs are acting to titrate repressible miRNA transcripts out of solution, allowing mRNA transcripts to accumulate. An increase in lncRNA expression should therefore correlate with an increase in mRNA transcript levels. All lncRNAs and mRNAs used in predicting the 9 networks are known to be differentially expressed. While miRNA expression patterns can be examined, they are not as informative. Firstly none of the miRNAs show significant differential expression (chapter 3) other than miRNA_41. Secondly the miRNA expression levels are for all miRNA transcripts present in the leaves, not a measure of miRNA availability remaining after decoy lncRNA binding. To validate the proposed networks, the decoy lncRNA and mRNA expression levels were therefore compared to the expected competitive endogenous RNA (ceRNA) model. No statistical testing for correlation was performed, and all comparisons are visual and qualitative in nature. Unlike the miRNAs, which have a staggered set of RWC time-points, the lncRNAs and mRNAs have shared RWCs meaning statistical testing could be done to validate these observations.

Another expression observation is that some of the networks have one or more RNA classes where the RNAs do not all share an expression profile, indicated by the separate expression clusters. It seems that a functional regulatory network could not regulate or be regulated by more than one set of transcripts showing vastly different, and at times opposite, patterns of expression. As such in cases where this occurs, and isn't simply a case of differing specificity resulting in a slightly delayed response by one transcript set, one or more of the expression clusters must not be a part of the true regulatory network. If this is the case the most feasible cluster should be identifiable by expression correlation with the other classes of RNAs. If none of the RNA classes or clusters appear to have expression profiles consistent with a functional biological interaction, then the network can be invalidated. It is unclear however where these unexplained interactions come from, or relate to the rest of the network. It does however highlight that the computational prediction of these networks is not infallible and experimental validation is ultimately required.

By annotating the PC genes involved in each proposed RNA network, the identity and functions of the genes being regulated can be identified and screened for relevance to the leaf desiccation response. Networks with fungal or bacterial PC genes, such as the 'miRNA_22' network, can be discarded. Furthermore, this allows not only the function of the PC genes to be identified, but also the function of the lncRNAs and the regulatory network as a whole to be inferred. Networks with no genes of any relevance are likely not of interest. Gene Ontology (GO) categories were also tested for statistically overrepresentation in each network or their individual expression clusters. The limitation of such a statistical test is that many of the smaller networks I predicted only have a single PC gene in the network or expression cluster. Where this was the case, all GO terms assigned to the gene are reported as being enriched, so enrichment analysis was not performed and the appropriate GO terms simply reported. Of the 9 networks two in particular stand out with regards to being enriched for GO terms biologically relevant to desiccation. The networks containing miRNA_3 and miRNA_4 both have PC genes, with expression consistent with function and lncRNA regulation, that show enrichment for GO terms associated with transcription factor (TF) activity, plant stress response, seed development/dormancy (orthodox seed is desiccation tolerant), chlorophyll (*X. humilis* is poikilochlorophyllous), floral development (for rapid flowering post desiccation) and meristem activity (Key tissue that needs protection during anhydrobiosis). The presence of protein TFs is of particular note as these in turn regulate a number of other processes. These networks are therefore of key interest.

4.4.5. Suggested improvements to the network analysis.

While interesting observations have been made, and a number of interesting networks show promise as possible regulators of VDT, it is also clear that a much more rigorous approach is required to adequately assess the validity and biological functions of the predicted regulatory ncRNA networks. I therefore propose that moving forward, the following steps be taken with regards to expression analysis and functional annotation:

Firstly, all predicted networks contain only a small number of interacting RNAs. As such expression clustering is not needed and genes should be evaluated independently. The most important correlation is that between the ceRNAs and mRNAs, as only miRNA availability and not miRNA transcript levels is important, which cannot be measured through RNA-SEq. As correlated expression is a key indicated of functional ceRNA activity, proper quantitative correlation analysis is important. Pearson correlation should be used to statistically test for significant expression correlation on an individual ceRNA-mRNA basis. Any mRNAs found not to significantly correlate with ceRNA expression, are likely not involved in functional ceRNA binding and should be discarded from the networks.

Furthermore, it is clear that the very small numbers of mRNAs in the predicted networks, severely hamper the ability to perform GO enrichment which is in any way significant or biologically meaningful. It seems logical to rather directly annotate each mRNA and examine the known functions of the mRNAs, which statistically correlate with ceRNA expression. It is important to identify the mRNAs involved in each network as expression changes in a single key gene can have marked physiological implications.

4.4.6. Promising regulatory networks worth further investigation.

Despite the shortfalls discussed, positive results were obtained. From the 9 predicted RNA networks, two appear most promising as playing key roles during the vegetative desiccation response. The networks centred on miRNA_3 and miRNA_4 both show good lncRNA-mRNA expression correlation, and GO term enrichment for biological processes directly related to desiccation tolerance, stress response and the processes immediately following desiccation. The latter is relevant as Xerophyta prepare and package silenced mRNA transcripts and cellular machinery in preparation for rapid rehydration). One of the PC genes part of the ‘miRNA_4 network’ is the Ethylene-responsive transcription factor RAP2-7. This transcription factor is a transcriptional activator which is known to

bind to the GCC-box pathogenesis-related promoter element to regulate the expression of stress factors and components of stress signal transduction pathways, as well as to prepress and confer a delay to flowering time (Aukerman & Sakai. 2003). The 'miRNA_28 network' meets the same criteria as the aforementioned networks, but with GO terms related to membrane structure and electron transport. While this may well be relevant to the desiccation response, the exact means still need to be identified. The networks centred around miRNA 11 and miRNA _24 both meet the required criteria, but the GO terms associated with the mRNA transcripts involved were extremely diverse making identification of any specific roles impossible from the annotations alone. Lastly the 'miRNA 41 network' codes for a known homeobox transcription factor BEL1, which controls ovule development through negative regulation of the AGAMOUS gene (Ray et al. 1994; Reiser et al. 1995). This is likely involved in preparation for the rapid onset of flowering following rehydration. These 6 predicted networks appear to be the most promising.

4.5. Conclusion.

The aims of this chapter and project as a whole were to identify the lncRNA and miRNA players involved in regulation of the vegetative desiccation response in the leaves of the resurrection plant *X. humilis*, as well as to predict and analyse the networks of interacting ceRNAs, miRNAs and mRNAs, including their predicted roles and functions within the desiccation response. Total RNA and sRNA sequencing, presented in chapters 2 and 3, allowed for core sets of high confidence lncRNA and miRNA expressed during desiccation to be identified. Mapping of these lncRNAs to the miRNA sequences, as well as the miRNA sequences to all DE sequences in the *X. humilis* desiccation transcriptome led to the identification of 9 putative regulatory networks of interacting RNAs. Of these 9 Networks 6 appear promising as true sets of regulatory interactions, possibly related to the mechanisms of vegetative desiccation response. Many also seem involved in floral repression which may correspond to a desiccation-induced vernalisation-like phenomenon. While the role of target lncRNAs in these networks has not been investigated, it would appear that decoy lncRNA regulation of miRNA activity is present and important in the regulation of the vegetative desiccation response in *X. humilis*.

While these networks have been identified and characterised with regards to protein identity and probable function, the presence of seemingly discordant pattern of expression within a RNA class of some networks, as well as the lack of correlated expression between classes in other networks, suggest not all predicted interactions represent true regulatory relationships. As such further experimental validation is required to confirm the predicted interactions as well as their proposed

functionality. This may be performed through a number of approaches including but not limited to electrophoretic mobility shift assays, fluorescent in situ hybridization (FISH), reporter assays or gene knockout, restoration and overexpression assays. None of these however are within the scope of this project and all may prove difficult to achieve given the unique phenotype and non-model organism in question. Furthermore, given the strict rules pertaining to the annotation and naming of new miRNAs and lncRNAs, until function is validated, these identified transcripts and networks will remain putative and cannot be officially named or submitted to their relevant repositories (Ambros et al. 2003; Griffiths-Jones et al. 2006; Meyers et al. 2008).

Overall ncRNAs have been shown to be involved in regulation of the vegetative desiccation response, and key lncRNA and miRNA players have been identified within their relative regulatory networks and roles, thereby laying a strong foundation for future studies into these networks and the role of ncRNA regulation in vegetative desiccation tolerance.

References

Abdel-Ghany S., 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. doi: 10.1074/jbc.M801406200.

Alpert P., 2005. The limits and frontiers of desiccation-tolerant life. *Integr Comp Biol*. 45(5):685-95. doi:10.1093/icb/45.5.685.

Alpert P., 2006. Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? *J Exp Biol.* 209(Pt 9):1575-84. doi:10.1242/jeb.02179.

Alpert P., 2006. Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? *J. Exp. Biol*. 209:1575–1584 (2006). doi:10.1242/jeb.02179

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., 1990. Basic local alignment search tool. *J Mol Biol.* 215(3):403-10. doi:10.1016/S0022-2836(05)80360-2.

Amaral P., Dinger M., Mercer T., Mattick J., 2008. The eukaryotic genome as an RNA machine. *Science.* 319(5871):1787-9. doi:10.1126/science.1155472.

Ambros V., Bartel B., Bartel D.P., Burge C.B., Carrington J.C., Chen X., Dreyfuss G., Eddy S.R., Griffiths-Jones S., Marshall M., Matzke M., Ruvkun G., Tuschl T., 2003. A uniform system for microRNA annotation. *RNA.* 29(3):277-9. doi:10.1261/rna.2183803.

Andronescu M., Zhang Z.C., Condon A., 2005. Secondary structure prediction of interacting RNA molecules. *J Mol Biol*. 345(5):987-1001. doi:10.1016/j.jmb.2004.10.082.

Aravin A., Tuschl T., 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett*. 579(26):5830-40. doi:10.1016/j.febslet.2005.08.009.

Atkinson S., Marguerat S., Bähler J., 2012. Exploring long non-coding RNAs through sequencing. *Semin Cell Dev Biol.* 23(2):200-5. doi:10.1016/j.semcdb.2011.12.003.

Aung K., Lin S., Wu C., Huang Y., Su C., Chiou T., 2006. pho2, a phosphate overaccumulator, is caused by a nonsense mutation in a microRNA399 target gene. *Plant Physiol.* 141(3):1000-11. doi:10.1104/pp.106.078063

Axtell M.J., 2008. Evolution of microRNAs and their targets: are all microRNAs biologically relevant? *Biochim Biophys Acta.* 1779(11):725-34. doi:10.1016/j.bbagrm.2008.02.007.

Axtell M.J., 2013. Classification and comparison of small RNAs from plants. *Annu Rev Plant Biol.* 64:137-59. doi:10.1146/annurev-arplant-050312-120043.

Axtell M.J., 2013. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA.* 19(6):740-51. doi:10.1261/rna.035279.112.

Axtell M.J., Bowman J.L., 2008. Evolution of plant microRNAs and their targets. *Trends Plant Sci.* 13(7):343-9. doi:10.1016/j.tplants.2008.03.009.

Banks I.R., Zhang Y., Wiggins B.E., Heck G.R., Ivashuta S., 2012. RNA decoys: an emerging component of plant regulatory networks? *Plant Signal Behav.* 7(9):1188-93. doi:10.4161/psb.21299.

Bartel D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 116(2):281-97. doi:10.1016/S0092-8674(04)00045-5.

Bartel D.P., 2009. MicroRNAs: target recognition and regulatory functions. *Cell.* 136(2):215-33. doi:10.1016/j.cell.2009.01.002.

Baumberger N., Baulcombe D.C., 2005. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proc Natl Acad Sci USA*. 102(33):11928-33. doi:10.1073/pnas.0505461102.

Bentwich I., Avniel A., Karov Y., Aharonov R., Gilad S., Barad O., Barzilai A., Einat P., Einav U., Meiri E., Sharon E., Spector Y., Bentwich Z., 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet.* 37(7):766-70. doi:10.1038/ng1590.

Berezikov E., Thuemmler F., van Laake L., Kondova I., Bontrop R., Cuppen E., Plasterk R., 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet.* 38(12):1375-7. doi:10.1038/ng1914.

Bernhart S.H., Tafer H., Mückstein U., Flamm C., Stadler P.F., Hofacker I.L., 2006. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Mol. Bio*. 1:3. doi:10.1186/1748-7188-1-3.

Bertolini E., Verelst W., Horner D.S., Gianfranceschi L., Piccolo V., Inze D., Pe M.E., Mica E., 2013. Addressing the role of microRNAs in reprogramming leaf growth during drought stress in *Brachypodium distachyon*. *Mol Plant*. 6:423-443. doi:10.1093/mp/sss160.

Bewley J., Oliver M., 1992. Desiccation-tolerance in vegetative plant tissues and seeds: Protein synthesis in relation to desiccation and a potential role for protection and repair mechanisms. In: Water and life: A comparative analysis of water relationships at the organismic, cellular and molecular levels, ed. by Osmond C., Somero G. and Bolis C. Springer-Verlag, Berlin. 141–160. doi:10.1007/978-3-642-76682-4_10.

Bo X., Wang S., 2005. TargetFinder: a software for antisense oligonucleotide target site selection based on MAST and secondary structures of target mRNA. *Bioinformatics*. 21(8):1401-2. doi:10.1093/bioinformatics/bti211.

Bonasio R., Lecona E., Narendra V., Voigt P., Parisi F., Kluger Y., Reinberg D., 2014. Interactions with RNA direct the Polycomb group protein SCML2 to chromatin where it represses target genes. *Elife.* 3:e02637. doi:10.7554/eLife.02637.

Bonasio R., Tu S., Reinberg D., 2010. Molecular Signals of Epigenetic States. *Science.* 330(6004): 612–616. doi: 10.1126/science.1191078.

Bonnet E., He Y., Billiau K., Van de Peer Y., 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics.* 26(12):1566-8. doi:10.1093/bioinformatics/btq233.

Brocard-Gifford I., Lynch T., Finkelstein R., 2003. Regulatory networks in seeds integrating developmental, abscisic acid, sugar, and light signaling. *Plant Physiol*. 131(1):78-92. doi:10.1104/pp.011916.

Brodersen P., Sakvarelidze-Achard L., Bruun-Rasmussen M., Dunoyer P., Yamamoto Y.Y, Sieburth L., Voinnet O., 2008. Widespread translational inhibition by plant miRNAs and siRNAs. *Science.* 320(5880):1185-90. doi:10.1126/science.1159151.

Bruinsma J., 2009. The Resource Outlook to 2050: By How Much Do Land, Water and Crop Yields Need to Increase by 2050? In: FAO Expert Meeting. How to Feed the World in 2050. Food and Agriculture Organization of the United Nations, Economic and Social Development Department, Rome, 2-16.

Buitink J., Leger J., Guisle I., Vu B., Wuillème S., Lamirault G., Le Bars A., Le Meur N., Becker A., Küster H., Leprince O., 2006. Transcriptome profiling uncovers metabolic and regulatory processes occurring during the transition from desiccation-sensitive to desiccation-tolerant stages in *Medicago truncatula* seeds. *Plant J.* 47(5):735-50. doi:10.1111/j.1365-313X.2006.02822.x

Buzas D., Robertson M., Finnegan E., Helliwell C., 2011. Transcription-dependence of histone H3 lysine 27 trimethylation at the Arabidopsis polycomb target gene *FLC*. *The Plant J*. 65:872–881 doi:10.1111/j.1365-313X.2010.04471.x.

Cabili M.N., Trapnell C., Goff L., Koziol M., Tazon-Vega B., Regev A., Rinn J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 25(18):1915-27. doi:10.1101/gad.17446611.

Carninci P., et al. 2005. The transcriptional landscape of the mammalian genome. *Science.* 309(5740):1559-63. doi:10.1126/science.1112014.

Cawley S., Bekiranov S., Ng H., Kapranov P., Sekinger E., Kampa D., Piccolboni A., Sementchenko V., Cheng J., Williams A., Wheeler R. Wong B., Drenkow J., Yamanaka M., Patel S., Brubaker S., Tammana H., Helt G., Struhl K., Gingeras T., 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cel*l. 116(4):499-509. doi;10.1016/S0092-8674(04)00127-8.

Cech T., Steitz J., 2014. The Noncoding RNA Revolution - Trashing Old Rules to Forge New Ones. *Cell.* 157(1):77-94. doi: 10.1016/j.cell.2014.03.008.

Chang Z., Li G., Liu J., Zhang Y., Ashby C., Liu D., Cramer C.L., Huang X., 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.* 16, 30. doi:10.1186/s13059-015-0596-2.

Chávez Montes R.A., de Fátima Rosas-Cárdenas F., De Paoli E., Accerbi M., Rymarquis L.A., Mahalingam G., Marsch-Martínez N., Meyers B.C., Green P.J., de Folter S., 2014. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat Commun*. 5:3722. doi:10.1038/ncomms4722.

Chiou T.J., Aung K., Lin S.I., Wu C.C., Chiang S.F., Su C.L., 2006. Regulation of phosphate homeostasis by MicroRNA in Arabidopsis. *Plant Cell.* 18(2):412-21. doi:10.1105/tpc.105.038943.

Comings et al. 1972. The structure and function of chromatin. *Adv Hum Genet*. 3:237-431.

Costa M., Artur M., Maia J., Jonkheer E., Derks M., Nijveen H., et al. 2017. A footprint of desiccation tolerance in the genome of *Xerophyta viscosa. Nat Plants*. 3:17038. doi:10.1038/nplants.2017.38.

Csorba T., Questa J., Sun Q., Dean C., 2014. Antisense COOLAIR mediates the coordinated switching of chromatin states at *FLC* during vernalization. *Proc. Natl. Acad. Sci. USA*, 111:16160-5. doi:10.1073/pnas.1419030111.

Cuperus J.T., Fahlgren N., Carrington J.C., 2011. Evolution and functional diversification of MIRNA genes. *Plant Cell*. 23(2):431-42. doi:10.1105/tpc.110.082784.

Dace H., Sherwin H., Illing N., Farrant J., 1998. Use of metabolic inhibitors to elucidate mechanisms of recovery from desiccation stress in the resurrection plant *Xerophyta humilis. Plant Growth Reg.* 24:171–177. doi:10.1023/A:1005883907800.

Davidson N.M., Oshlack A., 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol*. 15(7):410. doi: 10.1186/s13059-014-0410-6.

Debernardi J.M., Rodriguez R.E., Mecchia M.A., Palatnik J.F., 2012. Functional Specialization of the Plant miR396 Regulatory Network through Distinct MicroRNA–Target Interactions*. PLOS Genetics*. 8(1):e1002419. doi:10.1371/journal.pgen.1002419.

Deng F., Zhang X., Wang W., Yuan R., Shen F., 2018. Identification of *Gossypium hirsutum* long non-coding RNAs (lncRNAs) under salt stress. *BMC Plant Biol*. 18:23. doi: 10.1186/s12870-018-1238-0.

Derrien T., Johnson R., Bussotti G., et al., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression*. Genome Res*. 22(9):1775-89. doi:10.1101/gr.132159.111.

Dinakar C., Bartels D., 2013. Desiccation tolerance in resurrection plants: new insights from transcriptome, proteome and metabolome analysis. *Front Plant Sci*. 4:482. doi:10.3389/fpls.2013.00482.

Ding J., Zhou S., Guan J., 2011. miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM. *BMC Bioinformatics*. 12:216. doi:10.1186/1471-2105-12-216.

Dirks R.M., Pierce N.A., 2004. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J Comput Chem*. 25(10):1295-304. doi:10.1002/jcc.20057.

Djebali S., Davis C., Merkel A., et al. 2012. Landscape of transcription in human cells. *Nature.* 489(7414):101-8. doi:10.1038/nature11233.

Ebert M.S., Neilson J.R., Sharp P.A. 2007. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods. 24(9):721-6. doi:10.1038/nmeth1079.

Eddy S., 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol*. 22(21):R898-9. doi:10.1016/j.cub.2012.10.002.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 447(7146):799-816. doi:10.1038/nature05874.

Eulalio A., Huntzinger E., Izaurralde E., 2008. Getting to the root of miRNA-mediated gene silencing. *Cell*. 132(1):9-14. doi:10.1016/j.cell.2007.12.024.

Ezkurdia I., Juan D., Rodriguez J., Frankish A., Diekhans M., Harrow J., Vazquez J., Valencia A., Tress M., 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet.* 23(22):5866-78. doi: 10.1093/hmg/ddu309.

Fahlgren N., Howell M.D., Kasschau K.D., Chapman E.J., Sullivan C.M., Cumbie J.S., Givan S.A., Law T.F., Grant S.R., Dangl J.L., Carrington J.C., 2007. High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS One*. 2(2):e219. doi:10.1371/journal.pone.0000219.

Fan C., Hao Z., Yan J., Li G., 2015. Genome-wide identification and functional analysis of lincRNAs acting as miRNA targets or decoys in maize. *BMC Genomics*. 16:793 doi:10.1186/s12864-015-2024-0.

Fang Y., Spector D.L., 2007. Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living Arabidopsis plants. *Curr Biol.* 17(9):818-23. doi:10.1016/j.cub.2007.04.005.

Farrant J. 2000. A comparison of mechanisms of desiccation tolerance among three angiosperm resurrection plant species. *Plant Ecol*. 151(1):29–39. doi:10.1023/A:102653430

Farrant J., Cooper K., Hilgart A., Abdalla K., Bentley J., Thomson J., Dace H., Peton N., Mundree S., Rafudeen M., 2015. A molecular physiological review of vegetative desiccation tolerance in the resurrection plant *Xerophyta viscosa* (Baker). *Planta.* 242(2):407–426. doi:10.1007/s00425-015-2320-6.

Farrant J., Moore J., 2011. Programming desiccation-tolerance: from plants to seeds to resurrection plants. *Curr Opin Plant Biol* 14(3):340-5. doi: 10.1016/j.pbi.2011.03.018.

Fei Q., Xia R., Meyers B., 2013. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell*. 25(7):2400-15. doi: 10.1105/tpc.113.114652.

Felippes F.F., Wang J.W., Weigel D., 2012. MIGS: miRNA-induced gene silencing. *Plant J*. 70(3):541-7. doi: 10.1111/j.1365-313X.2011.04896.x.

Fickett J.W., 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*. 10(17): 5303–5318.

Franco-Zorrilla J.M., Valli A., Todesco M., Mateos I., Puga M.I., Rubio-Somoza I., Leyva A., Weigel D., García J.A., Paz-Ares J., 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet*. 39(8):1033-7. doi:10.1038/ng2079.

Friedländer M., Chen W., Adamidi C., Maaskola J., Einspanier R., Knespel S., Rajewsky N., 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*. 26(4):407-415. doi:10.1038/nbt1394.

Friedländer M., Lizano E., Houben A., Bezdan D., Báñez-Coronel M., Kudla G., Mateu-Huertas E., Kagerbauer B., González J., Chen K., LeProust E., Martí E., Estivill X., 2014. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol*. 15(4):R57. doi:10.1186/gb-2014-15-4-r57.

Friedländer M.R., Adamidi C., Han T., Lebedeva S., Isenbarger T.A., Hirst M., Marra M., Nusbaum C., Lee W.L., Jenkin J.C., Sánchez Alvarado A., Kim J.K., Rajewsky N., 2009. High-resolution profiling and discovery of planarian small RNAs. *Proc Natl Acad Sci USA*. 106(28):11546-51. doi:10.1073/pnas.0905222106.

Gaff D. F., Oliver M., 2013. The evolution of desiccation tolerance in angiosperm plants: a rare yet common phenomenon. *Funct. Plant Biol*. 40:315–328 doi:10.1071/FP12321.

Garcia D., 2008. A miRacle in plant development: role of microRNAs in cell differentiation and patterning. *Semin Cell Dev Biol*. 19(6):586-95. doi:10.1016/j.semcdb.2008.07.013.

Gascoigne D.K., Cheetham S.W., Cattenoz P.B., Clark M.B., Amaral P.P., Taft R.J., Wilhelm D., Dinger M.E., Mattick J.S., 2012. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics*. 28(23):3042-50. doi:10.1093/bioinformatics/bts582.

Gechev T.S., Benina M., Obata T., Tohge T., Sujeeth N., Minkov I., Hille J., Temanni M.R., Marriott A.S., Bergström E., Thomas-Oates J., Antonio C., Mueller-Roeber B., Schippers J.H., Fernie A.R., Toneva V., 2013. Molecular mechanisms of desiccation tolerance in the resurrection glacial relic *Haberlea rhodopensis. Cell Mol Life Sci*. 70(4):689-709. doi:10.1007/s00018-012-1155-6.

Gensel P., Andrews H., 1984. In: Plant Life Devonian. Praeger Scientific, New York. 380.

Gerlach W., Giegerich R., 2006. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics.* 22(6):762-4. doi:10.1093/bioinformatics/btk041.

Gilbert D., 2013. Gene-omes built from mRNA seq not genome DNA., in: 7th Annual Arthropod Genomics Symposium. Notre Dame.

Good M., Zalatan J., Lim W., 2011. Scaffold proteins: hubs for controlling the flow of cellular information. *Science*. 332(6030):680-6. doi:10.1126/science.1198701.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29: 644–52. doi:10.1038/nbt.1883.

Griffiths-Jones S., Saini H.K., van Dongen S., Enright A.J., 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 36(Database issue):D154–D158. doi:10.1093/nar/gkm952.

Guenther M., Levine S., Boyer L., Jaenisch R., Young R., 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell.* 130(1):77-88. doi:10.1016/j.cell.2007.05.042.

Guleria P., Mahajan M., Bhardwaj J., Yadav S., 2011. Plant small RNAs: biogenesis, mode of action and their roles in abiotic stresses. *Genomics Proteomics Bioinformatics.* 9(6):183-99. doi:10.1016/S1672-0229(11)60022-3.

Guo C., Xu Y., Shi M., Lai Y., Wu X., Wang H., Zhu Z., Poethig R.S., Wu G., 2017. Repression of miR156 by miR159 regulates the timing of the juvenile-to-adult transition in Arabidopsis. *Plant Cell.* 29(6):1293-1304. doi:10.1105/tpc.16.00975.

Guttman M., Amit I., Garber M., Frenc C., Lin M.F., Feldser D., Huarte M., et al., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 458(7235): 223–227. doi: 10.1038/nature07672.

Guttman M., Garber M., Levin J., Donaghey J., Robinson J., Adiconis X., Fan L., Koziol M., Gnirke A., Nusbaum C., Rinn J., Lander E., Regev A., 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*. 28(5):503-10. doi:10.1038/nbt.1633.

Hanson L., McMahon K.A., Johnson M.A.T., Bennet M.D., 2001. First Nuclear DNA C-values for 25 Angiosperm Families. Ann Bot. 87:251-258. doi:10.1006/anbo.2000.1325.

Hawkins P.G., Morris K.V., 2010. Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription*. 1(3):165–175. doi:10.4161/trns.1.3.13332.

Helliwell C., Robertson M., Finnegan E., Buzas D., Dennis E., 2011. Vernalization-repression of Arabidopsis *FLC* requires promoter sequences but not antisense transcripts. *PLoS One*. 6(6):e21513. doi:10.1371/journal.pone.0021513.

Heo J., Sung S., 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*. 331(6013):76-9. doi:10.1126/science.1197349.

Hofacker I.L., Fontana W., Stadler P.F., Bonhoeffer L.S., Tacker M., Schuster P., 1994. Fast folding and comparison of RNA secondary structures. *Chemie/Chemical Monthly.* 125(2):167–188. doi:10.1007/BF00818163.

Huang W., Peng S., Xian Z., Lin D., Hu G., Yang L., Ren M., Li Z., 2017. Overexpression of a tomato miR171 target gene SlGRAS24 impacts multiple agronomical traits via regulating gibberellin and auxin homeostasis. *Plant Biotechnol J*. 15(4): 472–488. doi:10.1111/pbi.12646.

Huang X., Madan A., 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868-77. doi:10.1101/gr.9.9.868.

Ietswaart R., Wu Z., Dean C., Flowering time control: another window to the connection between antisense RNA and chromatin. *Trends Genet*. 28(9):445-53. doi:10.1016/j.tig.2012.06.002.

Illing N., Denby K., Collett H., Shen A., Farrant J., 2005. The Signature of Seeds in Resurrection Plants: A Molecular and Physiological Comparison of Desiccation Tolerance in Seeds and Vegetative Tissues. *Integr Comp Biol.* 45(5):771–787. doi:10.1093/icb/45.5.771

Ivashuta S., Banks I.R., Wiggins B.E., Zhang Y., Ziegler T.E., Roberts J.K., Heck G.R., 2011. Regulation of Gene Expression in Plants through miRNA Inactivation. *PLoS One*. 6(6):e21330. doi:10.1371/journal.pone.0021330.

Jia H., McCarty D.R., Suzuki, M., 2013. Distinct roles of LAFL network genes in promoting the embryonic seedling fate in the absence of VAL repression. *Plant Physiol.* 163:1293–305. doi:10.1104/pp.113.220988.

Jin J., Liu J., Wang H., Wong L., Chua N.H., 2013. PLncDB: plant long non-coding RNA database. *Bioinformatics.* 29(8): 1068–1071. doi: 10.1093/bioinformatics/btt107.

Johnson N., Yeoh J., Coruh C., Axtell M., 2016. Improved Placement of Multi-mapping Small RNAs. *G3* 6(7):2103-11. doi:10.1534/g3.116.030452.

Jones-Rhoades M.W., Bartel D.P., 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell.* 14(6):787-99. doi:10.1016/j.molcel.2004.05.027.

Jones-Rhoades M.W., Bartel D.P., Bartel B., 2006. MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol.* 57:19-53. doi: 10.1146/annurev.arplant.57.032905.105218.

Joshi R.K., Megha S., Basu U., Rahman M.H., Kav N.N., 2016. Genome Wide Identification and Functional Prediction of Long Non-Coding RNAs Responsive to *Sclerotinia sclerotiorum* Infection in *Brassica napus*. *PLoS One*. 11:e0158784. doi:10.1371/journal.pone.0158784.

Jung J., Park C., 2007. MIR166/165 genes exhibit dynamic expression patterns in regulating shoot apical meristem and floral development in Arabidopsis. *Planta.* 225(6):1327-38. doi:10.1007/s00425-006-0439-1.

Kaczkowski B., Torarinsson E., Reiche K., Havgaard J.H., Stadler P.F., Gorodkin J., 2009. Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*. 25(3):291-4. doi:10.1093/bioinformatics/btn628.

Kang W., Friedländer M.R., 2015. Computational Prediction of miRNA Genes from Small RNA Sequencing Data. *Front Bioeng Biotechnol*. 3:7:1-14. doi: 10.3389/fbioe.2015.00007.

Kapranov P. Cheng J., Dike S., et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*. 316(5830):1484-8. doi:10.1126/science.1138341.

Kartha R., Subramanian S., 2014. Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation. *Front Genet*. 5:8. doi:10.3389/fgene.2014.00008.

Kawashima C.G., Yoshimoto N., Maruyama-Nakashita A., Tsuchiya Y.N., Saito K., Takahashi H., Dalmay T., 2009. Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. *Plant* J. 57(2):313-21. doi:10.1111/j.1365-313X.2008.03690.x.

Kenrick P., Crane P.R., 1997. The origin and early evolution of plants on land. *Nature*. 389:33-39.

Khvorova A., Reynolds A., Jayasena S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell*. 115(2):209-16. doi:10.1016/S0092-8674(03)00801-8.

Kino T., Hurt D., Ichijo T., Nader N., Chrousos G., 2010. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*. 3(107):ra8. doi: 10.1126/scisignal.2000568.

Kong L., Zhang Y., Ye Z.Q., Liu X.Q., Zhao S.Q., Wei L., Gao G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 35(Web Server issue):W345-9. doi:10.1093/nar/gkm391.

Kornienko A., Dotter C., Guenzl P., Gisslinger H., Gisslinger B., Cleary C., Kralovics R., Pauler F., Barlow D., 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol*. 17:14. doi: 10.1186/s13059-016-0873-8

Kozomara A., Griffiths-Jones S., 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39(Database issue):D152-7. doi: 10.1093/nar/gkq1027.

Kulcheski F.R., de Oliveira L.F., Molina L.G, Almerão M.P., Rodrigues F.A., Marcolino J., Barbosa J.F., Stolf-Moreira R., Nepomuceno A.L., Marcelino-Guimarães F.C., Abdelnoor R.V., Nascimento L.C, Carazzolle M.F., Pereira G.A., Margis R., 2011. Identification of novel soybean microRNAs involved in abiotic and biotic stresses. *BMC Genomics.* 10;12:307. doi:10.1186/1471-2164-12-307.

Kurihara Y., Takashi Y., Watanabe Y., 2006. The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. *RNA*. 12(2):206-12. doi:10.1261/rna.2146906.

Lagos-Quintana M., Rauhut R., Lendeckel W., Tuschl T., 2001. Identification of novel genes coding for small expressed RNAs. *Science*. 294(5543):853-8. doi:10.1126/science.1064921.

Lai E.C., 2002. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*. 30(4):363-4. doi:10.1038/ng865.

Lander E., Linton L., Birren B., et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409(6822):860-921. doi:10.1038/35057062.

Langmead B., Salzberg S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357-9. doi:10.1038/nmeth.1923.

Langmead B., Trapnell C., Pop M., Salzberg S., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10(3):R25. doi:10.1186/gb-2009-10-3-r25.

Lau N., Lim L., Weinstein E., Bartel D., 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science*. 294(5543):858-62. doi:10.1126/science.1065062.

Lee J., 2009. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome*. Genes Dev*. 23(16):1831-42. doi:10.1101/gad.1811209.

Lee R., Ambros V., 2001. An extensive class of small RNAs in *Caenorhabditis elegans. Science.* 294(5543):862-4. doi:10.1126/science.1065329.

Lee Y., Kim M., Han J., Yeom K.H., Lee S., Baek S.H., Kim N., 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* 23(20): 4051–4060. doi: 10.1038/sj.emboj.7600385.

Li H., Wang Y., Chen M., Xiao P., Hu C., Zeng Z., Wang C., Wang J., Hu Z., 2016. Genome-wide long non-

coding RNA screening, identification and characterization in a model microorganism *Chlamydomonas reinhardtii. Scientific Reports* 6:34109. doi:10.1038/srep34109.

Li J., Yang Z., Yu B., Liu J., Chen X., 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. *Curr Biol*. 15(16):1501-7. doi:10.1016/j.cub.2005.07.029.

Li L., Eichten S.R., Shimizu R., Petsch K., Yeh C.T., Wu W., 2014. Genome-wide discovery and characterization of maize long non-coding RNAs*. Genome Biol*. 15:R40. doi:10.1186/gb-2014-15-2-r40.

Li W., Li C., Li S., Peng M., 2017. Long noncoding RNAs that respond to *Fusarium oxysporum* infection in 'Cavendish' banana (*Musa acuminata*). *Scientific Reports* 7:16939. doi:10.1038/s41598-017-17179-3.

Li W., Oono Y., Zhu J., He X., Wu J., Iida K., Lu X., Cui X., Jin H., Zhu J., 2008. The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and post-transcriptionally to promote drought resistance. *Plant Cell.* 20(8):2238-51. doi:10.1105/tpc.108.059444.

Liang G., Yang F., Yu D., 2010. MicroRNA395 mediates regulation of sulfate accumulation and allocation in *Arabidopsis thaliana. Plant J.* 62(6):1046-57. doi:10.1111/j.1365-313X.2010.04216.x.

Lim L.P., Lau N.C., Weinstein E.G., Abdelhakim A., Yekta S., Rhoades M.W., Burge C.B., Bartel D.P., 2003. The microRNAs of *Caenorhabditis elegans. Genes Dev.* 17(8): 991–1008. doi:10.1101/gad.1074403.

Lin M.F., Jungreis I., Kellis M., 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* 27(13):i275-82. doi:10.1093/bioinformatics/btr209.

Lin S., Chiang S., Lin W, Chen J., Tseng C., Wu P., Choiu T., 2008. Regulatory Network of MicroRNA399 and PHO2 by Systemic Signaling. *Plant Physiol*. 147(2): 732–746. doi:10.1104/pp.108.116269.

Lingel A., Simon B., Izaurralde E., Sattler M., 2003. Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. *Nature*. 426(6965):465-9. doi:10.1038/nature02123.

Liu C.G., Calin G.A., Meloon B., et al., 2004. An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc Natl Acad Sci USA*. 101(26):9740-4. doi:10.1073pnas.0403293101.

Liu J., Jung C., Xu J., Wang H., Deng S., Bernad L., 2012. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant. *Cell*. 24:4333–4345. doi:10.1105/tpc.112.102855.

Liu J., Wang H., Chua1 N.H., 2015b. Long noncoding RNA transcriptome of plants. *Plant Biotechnol J.* 13(3):319-28. doi:10.1111/pbi.12336.

Liu T.Y., Huang T.K., Tseng C.Y., Lai Y.S., Lin S.I., Lin W.Y., Chen J.W., Chiou T.J., 2012b. PHO2-dependent

degradation of PHO1 modulates phosphate homeostasis in Arabidopsis. *Plant Cell.* 24(5):2168-83. doi:10.1105/tpc.112.096636.

Liu X., Hao L., Li D., Zhu L., Hu S., 2015a. Long Non-coding RNAs and Their Biological Roles in Plants. *Genomics Proteomics Bioinformatics*. 13(3):137–47. doi:10.1016/j.gpb.2015.02.003. doi:10.1016/j.gpb.2015.02.003.

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing. *Science* 327, 94–97.

Llave C., Xie Z., Kasschau K.D., Carrington J.C., 2002. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*. 297(5589):2053-6. doi:10.1126/science.1076311.

Lorenz R., Bernhart S., Hoener zu Siederdissen C., Tafer H., Flamm C., Stadler P., Hofacker I., 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 6:26. doi:10.1186/1748-7188-6-26.

Love M., Huber W., Anders S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15(12):550. doi:10.1186/s13059-014-0550-8.

Lu X., Chen X., Mu M., Wang J., Wang X., Wang D., Yin Z., Fan W., Wang S., Guo L., Ye W., 2016. Genome-Wide Analysis of Long Noncoding RNAs and Their Responses to Drought Stress in Cotton (Gossypium hirsutum L.). PLoS One. 11(6): e0156723. doi:10.1371/journal.pone.0156723.

Lyall R., (2016). Regulation of desiccation tolerance in Xerophyta seedlings and leaves (Unpublished doctoral dissertation). University of Cape Town, Cape Town.

Ma Z., Hu X., Cai W., Huang W., Zhou X., Lio Q., Yang H., Wang J., Huang J., 2014. Arabidopsis miR171-Targeted Scarecrow-Like Proteins Bind to GT cis-Elements and Mediate Gibberellin-Regulated Chlorophyll Biosynthesis under Light Conditions. *PLOS Genetics.* 10(8):e1004519. doi:10.1371/journal.pgen.1004519.

Maere S., Heymans K, Kuiper M., 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics.* 21(16):3448-9. doi:10.1093/bioinformatics/bti551.

Maia J., Dekkers B.J., Dolle M.J., Ligterink W., Hilhorst H., 2014. Abscisic acid (ABA) sensitivity regulates desiccation tolerance in germinated Arabidopsis seeds. *New Phytol*. 203(1):81-93. doi:10.1111/nph.12785.

Maia J., Dekkers B.J., Provart N., Ligterink W., Hilhorst H., 2011. The Re-Establishment of Desiccation Tolerance in Germinated Arabidopsis thaliana Seeds and Its Associated Transcriptome. *PLoS One*. 6(12): e29123. doi: 10.1371/journal.pone.0029123.

Maoa H., Yu L., Li Z., Yan Y., Han R., Liu H., Ma M., 2016. Genome-wide analysis of the SPL family transcription factors and their responses to abiotic stresses in maize. *Plant Gene*. 10, 1-12. doi:10.1016/j.plgene.2016.03.003.

Marques A., Ponting C., 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol*. 10(11):R124. doi:10.1186/gb-2009-10-11-r124.

Martianov I., Ramadass A., Serra Barros A., Chow N., Akoulitchev A., 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*. 445(7128):666-70. doi:10.1038/nature05519.

Mathews D.H., 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA.* 10(8):1178-90. doi:10.1261/rna.7650904.

Mathews D.H., Sabina J., Zuker M., Turner D.H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol.* 288(5):911-40. doi:10.1006/jmbi.1999.2700.

Matranga C., Tomari Y., Shin C., Bartel D.P., Zamore P.D., 2005. Passenger-Strand Cleavage Facilitates Assembly of siRNA into Ago2-Containing RNAi Enzyme Complexes. *Cell.* 123(4):543-5. doi:10.1016/j.cell.2005.08.044.

Mattick J.S., 2011. The central role of RNA in human development and cognition. *FEBS Lett.* 585(11):1600-16. doi:10.1016/j.febslet.2011.05.001.

Mattick J.S., Rinn J.L., 2015. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol.* 22(1):5-7. doi:10.1038/nsmb.2942.

Mehler M.F., Mattick J.S., 2007. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. *Physiol Rev.* 87(3):799-823. doi:10.1152/physrev.00036.2006.

Meng Y., Shao C., Wang H., Jin Y., 2012. Target mimics: an embedded layer of microRNA-involved gene regulatory networks in plants. *BMC Genomics.* 13:197. doi: 10.1186/1471-2164-13-197.

Mercer T.R., Dinger M.E., Mattick J.S., 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet.* 10(3):155-9. doi: 10.1038/nrg2521.

Metpally R., Nasser S., Malenica I., Courtright A., Carlson E., Ghaffari L. Villa S., Tembe W., Van Keuren-Jensen K., 2013. Comparison of Analysis Tools for miRNA High Throughput Sequencing Using Nerve Crush as a Model. *Front Genet*. 4:20:1-13. doi:10.3389/fgene.2013.00020.

Meyers B., Axtell M., Bartel B., Bartel D., Baulcombe D., Bowman J., Cao X., Carrington J., Chen X., Green P., Griffiths-Jones S., Jacobsen S., Mallory A., Martienssen R., Poethig R., Qi Y., Vaucheret H., Voinnet O., Watanabe Y., Weigel D., Zhu J., 2008. Criteria for annotation of plant MicroRNAs. *Plant Cell*. 20(12):3186-90. doi:10.1105/tpc.108.064311.

Mishra A.K., Duraisamy G.S., Matoušek J., 2015. Discovering microRNAs and their targets in plants. *Crit. Rev. Plant Sci*. 34: 553–571. doi:10.1080/07352689.2015.1078614.

Morea, E., da Silva, E., e Silva. G., Valente, G., Rojas, C., Vincentz, M., Nogueira, F., 2016. Functional and evolutionary analyses of the miR156 and miR529 families in land plants. *BMC Plant Biol.* 16: 153. doi:10.1186/s12870-016-0802-8.

Morriss G.R., Cooper T.A., 2017. Protein sequestration as a normal function of long noncoding RNAs

and a pathogenic mechanism of RNAs containing nucleotide repeat expansions. *Hum Genet.* 136(9):1247-1263. doi:10.1007/s00439-017-1807-6.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 5(7):621-8. doi: 10.1038/nmeth.1226.

Moxon S., Jing R., Szittya G., Schwach F., Rusholme Pilcher R.L., Moulton V., Dalmay T., 2008. Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res.* 18(10):1602-9. doi: 10.1101/gr.080127.108.

Mückstein U., Tafer H., Bernhart S.H., Hernandez-Rosales M., Vogel J., Stadler P.F., Hofacker I.L., 2008. Translational Control by RNA-RNA Interaction: Improved Computation of RNA-RNA Binding Thermodynamics. In: Elloumi M., Küng J., Linial M., Murphy R.F., Schneider K., Toma C. (eds) Bioinformatics Research and Development. Communications in Computer and Information Science, vol 13. Springer, Berlin, Heidelberg.

Mückstein U., Tafer H., Hackermüller J., Bernhart S.H., Stadler P.F., Hofacker I.L., 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics*. 22(10):1177-82. doi:10.1093/bioinformatics/btl024.

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., Snyder M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 320(5881):1344-9. doi:10.1126/science.1158441.

Nakano T., Suzuki K., Fujimura T., Shinshi H., 2006. Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiol*. 140(2):411-32. doi:10.1104/pp.105.073783

Ng S.Y., Johnson R., Stanton L.W., 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J*. 31(3):522-33. doi:10.1038/emboj.2011.459.

Ni Z., Hu Z., Jiang Q., Zhang H., 2013. GmNFYA3, a target gene of miR169, is a positive regulator of plant tolerance to drought stress. *Plant Mol Biol*. 82(1-2):113-29. doi:10.1007/s11103-013-0040-5.

Ohno S., 1972. So much "junk" DNA in our genome. In: Evolution of Genetic Systems (ed. H.H. Smith), pp. 366-370. Gordon and Breach, New York.

Okazaki Y., Furuno M., Kasukawa T., et al., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* 420(6915):563-73. doi:10.1038/nature01266.

Oliver M.J., Dowd S.E., Zaragoza J., Mauget S.A., Payton P.R., 2004. The rehydration transcriptome of the desiccation-tolerant bryophyte *Tortula ruralis*: transcript classification and analysis. *BMC Genomics*. 5:89. doi: 10.1186/1471-2164-5-89.

Oliver M.J., Tuba Z., Mishler B.D., 2000. The Evolution of Vegetative Desiccation Tolerance in Land Plants. *Plant Ecol*. 151:85–100. doi:10.1023/A:1026550808557.

Palazzo A.F., Lee E.S., 2015. Non-coding RNA: what is functional and what is junk? *Front Genet*. 6:2. doi:10.3389/fgene.2015.00002.

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):1484-95. doi:10.1016/S0960-9822(02)01017-5.

Pearson W.R., Lipman D.J., 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*. 85(8):2444-8.

Pedersen I., David M., 2008. MicroRNAs in the immune response. *Cytokine*. 43(3):391-4. doi:10.1016/j.cyto.2008.07.016.

Phillips J., Dalmay T., Bartels D., 2007. The role of small RNAs in abiotic stress*. FEBS Lett*. 581(19):3592-7. doi:10.1016/j.febslet.2007.04.007.

Ponjavic J., Ponting C.P., Lunter G., 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*. 17(5): 556–565. doi: 10.1101/gr.6036807.

Ponting C.P., Oliver P.L., Reik W., 2009. Evolution and functions of long noncoding RNAs. *Cel*l. 136(4):629–641. doi:10.1016/j.cell.2009.02.006.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., et al., 2012. The Pfam protein families database. *Nucleic Acids Res*. 40(Database issue): D290–D301. doi: 10.1093/nar/gkr1065.

Qin T., Zhao H., Cui P., Albesher N., Xiong L., 2017. A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant Physiol*. 175(3):1321-1336. doi:10.1104/pp.17.00574.

Qin Z., Li C., Mao L., Wu L., 2014. Novel insights from non-conserved microRNAs in plants. *Front Plant Sci*. 5:586. doi: 10.3389/fpls.2014.00586.

Rajagopalan R., Vaucheret H., Trejo J., Bartel D.P., 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana. Genes Dev*. 20(24):3407-25. doi:10.1101/gad.1476406.

Ray A., Robinson-Beers K., Ray S., Baker S.C., Lang J.D., Preuss D, Milligan S.B., Gasser C.S., 1994. Arabidopsis floral homeotic gene BELL (BEL1) controls ovule development through negative regulation of *AGAMOUS* gene (AG). *Proc Natl Acad Sci USA*. 91(13):5761-5.

Rehmsmeier M., Steffen P., Hochsmann M., Giegerich R., 2004. Fast and effective prediction of microRNA/target duplexes. *RNA*. 10(10):1507-17. doi:10.1261/rna.5248604.

Reinhart B.J., Weinstein E.G., Rhoades M.W., Bartel B., Bartel D.P., 2002. MicroRNAs in plants. *Genes Dev.* 16(13):1616-26. doi:10.1101/gad.1004402.

Reiser L., Modrusan Z., Margossian L., Samach A., Ohad N., Haughn G.W., Fischer R.L., 1995. The BELL1 gene encodes a homeodomain protein involved in pattern formation in the Arabidopsis ovule primordium. *Cell.* 83(5):735-42.

Rhoades M.W., Reinhart B.J., Lim L.P., Burge C.B., Bartel B., Bartel D.P., 2002. Prediction of plant microRNA targets*. Cell*. 110(4):513-20. doi:10.1016/S0092-8674(02)00863-2 .

Riechmann J.L., Meyerowitz E.M., 1998. The AP2/EREBP family of plant transcription factors*. Biol Chem*. 379(6): 633-46. doi:10.1515/bchm.1998.379.6.633.

Rinn J.L., Chang H.Y., 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem*. 81:145-66. doi:10.1146/annurev-biochem-051410-092902.

Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S.D., Mungall K., et al., 2010. De novo assembly and analysis of RNA-seq data*. Nat. Methods* 7: 909–12. doi:10.1038/nmeth.1517.

Rodriguez M., Edsgärd D., Hussain S, Alquezar D., Rasmussen M., Gilbert T., Nielsen B., Bartels D., Mundy J., 2010a. Transcriptomes of the desiccation-tolerant resurrection plant *Craterostigma plantagineum. Plant J*. 63(2):212-28. doi:10.1111/j.1365-313X.2010.04243.x.

Rodriguez R.E., Mecchia M.A., Debernardi J.M., Schommer C., Weigel D., Palatnik J.F., 2010b. Control of cell proliferation in *Arabidopsis thaliana* by microRNA miR396. *Development*. 137(1):103-12. doi:10.1242/dev.043067.

Rubio-Somoza I., Weigel D., 2011. MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci*. 16(5):258-64. doi:10.1016/j.tplants.2011.03.001.

Ruby J., Jan C., Player C., Axtell M., Lee W., Nusbaum C., Ge H., Bartel D., 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans. Cell.* 127(6):1193-207. doi:10.1016/j.cell.2006.10.040.

Rymarquis L.A., Kastenmayer J.P., Hüttenhofer A.G., Green P.J., 2008. Diamonds in the rough: mRNA-like non-coding RNAs. *Trends Plant Sci.* 13(7):329-34. doi:10.1016/j.tplants.2008.02.009.

Salmena L., Poliseno L., Tay Y., Kats L., Pandolfi P.P., A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell.* 146(3):353-8. doi:10.1016/j.cell.2011.07.014.

Samani N.J., Erdmann J., Hall A.S., et al., 2007. Genomewide association analysis of coronary artery disease. *N Engl J Med.* 357(5):443-53. doi:10.1056/NEJMoa072366.

Santos-Mendoza M., Dubreucq B., Baud S., Parcy F., Caboche M., Lepiniec L., 2008. Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *Plant J.* 54(4):608-20. doi:10.1111/j.1365-313X.2008.03461.x.

Schwarz D.S., Hutvágner G., Du T., Xu Z., Aronin N., Zamore P.D., 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*. 115(2):199-208. doi:10.1016/S0092-8674(03)00759-1.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D, Amin N., Schwikowski B., Ideker T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13(11):2498-504. doi:10.1101/gr.1239303.

Sheldon C.C., Rouse D.T., Finnegan E.J., Peacock W.J., Dennis E.S., 2000. The molecular basis of vernalization: the central role of *FLOWERING LOCUS C* (*FLC*). *Proc Natl Acad Sci USA.* 97(7):3753-8. doi:10.1073/pnas.060023597.

Sherwin H., Farrant J., 1998. Protection mechanisms against excess light in the resurrection plants *Craterostigma wilmsii* and *Xerophyta viscosa*. *Plant Growth Regul*. 24(3):203–210. doi:10.1023/A:100580161

Shuai P., Liang D., Tang S., Zhang Z., Ye C.Y., Su Y., Xia X., Yin W., 2014. Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in *Populus trichocarpa. J Exp Bot*. 65(17):4975-83. doi:10.1093/jxb/eru256.

Song J.J., Liu J., Tolia N.H., Schneiderman J., Smith S.K., Martienssen R.A., Hannon G.J., Joshua-Tor L., 2003. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol*. 10(12):1026-32. doi:10.1038/nsb1016.

Song J.J., Smith S.K., Hannon G.J., Joshua-Tor L., 2004. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science*. 305(5689):1434-7. doi:10.1126/science.1102514.

Spitale R.C., Tsai M.C., Chang H.Y., 2011. RNA templating the epigenome: long noncoding RNAs as molecular scaffolds. *Epigenetics*. 6(5):539-43. doi:10.1016/j.molcel.2011.08.018.

Struhl K., 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol.* 14(2):103-5. doi:10.1038/nsmb0207-103.

Sun K., Chen X., Jiang P., Song X., Wang H., Sun H., 2013b. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics*. 14(Suppl 2): S7. doi:10.1186/1471-2164-14-S2-S7.

Sun L., Luo H., Bu D., Zhao G., Yu K., Zhang C., Liu Y., Chen R., Zhao Y., 2013a. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41(17):e166. doi:10.1093/nar/gkt646.

Sunkar R., Chinnusamy V., Zhu J., Zhu J.K., 2007. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends Plant Sci.* 12(7):301-9. doi:10.1016/j.tplants.2007.05.001

Sunkar R., Li Y.F., Jagadeeswaran G., 2012. Functions of microRNAs in plant stress responses. *Trends Plant Sci.* 17(4):196-203. doi:10.1016/j.tplants.2012.01.010.

Sunkar R., Zhu J.K., 2004. Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell.* 16(8):2001-19. doi:10.1105/tpc.104.022830.

Suzek B.E., Wang Y., Huang H., McGarvey P.B., Wu C.H., UniProt Consortium., 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics.* 31(6): 926–932. doi: 10.1093/bioinformatics/btu739.

Swiezewski S., Liu F., Magusin A., Dean C., 2009. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*. 462(7274):799-802. doi:10.1038/nature08618.

Tafer H., Hofacker I.L., 2008. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*. 24(22):2657-63. doi:10.1093/bioinformatics/btn193.

Taft R.J., Pheasant M., Mattick J.S., 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays.* 29(3):288-99. doi:10.1002/bies.20544.

Tang Y., Wang F., Zhao J., Xie K., Hong Y., Liu Y., 2010. Virus-Based MicroRNA Expression for Gene Functional Analysis in Plants. *Plant Physiol*. 153(2): 632–641. doi: 10.1104/pp.110.155796.

Taylor R.S., Tarver J.E., Hiscock S.J., Donoghue P.C., 2014. Evolutionary history of plant microRNAs. *Trends Plant Sci.* 19(3):175-82. doi:10.1016/j.tplants.2013.11.008.

Terrasson E., Buitink J., Righetti K., Ly Vu B., Pelletier S., Zinsmeister J., Lalanne D., Leprince O., 2013. An emerging picture of the seed desiccome: confirmed regulators and newcomers identified using transcriptome comparison. *Front. Plant Sci.* 4:497. doi:10.3389/fpls.2013.00497.

Thiebaut F., Grativol C., Tanurdzic M., Carnavale-Bottino M., Vieira T. et al. 2014. Differential sRNA Regulation in Leaves and Roots of Sugarcane under Water Depletion. *PLoS One*. 9(4):e93822. doi: 10.1371/journal.pone.0093822.

Todesco M., Rubio-Somoza I., Paz-Ares J., Weigel D., 2010. A Collection of Target Mimics for Comprehensive Analysis of MicroRNA Function in Arabidopsis thaliana. *PLoS Genet*. 6(7):e1001031. doi: 10.1371/journal.pgen.1001031.

Tomari Y., Matranga C., Haley B., Martinez N., Zamore P.D., 2004. A protein sensor for siRNA asymmetry. *Science.* 306(5700):1377-80. 10.1126/science.1102755.

Trindade I.M, Capitão C., Dalmay T., Fevereiro M., Santos D., 2010. miR398 and miR408 are up-regulated in response to water deficit in *Medicago truncatula*. *Planta*. 231(3):705-16. doi:10.1007/s00425-009-1078-0.

Tripathi R., Patel S., Kumari V., Chakraborty P., Varadwaj P.K., 2016. DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Netw Model Anal Health Inform Bioinform*. 5:21. doi:10.1007/s13721-016-0129-2.

Turner D.H., Sugimoto N., Freier S.M., 1988. RNA structure prediction. *Annu Rev Biophys Biophys* Chem. 17:167-92. doi:10.1146/annurev.bb.17.060188.001123.

Ulitsky I., Bartel D.P., 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*. 154(1): 26–46. doi:10.1016/j.cell.2013.06.020.

Ulitsky I., Shkumatava A., Jan C.H., Sive H., Bartel D.P., 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell.* 147(7):1537-50. doi:10.1016/j.cell.2011.11.055.

Valadkhan S., Valencia-Hipólito A., 2017. lncRNAs in Stress Response. *Curr Top Microbiol Immunol*. 394:203-36. doi:10.1007/82_2015_489.

Valencia-Sanchez M.A., Liu J., Hannon G.J., Parker R., 2006. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev*. 20(5):515-24. doi:10.1101/gad.1399806.

Vaucheret H., 2008. Plant ARGONAUTES. *Trends Plant Sci*. 13(7):350-8. doi:10.1016/j.tplants.2008.04.007.

Vaucheret H., Mallory A.C., Bartel D.P., 2006. AGO1 homeostasis entails coexpression of MIR168 and AGO1 and preferential stabilization of miR168 by AGO1. *Mol Cell*. 7;22(1):129-36. doi: 10.1016/j.molcel.2006.03.011.

Vazquez F., Blevins T., Ailhas J., Boller T., Meins F. Jr., 2008. Evolution of Arabidopsis MIR genes generates novel microRNA classes. *Nucleic Acids Res*. 36(20):6429–6438. doi:10.1093/nar/gkn670.

Voinnet O., 2008. Use, tolerance and avoidance of amplified RNA silencing by plants. *Trends Plant Sci.* 13(7):317-28. doi:10.1016/j.tplants.2008.05.004.

Voinnet O., 2009. Origin, biogenesis, and activity of plant microRNAs. *Cell.* 136(4):669-87. doi:10.1016/j.cell.2009.01.046.

Wang H., Chung P.J., Liu J., Jang I.C., Kean M.J., Xu J., Chua N.H., 2014. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res*. 24(3):444-53. doi:10.1101/gr.165555.113.

Wang J., Meng X., Dobrovolskaya OB., Orlov YL., Chen M., 2017. Non-coding RNAs and Their Roles in Stress Response in Plants. *Genomics Proteomics Bioinformatics.* 15(5):301-312. doi:10.1016/j.gpb.2017.01.007.

Wang J., Yu W., Yang Y., Li X., Chen T., Liu T., Ma N., Yang X., Liu R., Zhang B., 2015. Genome-wide analysis of tomato long non-coding RNAs and identification as endogenous target mimic for microRNA in response to TYLCV infection. *Scientific Reports*. 18;5:16946. doi:10.1038/srep16946.

Wang K.C., Chang H.Y., 2011. Molecular mechanisms of long noncoding RNAs*. Mol Cell.* 43(6): 904–914. doi: 10.1016/j.molcel.2011.08.018.

Wang L., Park H., Dasari S., Wang S., Kocher J.P., Li W., 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41(6):e74. doi: 10.1093/nar/gkt006. Epub 2013 Jan 17.

Wang M., Yuan D., Tu L., Gao W., He Y., Hu H., et al., 2015. Long noncoding RNAs and their proposed functions in fibre development of cotton (*Gossypium spp*.). *New Phytol.* 207(4):1181–1197. doi:10.1111/nph.13429.

Wark A., Lee H., Corn R., 2008. Multiplexed detection methods for profiling microRNA expression in biological samples. *Angew Chem Int Ed Engl.* 47(4):644-52. doi:10.1002/anie.200702450.

Wilhelm B.T., Marguerat S., Watt S., Schubert F., Wood V., Goodhead I., Penkett C.J., Rogers J., Bähler J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 453(7199):1239-43. doi:10.1038/nature07002.

Williamson V., Kim A., Xie B., McMichael G., Gao Y., Vladimirov V., 2013. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief Bioinform*. 14(1): 36–45. doi:10.1093/bib/bbs010.

Wu H.J., Wang Z.M., Wang M., Wang X.J., 2013. Widespread Long Noncoding RNAs as Endogenous Target Mimics for MicroRNAs in Plants. *Plant Physiol*. 161(4):1875–1884. doi:10.1104/pp.113.215962.

Wu J., Liu Q., Wang X., Zheng J., Wang T., You M., Sheng Sun Z., Shi Q., 2013. mirTools 2.0 for non-

coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol*. 10(7):1087-92. doi:10.4161/rna.25193.

Xiao L., Yang G., Zhang L., Yang X., Zhao S., Ji Z., Zhou Q., et al., 2015. The resurrection genome *of Boea hygrometrica*: A blueprint for survival of dehydration. *Proc. Natl. Acad. Sci. U. S. A.* 112:5833–7. doi:10.1073/pnas.1505811112.

Xin M., Wang Y., Yao Y., Song N., Hu Z., Qin D., Xie C., et al., 2011. Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biol*. 11:61. doi:10.1186/1471-2229-11-61.

Xu M.Y., Zhang L., Li W.W., Hu X.L., Wang M.B., Fan Y.L., Zhang C.Y., Wang L., 2013. Stress-induced early flowering is mediated by miR169 in *Arabidopsis thaliana. J Exp Bot*. 65(1):89-101. doi:10.1093/jxb/ert353.

Yamada K., Lim J., Dale J.M., et al., 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*. 302(5646):842-6. doi:10.1126/science.1088305.

Yan K., Liu P., Wu C.A., Yang G.D., Xu R., Guo Q.H., Huang J.G., Zheng C.C., 2012. Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in *Arabidopsis thaliana. Mol Cell*. 48(4):521-31. doi:10.1016/j.molcel.2012.08.032.

Yan K.S., Yan S., Farooq A., Han A., Zeng L., Zhou M.M., 2003. Structure and conserved RNA binding of the PAZ domain. *Nature.* 426(6965):468-74. doi:10.1038/nature02129.

Yang J., Qu L., 2012. DeepBase: annotation and discovery of microRNAs and other noncoding RNAs from deep-sequencing data. *Methods Mol Biol* 822:233-48. doi:10.1007/978-1-61779-427-8_16.

Yang X., Li L., 2011. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*. 27(18):2614-5. doi:10.1093/bioinformatics/btr430.

You C., Cui J., Wang H., Qi X., Kuo L.Y., Ma H., Gao L., Mo B., Chen X., 2017. Conservation and divergence of small RNA pathways and microRNAs in land plants. *Genome Biol.* 18(1):158. doi:10.1186/s13059-017-1291-2.

Yu B., Yang Z., Li J., Minakhina S., Yang M., Padgett R.W., Steward R., Chen X., 2005. Methylation as a crucial step in plant microRNA biogenesis. *Science.* 307(5711):932-5. doi:10.1126/science.1107130.

Zhang B., Pan X., Cannon C.H., Cobb G..P, Anderson T.A., 2006. Conservation and divergence of plant microRNA genes. *Plant J*. 46(2):243-59. doi:10.1111/j.1365-313X.2006.02697.x.

Zhang J., Xu Y., Huan Q., Chong K., 2009. Deep sequencing of *Brachypodium* small RNAs at the global genome level identifies microRNAs involved in cold stress response. *BMC Genomics*. 10:449. doi:10.1186/1471-2164-10-449.

Zhang Y., Liao J., Li Z., Yu Y., Zhang J., Li Q., Qu L., Shu W., Chen Y., 2014. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol*. 15:512. doi:10.1186/s13059-014-0512-1

Zhao B., Liang R., Ge L., Li W., Xiao H., Lin H., Ruan K., Jin Y., 2007. Identification of drought-induced microRNAs in rice. *Biochem Biophys Res Commun*. 354(2):585-90. doi:10.1016/j.bbrc.2007.01.022.

Zhu E., Zhao F., Xu G., Hou H., Zhou L., Li X., Sun Z., Wu J., 2010. mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res*. 38:392-397. doi:10.1093/nar/gkq393.

Zhu Q.H., Stephen S., Taylor J., Helliwell C A., Wang M.B., 2014. Long noncoding RNAs responsive to *Fusarium oxysporum* infection in *Arabidopsis thaliana*. *New Phytol*. 201:574–584. doi:10.1111/nph.12537.

Zuker M., 2000. Calculating nucleic acid secondary structure. *Curr Opin Struct Biol*. 10(3):303-10. doi:10.1016/S0959-440X(00)00088-9.

Zuker M., Stiegler P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*. 9(1):133-48. doi:10.1093/nar/9.1.133.