Dissertations and Theses in Biological Sciences   Biological Sciences, School of

10-2014

Investigating the Role of MicroRNAs in the Response to Nitrogen Deprivation in the Green Alga *Chlamydomonas reinhardtii*

Adam Voshall

University of Nebraska-Lincoln, avoshal1@unl.edu

Investigating the Role of MicroRNAs in the Response to Nitrogen Deprivation in the

Green Alga *Chlamydomonas reinhardtii*

by

Adam Voshall

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

(Genetics, Cellular and Molecular Biology; Bioinformatics)

Under the Supervision of Professor Heriberto Cerutti

Lincoln, Nebraska

October, 2014

Investigating the Role of MicroRNAs in the Response to Nitrogen Deprivation in the

Green Alga *Chlamydomonas reinhardtii*

Adam Voshall, Ph.D.

University of Nebraska, 2014

Adviser: Heriberto Cerutti

Microalgae are gaining attention as a potential feedstock for the production of biodiesel,

mainly derived from triacylglycerols (TAG). In many algae, TAG synthesis increases

dramatically upon certain stresses but this is often accompanied by growth retardation.

Rational improvements to strain productivity are limited by the scant knowledge on algal

lipid metabolism and gene regulatory mechanisms. In this context, systems-level

approaches aimed at understanding and modeling metabolic and regulatory networks may

enable hypothesis-driven genetic engineering strategies. The green microalga

Chlamydomonas reinhardtii accumulates significant amounts of TAGs under nutrient

starvation and provides a genetically tractable model for manipulating biosynthetic

pathways. In order to gain insight into *Chlamydomonas* TAG metabolism and regulation,

we have examined both the transcriptome of strain CC-125 grown photoautotrophically

in nutrient-replete or nitrogen-depleted media and the corresponding changes in

microRNA population. While the production of microRNAs (miRNAs) by

Chlamydomonas reinhardtii has been established for several years, little is known about

how they target transcripts for regulation or what roles they play in cellular processes, in

particular whether they play a role in regulating the accumulation of TAG in nitrogen-depleted media. To characterize functional miRNAs in *Chlamydomonas*, we identified small RNAs associated with Flag-tagged-AGO3 by affinity purification and deep sequencing in cells grown heterotrophically and cells grown photoautotrophically in nitrogen-replete and nitrogen-deplete media and used these small RNAs to determine changes in the miRNA populations across these three conditions. We determined the role that these miRNAs play in regulating the response to nitrogen-deplete media by searching the genes that are differentially expressed in that condition for potential targets of these miRNA.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Heriberto Cerutti, for his support and guidance throughout this project. Besides introducing me to the project, and the field of microRNAs in general, his guidance helped the project develop in new and exciting ways none of us expected when we started.

The guidance and understanding of my committee member Dr. Etsuko N. Moriyama was also indispensible for this project. Without her expertise, I would have had no idea where to begin or how to perform much of the work presented here.

I would also like to thank my collaborators Dr. Joseph Msanne and Eun-Jeong Kim, whom performed the bulk of the experimental work this project is based on, and without whom this project would not have been possible. They, along with the other collaborators for each part, also helped to verify and refine many of the computational predictions made for each part of this project, helping to ensure that the final product reflects what is actually happening inside the cells.

To all of those listed here, and everyone else that has either contributed to this work or helped to support me throughout it, I thank you.

Table of Contents

List of Tables

List of Figures

List of Appendices

CHAPTER 1- INTRODUCTION

Overview

The work for this dissertation is to examine what biological roles microRNAs (miRNAs) perform in the green alga *Chlamydomonas reinhardtii.* It is divided into three main parts. The first part (Chapter 2) investigates the changes in gene expression across the transcriptome between *C. reinhardtii* cells grown photoautotrophically in minimal media with and without a source of nitrogen. The second part (Chapter 3) uses small RNAs (sRNAs) associated with Argonaute-family protein AGO3 to predict canonical miRNAs expressed in cells grown in heterotrophic conditions and to predict the targets of those canonical miRNAs. The third part (Chapter 4) uses the criteria for canonical miRNAs and target prediction established in Chapter 2 to determine the changes in miRNA populations in the conditions used for part one and determine which of the genes that are differentially expressed in those conditions are potential targets for miRNAs.

The work on the transcriptomic changes between cells grown in nitrogen-replete and nitrogen-deplete media focuses primarily on the pathways relating to triacylglycerol (TAG) biosynthesis and degradation, and the competing pathway of starch biosynthesis and degradation for four time points (0h, 24h, 48h, and 144h after being transferred to nitrogen-deplete media). TAG, which *C. reinhardtii* accumulates under certain stress conditions including nitrogen starvation, is a very energy dense neutral lipid that can be converted into diesel fuel. While *C. reinhardtii* itself is not an ideal candidate to use for bio-diesel production because of its relatively low lipid content even under conditions in which it accumulates lipids, as a model algae insights gained into how algae accumulate

TAG will allow more efficient engineering of naturally high oil content algae to make bio-diesel production more economically viable. The previous work on this topic has been performed on cells grown in heterotrophic conditions in which they can use acetate as a carbon source without the need for photosynthesis. The work presented here, however, is performed on cells grown photoautotrophically, restricting the amount of carbon available for the production of lipids or starch. The metabolic pathways analyzed in this part were constructed by combining existing annotations for all of the pathways and using manually curated similarity-based searches from fungus and land plants to identify unnanotated genes in each of the pathways. The gene expression levels for all of the genes were compared between each of the time points to determine which genes were differentially expressed and when, revealing when the cells are preferentially storing carbon and energy in either starch or TAG.

miRNAs, which are part of the RNA interference (RNAi) machinery, are a widely conserved eukaryotic mechanism for controlling gene expression at a transcript level by either triggering the degradation of the target transcript through cleavage or by blocking the production of protein from the transcript through translational repression. While the biosynthesis and mechanism of action differs slightly between plants, animals, and fungus, the core machinery is present in nearly all eukaryotic lineages. The machinery has been confirmed to be absent from relatively few unicellular and simple multicellular species. In contrast, there are no known cases of the machinery being absent from complex multicellular species with multiple tissue types, where they appear to play a critical role in maintaining cell specialization and tissue differentiation. In many

unicellular organisms, such as *C. reinhardtii*, however, it remains unclear what role they play in the cells, despite their widespread conservation.

The work presented here attempts to determine whether miRNA-associated gene regulation plays a role in the response to nitrogen starvation that leads to the accumulation of TAG by predicting the miRNAs present in the cells and predicting which transcripts they are targeting. We based these predictions on small RNAs that were already associated with AGO3, reducing the possibility of sequenced reads being from unrelated degradation products. To determine which of the AGO3-associated sRNAs are cannonical miRNAs as opposed to other types of small interfering RNAs (siRNAs), all of the sequences sRNAs were mapped to the genome to attempt to reconstruct the miRNA precursors (pre-miRNAs). To determine which genes in *C. reinhardtii* are being regulated by miRNAs in a given condition, this work searches the transcriptome for potential binding for both cleavage and translational repression using the predominant reads from the pre-miRNAs that meet the criteria of canonical miRNAs. This work determines how miRNAs are involved in the regulation of the response to nitrogen starvation by comparing the changes in the miRNA populations and the corresponding change in the genes they are predicted to target between the nitrogen-replete and nitrogen-deplete conditions used to TAG accumulation.

The remainder of this chapter provides an introduction to the biosynthesis and function of miRNAs, current approaches for the prediction of miRNAs and their targets, and presents what is currently known about miRNAs in algae.

Review of Literature

Summary of microRNA function

RNA interference (RNAi) provides cells with a delicate means of translational repression through a variety of mechanisms. MicroRNAs (miRNAs) are one component of RNAi machinery, which consists of approximately 21 nucleotide (nt) RNA fragments [1]. These fragments are derived from small hairpin loops up to ~300 base pairs (bp) in length that are processed by the Dicer family of proteins. This short fragment is in turn used in the RNA-induced silencing complex (RISC) machinery to repress translation of target mRNA, either by preventing translation, or by cleaving the mRNA [2]. In order for this repression to occur, the RISC complex, with the miRNA present within it, binds to the target mRNA, using the miRNA to bind in a sequence specific manner. The extent of the complementarity between the miRNA and the mRNA plays a major role in determining whether silencing occurs through cleavage or translational repression. However, this correlation is not absolute [2]. Each step is described in detail in the following sections.

microRNA processing

Mature miRNAs can be derived from two sources: expressed endogenously within cells, or externally introduced in the form of double stranded RNA (dsRNA) [1]. The focus here will be on endogenously expressed miRNAs, though the final steps in processing and function are shared between both sources, as well as other types of small interfering RNAs (siRNAs). miRNAs are differentiated from other types of siRNAs by the characteristic, imperfect hairpin loop they form prior to processing, due to sequences

containing inverted repeats, also known as reverse compliments, allowing the single stranded RNA (ssRNA) to fold back onto itself, forming the loop structure [3]. These inverted repeats are reflected in the genomic sequence, which plays an important role in computational detection of *de novo* miRNAs [4].

For all of the organisms that produce miRNAs, he genomic sequences for miRNA are located in a variety of regions throughout their genomes [5]. While most are found either in intergenic regions or introns (in either sense or antisense orientations) of protein-coding genes, they can also be found within introns and exons on non-protein-coding genes and, rarely, in the exons of protein-coding genes [1]. Additionally, in many organisms, certain regions of the genome contain several miRNA loci in disproportionately high numbers. These regions are known as miRNA clusters, and the miRNAs contained in them may be controlled by the same or similar regulating factors.

miRNA precursors are transcribed into what is called the primary microRNA (pri-miRNA) (Figures 1.1, 1.2). The pri-miRNA structures, which can be several kilobases long in some species, are processed within the nucleus in a similar manner to mRNAs [5]. As with mRNA, most pri-miRNA are transcribed by RNA polymerase II, although some instances of transcription by RNA polymerase III (used primarily for transcription of ribosomal RNAs) have been identified [3]. For intronic pri-miRNAs, the mRNA containing them must first be spliced to release the pri-miRNA before miRNA-specific processing can continue.

In animals, the pri-miRNA is processed and cleaved by the Drosha protein, which contains both RNase III domains and a double stranded RNA-binding domain (dsRBD) (Figures 1.1, 1.2) [1]. Drosha forms a protein complex with either Pasha or DGCR8

proteins, depending on the species. These proteins also contain two dsRBD domains, and

are thought to play a crucial role in sequence recognition for the activity of Drosha.

Additionally, these cofactors help to determine where on the pri-miRNA structure Drosha

cleaves. This complex functions by cleaving the pri-miRNA into a smaller hairpin loop

structure, ~60-100bp in length, termed the preliminary miRNA (pre-miRNA). During

cleavage, Drosha leaves two unpaired nucleotides at the 3' end of the pre-miRNA. Once

processed by Drosha, the pre-miRNA is exported from the nucleus into the cytoplasm by

the nuclear export protein exportin-5 [1].

In the final step of miRNA processing prior to the sequence being loaded into the

effector complex that performs the gene silencing, the pre-miRNA is cleaved by Dicer

(Figures 1.1, 1.2) [1]. The Dicer protein is another RNase III protein, which functions

only in the cytoplasm of animals, cleaves the loop section of the pre-miRNA, leaving

only the small, dsRNA fragment containing the actual miRNA sequence and its

compliment (often referred either as the guide and passenger miRNAs or miRNA and

miRNA*, respectively). The Dicer protein acts as a measuring stick to ensure the final

product is the proper length (~21nt). As with the Drosha, cleavage by Dicer leaves two

unpaired nucleotides at the 3' end of the opposite strand, leaving two nucleotides

available for both of the strands in the duplex. The short duplex formed by Dicer is not

highly stable, with the miRNA* strand quickly separating and degrading while the

miRNA itself is loaded into the effector complex [1].

Contrary to miRNA processing in animals, plants, including *Chlamydomonas* and

Volvox, lack the Drosha protein and its cofactors (see Figures 1.3, 1.5 and 1.6) [1].

Instead, Dicer-like proteins, predominantly DCL1, process the pri-miRNA within the

nucleus, cutting the pri-miRNA in two locations instead of the one in animals, allowing

the immediate release of the miRNA/miRNA* duplex within the nucleus [6]. In this step,

the protein HYL1, which is also localized to the nucleus, binds to the RNA and fills the

role of Pasha/DCR1. Finally, HASTY and other nuclear exporters export the duplex out

of the nucleus. Once in the cytoplasm, the duplex is unwound, and the miRNA is loaded

into the effector complex, again leaving the miRNA* to be rapidly degraded [6].   Figure

1.1 summarizes the biosynthesis pathway in *C. reinhartii*.

RISC complex structure and function

     For miRNAs, as with other siRNAs, the effector complex within which the

miRNAs function is the RNA-induced silencing complex (RISC) [7]. The complex

consists of the Argonaute (AGO) protein into which the miRNA is fitted (Figure 1.3) [8].

AGO is a widely conserved family of proteins found in almost all eukaryotes, as well as

some archea and eubacteria. The AGO protein itself consists of several domains,

including the PAZ domain, into which the 3'-end of the miRNA is loaded, the MID

domain, which secures the 5'-end, and the PIWI domain, which contains the catalytic site

for mRNA binding and cleavage [8].

     In its inactive state prior to the miRNA binding, the AGO protein has a very closed

structure, concealing the catalytic sites [9]. In order to facilitate the binding of the

miRNA to the PAZ domain, the structure hinges open in the N-terminal domain and

pivots slightly, exposing the region of the PAZ domain that contains the RNA-binding

sites for the 3'-end of the miRNA (corresponding to the two unpaired nucleotides in the

miRNA/miRNA* duplex) [7]. The first nucleotide of the 5'-end of the miRNA is

secured within the MID domain, with the binding sites turned in towards the protein, making this nucleotide inaccessible to bind with target mRNAs [8]. In this conformation, nucleotides 2-8 from the 5'-end of the miRNA (known as the seed region for its role in seeding the catalytic reaction between AGO and the mRNA) face outward into the channel that will eventually contain the mRNA, while the remaining nucleotides are initially inaccessible and do not play an initial role in target selection [7].

In order to facilitate introduction of the mRNA strand into the central channel of the AGO protein, the AGO protein goes through an additional conformational change similar to the first [8]. In this stage, it hinges open further, creating a space in the central channel wide enough for the mRNA to pass through, and to facilitate bulges formed by imperfect pairing or gaps between the miRNA and mRNA sequence [10]. Because the miRNA structure is fixed against the AGO protein, however, bulges can only be present within the mRNA sequence, and not within the miRNA [8]. Within the bound complex, cleavage is depending on perfect base pairing between nucleotides 10 and 11, and without this binding, silencing can only occur through translational repression (Figure 1.3). It has been proposed that the binding of the seed region to the mRNA triggers the release of the remaining nucleotides of the miRNA and bringing the mRNA into the catalytic site of the PIWI domain [2]. However, the crystal structure for such a configuration has not yet been resolved, leaving open the question of how, mechanistically, the complementarity between the miRNA and target mRNA affects target recognition and cleavage.

In addition to the binding between the miRNA and target sequence, the type of repression that occurs also depends on the specific AGO protein containing the miRNA

[11]. In humans, there are at least 4 main AGO proteins: hAGO1, hAGO2, hAGO3, and hAGO4. Both hAGO2 and hAGO3 associate with miRNA, but only hAGO2 displays cleaving activity. Instead, hAGO3 functions by localizing mRNA to p-bodies within the cytoplasm, regardless of complementarity. In this case, the gene is still silenced through translational repression, and still leads to target degradation within the p-bodies. However, such localization effects leave open the possibility that other AGO proteins could localize targeted mRNAs to different locations in the cells or could repress translation in a manner that does not negatively affect the stability of the mRNA [11].

Evolution of the miRNA machinery

The machinery for the RNAi machinery is nearly ubiquitous in eukaryotic cells, with the core machinery likely present in the last eukaryotic common ancestor (LECA) [12, 13]. The core machinery for RNAi, a PIWI superfamily protein, a DICER-like protein, and an RNA dependent RNA polymerase (RdRP), appear to derive from a variety of prokaryotic sources, including both bacteria and archaea, but seem to have all been present in the last eukaryotic common ancestor (LECA) (Figure 1.5) [12-15]. The PIWI superfamily protein, of which AGO is a member, forms the core of the RNAi machinery in all of its forms, and is classically defined as containing a PIWI, MID, and PAZ domain, as with AGO described above (Figure 1.5) [13, 15]. These proteins are most similar to the prokaryotic PIWI (pPIWI) and pPIWI-RE families of proteins, which primarily associate with DNA/RNA hybrids and are thought to play a role in defending the cells against foreign DNA and mobile genetic elements [16-20]. These pPIWI proteins are found primarily in Archaea, however, the main catalytic feature of the PIWI

superfamily proteins, the RNaseH fold found within the PIWI domain, closely resembles the Endonucleos V family of proteins, including EndoV in bacteria and UvrC in archaea, which are highly conserved in both lineages [13, 21]. This distribution of domains suggest that the PIWI family of proteins originated from the UrvC family of proteins, and that the classic PIWI structure found in the LECA was already largely present in the archaeal pPIWI proteins [13, 22].

In contrast, the DICER-like proteins, which consist of two RNaseIII domains and a dsRNA-binding domain (dsRBD), is much more similar to the bacterial Type II CRISPR protein, which also processes dsRNA in a similar manner to eukaryotic Dicer-like proteins [23]. A phylogenetic analysis performed by Burroughs *et al.* [13] suggests that all eukaryotic RNaseIII domains share a common origin that is closely related to the bacterial RNaseIII domains, such as the one in CRISPR. Additionally, the transition from domain distribution in CRISPR, with one RNaseIII domain and a dsRBD, to that of Dicer-like proteins with two concurrent RNaseIII domains and a dsRBD, appears to have occurred at least as early as the LECA [13]. The final piece of the core RNAi machinery, RdRP, appears instead to have originated from a DNA dependent RNA polymerase (DdRP) in bacteriophages which was adapted to use an RNA template [24, 25]. How and when these proteins from a wide range of prokaryotic sources first coalesced into a single organism remains an open question, but follows a trend of other eukaryotic specific processes such as the acquisition of the spliceosome and the formation of pores in the nuclear envelope [26, 27]

The earliest RNAi machinery, however, likely processed sense-antisense (s-as) transcripts that form double stranded RNA (dsRNA) rather than the small hairpin

structures characteristic of miRNA, based on the presence of the former exclusive of the latter among some basal eukaryotes, including the chromalveoalte ciliate *Tetrahymena thermophila* and the kinetoplastid *Trypanosoma brucei* (Figuress. 1.5, 1.6) [13, 28-32]. The processing of independently transcribed miRNAs appears to have arisen sometime between the branching of ciliates and the branching of plants from the fungal and metazoan lineages given their presence in the later lineages as well as the apicomplexa *Toxoplasma gondii* and the amoebazoan mycetozoa *Dictyostelium discoideum* [32-34]. It is important to note, however, that the diplomonad *Giardia lamblia* and the parabasalid *Trichomonas vaginalis* also appear to process independently transcribed hairpin structures that resemble miRNAs, at least one of which has also been shown to perform post-transcriptional gene silencing [35-37]. While remains unclear how structurally and functionally similar the *Giardia* and *Trichomonas* hairpin-derived small RNAs are to plant and animal miRNAs (which in themselves have a number of differences in their biogenesis and function, see Figure 1.4 [38]) they may either indicate the emergence of miRNAs occurred closer to the LECA than previously thought or convergent evolution of miRNA-like regulation in the two lineages. The number of loss events necessary for the former scenario, however, make the latter scenario much more probable.

Loss of miRNAs

While the processing of miRNAs from independently transcribed hairpins appears to have evolved before the split between plant lineages from fungal and metazoan lineages, there have been a number of species in the these lineages that have lost some or all of the RNAi machinery (Figures 1.6, 1.7, 1.8) [12, 13, 39-41]. In general, these loss

events occur at early branches of these lineages, and occur in unicellular species such as the Choanoflagellida *Monosiga brevicollis* [41], several algae, including *Osterococcus* and *Micromonas* [42], and several fungi, including the Unikonta *Saccharomyces cerevisiae* [39, 40]. Additionally, these loss events can either involve the loss of the entire RNAi machinery, as in the cases mentioned above, or only of miRNA processing specifically, as in the Placazoa *Trichoplax adherens*, which, while multicellular, does not differentiate cell types [41, 43].

The loss of RNAi in fungi is particularly interesting, as, contrary to the earlier belief that the loss of RNAi was a random occurrence, the presence or absence of RNAi greatly impacts the ability of the cells to maintain the killer system [40]. The killer system, which allows the host organism to produce and excrete toxins to kill competing cells in their environment, is comprised of two viral components: the satellite dsRNA M (which are separate from the viral particles) and the L-A dsRNA mycovirus [44]. The M dsRNAs encode the toxins used in the killer system as well as the proteins necessary to protect the host cell from the toxin produced, while the L-A mycovirus encodes the toxins that allow the cells to kill other cells in the surrounding environment as well as the proteins necessary to protect itself against the toxins produced. However, the M dsRNAs cannot reproduce independently of the L-A mycovirus, and loss of the L-A virus leads to loss of the killer system as a whole. In cells that contain the killer system, both the M dsRNAs and the L-A mycovirus are maintained and replicated in the cytoplasm of the cells [44]. This system is found in many of the fungal species that have lost RNAi (see Figure 1.7). For *Saccharomyces cerevisiae*, the RNAi system was experimentally restored by introducing Dicer and Ago from *Saccharyomyces castelli*, a close relative of

S. cerevisiae that retains the RNAi pathway [40]. While the reintroduced pathway functionally silenced endogenous transposons, it had no observable impact on the growth rate or survivability of the cells in lab conditions. However, in addition to silencing transposons, the RNAi machinery also processed the dsRNA in both the M satellites and L-A mycovirus, similar to the role of RNAi in viral defense for other organisms. By silencing the L-A mycovirus, the RNAi system prevented the replication of the M satellites, and the killer system was quickly lost, rendering the cells susceptible to the toxins produced by cells without the reintroduced RNAi machinery [40]. For these fungal cells, the evolutionary benefits of harboring and maintaining the killer system to reduce competition appear to outweigh the cost of losing RNAi.

As mentioned previously, the loss of miRNAs generally occurs in unicellular or simple multicellular (multicellular organisms without cell differentiation) organisms, and there appears to be a positive correlation, at least within the plant and animal lineages, between the increased complexity of an organism an increase in the number of miRNAs contained within the genome (See Figures 1.8, 1.9) [41, 45]. Within two distantly related algal lineages, green alga (Chlorophyceae) and brown alga (Phaeophyceae), miRNAs have been categorized for two species that have independently evolved multicellularity with different cell types, *Volvox carteri* and *Ectocarpus siliculosus*, respectively [46, 47]. For both algae, target prediction of the miRNAs identified showed an over-representation of proteins involved in triggering and maintaining the different cell types. This mirrors the role that a subset of miRNAs play in higher plants and animals [6, 48-54]. The targeting of developmental genes in these distantly related algae suggests that, at least for some lineages, miRNAs became entwined in the regulation of developmental genes very

early on in the independent evolution of multicellularity, rendering them essential for the organism's proper development and survival [12, 13]. Outside of these lineages with strong conservation of a subset of miRNAs (see Figures 1.8, 1.9), however, loss events across several lineages (including some algae, but not Chlamydomonas, see Figure 1.6) suggest that RNAi may not be essential for survival [12, 13, 55].

Evolution of miRNAs

There have been several studies recently that have looked at the population genetics in animal and certain plant miRNAs to try to identify how the miRNAs and their targets are coevolving and what constraints are placed upon mutations within the miRNAs and their targets [56-63]. Studies about the polymorphisms in miRNAs have focused primarily on the relationship between the miRNA and their targets [56-58]. A study of polymorphisms in human miRNAs found ~400 single nucleotide polymorphisms (SNPs) across all of the miRNA and miRNA-binding sites investigated, of which the majority (~250) created the potential for novel targets for those miRNAs [56]. While the majority of the SNPs had distributions consistent with neutral evolution, SNP densities in both the miRNA sequence itself and the sequence of the miRNA binding sites show that polymorphisms capable of preventing regulation are subject to purifying selection. Interestingly, two SNPs in potential miRNA binding sites that allow the miRNA to target additional genes (one located in a taurine receptor and the other a subunit for the actin related protein 2/3 complex) show evidence positive selection when comparing the occurrence of the SNPs in humans and chimps compared to other primates. It is not clear what functional advantages regulating these targets could provide.

Similar evidence of purifying selection is found in higher plants. In the angiosperm *Arabidopsis*, along with strong constraints against sequence variation in either the miRNA or the target, with polymorphism rates much lower than in the regions surrounding the binding sites [57]. Additional research in the angiosperm *Oryza sativa* also showed that certain miRNAs are maintained through purifying selection during evolution protecting against loss of regulation for certain genes or the deleterious regulation of others, allowing only variation that produces no functional change, results neutral *de novo* targets, or introduces a selective advantage [58].

Despite the constraints on variation for existing miRNAs, studies looking at the wider evolution of miRNA populations between species have found that new miRNAs can arise spontaneously, giving rise to a plasticity in the overall miRNA populations [59-63]. In plants, this can be seen in the number of miRNAs that are either species-specific or that are contained in only a few very closely related species [59, 60]. Within *Arabidopsis*, there is evidence, based on the lack of similarity in the pre-miRNA sequence to any other region of the genome, that these sequences can arise from random sequences that form small or partial hairpin forming sequences in transcripts [61]. The spontaneous production of novel miRNAs from random sequences could account for large number of unique miRNA sequences between *A. thaliana* and *A. lyrata*, roughly 13% of the total miRNA sequences for each species, despite the relatively recent speciation [62].

There are many possible mechanisms for novel miRNAs to evolve in both plants and animals. In plants, where it is common for not only the binding site of the miRNA to be highly complementary to the mature miRNA, but also for the region surrounding the

binding site to be complementary to part of the pre-miRNA, one possibility is a partial

inverted duplication of part of the target sequence (Figure 1.10a) [64, 65]. Such inverted

duplication can occur as the result of the genomic rearrangements and segmental

duplications the evolution of plants [66]. Another similar possibility, which appears

possible in both plants and animals, is the insertion of transposable elements creating an

inverted repeat (Figure 1.10b) [67-69]. In both of these cases, in addition to formation of

inverted repeats in the genome, the sequences must also contain sufficient promoter

regions to be transcribed, giving rise to a number of potential hairpin sequences in the

genomes that are either too weakly transcribed to be functional, or not transcribed at all,

which has a significant impact on the prediction of non-conserved miRNAs that evolve in

such a manner (see *miRNA prediction* below). One final possibility, which is more likely

in animals with their low degree of complementarity between the miRNAs and their

targets than in plants, is that any transcript that forms a hairpin structure could be

processed by the miRNA machinery (Figure 1.10c) [70, 71]. In this model, a huge

variety of small RNAs could be processed and loaded into the RISC complex, regardless

of whether they are able to target a transcript. When these new sequences do arise, the

level of expression for each is also subjected to the same selective pressure as

polymorphisms, allowing tight control of the regulatory roles these miRNA play within

the cells [63].

miRNA prediction

A variety of approaches have been developed to predict miRNAs, both purely

computationally and by using a combination of computational and experimental methods,

most notably using high throughput RNA sequences (RNAseq or NGS) [72-77]. A brief

description of a few of these methods can be found in Table 1.1. The purely

computational methods can be further divided into similarity-based methods, which look

at the conservation of the predicted pre-miRNA sequence or structure, and *ab initio*

methods that solely on the features of the sequence being considered [72, 73]. These

computational methods, regardless of whether or not they depend on homology, include

several approaches that are either entirely heuristic or depend on machine learning for

making and filtering predictions [73, 78]. Determining the accuracy of methods is

difficult, especially in multi-cellular organisms with specialized tissues. While numerous

regions of the genomes are capable of being folded into pre-miRNA like hairpin

structures, not every hairpin produces an miRNA. Additionally, many miRNAs are

tissue specific, making the determination of true negative examples extremely difficult,

which has a significant impact on the training of machine-learning based approaches, as

well as the testing for all approaches.

One feature that is shared across all prediction methods, regardless of whether

they are similarity-based or *ab initio* based approaches, utilize machine learning, or

incorporate experimental data, is the prediction of the secondary structure of the pre-

miRNA transcript [72, 73, 78, 79]. The basis for this step is to determine whether

Dicer/Drosha in animals or DCL proteins in plants could process the sequence of the

predicted miRNA into the duplex containing the mature miRNA. The size of the hairpin

generated differs between plants (roughly 250nt in length [80, 81]) and animals (roughly

110nt in length [74, 82, 83]) to reflect the differences in miRNA biogenesis between the

two lineages. At a minimum, these secondary structure predictions are used solely to

determine whether the potential pre-miRNA would be capable of forming the requisite hairpin structure before using other methods, such as similarity-based searches, to determine the likelihood of being a real pre-miRNA, such as in C-mii [84]. In other cases, especially *ab initio* prediction methods, the minimum free energy (MFE) of the predicted structure, number and location of mismatches, or the relationship between the sequence and predicted structure are used for prediction, such as in miRNAminer, MiRFinder, MiPred, or NOVOMIR [80, 82, 83, 85]. As a result, the quality of the secondary structure prediction is essential for the accurate prediction of miRNAs, regardless of the method used. In cases where the secondary structure is predicted using different conditions than the organisms normally live, for instance differences in temperature, salinity, or pH, the predicted structure could differ wildly from what would occur in the cell, leading in turn to inaccurate miRNA predictions regardless of the method used [86].

The similarity-based methods for predicting miRNAs are based on the assumption that mature miRNAs that are essential for the organism's survival will be conserved between closely related species [82, 84, 87, 88]. These methods usually involve using known miRNAs to search a new genome either using only the mature miRNA sequence, such as MapMi and miRminer [82, 87], or by using the entire pre-miRNA sequence, such as in ProMir II or C-mii [84, 88]. In the cases where only the mature miRNA is used, the sequence is extended upstream and downstream to attempt to recover the pre-miRNA sequence. In both cases, sequences that meet the conservation threshold are folded as described above. Because there are a number of conserved hairpin structures, all of these methods use additional filters, either on the structure of the hairpin itself or in the patterns

of conservation to ensure positive predictions [89]. Algorithms such as miRNAminer that use the entire pre-miRNA for the initial search will generally also include a threshold (in this case 80% homology) for the conservation of the mature miRNA itself [82]. For algorithms that use the pattern of conservation within the hairpin, the number of substitutions in the sequences of the mature miRNA, miRNA*, remaining stem, and loop structure are compared under the assumption that the selective pressure to remain unchanged will be highest for the mature miRNA sequence and be progressively lower for the remaining sequences [73]. It is important to note, however, that these approaches cannot detect novel miRNAs, including any species-specific miRNAs.

Another approach, used by RNAmicro, allows the detection of conserved miRNAs without the need for already known miRNAs or pre-miRNA structures by leveraging comparative genomics to align and search entire genomes for possible pre-miRNA structures [90]. In this approach, the entire genomes are aligned against each other using RNAz [91, 92] (it is designed to detect a wide range of conserved non-coding RNAs (ncRNAs), of which miRNAs are a subset) or a similar alignment method. The resulting alignment is folded using RNAalifold, which predicts the conserved secondary structure instead of the secondary structure of the individual sequences, as is the case for the other similarity-based methods. If the conserved secondary structure resembles a hairpin, which the authors term almost hairpins, the conserved structure and the individual sequences are analyzed by a support vector machine (SVM). This SVM, which is trained on data from known metazoan miRNAs, classifies the sequences as either pre-miRNAs or other ncRNA based on the sequence composition (predominately the G/C content of the sequence) and the thermodyanimcs of the predicted hairpin

structure. However, because the SVM is trained on only metazoan miRNAs, this algorithm would be unreliable for prediction of miRNAs in other lineages [90]. Additionally, even though this approach does not require sequence similarity to any known miRNAs and is thus able to detect novel conserved miRNAs, because it depends on the alignment and conserved secondary structure of two different genomes, it still shares the limitation of other similarity-based methods that it cannot detect any species-specific miRNAs.

In order to overcome the limitations of similarity-based searches and find species-specific or novel miRNAs, *ab initio* methods make no assumptions about conservation of miRNAs between, instead basing their predictions solely on the features of the potential miRNA in organism being searched [75, 80, 83, 85, 93, 94]. These methods primarily focus on the sequence itself, the distribution and location of gaps and mismatches in the hairpin structure, the combination of the sequence and structure of the hairpin and the MFE of the hairpin (the MFE of miRNA containing hairpins has been reported to be lower than those of most non-miRNA hairpins) [85]. Because the differences in sequence and structure between true and false positive predictions are often very subtle, most recently developed *ab initio* methods rely on machine-learning based approaches to distinguish between the two cases which are generally trained on published data (see Table 1.1). However, because only a subset of miRNAs are probably known for each organism and very few known true negative miRNA-like hairpins are known, it remains unclear to what extent overfitting is occurring in these methods. Studies that have used these algorithms on separate data sets from which they were trained, whether within the same lineage or different lineages, often report much lower performance than the original

publication [95]. An additional source of problems for these algorithms comes from the

data used for training itself. Many of these algorithms use miRNA sequences deposited

in miRBase [96]. However, the validity of some of the miRNAs and pre-miRNA

structures deposited in the database (and by extension used to train these algorithms) is of

dubious quality and may not represent real miRNAs [97-99]. Some work has been done

to strengthen the reliability of these annotated sequences, but such work is still being

done only on a species by species basis as additional experimental evidence becomes

available [100, 101].

To overcome some of the issues of overtraining within an individual method,

ensemble methods such as MiRenSVM produce multiple machine-learning classifiers

each based on subsets of the training data available and combining the results of each

classifier to make an overall prediction [93]. In the case of MiRenSVM, each SVM is

trained using similar features to other *ab initio* methods (a combination of sequence and

structural information, the sequence composition, and thermodynamic characteristics of

the secondary structure) using a random sampling of positive and negative examples from

the training set proportional to the number of SVMs used. To combine the predictions of

each individual SVM, MiRenSVM uses a combination of Majority Vote, in which the

overall classification is whichever classification majority of the individual SVMs

predicts, and Mean Distance, where each example being tested is classified only by the

SVM that most closely matches that sample. However, the authors do not explain how

the algorithm handles samples where the two aggregation schemes disagree. In their

tests, which are also performed on sequences from miRBase and are thus also subjected

to some of the quality control issues described above, this ensemble approach does have a

overall higher accuracy geometric mean of classifications than individual classification methods used for comparison [93]. One important note, however, is that this algorithm was designed and tested only on animal miRNAs. As such, it remains unclear how reliable it would be for predicting miRNAs from other lineages.

More recently, miRNA prediction programs have started leveraging high-throughput sequencing technology to combine experimental small RNA data with other computational prediction methods to identify expressed miRNAs [74, 76]. These algorithms depend on sequence data of all of the small RNAs in the cell, which can then be mapped to the organism's genome to strengthen predictions of miRNA producing hairpins. For both of these algorithms, all of the reads with no more than 30nt between them following mapping are grouped. These groups are then extended in either direction to form two putative pre-miRNAs, one for the upstream extenstion and one for the downstream extension of the appropriate length for the lineage being searched. As mentioned above, this is 110nt for metazoans [74, 76] and 250nt for plants [81]. A small extension is added to the opposing side of the miRNA to balance the two arms of the hairpin in true pre-miRNAs. After the extensions are added, both algorithms perform a similar secondary structure analysis to the other methods listed above, with added constraints about the distributions of the reads around the hairpin. In both cases, the most highly expressed read from the hairpins that meet all of the criteria are considered the mature miRNA, on the assumption that each hairpin will only produce on mature sequence [74, 76]. For organisms in which antibodies for the Argonaute proteins are available, the small RNA isolation can be further refined by cross-linking the miRNAs to the argonaute protein of interest and immunoprecipitating the RISC complex [102, 103].

By isolating only small RNAs that are associated with the argonaute proteins, many of

the extraneous small RNAs, particularly degradation products that are roughly the same

size as miRNAs, are removed from the library, in turn further strengthening predictions

that the hairpins identified in the analysis are genuine miRNAs. However, because these

methods require miRNAs to be expressed to be predicted, they cannot identify tissue- or

condition-specific miRNAs not expressed in the sample.

miRNA target prediction

The prediction of miRNA targets can be done experimentally or computationally.

However, experimental detection of targets for novel miRNA sequences is extremely

limited, and current computational prediction techniques differ widely in both predictions

and accuracy, requiring additional experimental testing to verify predictions [104].

Computational prediction of miRNA targets is particularly challenging given their small

size and because only a portion of the miRNA needs to be complimentary to suppress

translation [77]. As described before (section 2), there are two modes for gene silencing,

one found predominantly in plants and the other found predominantly in animals, which

coincide to the extent of complementarity present between the miRNA and its targets

(Figures 1.2, 1.4). In the mode found predominantly in plants, interaction between the

miRNA and its targets requires complete or near complete complementarity, allowing for

the exclusion of any sequences with more than a very few mismatches. These interactions

almost exclusively lead to cleavage of the miRNA. The miRNA binding sites for this

type of interaction are most commonly found within the coding region of the genes.

Current models suggest that silencing through translational repression has been almost

entirely lost in plants [2]. However, there is a growing body of evidence for this type of silencing [105].

In contrast, the mechanism of regulation found predominantly in animals allows for considerable mismatching (Figures 1.2, 1.4), requiring more complicated rules for differentiating true interactions from false ones. These additional rules often focus on dividing the microRNA target interaction into two regions; the "seed" region corresponding to nucleotides 2–8 from the 5'-end of the miRNA initially exposed in the RISC structure, and the "non-seed" region, spanning the remainder of the sequence. Most computational models require either complete or near complete complementarity within the seed region [106]. The rest of the sequence requires considerably lower, if any, complementarity, depending on the model used and type of interaction [2].

In some models complementarity in the seed region is sufficient for miRNA/target interactions [2]. Most new models, however, include the need for some additional feature, either within the non-seed region, or in the mRNA regions flanking the miRNA-binding site [104, 107, 108]. These additional features include binding in the non-seed region and cobinding sites either for the same or other miRNAs [108], the mRNA secondary structure [107], and potential protein-binding sites along the surrounding region [104]. Also unlike the interactions found most often in plants, these interactions lead almost exclusively to silencing through translational repression, with little or no silencing through mRNA cleavage, and the binding sites for these interactions are found most often in the 3'-UTRs of the mRNAs they target.

Especially for this type of interaction, computational prediction of miRNA targets is extremely difficult due to how little specificity it requires. Using only 7nt to query a

large number of mRNA sequences can result in several thousand predicted matches, while, under even the highest estimates, each miRNA probably does not affect more than a few hundred targets, with many estimates well under 100 targets each [2]. As such, most computational methods require additional constraints during the search [109]. One of the most common of these constraints not based on the features of the interaction itself is the inclusion of target site conservation among closely related species. This conservation implies a functional significance, which, if a miRNA is predicted to bind in that location, strengthens the probability that that prediction is valid [109].

While most of the current prediction programs are heuristic based, machine-learning approaches represent a steadily growing approach to target prediction. These methods include TargetBoost, which relies on genetic programming to spawn and evolve pattern sequences in order to discern previously unknown rules for miRNA/target interactions and several support vector machine approaches [106]. Both of these methods function independently, in that they do all of the searching and matching without external programs, and predict the use of the heuristics used above. Instead, they rely on determining relevant features within the interaction itself. For TargetBoost, the genetic algorithm part attempts to find any necessary features such as stretches of complimentary pairings, such as the seed region above, or gaps or mismatches in certain regions of the interaction. Once all of these features are determined, they are combined and weighted to give the final classifier. However, because this approach is trained to determine the specific characteristics of certain miRNA, it performs very poorly on detecting any interactions that deviate from those specific patterns.

There are two basic approaches to use SVMs for screening microRNAs. They can either be used on their own, in which they search for potential target sites and model the interactions to determine to classify the interactions as real or false without any outside program, or they can take the results of another target prediction program and analyze only those sequences that are predicted to be interactions to determine which are false positives [109]. One example of the first approach is the miTarget SVM, which searches up to 10 mRNAs for possible hybridization sites for the miRNA, then builds a feature vector based on the first and last half of the miRNA/target hybridization site including thermodynamics and the position along the UTR of the hybridization [106]. However, the extremely limited number of inputs makes this specific algorithm very ill suited for genome wide analyses, using such an approach is very inflexible for comparing the results of several different methods, all of which are prone to false positives and could benefit from machine learning [109]. By using the second approach, where the SVM is used as a post-prediction filter, the results of any prediction method, using any of the heuristics listed above or any others, could still benefit from the machine learning.

miRNAs in green algae

In the green algae *Chlamydomonas reinhardtii* and *Volvox carteri*, microRNAs were first described only recently [110-112]. While it would initially seem that miRNA target selection in *C. reinhardtii* would function in a manner consistent with higher plants, very few of the miRNAs previously described have identifiable targets searching for near complete complementarity. This lack of targets leaves the possibility that these miRNAs instead function either largely or primarily through translational repression, in a

manner more akin to animals. This possibility is strengthened by growing amounts of evidence that in plants, in addition to the silencing by cleavage, plant genes are silenced through translation repression, as in animals [105]. Additionally, recent work using artificial constructs has shown that *C. reinhardtii* is capable of regulating protein levels through both mechanisms, determined by the complementarity of the seed region and the region around the catalytic site [113]. However, it is still unclear how prevalent each of the mechanisms are for regulating targeted genes in this species, and almost nothing is known about what kinds of genes and pathways these miRNA would be targeting under such a model.

While little is known about the miRNA/taret interactions in algae, studies have characterized both the machinery that processes miRNAs [114] and aids in quality control of the miRNAs produced [115] in *C. reinhardtii*. *C. reinhardtii* shares the same core processing machinery as higher plants (as shown in Figure 1.1) [115]. This machinery includes three AGO proteins, of which AGO-3 appears to be the most extensively involved in miRNA function, along with three DCL proteins. AGO1 and DCL1 appear to be primarily involved in the silencing of transposable elements. The quality control of miRNAs is at least partially controlled by the Mut68 protein, which appears to function by placing untemplated nucleotides (generally uridyls) at the 3'-end of the miRNA molecules, flagging them for degradation [114]. Analysis of the miRNA sequences produced by the Mut68 deficient mutant suggest that the DCL proteins are much less accurate in processing the miRNAs than previously though, but that the misprocessed molecules are rapidly degraded. However, it is possible that not all of the misprocessed miRNAs are immediately degraded, leaving open the possibility that new

miRNAs could be produced that provide selective advantages for the cells, allowing for

rapid changes in the overall miRNA populations in these cells.

Literature Cited

1. Kim, V.N., *MicroRNA biogenesis: coordinated cropping and dicing.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 376-85.
2. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
3. Schmittgen, T.D., *Regulation of microRNA processing in development, differentiation and cancer.* J Cell Mol Med, 2008. **12**(5B): p. 1811-9.
4. Brameier, M. and C. Wiuf, *Ab initio identification of human microRNAs based on structure motifs.* BMC Bioinformatics, 2007. **8**: p. 478.
5. Liu, X., K. Fortin, and Z. Mourelatos, *MicroRNAs: biogenesis and molecular functions.* Brain Pathol, 2008. **18**(1): p. 113-21.
6. Willmann, M.R. and R.S. Poethig, *Conservation and evolution of miRNA regulatory programs in plant development.* Curr Opin Plant Biol, 2007. **10**(5): p. 503-11.
7. Wang, Y., et al., *Structure of the guide-strand-containing argonaute silencing complex.* Nature, 2008. **456**(7219): p. 209-13.
8. Wang, Y., et al., *Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex.* Nature, 2008. **456**(7224): p. 921-6.
9. Ming, D., M.E. Wall, and K.Y. Sanbonmatsu, *Domain motions of Argonaute, the catalytic engine of RNA interference.* BMC Bioinformatics, 2007. **8**: p. 470.
10. Rashid, U.J., et al., *Structure of Aquifex aeolicus argonaute highlights conformational flexibility of the PAZ domain as a potential regulator of RNA-induced silencing complex function.* J Biol Chem, 2007. **282**(18): p. 13824-32.
11. Azuma-Mukai, A., et al., *Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing.* Proc Natl Acad Sci U S A, 2008. **105**(23): p. 7964-9.
12. Shabalina, S.A. and E.V. Koonin, *Origins and evolution of eukaryotic RNA interference.* Trends Ecol Evol, 2008. **23**(10): p. 578-87.
13. Burroughs, A.M., Y. Ando, and L. Aravind, *New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing.* Wiley Interdiscip Rev RNA, 2014. **5**(2): p. 141-81.
14. Muljo, S.A., C. Kanellopoulou, and L. Aravind, *MicroRNA targeting in mammalian genomes: genes and mechanisms.* Wiley Interdiscip Rev Syst Biol Med, 2010. **2**(2): p. 148-61.
15. Anantharaman, V., E.V. Koonin, and L. Aravind, *Comparative genomics and evolution of proteins involved in RNA metabolism.* Nucleic Acids Res, 2002. **30**(7): p. 1427-64.
16. Iyer, L.M., et al., *Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA.* Nucleic Acids Res, 2013. **41**(16): p. 7635-55.
17. Ma, J.B., et al., *Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein.* Nature, 2005. **434**(7033): p. 666-70.

18. Yuan, Y.R., et al., *Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage.* Mol Cell, 2005. **19**(3): p. 405-19.

19. Cerutti, L., N. Mian, and A. Bateman, *Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain.* Trends Biochem Sci, 2000. **25**(10): p. 481-2.

20. Makarova, K.S., et al., *Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements.* Biol Direct, 2009. **4**: p. 29.

21. Aravind, L., D.R. Walker, and E.V. Koonin, *Conserved domains in DNA repair proteins and evolution of repair systems.* Nucleic Acids Res, 1999. **27**(5): p. 1223-42.

22. Sasaki, H.M. and Y. Tomari, *The true core of RNA silencing revealed.* Nat Struct Mol Biol, 2012. **19**(7): p. 657-60.

23. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.

24. Rogozin, I.B., et al., *Origin and evolution of spliceosomal introns.* Biol Direct, 2012. **7**: p. 11.

25. Taliaferro, J.M., et al., *Two new and distinct roles for Drosophila Argonaute-2 in the nucleus: alternative pre-mRNA splicing and transcriptional repression.* Genes Dev, 2013. **27**(4): p. 378-89.

26. Koonin, E.V. and L. Aravind, *Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex.* Cell Cycle, 2009. **8**(13): p. 1984-5.

27. Collins, L. and D. Penny, *Complex spliceosomal organization ancestral to extant eukaryotes.* Mol Biol Evol, 2005. **22**(4): p. 1053-66.

28. Shi, H., C. Tschudi, and E. Ullu, *An unusual Dicer-like1 protein fuels the RNA interference pathway in Trypanosoma brucei.* RNA, 2006. **12**(12): p. 2063-72.

29. Zhang, H., et al., *Small RNA pyrosequencing in the protozoan parasite Entamoeba histolytica reveals strain-specific small RNAs that target virulence genes.* BMC Genomics, 2013. **14**: p. 53.

30. Couvillion, M.T., et al., *Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of Tetrahymena thermophila.* Genes Dev, 2009. **23**(17): p. 2016-32.

31. Tschudi, C., et al., *Small interfering RNA-producing loci in the ancient parasitic eukaryote Trypanosoma brucei.* BMC Genomics, 2012. **13**: p. 427.

32. Avesson, L., et al., *MicroRNAs in Amoebozoa: deep sequencing of the small RNA population in the social amoeba Dictyostelium discoideum reveals developmentally regulated microRNAs.* RNA, 2012. **18**(10): p. 1771-82.

33. Hinas, A., et al., *The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway.* Nucleic Acids Res, 2007. **35**(20): p. 6714-26.

34. Braun, L., et al., *A complex small RNA repertoire is generated by a plant/fungal-like machinery and effected by a metazoan-like Argonaute in the single-cell human parasite Toxoplasma gondii.* PLoS Pathog, 2010. **6**(5): p. e1000920.

35. Chen, X.S., et al., *High throughput genome-wide survey of small RNAs from the parasitic protists Giardia intestinalis and Trichomonas vaginalis.* Genome Biol Evol, 2009. **1**: p. 165-75.

36. Huang, P.J., et al., *Identification of putative miRNAs from the deep-branching unicellular flagellates.* Genomics, 2012. **99**(2): p. 101-7.

37. Saraiya, A.A., W. Li, and C.C. Wang, *A microRNA derived from an apparent canonical biogenesis pathway regulates variant surface protein gene expression in Giardia lamblia.* RNA, 2011. **17**(12): p. 2152-64.

38. Millar, A.A. and P.M. Waterhouse, *Plant and animal microRNAs: similarities and differences.* Funct Integr Genomics, 2005. **5**(3): p. 129-35.

39. Drinnenberg, I.A., et al., *RNAi in budding yeast.* Science, 2009. **326**(5952): p. 544-50.

40. Drinnenberg, I.A., G.R. Fink, and D.P. Bartel, *Compatibility with killer explains the rise of RNAi-deficient fungi.* Science, 2011. **333**(6049): p. 1592.

41. Berezikov, E., *Evolution of microRNA diversity and regulation in animals.* Nat Rev Genet, 2011. **12**(12): p. 846-60.

42. Cerutti, H., et al., *RNA-mediated silencing in Algae: biological roles and tools for analysis of gene function.* Eukaryot Cell, 2011. **10**(9): p. 1164-72.

43. Hertel, J., et al., *Non-coding RNA annotation of the genome of Trichoplax adhaerens.* Nucleic Acids Res, 2009. **37**(5): p. 1602-15.

44. Magliani, W., et al., *Yeast killer systems.* Clin Microbiol Rev, 1997. **10**(3): p. 369-400.

45. Nozawa, M., S. Miura, and M. Nei, *Origins and evolution of microRNA genes in plant species.* Genome Biol Evol, 2012. **4**(3): p. 230-9.

46. Li, J., Y. Wu, and Y. Qi, *MicroRNAs in a multicellular green alga Volvox carteri.* Sci China Life Sci, 2014. **57**(1): p. 36-45.

47. Billoud, B., et al., *Computational prediction and experimental validation of microRNAs in the brown alga Ectocarpus siliculosus.* Nucleic Acids Res, 2014. **42**(1): p. 417-29.

48. Carrington, J.C. and V. Ambros, *Role of microRNAs in plant and animal development.* Science, 2003. **301**(5631): p. 336-8.

49. Garcia, D., *A miRacle in plant development: role of microRNAs in cell differentiation and patterning.* Semin Cell Dev Biol, 2008. **19**(6): p. 586-95.

50. Hake, S., *MicroRNAs: a role in plant development.* Curr Biol, 2003. **13**(21): p. R851-2.

51. Llave, C., *MicroRNAs: more than a role in plant development?* Mol Plant Pathol, 2004. **5**(4): p. 361-6.

52. Niwa, R. and F.J. Slack, *The evolution of animal microRNA function.* Curr Opin Genet Dev, 2007. **17**(2): p. 145-50.

53. Wang, D., et al., *Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets.* PLoS One, 2010. **5**(9).

54. Bhaskaran, M. and M. Mohan, *MicroRNAs: History, Biogenesis, and Their Evolving Role in Animal Development and Disease.* Vet Pathol, 2013. **51**(4): p. 759-774.

55.   Cerutti, H. and J.A. Casas-Mollano, *On the origin and functions of RNA-mediated silencing: from protists to man.* Curr Genet, 2006. **50**(2): p. 81-99.

56.   Saunders, M.A., H. Liang, and W.H. Li, *Human polymorphism at microRNAs and microRNA target sites.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3300-5.

57.   Ehrenreich, I.M. and M.D. Purugganan, *Sequence variation of MicroRNAs and their binding sites in Arabidopsis.* Plant Physiol, 2008. **146**(4): p. 1974-82.

58.   Guo, X., et al., *Selection and mutation on microRNA target sequences during rice evolution.* BMC Genomics, 2008. **9**: p. 454.

59.   Loh, Y.H., S.V. Yi, and J.T. Streelman, *Evolution of microRNAs and the diversification of species.* Genome Biol Evol, 2011. **3**: p. 55-65.

60.   Barakat, A., et al., *Conservation and divergence of microRNAs in Populus.* BMC Genomics, 2007. **8**: p. 481.

61.   Felippes, F.F., et al., *Evolution of Arabidopsis thaliana microRNAs from random sequences.* RNA, 2008. **14**(12): p. 2455-9.

62.   Fahlgren, N., et al., *MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana.* Plant Cell, 2010. **22**(4): p. 1074-89.

63.   Takuno, S. and H. Innan, *Selection fine-tunes the expression of microRNA target genes in Arabidopsis thaliana.* Mol Biol Evol, 2011. **28**(9): p. 2429-34.

64.   Allen, E., et al., *Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana.* Nat Genet, 2004. **36**(12): p. 1282-90.

65.   Rajagopalan, R., et al., *A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.* Genes Dev, 2006. **20**(24): p. 3407-25.

66.   Li, A. and L. Mao, *Evolution of plant microRNA gene families.* Cell Res, 2007. **17**(3): p. 212-8.

67.   Piriyapongsa, J. and I.K. Jordan, *A family of human microRNA genes from miniature inverted-repeat transposable elements.* PLoS One, 2007. **2**(2): p. e203.

68.   Piriyapongsa, J., L. Marino-Ramirez, and I.K. Jordan, *Origin and evolution of human microRNAs from transposable elements.* Genetics, 2007. **176**(2): p. 1323-37.

69.   Piriyapongsa, J. and I.K. Jordan, *Dual coding of siRNAs and miRNAs by plant transposable elements.* RNA, 2008. **14**(5): p. 814-21.

70.   Svoboda, P. and A. Di Cara, *Hairpin RNA: a secondary structure of primary importance.* Cell Mol Life Sci, 2006. **63**(7-8): p. 901-8.

71.   Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs.* Nat Genet, 2005. **37**(7): p. 766-70.

72.   Ghosh, Z., J. Chakrabarti, and B. Mallick, *miRNomics-The bioinformatics of microRNA genes.* Biochem Biophys Res Commun, 2007. **363**(1): p. 6-11.

73.   Allmer, J., *Computational and Bioinformatics Methods for MicroRNA Gene Prediction.* Methods Mol Biol, 2014. **1107**: p. 157-75.

74.   An, J., et al., *miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data.* Nucleic Acids Res, 2013. **41**(2): p. 727-37.

75.   Wu, Y., et al., *MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences.* BMC Bioinformatics, 2011. **12**: p. 107.

76. Hendrix, D., M. Levine, and W. Shi, *miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data.* Genome Biol, 2010. **11**(4): p. R39.

77. Yoon, S. and G. De Micheli, *Computational identification of microRNAs and their targets.* Birth Defects Res C Embryo Today, 2006. **78**(2): p. 118-28.

78. Sacar, M.D. and J. Allmer, *Machine learning methods for microRNA gene prediction.* Methods Mol Biol, 2014. **1107**: p. 177-87.

79. Mitra, C.K. and K. Korla, *Functional, structural, and sequence studies of microRNA.* Methods Mol Biol, 2014. **1107**: p. 189-206.

80. Teune, J.H. and G. Steger, *NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome.* J Nucleic Acids, 2010. **2010**.

81. Yang, X. and L. Li, *miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants.* Bioinformatics, 2011. **27**(18): p. 2614-5.

82. Artzi, S., A. Kiezun, and N. Shomron, *miRNAminer: a tool for homologous microRNA gene search.* BMC Bioinformatics, 2008. **9**: p. 39.

83. Jiang, P., et al., *MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W339-44.

84. Numnark, S., et al., *C-mii: a tool for plant miRNA and target identification.* BMC Genomics, 2012. **13 Suppl 7**: p. S16.

85. Huang, T.H., et al., *MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans.* BMC Bioinformatics, 2007. **8**: p. 341.

86. Keshavan, R., et al., *Computational identification of Ciona intestinalis microRNAs.* Zoolog Sci, 2010. **27**(2): p. 162-70.

87. Guerra-Assuncao, J.A. and A.J. Enright, *MapMi: automated mapping of microRNA loci.* BMC Bioinformatics, 2010. **11**: p. 133.

88. Nam, J.W., et al., *ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W455-8.

89. Lai, E.C., et al., *Computational identification of Drosophila microRNA genes.* Genome Biol, 2003. **4**(7): p. R42.

90. Hertel, J. and P.F. Stadler, *Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.* Bioinformatics, 2006. **22**(14): p. e197-202.

91. Gruber, A.R., et al., *The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W335-8.

92. Gruber, A.R., et al., *RNAz 2.0: improved noncoding RNA detection.* Pac Symp Biocomput, 2010: p. 69-79.

93. Ding, J., S. Zhou, and J. Guan, *MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features.* BMC Bioinformatics, 2010. **11 Suppl 11**: p. S11.

94. Leung, W.S., et al., *Filtering of false positive microRNA candidates by a clustering-based approach.* BMC Bioinformatics, 2008. **9 Suppl 12**: p. S3.

95. Sebastian, B. and S.E. Aggrey, *Specificity and sensitivity of PROMIR, ERPIN and MIR-ABELA in predicting pre-microRNAs in the chicken genome.* In Silico Biol, 2008. **8**(5-6): p. 377-81.

96. Griffiths-Jones, S., *miRBase: microRNA sequences and annotation.* Curr Protoc Bioinformatics, 2010. **Chapter 12**: p. Unit 12 9 1-10.

97. Tarver, J.E., P.C. Donoghue, and K.J. Peterson, *Do miRNAs have a deep evolutionary history?* Bioessays, 2012. **34**(10): p. 857-66.

98. Sacar, M.D., H. Hamzeiy, and J. Allmer, *Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins?* J Integr Bioinform, 2013. **10**(2): p. 215.

99. Meng, Y., et al., *Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants.* RNA Biol, 2012. **9**(3): p. 249-53.

100. Hansen, T.B., J. Kjems, and J.B. Bramsen, *Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis.* RNA Biol, 2011. **8**(3): p. 378-83.

101. Wang, X. and X.S. Liu, *Systematic Curation of miRBase Annotation Using Integrated Small RNA High-Throughput Sequencing Data for C. elegans and Drosophila.* Front Genet, 2011. **2**: p. 25.

102. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.

103. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.

104. Didiano, D. and O. Hobert, *Molecular architecture of a miRNA-regulated 3' UTR.* RNA, 2008. **14**(7): p. 1297-317.

105. Lanet, E., et al., *Biochemical evidence for translational repression by Arabidopsis microRNAs.* Plant Cell, 2009. **21**(6): p. 1762-8.

106. Mendes, N.D., A.T. Freitas, and M.F. Sagot, *Current tools for the identification of miRNA genes and their targets.* Nucleic Acids Res, 2009. **37**(8): p. 2419-33.

107. Long, D., et al., *Potent effect of target structure on microRNA function.* Nat Struct Mol Biol, 2007. **14**(4): p. 287-94.

108. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

109. Yang, Y., Y.P. Wang, and K.B. Li, *MiRTif: a support vector machine-based microRNA target interaction filter.* BMC Bioinformatics, 2008. **9 Suppl 12**: p. S4.

110. Molnar, A., et al., *miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii.* Nature, 2007. **447**(7148): p. 1126-9.

111. Zhao, T., et al., *A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii.* Genes Dev, 2007. **21**(10): p. 1190-203.

112. Nematollahi, G., A. Kianianmomeni, and A. Hallmann, *Quantitative analysis of cell-type specific gene expression in the green alga Volvox carteri.* BMC Genomics, 2006. **7**: p. 321.

113. Yamasaki, T., et al., *Complementarity to an miRNA seed region is sufficient to induce moderate repression of a target transcript in the unicellular green alga Chlamydomonas reinhardtii.* Plant J, 2013. **76**(6): p. 1045-56.

114. Casas-Mollano, J.A., et al., *Diversification of the core RNA interference machinery in Chlamydomonas reinhardtii and the role of DCL1 in transposon silencing.* Genetics, 2008. **179**(1): p. 69-81.

115. Ibrahim, F., et al., *Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase promotes their degradation in Chlamydomonas.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3906-11.

116. Watson, J.D., *Molecular biology of the gene.* 6th ed. 2008, San Francisco Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ; Cold Spring Harbor Laboratory Press. xxxii, 841 p.

117. Friedlander, M.R., et al., *Discovering microRNAs from deep sequencing data using miRDeep.* Nat Biotechnol, 2008. **26**(4): p. 407-15.
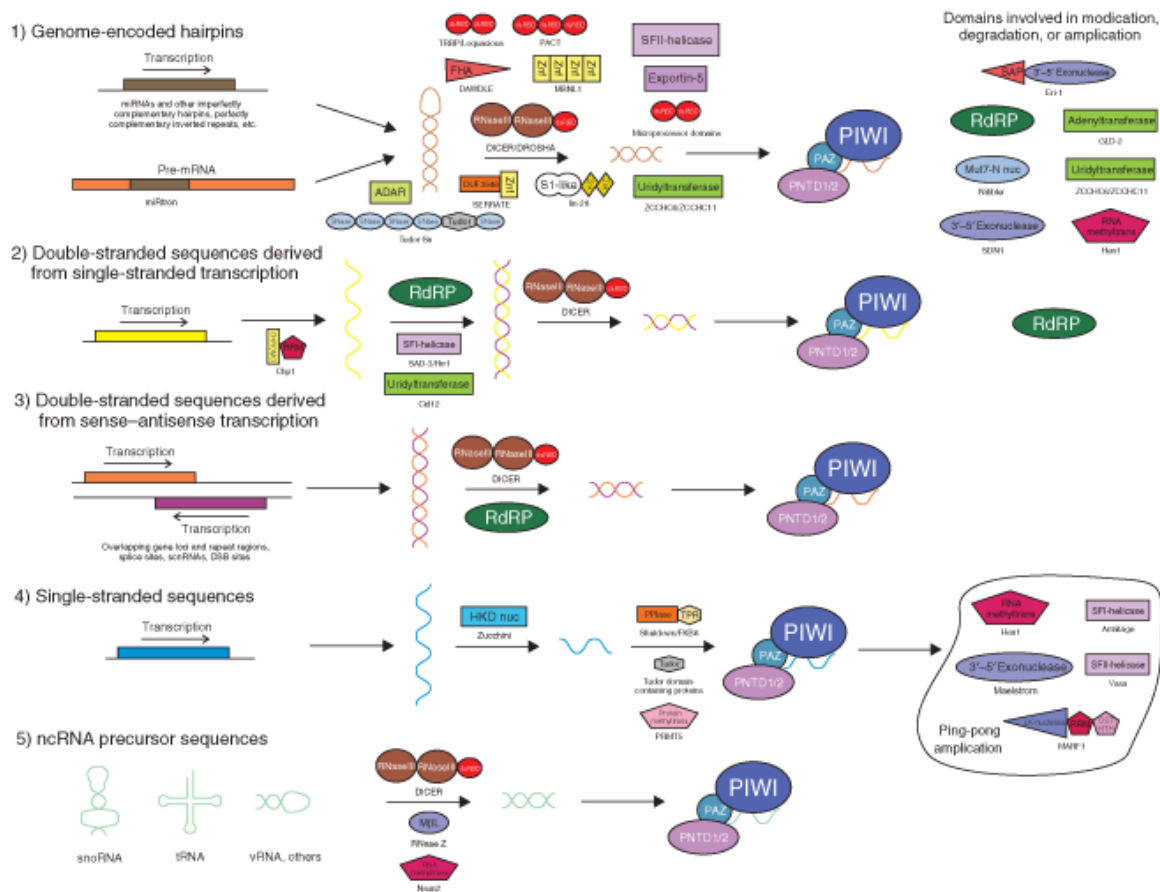
Figure 1.1: Overview of the standard biosynthesis pathways for the most common RNAi pathways

Taken from [13]. The major biosynthetic pathways are shown, along with the canonical molecular machinery to process the initial structures into the mature sequence and, when known, the accessory proteins necessary for loading the mature sequence into the RISC complex or that associate with the complex after activation.
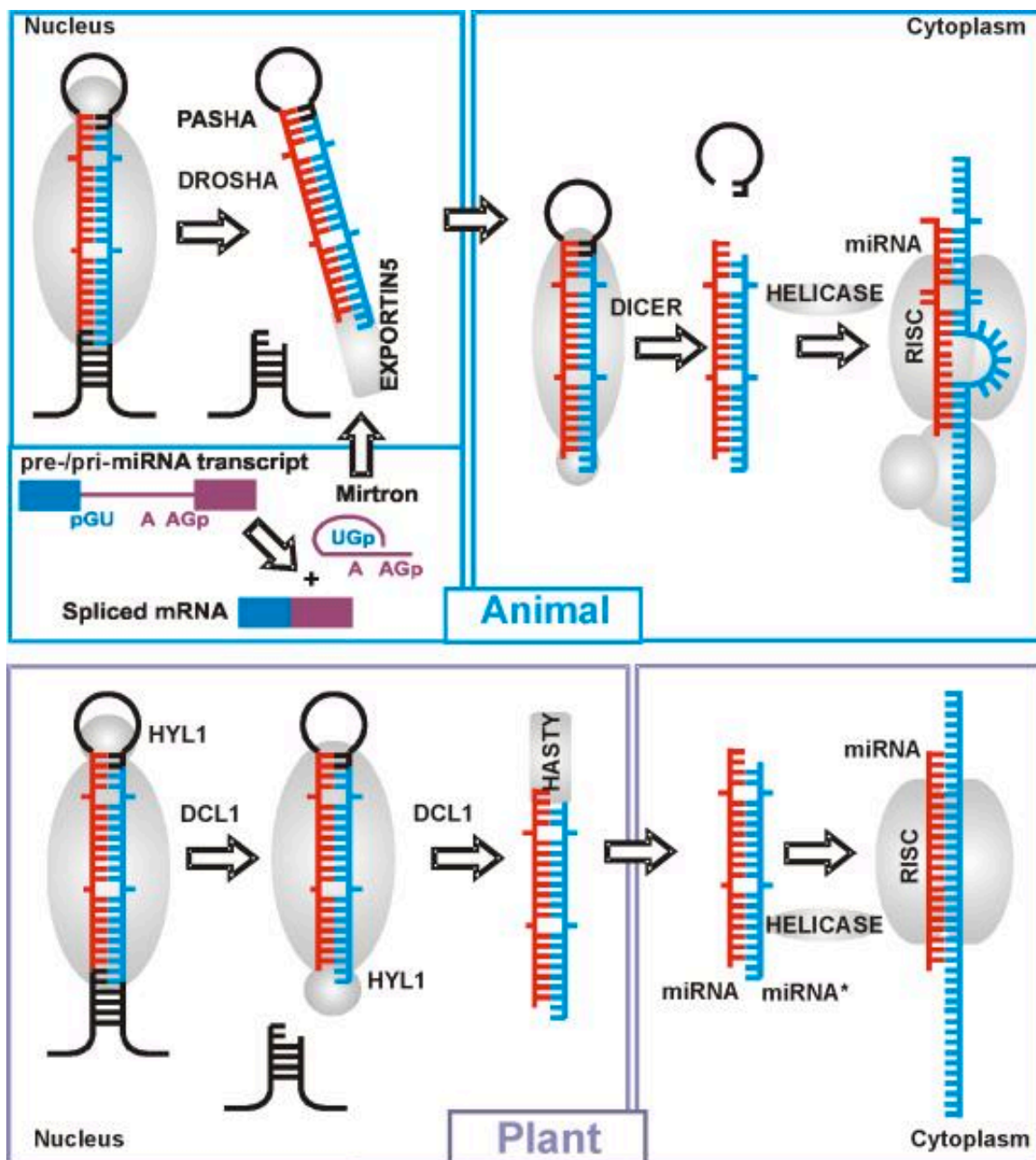
Figure 1.2: Comparison of miRNA processing in plants and animals

Taken from [12]. The processing of animal miRNAs from either independently

transcribed hairpins or from the introns of protein-coding genes is split between the

processing of the pre-miRNA by Pasha and Drosha in the nucleus and processing by

Dicer in the cytoplasm. In plants, the processing occurs in the nucleus by consequence

processing by DCL1 and only the guide and passenger strands are exported into the

cytoplasm.
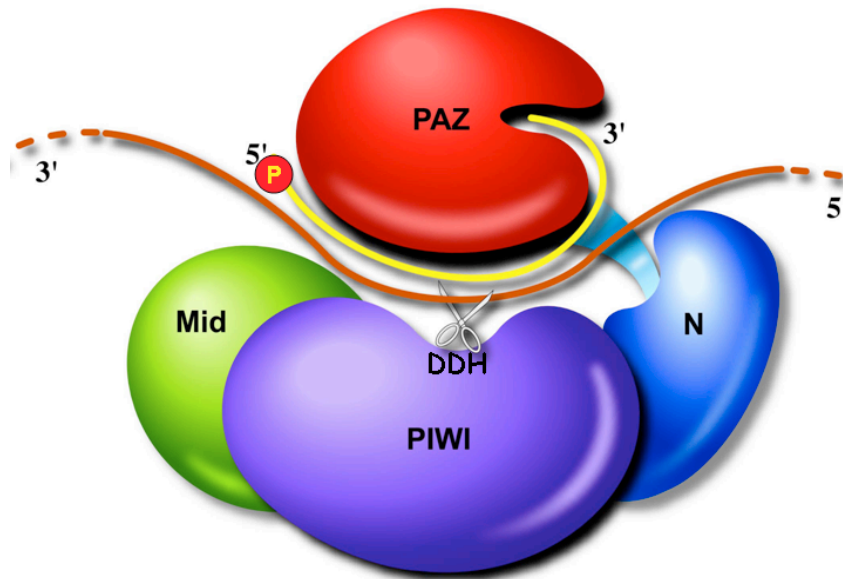
Figure 1.3: Structure of the RISC complex, showing the central channel with associated miRNA and target, and the AGO catalytic site

Taken from [116].
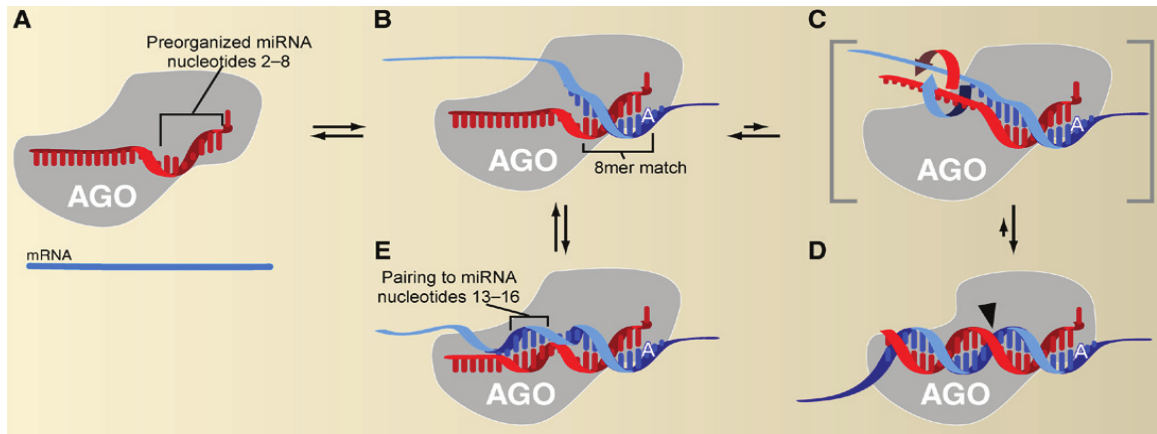
Figure 1.4: Overview of miRNA function

Taken from [2]. A) Initial seed region of miRNA is exposed to mRNA. B) Seed region pairs with target. C and D) If the miRNA and target are highly complimentary, binding continues down the channel in the AGO complex, holding the target close to the AGO catalytic site and allowing cleavage of the target. E) If there is not extensive complimentarity between the miRNA and its target the expression of the target will be repressed through translational repression.
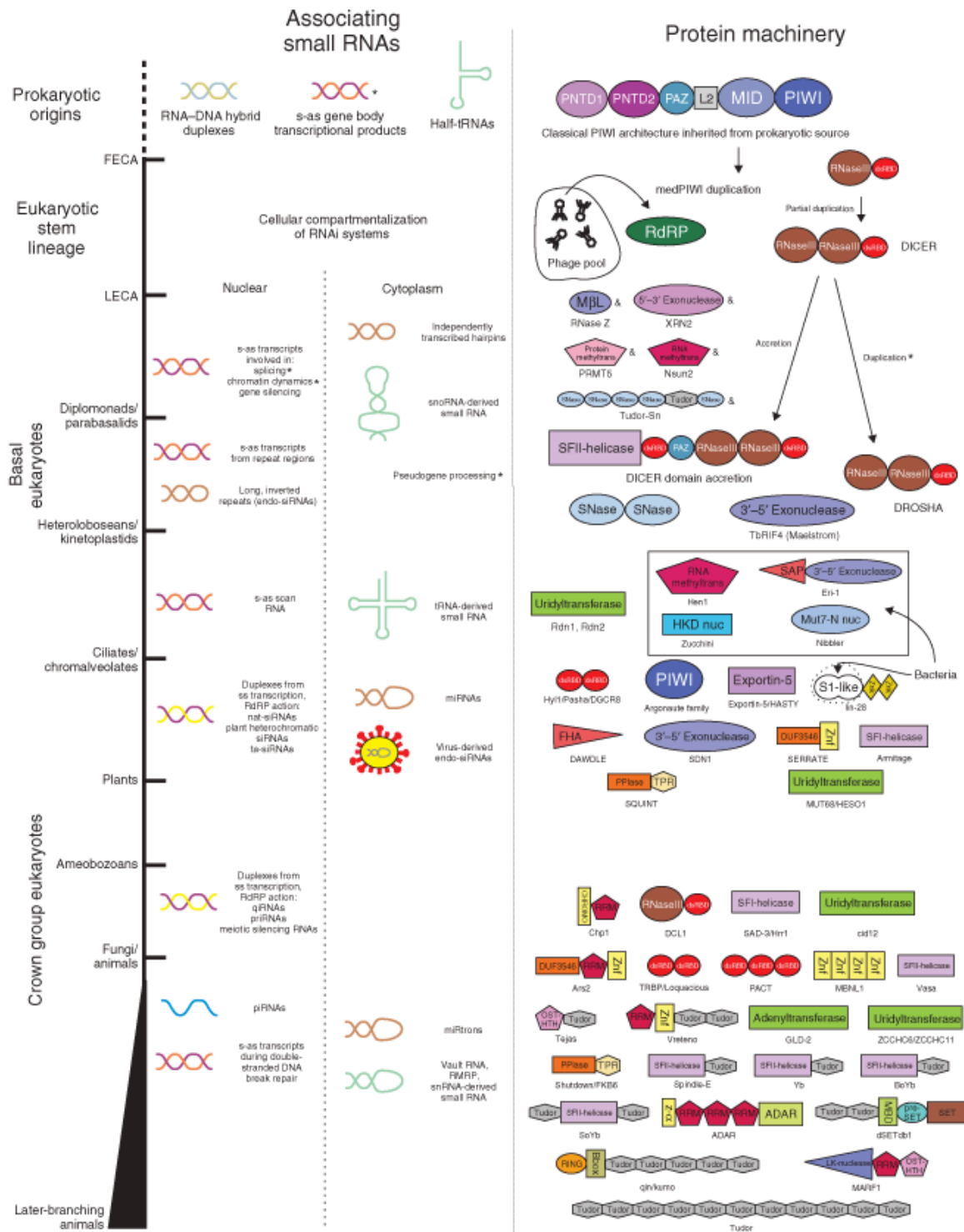
Figure 1.5: Overview of the evolution of the RNAi pathways and the proliferation of the RNAi machinery

Taken from [13]. The timeline for the development of different types of RNAi and the related machinery from the prokaryotic origins through later branching animals. Ampersands next to the protein structure indicates that the protein has originated early in the timeline, but the timing of when it became involved in an RNAi pathway is unclear. Asterisks next to the even name indicate uncertainty about the timing of the event labeled.
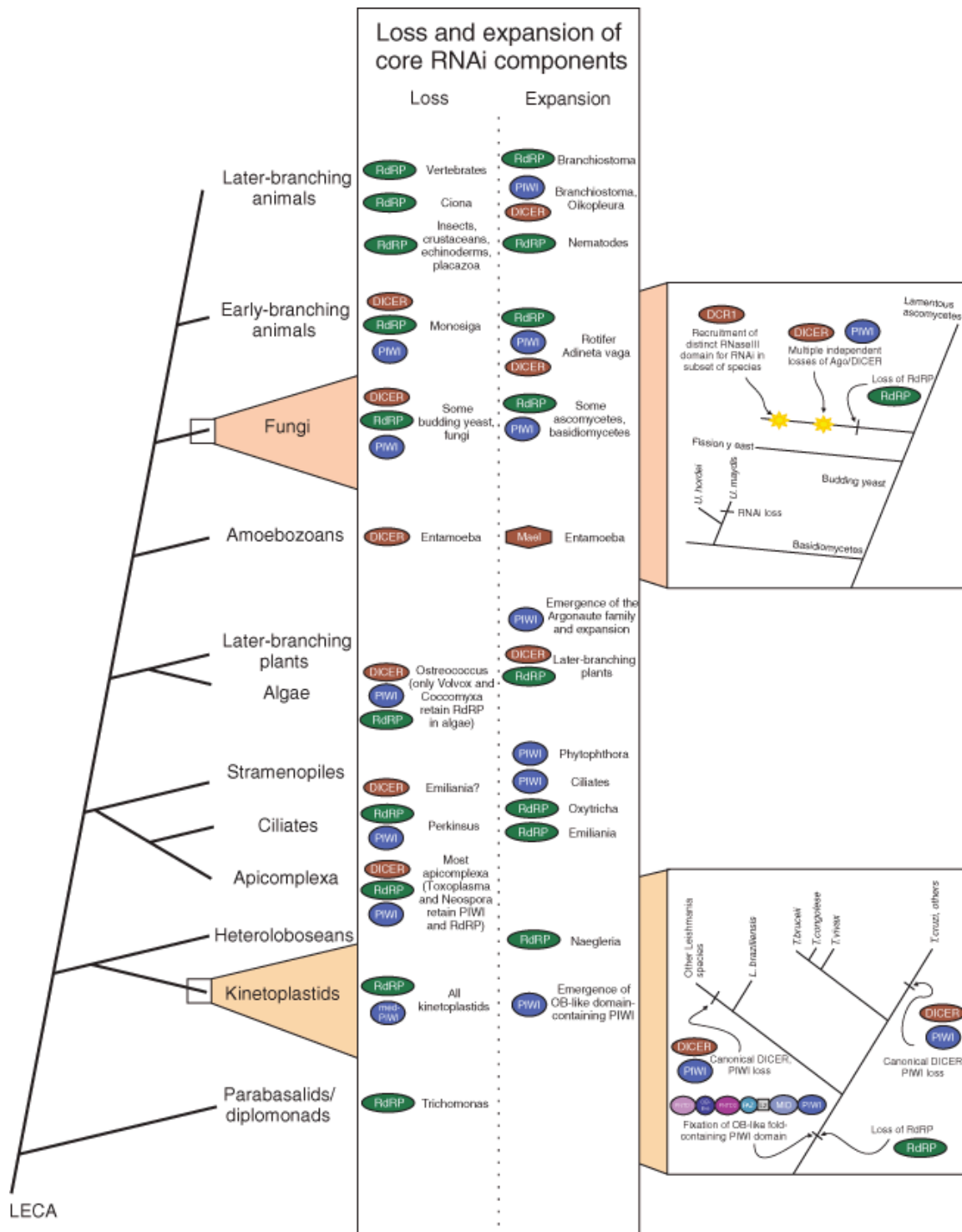
Figure 1.6: Time line for the loss and expansion of the core RNAi proteins across several lineages

Taken from [13]. Additional details of the loss events for kinetoplastid (bottom) and

fungi (top) are shown in boxes.
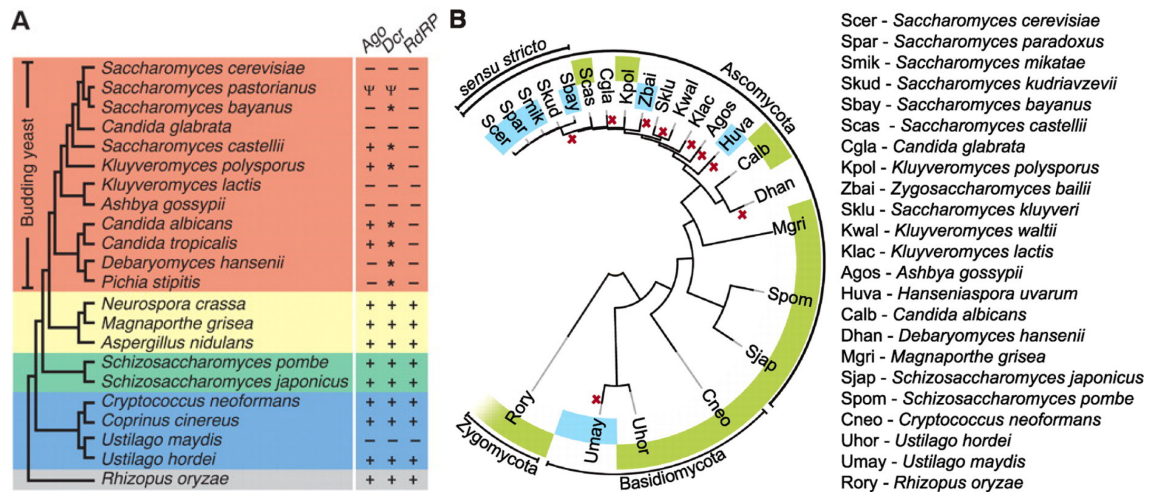
Figure 1.7: Phylogeny of the RNAi machinery and compatibility with the endogenous dsRNA virus Killer in budding yeasts

Parts taken from [39, 40] and merged. A) Phylogeny of three major fungal lineages showing Zygomycota (gray), Basidiomycota (blue), and Ascomycota, which is further divided into Taphrinomycotina (green), Pezizomycotina (yellow), and the buding yeasts Saccharomycotina (orange) [39]. The presence or absence of the core RNAi machinery within each species is indicated by (+) if present and function, (-) if absent, and (ψ) for pseudogenes. Species which contain additional RNaseIII containing genes are indicated with (*) (See Figure 1.6). B) Phylogeny of the same three fungal lineages showing the presence of the RNAi machinery (green highlight), and compatibility with the endogenously produced, cytoplasmically inherited dsRNA virus Killer (blue highlight) [40].
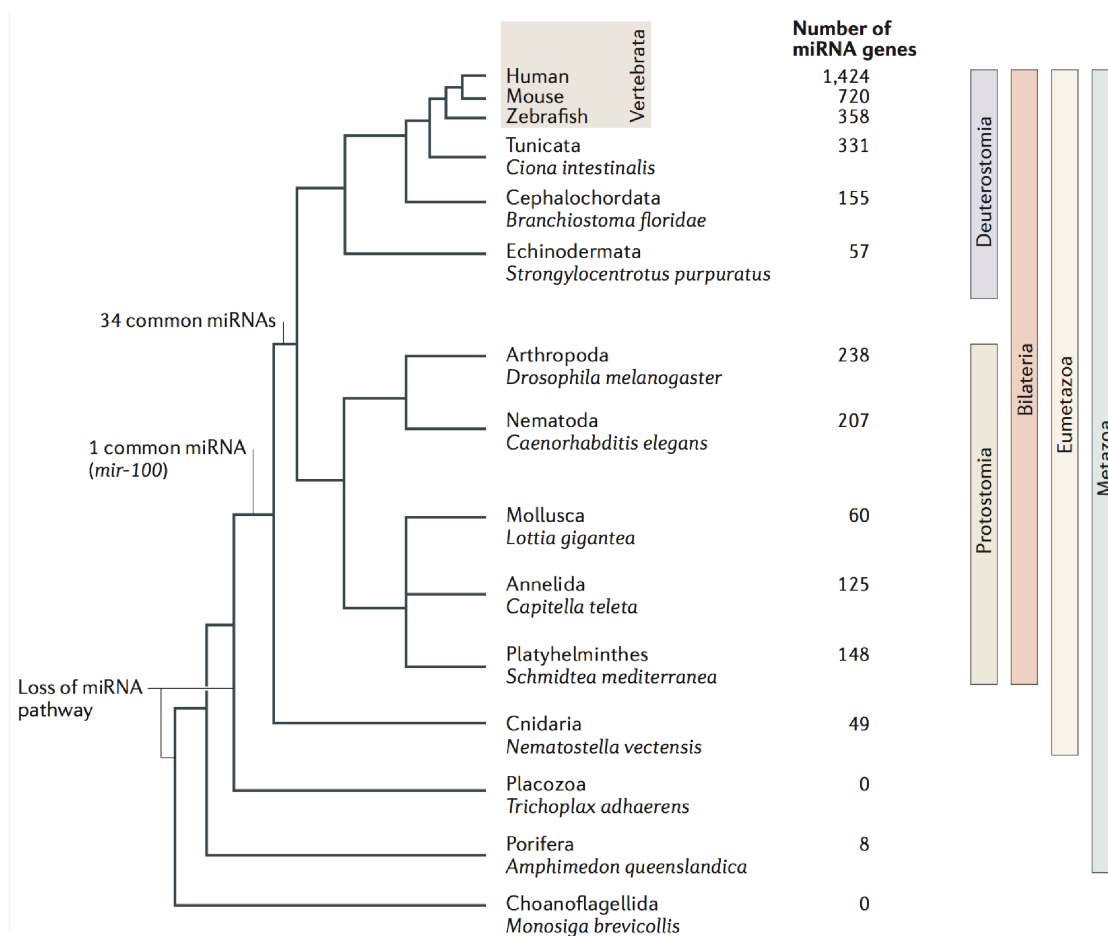
Figure 1.8: Distribution of miRNAs in metazoa

Taken from [41]. The phylogeny of a subset of metazoa is shown, along with the number of predicted miRNA genes in the genomes of each organism. Also shown are the predicted loss of miRNAs in the Placazoa *Trichoplax adherens*, which does maintain the core machinery for other types of RNAi, and the complete loss of RNAi in the Choanoflagellida *Monosiga brevicollis*.
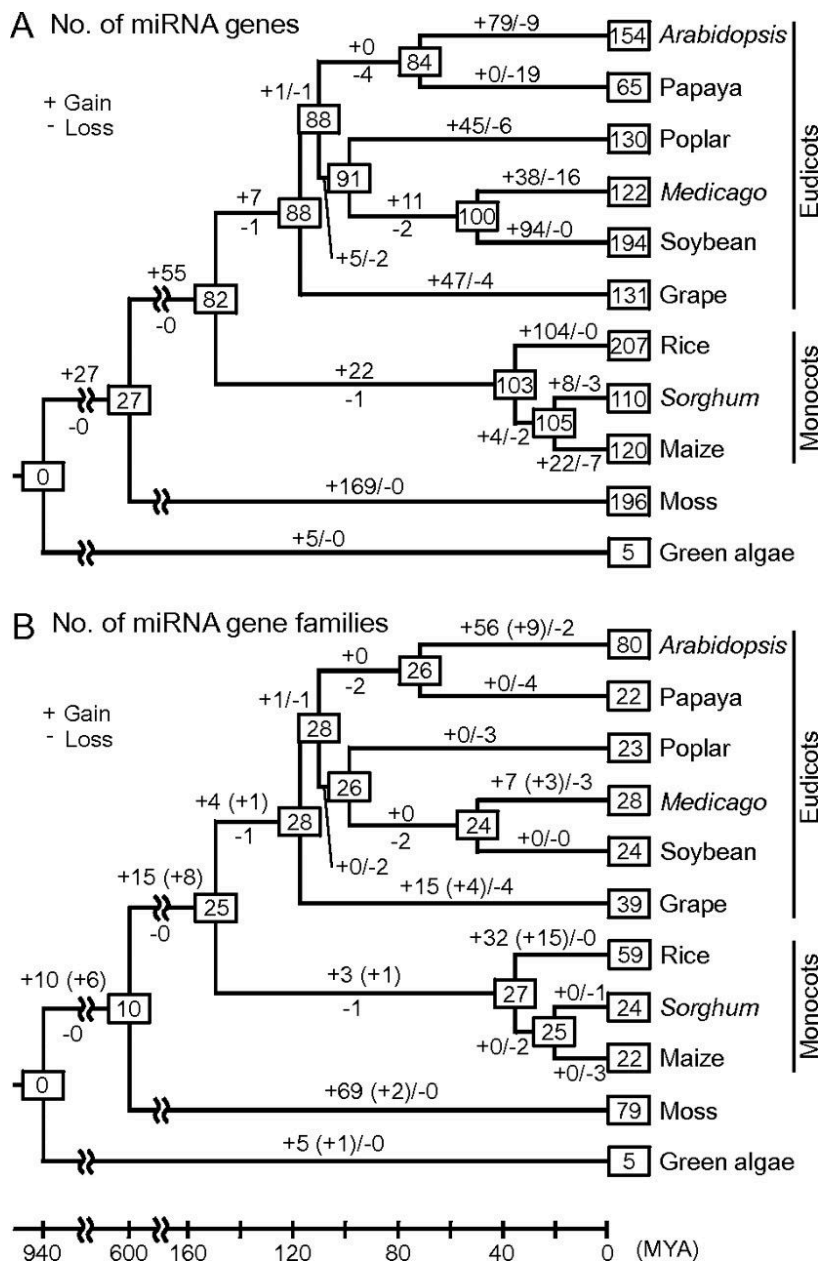
Figure 1.9: Number of miRNA genes and miRNA gene families in plants

Taken from [45]. A) The estimated number of distinct miRNA encoding genes within genomes of certain plants. B) The estimated number of miRNA gene families coding for similar or identical mature miRNAs. The number within the boxes at the branch points indicate the estimated number of miRNA genes or gene families contained within the

ancestral species that are shared between the two branches. The (+/-) indicate the lineage

or species specific gains or losses of those ancestral conserved miRNAs.
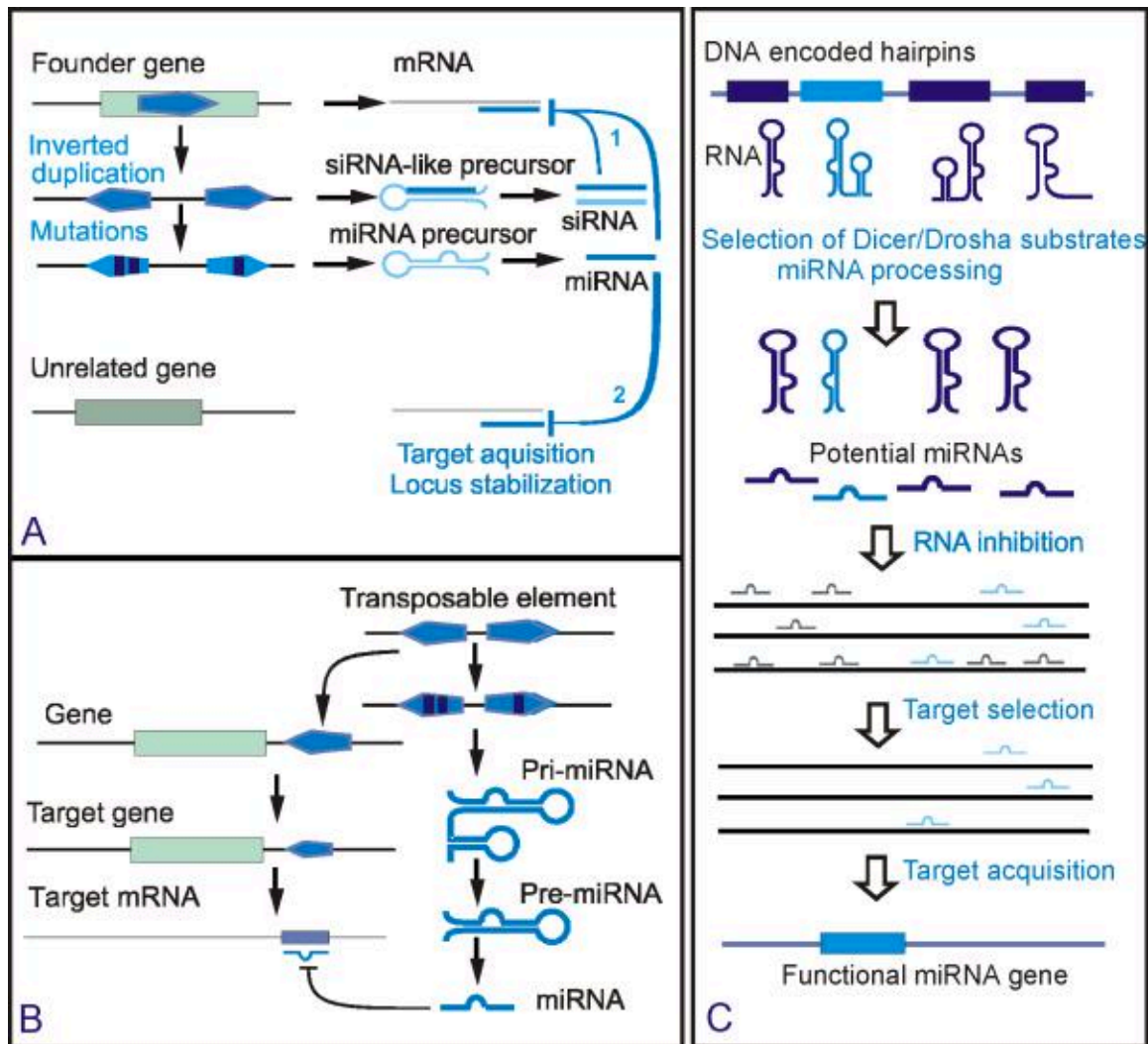
Figure 1.10: Models of evolution for novel miRNAs

Taken from [12].   A) The inverted duplication model in plants, where a segment of the target sequence is duplicated in tandem, with one duplication inverted, forming a hairpin structure when transcribed. B) The transposable element model. Similar to the inverted duplication model in A, when transposable elements insert into the genome, they can form inverted repeats, which can then form into hairpins processed by the RNAi machinery. C) Random selection model of evolution, where any transcribed regions that form hairpins may be processed by the RNAi machinery. Many of these randomly

processed miRNA may have no targets until random mutations render them sufficiently

complimentary to a target transcript.

Table 1.1: Summary of selected miRNA prediction algorithms

Method	Approach	Predominant lineage	Sequence selection	Major features	Ref
miRNAminer	Similarity-based heuristic	Mammalian	Results of BLAST search for query pre-miRNA	Folding energy, minimal base-pairing, hairpin-shape, position of mature miRNA, mismatches in hairpin within mature miRNA	[82]
miRTRAP	sRNA sequencing based heuristics	*C. intestinalis*	Contiguous read sequence	Size of contiguous read region, read distribution, hairpin structure	[76]
MiRFinder	*ab initio* SVM	*D. melanogaster*	Genomic search for short hairprins	Folding energy, pairing of hypothetical miRNA, pairing of hairpin branches	[85]
miRPara	*ab initio* SVM	Plant or animal	Sliding window of genomic sequence	Sequence composition, size of hairpin, size of loop, percent unpaired	[75]
MiPred	*ab initio* Random Forest	Human	User provided	Probability of MFE for sequence, MFE, nucleotide and structural features	[83]
MapMi	Similarity-based heuristic	Metazoa	Extended genomic sequence around mapped reads	Hairpin structure, mismatches in hairpin, MFE	[87]
ProMiR II	Similarity-based or *ab initio* machine learning	Metazoa	Sliding window of genomic sequence	Proximity to known miRNAs, conservation of pre-miRNA, MFE, G/C ratio, entropy	[88]
NOVOMIR	*ab initio* HMM	Plant	Sliding window of genomic sequence	Size and shape of hairpin, MFE, nucleotide and structural features	[80]
MiRenSVM	*ab initio* ensemble SVM	Metazoa	User provided	Hairpin structure (allows multiple loops), MFE, nucleotide and structural features	[93]
miRDeep	sRNA	Plant or animal	Contiguous read	Read distribution,	[74,

	sequencing based heuristics		sequence	hairpin structure	81, 117]
C-mii	Similarity-based heuristic	Plant	User provided	Conservation of pre-miRNA, hairpin structure	[84]
RNAmicro	Similarity-based SVM	Metazoa	Genomic alignment	Sequence conservation, hairpin structure, G/C ratio, MFE, thermodynamic stability	[90]

CHAPTER 2- TRANSCRIPTOME ANALYSIS OF TRIACYLGLYCEROL BIOSYNTHESIS PATHWAYS IN NITROGEN-DEPRIVED PHOTOAUTOTROPHICALLY GROWN *CHLAMYDOMONAS REINHARDTII*

Adam Voshall[1*], Joseph Msanne[1*], Insun Kook[1], Tala Awada[1], Etsuko N. Moriyama[1,2] and Heriberto Cerutti[1,2]

[1]School of Biological Sciences, University of Nebraska-Lincoln [2]School of Biological

[*]These authors contributed equally to this work

Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln

Abstract

While the existence of small RNAs in *C. reinhardtii* has been established for several years, little is known about how they target transcripts for regulation or what role(s) they play in cellular processes. To define functional microRNAs (miRNAs) in Chlamydomonas, we characterized small RNAs associated with an argonaute protein, AGO3, by affinity purification and deep sequencing. Using a stringent set of criteria for canonical miRNA annotation, we identified 39 suitable precursor miRNAs, which produce 46 unique, AGO3-associated miRNA sequences including 11 previously reported *C. reinhardtii* miRNAs and 35 novel ones. We also attempted to identify miRNA target transcripts, based on the complementarity of predicted miRNAs with the

respective binding sites. Recent results with reporter constructs have indicated that Chlamydomonas miRNAs may regulate target genes through either transcript cleavage or translation repression. Thus, potential targets were divided into two categories depending on the extent of complementarity to a given miRNA, those likely to be regulated through cleavage and those likely to be regulated through translational repression. The search for cleavage targets identified 75 transcripts. However, only 6 of them showed at least a 2-fold up-regulation of mRNA levels in a mutant strain almost devoid of miRNAs. The search for translational repression targets, which used complementarity criteria more stringent than those experimentally required for a reduction in target protein levels, identified 493 transcripts. Additionally, we experimentally validated endogenous targets regulated through cleavage and translational repression *in vivo*. Our results emphasize the difficulty of identifying genuine miRNAs and miRNA targets in *Chlamydomonas reinhardtii* and suggest that at least some miRNAs might regulate endogenous genes primarily through translational repression.

Introduction

Decreasing fossil fuel and its impact on global warming led to an increasing demand for its replacement by sustainable renewable biofuels. Microalgae may offer a potential feedstock for renewable biofuels. They have fast growth rate, permit the use of non-arable land and non-potable water, and do not affect the supply of food and other crop products [1-5]. The interest in microalgae as potential source for biofuel production is due to the high lipid contents in some species, which is of great importance for the food and energy industries [6].

The unicellular photosynthetic microalga *Chlamydomonas reinhardtii* is regarded as non-oleaginous [7]. While under optimal growth conditions, it synthesizes fatty acids principally for esterification into membrane lipids [8], under conditions limiting to growth, particularly nutrient deprivation, it slows down the cell proliferation and alters the lipid biosynthetic pathways towards the formation and accumulation of the neutral lipid triacylglycerols (TAGs) [8]. TAGs are one of the most energy-rich forms of reduced carbon available from nature [9]. They can be converted to biodiesel by trans-esterification with methanol in the presence of an acid or alkali catalyst [6, 9, 10]. Nutrient deprivation also stimulates a considerable increase in the biosynthesis of starch granules [5, 8, 11-14] accompanied by a reduction in chlorophyll levels [15-17]. Since their total lipid accumulation is only 20% in wild-type strains and 40% for starchless mutants [18-20], *C. reinhardtii* is unlikely to be chosen for biofuel production. However, it is a model alga whose complete genome is available and used as a reference to understand metabolic and regulatory networks involved in lipid metabolism and TAG accumulation.

For economic viability and sustainability of algal biofuels, it is essential to understand the basic biology of these microalgae, particularly the effects of environmental stresses on cellular metabolisms and the regulation of biosynthetic pathways of fatty acids and TAGs, as well as carbon fixation and allocation [20]. Neutral lipid biosynthesis and turnover have been studied in higher eukaryotes particularly yeast, mice, and the model plant *Arabidopsis thaliana* where extensive knowledge exists on lipid metabolism and several key enzymes identified [21-26]. However, these pathways are not fully documented in microalgae and little is known about the enzymes involved in the formation, accumulation, or degradation of TAGs. More research is needed to understand signal perception and transduction under stress and different molecular mechanisms leading to TAG and starch accumulations. In this context, rapidly developing systems-level "omics" analyses (transcriptomics, proteomics, metabolomics), which are both sensitive and quantitative [27], and the availability of annotated *C. reinhardtii* genome (http://genome.jgi-psf.org/chlamy/chlamy.home.html) allow us to investigate changes in response to nutrient deprivation. It may provide useful hypotheses for microalgal strain improvement to optimize their biofuel production.

Since nitrogen deprivation is most effective in inducing TAG accumulation, using RNAseq technology, we analyzed gene expression changes occurring in *C. reinhardtii* in response to N-depletion, under strictly photoautotrophic conditions, over an extended treatment up to 144 hours. Other groups have been using high-throughput approaches to analyze the transcriptome or the proteome of *C. reinhardtii* in response to environmental stress conditions [20, 27-31]. However, all their analyses were performed under photoheterotrophic growth conditions. The objectives of this study are to analyze the

transcript levels from *C. reinhardtii* cultures growing in both N-replete and N-depleted media, identify relevant enzyme-encoding genes, and reconstruct the metabolic pathways involved in the biosynthesis and degradation of precursor molecules that may have potential for biofuel production.

Results

Global expression profiles and reads assembly

Chlamydomonas reinhardtii cells grown photoautotrophically under both N-replete and N-deplete conditions were harvested from cultures at various time points during the course of N-depletion corresponding to 24, 48, and 144 hours, as well as from cells growing in N-replete medium, in order to isolate total RNA for transcriptomic analyses. Assembled sequences were subjected to BLAST similarity searches and the components of the different metabolic pathways were identified based on orthology relationships to known lipid and starch biosynthesis and catabolism pathways from land plants, mainly *Arabidopsis thaliana*, and fungi. Stringent analysis of the transcriptomic data identified 907 genes that showed at least 3-fold differential transcript change with a q-value ≤ 0.05 in one or more time points after N-depletion compared to the N-replete condition (Figure 2.1a). Among them, 40 genes could be assigned into the functional groups described in this study (Figure 2.1b). Genes with known putative functions were assigned to different metabolic pathways associated with biosynthesis and catabolism of TAG, starch, and proteins. These genes are listed in Table 2.1. Previous analysis showed biochemical changes in these metabolic pathways during N-depletion, with an overall decrease in the

total and soluble protein content, and a drastic increase in the accumulation of starch and

TAGs under photoautotrophic conditions [17].

Fatty acid and triacylglycerol metabolism pathways

Nutrient starvation in *C. reinhardtii*, particularly N-depletion, led to the production and

accumulation of high amounts of TAGs accompanied by a breakdown of proteins,

ribosomes, and cell membranes [17, 31]. Recycled compounds might be used for *de novo*

TAG biosynthesis [25]. These changes in metabolic pathways of N-depleted

photoautotrophically-grown *Chlamydomonas* cells were highlighted by our

transcriptomic analysis, which allowed determining the transcript abundance of the genes

and gene families encoding for key enzymes involved in fatty acid and TAG metabolic

pathways. These pathways with major key enzymes are represented in Figure 2.2.

In microalgae as in plants, fatty acids (FAs), the building blocks of various types

of lipids, are synthesized in the chloroplast by the fatty acid synthase II (FAS II)

complex, which represents the major pathway of plant FA synthesis [32], leading to the

production of palmitic acid (C16:0) and stearic acid (C18:0), the precursors for the

synthesis of cellular membranes and neutral lipids [33]. Subcellular localization analyses

of FAS II enzymes using PredAlgo [34] show that these are mainly targeted to the

chloroplast (Table 2.1). In photoautotrophically-grown *Chlamydomonas* cells, a general

trend toward downregulation of the genes encoding enzymes of the chloroplastic FAS II

complex was observed under N-depletion conditions. At least 3-fold decrease in

transcript levels was observed for two putative pyruvate dehydrogenase complex (PDH)-

encoding genes (au5.g11028_t1 and au5.g1301_t1, Figure 2.2 and Table 2.1). PDH

catalyzes the conversion of pyruvate into acetyl-CoA by a process called pyruvate

decarboxylation. Acetyl-CoA is then converted to malonyl-CoA through an irreversible

reaction catalyzed by acetyl-CoA carboxylase (ACCase). ACCase is the key enzyme in

FA biosynthesis; we noted that two putative ACCase-encoding genes were significantly

downregulated (au5.g1722_t1 and au5.g1722_t2, Figure 2.2 and Table 2.1) in N-depleted

Chlamydomonas cells. ACP-S-malonyl transferase (MAT) catalyzes the transfer of

malonyl-CoA to malonyl-ACP, the carbon donor for FA elongation reactions. Isoforms

of the condensing enzyme β-ketoacyl-ACP synthase (KAS) carry out all elongation

reactions using malonyl-ACP. In this study, we found a putative *KAS I* (au5.g9840_t1,

Figure 2.2 and Table 2.1) that was significantly downregulated. After each condensation

reaction, a sequence of reduction, dehydration, and another reduction carried out

respectively by three enzymes, β-ketoacyl-ACP reductase (KAR), β-hydroxyacyl-ACP

dehydrase (HD), and enoyl-ACP reductase (EAR), leads to the production of C16:0 and

C18:0 fatty acids. Two putative KAR-encoding genes (au5.g10590_t1 and au5.g2990_t1,

Figure 2.2 and Table 2.1) and one putative EAR-encoding gene (au5.g13296_t1, Figure

2.2 and Table 2.1) were significantly downregulated in this study. We also found that a

putative Δ9 acyl-ACP desaturase (AAD)-encoding gene (au5.g7079_t1, Figure 2.2 and

Table 2.1), involved in the production of unsaturated FAs, was significantly

downregulated, while putative fatty acyl-ACP thioesterase A and B (*FATA* and *FATB*)

that cleave the acyl chain from acyl carrier protein (ACP) and release the free FA

remained unchanged. In *C. reinhardtii*, a distinct pathway uses acyltransferases in the

chloroplast to allow the transfer of FAs directly from ACP to glycerol-3-phosphate (G3P)

producing lysophosphatidic acid (LPA) and phosphatidic acid (PA) [25, 35, 36]. PA is

then dephosphorylated into DAG, which is converted to TAG by the transfer of a third

acyl group. In this analysis, we found that major putative chloroplastic genes involved in

TAG assembly such as glycerol 3-phosphate acyltransferase (*GPAT*), lysophosphatidyl

acyltransferase (*LPAT*), and phosphatidic acid phosphatase (*PAP*) remained unchanged

while a putative diacylglycerol acyltransferase-like gene (*DGAT-like*, au5.g14782_t1,

Figure 2.2 and Table 2.1) showed significant upregulation in response to N-depletion.

Chlamydomonas DGAT-like shows high similarity to the phytyl ester synthase 1 and 2

(*PES1* and *PES2*) genes of the esterase/lipase/thioesterase family of acyltransferases in *A.*

thaliana, which have diacylglycerol acyltransferase and phytyl ester synthesis activities

in the chloroplast [37]. *PES1* and *PES2* are highly upregulated in *A. thaliana* during

senescence and N-depletion [37].

On the other hand, TAG assembly and accumulation in the cytosol occur through

the Kennedy pathway using acyl-CoA and glycerol-3-phosphate (G3P) as precursors. The

acyl-CoA is generated in the cytosol by esterification of free FAs to coenzyme A carried

out by acyl-CoA synthetase (ACS). Free FAs are imported from the chloroplast. G3P is

produced by the action of several enzymes namely the glycerol 3-phosphate

dehydrogenase (GPDH) converting the cytosolic dihydroxyacetone phosphate (DHAP) to

G3P and by the action of glycerol kinase (GK) on free glycerol, both DHAP and glycerol

are imported from the chloroplast. We found here significant upregulation in transcripts

encoding two putative cytosolic *GPDH* genes (au5.g2028_t1 and au5.g2031_t1, Figure

2.2 and Table 2.1) with N-depletion. A strong increase in transcript levels was also

observed for a few *DGAT* genes encoding the final step of TAG synthesis by the transfer

of a third fatty acid to the vacant position 3 of DAG. By similarity searches, we identified

six genes encoding for putative type-2 *DGATs* in the transcriptome of photoautotrophic *Chlamydomonas* cells. One of these genes (*DGTT1*, au5.g4218_t1, Figure 2.2 and Table 2.1) was highly responsive to N-depletion showing a significant increase in transcript abundance. This was consistent with findings from other groups that analyzed *Chlamydomonas* response to N-depletion under photoheterotrophic conditions (Miller *et al.,* 2010; Boyle *et al.,* 2012). Transcriptome analysis also allowed identification of other genes involved in the TAG biosynthesis pathways that significantly increased with N-depletion, particularly the putative major lipid droplet protein-encoding gene (*MLDP*, au5.g15585_t1, Figure 2.2 and Table 2.1) and a lysophospholipid acyltransferase (*LPLAT*, au5.g13569_t1, Figure 2.2 and Table 2.1). *MLDP* is unique to green algae and when suppressed by RNA interference, Moellering and Benning [31] observed an increase in lipid droplet size but not in TAG content. LPLAT functions by converting lysophopholipids (LPL) to phospholipids (PL). Overall, a general trend toward upregulation of the genes encoding enzymes involved in TAG assembly and accumulation was observed under N-depletion conditions. Gene expression levels presented as Reads per Kilobase per Million (RPKM) and log_2(Fold Change) of the *GPDH*, *LPLAT*, *DGTT1*, and *MLDP* transcripts showing significant upregulation with N-depletion are represented in Figure 2.3.

Additionally, we analyzed the transcript abundance of genes encoding for enzymes involved in the catabolism of TAGs and FAs recycling by β-oxidation (Figure 2.2). We identified several candidate lipases involved in the breakdown of TAGs by hydrolysis of ester bonds that link fatty acyl chains to the glycerol backbone. We found two transcripts encoding for putative TAG/DAG lipases (TAGL, au5.g7029_t1 and

au5.g12306_t1, Figure 2.2 and Table 2.1) that showed significant downregulation in response to N-depletion. Surprisingly, other candidate TAGLs showed an increase in transcripts abundance. These lipases might be involved in membrane remodeling and lipid recycling producing additional FAs that may be converted to TAGs [30]. In *Chlamydomonas*, β-oxidation pathway involves the enzymes acyl-CoA oxidase (AOX), multi-functional protein (dienoyl-CoA reductase/enoyl-CoA dehydrogenase) (MFP), and ketoacyl-CoA thiolase (KAT). These remove two carbons from the acyl chain during each cycle and release an acetyl-CoA that is recycled in the chloroplast or mitochondria, transcripts encoding these enzymes remained mostly unchanged.

qPCR, RNAi, and semi-quantitative RT-PCR

To confirm the differential gene expression observed with RNAseq data, we analyzed the abundance of selected transcripts by quantitative real-time PCR. The steady-state mRNA abundance of glycerol 3-phosphate dehydrogenase (*GPDH*, au5.g2028_t1), lysophospholipid acyltransferase (*LPLAT*, au5.g13569_t1), major lipid droplet protein (*MLDP*, au5.g15585_t1), and diacylglycerol acyltransferase (*DGTT1*, au5.g4218_t1) was determined by qPCR in cells grown in N-replete medium and in cells depleted of N for 24, 48, and 144 hours as shown in Figure 2.4. These results confirm the marked difference in relative transcript abundance of these genes between N-replete and N-depleted cells. Moreover, given the considerable increase of *DGTT1* mRNA content during N-depletion (Figure 2.3 and 2.4), it can be concluded that DGTT1 significantly contribute to the TAG production and accumulation in N-depleted *Chlamydomonas* cells. We tested this hypothesis by RNA interference used to knockdown the expression of

DGTT1 in the wild-type strain CC-124. TAG accumulation in one RNAi strain was reduced by 20% (Figure 2.5) when cells were subjected to 48 hours N-depletion.

RNAseq was performed independently twice. A few genes involved in TAG metabolism pathways encoding for acyl carrier protein (*ACP*), lysophosphatidyl acyltransferase (*LPAT*), phosphatidic acid phosphatase (*PAP*), betaine lipid synthase (*BTA*), acyl-CoA synthase (*ACS*), acyl-CoA oxidase (*AO*), and enoyl-CoA reductase (*EH*) showed a marked difference with the mRNA level at the control time points between the two experiments. We analyzed the expression of these genes by semi-quantitative RT-PCR using cells grown under N-replete (control) and N-depleted conditions to confirm the biological non-technical aspects of these differences. Some of these genes such as *ACP*, *LPAT*, and *AO* show clear difference in the expression patterns knowing that the same conditions were used for the two experiments (Figure A.1).

Starch metabolism pathways

In parallel with TAG accumulation, N-deprivation also increased the starch content in *Chlamydomonas* cells grown in photoautotrophic medium. This observation was confirmed previously by enzymatic determination of starch contents that reached a peak after 72 hours N-depletion, and then decreased after 144 hours [17]. Based on the transcriptome analysis, we identified genes and gene families encoding for enzymes involved in the biosynthesis and degradation of starch in N-depleted photoautotrophically-grown *Chlamydomonas* cells. These metabolic pathways are represented in Figure 2.6.

Starch accumulating in *Chlamydomonas* cells is composed of two types of molecules, amylose and amylopectin. Amylopectin forms the major fraction and accounts for 65-90% of the total starch granule [38, 39]. The pathway for starch biosynthesis involves the enzyme fructose bisphosphate aldolase (FBA) that catalyzes a reversible reaction generating fructose 1,6-phosphate from dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP). Fructose 1,6-bisphosphatase (FBP) converts fructose 1,6-bisphosphate to fructose 6-phosphate (F6P) involved in gluconeogenesis and Calvin cycle. F6P is converted to glucose 6-phosphate (G6P) through a reversible reaction catalyzed by the enzyme phosphoglucose isomerase (PGI). Phosphoglucomutase (PGM) generates glucose 1-phosphate (G1P) from G6P. ADP-glucose pyrophosphorylase (AGPase) uses G1P and ATP to generate ADP-glucose and inorganic pyrophosphate. AGPase catalyzes the rate-limiting step in the biosynthesis of starch. When starch biosynthesis was blocked by mutation of one AGPase subunit in *Chlamydomonas*, TAG content increased 30-fold compared to wild-type cells after 48 hours N-depletion in the presence of acetate [14, 40]. Soluble starch synthase (SSS) uses ADP-glucose to generate amylose, an essentially unbranched α-1,4-linked polyglucan, and starch branching enzyme (SBE) introduces 1,6-glycosidic bonds to the elongated α-1,4-linked polyglucan generating starch. We identified several candidate genes in the transcriptome data involved in the biosynthesis of starch. Despite the significant increase in starch content, we found that the transcripts abundance remained unchanged in response to N-depletion. Moreover, two transcripts encoding for putative *SSS* (au5.g10852_t1 and au5.g7473_t1, Figure 2.6 and Table 2.1) showed significant downregulation.

In photosynthetic eukaryotic cells, the degradation of starch occurs mainly in the chloroplast *via* α-amylases (AMYA), debranching enzymes (DBEs), glucan water dikinase (GWD), β-amylases (AMYB), and results in the formation of two neutral sugars, glucose and maltose. α-amylases (AMYA) catalyze the hydrolysis of starch to large-branched glucans. We identified three transcripts that encode for putative α-amylases in *Chlamydomonas*; one of them (au5.g15170_t1, Figure 2.6 and Table 2.1) was significantly downregulated in this study. DBEs, also called isoamylases (ISAs), catalyze the hydrolysis of 1,6-glycosidic bonds in starch yielding long linear glucan chains. Starch phosphorylase (PHOB) catalyzes the reaction that generates G1P. On the other hand, GWD is a glucan-phosphorylating enzyme that stimulates breakdown of starch granules by β-amylases. AMYB cleave the non-reducing end of α-1,4-glucan chains yielding maltose. We also identified three transcripts that encode for putative β-amylases, one of them (au5.g12805_t1, Figure 2.6 and Table 2.1) was significantly downregulated. Transcript encoding for putative α-glucosidase (GAA) enzyme (au5.g11038_t1, Figure 2.6 and Table 2.1) was significantly upregulated in this analysis. GAA hydrolyses maltose and maltodextrin to α-D-glucose, which may be degraded in the mitochondria to produce energy. With exception to amylases and α-glucosidase, transcripts abundance for putative genes encoding enzymes involved in starch degradation remained unchanged in response to N-depletion (Figure 2.6).

Proteolysis, amino acid degradation, and carbon metabolism genes

The total and soluble proteins decreased substantially in N-depleted *Chlamydomonas* cells [17]. Under stress conditions, proteins with chaperone activity are responsible for

recognizing the protein substrate in its unfolded/unstable form, and introduce it into the proteolytic chamber for degradation [41]. Most of the cell compartments have proteases involved in protein biogenesis and degradation, starting from maturation of precursors, to activation of signaling components, and finally to degradation of damaged proteins [41]. We identified in our transcriptome analysis few transcripts encoding for putative enzymes involved in different proteolytic metabolic pathways that were differentially regulated under N-depletion in photoautotrophically-grown *C. reinhardtii* cells (Table 2.1). The transcripts encoding for putative GMP synthetase (GUA, au5.g13735_t1, Figure 2.7 and Table 2.1), metalloproteases like deiminase/IAA-amino acid hydrolase/aminoacylase (M20, au5.g4278_t1, au5.g4678_t1, au5.g12930_t1, Figure 2.7 and Table 2.1), and serine carboxypeptidase (S10, au5.g9131_t1, Figure 2.7 and Table 2.1) were significantly upregulated.

Amino acids are broken down into metabolites that can be either oxidized into CO_2 and H_2O or used for gluconeogenesis. Amino acids may be glucogenic, ketogenic or both,these are degraded to TCA intermediates like pyruvate, α-ketoglutarate, succinyl-CoA, fumarate, oxaloacetate, acetyl-CoA, or acetoacetate. Our transcriptome analysis showed no change in the expression levels of the transcripts encoding for the enzymes involved in the TCA cycle except for one putative succinate dehydrogenase (SDHA, au5.g5444_t1, Figure 2.7 and Table 2.1) that catalyzes the conversion of succinyl-CoA to fumarate and was significantly downregulated. Glucogenic amino acids like alanine, cysteine, glycine, serine, and threonine are broken down to yield pyruvate. We identified one transcript in *Chlamydomonas* encoding for putative alanine-glyoxylate transaminase (AGT, au5.g9130_t1, Figure 2.7 and Table 2.1) that was significantly upregulated in this

metabolic pathway (Figure 2.7). Pyruvate can be converted into oxaloacetate or acetyl-CoA to enter the TCA cycle. Asparagine and aspartate are degraded to oxaloacetate. Our data showed upregulation of one transcript encoding for putative enzyme aspartate amino transferase (AAT, au5.g12556_t1, Figure 2.7 and Table 2.1) under N-depletion conditions. Arginine, glutamate, glutamine, histidine, and proline are all degraded by conversion to glutamate, which in turn is oxidized to α-ketoglutarate by glutamate dehydrogenase. In our analysis, we found one transcript encoding for putative agmatine deiminase (ADI, au5.g1160_t1, Figure 2.7 and Table 2.1) that was significantly upregulated. Conversion of certain amino acids such as arginine through the agmatine deiminase (ADI) pathway, generates ATP, which may help *Chlamydomonas* cells survive the N-depletion stress, mostly common in some bacterial systems [42]. The arginase enzyme, which is common in eukaryotes, converts arginine to ornithine and urea. This enzyme is absent in *Chlamydomonas*, suggesting that urea cycle is not present, and this is in contrast to plants and the diatom *Thalassiosira* [43]. Our analysis showed upregulation of two putative enzymes argininosuccinate synthase (AS, au5.g15792_t1, Figure 2.7 and Table 2.1) and argininosuccinate lyase (ASL, au5.g1397_t1, Figure 2.7 and Table 2.1). In mammalian cells, argininosuccinate synthase catalyzes the condensation of citrulline and aspartate to form argininosuccinate, the immediate precursor of arginine. Argininosuccinate lyase then splits argininosuccinate to release fumarate and arginine [44]. In *Chlamydomonas*, glyoxylate cycle can generate one molecule of succinate as the net product from two molecules of acetate. In this study, our transcriptome analysis showed two putative enzymes, isocitrate lyase (ICL, au5.g13056_t1, Figure 2.7 and Table 2.1) and malate synthase (MLS, au5.g2123_t1, Figure 2.7 and Table 2.1) that were

significantly downregulated in N-depleted photoautotrophically-grown *Chlamydomonas*

cells. Transcript abundance for the putative enzyme, phosphoenolpyruvate carboxykinase

(PCK, au5.g9626_t1, Figure 2.7 and Table 2.1), which catalyzes the committed reaction

of gluconeogenesis, decreased more than 25-fold compared to the N-replete cells under

photoautotrophic conditions. Similar studies by Miller *et al*. [30] showed that changes in

transcript abundance of genes encoding enzymes of primary metabolism such as

gluconeogenesis, glyoxylate cycle, and the photosynthetic carbon fixation cycle were

markedly decreased following N-deprivation under photoheterotrophic conditions.

Discussion

In this study, using RNAseq technology, we analyzed the transcriptome profiles

of nitrogen-deprived *C. reinhardtii* cells, and provided genome-wide insights into the

changes in metabolic pathways that led to the accumulation of significant amounts of

TAGs. Neutral lipid analysis in the model alga *C. reinhardtii* has been motivated by the

interest in green microalgae as a potential feedstock for the production of renewable

biofuels [45]. Comparison of transcriptomic changes of strain CC-125 grown

photoautotrophically in N-replete or N-depleted media, by parallel sequencing of cDNA

libraries, identified the changes in gene expression and cellular metabolism under N-

depletion conditions. In order to gain a better understanding of the processes controlling

the formation and accumulation of TAG in microalgae, we identified relevant enzyme-

encoding genes involved in the biosynthesis and degradation of TAG, starch, proteins,

and amino acids.

In *C. reinhardtii,* carbon metabolism pathways changed significantly after inoculation of the cells in N-depleted media; starch and TAG synthesis increased, while protein and ribosome synthesis decreased. Moreover, major cellular activities such as photosynthesis and cell division were reduced or completely inhibited. Energy-rich compounds such as TAG and starch are synthesized by competing metabolic pathways [46]. They share common precursor molecules (*e.g.* pyruvate, acetyl-CoA). Moreover, turnover of some compounds, such as proteins and ribosomes, might provide metabolites for biosynthesis of TAG and starch under nutrient deprivation conditions. In *A. thaliana,* lipid metabolism pathways account for over 100 enzymatic reactions and 600 proteins [47]. Despite major progress in recent years, fatty acids and TAG biosynthesis and degradation pathways are still not well understood in microalgae. In the present study, we found that the majority of transcripts encoding putative enzymes involved in the fatty acid biosynthetic pathway were downregulated or remain unchanged under N-depletion conditions over an extended period up to 144 hours. This was counterintuitive since *Chlamydomonas* cells were in a high neutral lipid accumulation state and acyl groups synthesized in the chloroplast by the FAS II complex form major components of the TAG molecules. Moreover, only a few transcripts involved in TAG biosynthetic machinery were significantly upregulated under the same conditions. However, on a genomic scale, transcripts expression level does not always correlate with protein abundance, which may explain the increase in TAG content under N-depletion.

Although TAG metabolism pathways were not extensively studied at the biochemical or molecular levels, it is generally believed that TAG accumulation in microalgae, under growth-limiting conditions, is due to DAG acylation by the

endoplasmic reticulum (ER)-resident acyltransferases, particularly the diacylglycerol

acyltransferases (DGATs), or by recycling of polar membrane lipids, which may be

catalyzed by phospholipid: diacylglycerol acyltransferase (PDAT) [5, 48]. In *A. thaliana*,

PDAT is critical for neutral lipid accumulation in seeds [49]. We found here a putative

type-2 *DGAT*, the *DGTT1* that was highly responsive to N-depletion at the transcript

level. *DGTT1* was almost undetectable under N-replete control conditions and showed

many fold increase in N-depleted cells (Figure 2.3 and 2.4). Therefore, DGTT1 might be

a key enzyme in TAG biosynthesis under photoautotrophic growth conditions. When we

used RNAi to downregulate the expression of *DGTT1*, we found a slight reduction in

TAG content. This slight decrease in TAG content may be due to the presence of other

acyltransferases in the *Chlamydomonas* genome that can counterbalance the loss of

DGTT1. While Boyle *et al.* [48] found that *PDAT* was significantly upregulated at the

transcript level in N-depleted *Chlamydomonas* cells grown under photoheterotrophic

conditions, our data showed that the transcript levels of putative *PDAT* (au5.g8928_t1)

remained unchanged throughout the experiment. Also in this study, glycerol 3-phosphate

dehydrogenase (GPDH), which catalyzes the formation of glycerol 3-phosphate (G3P)

and the lysophospholipid acyltransferase (LPLAT), which adds an acyl group to

lysophospholipids (LPL) to produce phospholipids (PL), showed significant increase in

expression with *GPDH* showing 10-fold increase in transcript level after 24 hours and

30-fold increase after 144 hours, while *LPLAT* showed 12-fold increase in transcript level

after 144 hours N-depletion (Figure 2.4). Genes encoding these enzymes may represent

potential targets for the biotechnological improvement of TAG accumulation in

microalgae since they seem to contribute extensively to TAG accumulation in

Chlamydomonas. We also found that the transcript encoding the structural protein named major lipid droplet protein (MLDP) was significantly upregulated in N-depleted *Chlamydomonas* cells. Moellering and Benning [31] used RNAi to downregulate the expression of *MLDP* and found that the lipid bodies increase in size with N-depletion but no change in TAG content in photoheterotrophically-grown *Chlamydomonas* cells. They suggested that MLDP work by preventing lipid body fusion [31].

Starch accumulates in photosynthesis-capable organisms through redirection of fixed carbon into storage [50]. In photosynthetic organisms, starch biosynthesis pathways share common precursor molecules with lipid biosynthesis pathways, and redirection of these precursors from the starch production by blocking the step catalyzed by AGPase, led to an increase in TAG accumulation [40]. N-deprivation under photoautotrophic conditions stimulated high levels of starch accumulation that occurred very rapidly upon incubation in N-depleted medium. Starch increased ~14-fold after 2 days incubation of *Chlamydomonas* cells in N-depleted medium compared to cells growing in N-replete medium [17]. Despite this considerable starch accumulation, we found no significant increase in transcripts encoding for enzymes involved in starch synthesis after transcriptomic analyses. In *C. reinhardtii*, starch accumulates as particles associated with the pyrenoid or as a pool of starch granules located in the chloroplast stroma [38]. Therefore, diverse metabolic pathways control starch biosynthesis and degradation [39]. One starch catabolic pathway involves the β-amylases, which hydrolyze starch into maltose that is enzymatically converted by α-glucosidase (GAA) to D-glucose, which may be used to produce the central metabolite pyruvate during glycolysis. Pyruvate may then be converted into acetyl-CoA, the precursor for fatty acids biosynthesis. While

transcripts encoding for β-amylases were downregulated, transcript encoding for GAA (au5.g11038_t1) was the only one in starch metabolism that showed significant upregulation in N-depleted *Chlamydomonas* cells. GAA showed many fold increase in transcript level at later time points, after 144 hours incubation in N-depleted media, when starch decreased and TAG continued accumulating. GAA, involved in starch degradation through the β-amylases pathways, may play a major role in starch recycling and providing the metabolites for the production of TAGs.

The protein content of N-depleted *Chlamydomonas* culture decreased steadily over time. Our transcriptomic data showed that cells subjected to N-depletion display a downregulation of transcripts encoding for enzymes required for amino acid biosynthesis and an upregulation of few transcripts encoding for proteases involved in protein degradation. Genes encoding for enzymes of the glyoxylate pathway, TCA cycle, and the gluconeogenesis show decrease or no change in expression following N-depletion. Amino acids recycled through the TCA cycle generate intermediates that might be converted to oxaloacetate (OAA), pyruvate, and acetyl CoA, the precursor for the fatty acid biosynthesis. Therefore, under N-depletion conditions, amino acids degradation yields intermediates that might be either directed towards fatty acid biosynthesis through acetyl CoA or used by the TCA cycle to produce ATP.

The high throughput transcriptomics-based data described in this work and the identification of major enzymes-encoding genes involved in TAG, starch, and protein metabolic pathways provide major insights into the biochemical and molecular changes that take place in N-depleted *C. reinhardtii* cells grown in photoautotrophic media. Mining the RNAseq data allowed us to identify targets for metabolic engineering that

might have the potential for increasing biofuel production from microalgae. In this study, few genes involved in the Kennedy pathway for TAG accumulation showed significant increase in the expression level. There are a number of potential explanations for why few genes change in expression despite the marked accumulation of both starch and TAG. One distinct possibility is posttranslational modification of the proteins involved in these pathways. These changes could either to activate proteins involved in these pathways, or block critical enzymes in competing pathways. Such modifications would allow for dramatic changes in enzyme function without any change in transcript levels. Further testing would be required to determine whether these modifications play a role in any of these pathways. However, the small number of genes with significant changes in expression provides promising targets for genetic engineering to increase TAG accumulation in these cells. In particular, *DGTT1* would make a promising target for overexpression. These findings provide important knowledge required for biotechnological improvement of next-generation biofuel production in *C. reinhardtii* or in other green microalgae.

Materials and methods

Strain and culture conditions

Chlamydomonas reinhardtii strain CC-125 was initially grown photoautotrophically to the middle of the logarithmic phase ($\sim$3 x 10^6 cells mL^{-1}) in nitrogen-replete high salt (HS) medium (Sueoka, 1960) under continuous illumination (250 μmol m^{-2} s^{-1} photosynthetically active radiation) on an orbital shaker (190 rpm) at 25 $^{\circ}$C and ambient levels of CO_2. These pre-cultured cells were collected by

centrifugation and resuspended at a density of 5 x 10^5 cells mL^{-1} in regular HS media or in the same media lacking nitrogen (HS-N). Cells used for different analyses were harvested after resuspension at regular time courses (0, 24, 48, and 144 hours).

Total RNA isolation

Total RNA was isolated with TriReagent (Molecular Research Center, Inc.) in accordance with the manufacturer's instructions from *C. reinhardtii* grown under N-replete and N-deplete conditions. Agarose gel electrophoresis (1.5%) was used to monitor the quality of the extracted RNA by checking the integrity of the RNA bands. The purified total RNA concentration was measured using a Thermo Scientific NanoDrop 2000c spectrophotometer. Total RNA isolated from these cells was used in transcriptomic, real-time quantitative polymerase chain reaction, and semi-quantitative polymerase chain reaction experiments.

Transcriptome analysis

Transcriptomic changes in *C. reinhardtii* grown photoautotrophically under N-replete and N-deplete conditions were measured using RNAseq. Experiments were performed independently twice and libraries were sequenced with the Illumina GAIIx analyzer. The Illumina reads from both experiments were mapped to the Augustus v5.0 transcript models for *C. reinhardtii* (available from http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html). The reads were first trimmed to 36 nucleotides in length to remove low quality 3′-ends before mapping. The alignment was performed using Burrows-Wheeler Aligner (BWA; v0.5.7) [51] with a seed length of 25, allowing 2

mismatches. A perl script was used to ensure that only reads that uniquely matched with a single transcript were counted. Raw gene counts were determined by counting the number of reads aligned to each transcript. To examine transcript abundance, gene expression was analyzed as Reads per Kilobase per Million (RPKM), which normalizes the read counts based on both transcript length and total number of reads, using the following formula:

$$RPKM = \frac{10^9 C}{NL}$$

where C is the number of reads mapped to the transcript, N is the total number of mapped reads in the library, and L is the length of the transcript in nucleotides (Mortazavi *et al.,* 2008). To determine the change in expression, each of the three time points during the N-deplete condition were compared pairwise with the control (N-replete) condition for each experiment. Differences in gene expression were examined as log_2(Fold Change), where Fold Change refers to the ratio of RPKM values, at each time point, between treatment and control. Statistical analysis of the data was performed using the DESeq package [52]. The q-values for each time point were calculated using both replicates. All transcripts that have at least one time point with a q-value ≤ 0.05 and at least a 3-fold change in expression (a log_2(Fold Change) of 1.585) are considered significant.

The functional annotations for each gene were assigned based on a combination of homology with genes in *Arabidopsis thaliana* involved in each of the pathways and conserved domains. Homology with *A. thaliana* was determined by reciprocal blastp searches [53] between the Augustus v5 protein models and the *A. thaliana* proteins from the TAIR database [54]. Additional putative genes were identified using a PFAM search for conserved domains consistent with each enzyme [55]. All functional annotations were

curated by manually inspecting the existing annotations for each gene, removing any genes already assigned to other pathways. Subcellular localization was determined by the localization of the top hit in *A. thaliana* with further analysis performed using the PredAlgo subcellular localization prediction algorithm for each enzyme with a significant change in expression [34].

Real-time quantitative polymerase chain reaction

Real-time PCR (qPCR) was performed on the Light-Cycler Instrument (iCycler iQ real-time PCR detection system; Bio-Rad) using RT^2 SYBR Green mastermix (Qiagen). qPCR data are represented as the fold change in mRNA abundance relative to the nitrogen-replete RNA sample and are normalized to the endogenous reference gene *Actin* using the cycle threshold (C_t) $2^{-\Delta\Delta Ct}$ method, where $\Delta\Delta C_t$ is the difference in the threshold cycles [56], according to the manufacturer's software (Bio-Rad). All experiments were performed in experimental replicates, and each sample was analyzed in technical triplicate. All our real-time PCR procedures and analysis follow the MIQE guidelines [57].

Semi-quantitative RT-PCR

Complementary DNA (cDNA) synthesis was performed using the Invitrogen first-strand synthesis kit with the oligo(dT) primer. Polymerase chain reaction was performed for 35 cycles at 94 °C for 30 s, 58 °C for 30 s, 72 °C for 30s in a reaction containing 2.5 mM dNTPs, and 2.5 μM oligonucleotide primers. Gene-specific oligonucleotide primers were designed to amplify transcripts and these were as follows:

522363-PAP-F1: 5'-ATCACCAACTGCCTCAAGCTG-3' and 522363-PAP-R1: 5'-

AGCAGCCAGAAGGTCAGGAA-3'; 522556-PAP-F1: 5'-

CTCTGTACGTGCTGGTGTTCGT-3' and 522556-PAP-R1: 5'-

GTACCTGCCCACCAGGTTGAT-3'; 514476-ACP-F: 5'-

CTGTCGATGATCCGCAAGTC-3' and 514476-ACP-R: 5'-

TTCTCCTCCAGAGCCATCAT-3'; 524437-BTA-F: 5'-

GCTCGTACCTGTCCCAAGAC-3' and 524437-BTA-R: 5'-

CGGCCAATCCACACGTAGTA-3'; 522727-LPAT-F: 5'-

GTGTGAACGACCCCACCTAC-3' and 522727-LPAT-R: 5'-

TGATGATTTCCAGCTCGTTG-3'; 513002-ACS-F: 5'-

CACGACAACCTGGAGTATGC-3' and 513002-ACS-R: 5'-

CGTACAGCGGCACAGACACC-3'; 516987-AO1-F: 5'-

ATGGAGTTCAGCCAGTGCTT-3' and 516987-AO1-R: 5'-

CCTCCTCGTACAGCTTGCTC-3'; 516819-AO3-F: 5'-

TCATGTACGCCGACTACAGC-3' and 516819-AO3-R: 5'-

CAGTACTTGCTGGCCTCGTC-3'; 512138-AO4-F: 5'-

GGCCCAGAAGTACTGGATCA-3' and 512138-AO4-R: 5'-

CACACGCACGTTGTCAAACC-3'; 516986-EH-F: 5'-

AGGCCCAGTATCGGGTAGAT-3' and 516986-EH-R: 5'-

CATCGTTGATGGAGTTGTCG-3'. Actin semi-quantitative RT-PCR was performed as

a positive control for all reactions. The primers were as follows: ACT-Cod-F: 5'-

GACATCCGCAAGGACCTCTAC-3' and ACT-Cod-R: 5'-

GATCCACATTTGCTGGAAGGT-3'. Reaction products were visualized after

electrophoresis on a TBE plus 1.5 % agarose gel containing ethidium bromide.

Literature Cited

1. Chisti, Y., *Biodiesel from microalgae beats bioethanol.* Trends Biotechnol, 2008. **26**(3): p. 126-31.
2. Dismukes, G.C., et al., *Aquatic phototrophs: efficient alternatives to land-based crops for biofuels.* Curr Opin Biotechnol, 2008. **19**(3): p. 235-40.
3. Nguyen, A.V., et al., *Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga Chlamydomonas reinhardtii.* Eukaryot Cell, 2008. **7**(11): p. 1965-79.
4. Gouveia, L. and A.C. Oliveira, *Microalgae as a raw material for biofuels production.* J Ind Microbiol Biotechnol, 2009. **36**(2): p. 269-74.
5. Radakovits, R., et al., *Genetic engineering of algae for enhanced biofuel production.* Eukaryot Cell, 2010. **9**(4): p. 486-501.
6. Rodolfi, L., et al., *Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor.* Biotechnol Bioeng, 2009. **102**(1): p. 100-12.
7. Sheehan, J., et al., *Look Back at the U.S. Department of Energy's Aquatic Species Program: Biodiesel from Algae; Close-Out Report*, in *Other Information: PBD: 1 Jul 1998*. 1998. p. Medium: ED; Size: 325 pages.
8. Hu, Q., et al., *Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances.* Plant J, 2008. **54**(4): p. 621-39.
9. Durrett, T.P., C. Benning, and J. Ohlrogge, *Plant triacylglycerols as feedstocks for the production of biofuels.* Plant J, 2008. **54**(4): p. 593-607.
10. Roessler, P.G., *ENVIRONMENTAL CONTROL OF GLYCEROLIPID METABOLISM IN MICROALGAE: COMMERCIAL IMPLICATIONS AND FUTURE RESEARCH DIRECTIONS.* Journal of Phycology, 1990. **26**(3): p. 393-399.
11. Sager, R. and S. Granick, *Nutritional control of sexuality in Chlamydomonas reinhardi.* J Gen Physiol, 1954. **37**(6): p. 729-42.
12. Martin, N.C. and U.W. Goodenough, *Gametic differentiation in Chlamydomonas reinhardtii. I. Production of gametes and their fine structure.* J Cell Biol, 1975. **67**(3): p. 587-605.
13. Mouille, G., et al., *Preamylopectin Processing: A Mandatory Step for Starch Biosynthesis in Plants.* Plant Cell, 1996. **8**(8): p. 1353-1366.
14. Wang, Z.T., et al., *Algal lipid bodies: stress induction, purification, and biochemical characterization in wild-type and starchless Chlamydomonas reinhardtii.* Eukaryot Cell, 2009. **8**(12): p. 1856-68.
15. Scott, S.A., et al., *Biodiesel from algae: challenges and prospects.* Curr Opin Biotechnol, 2010. **21**(3): p. 277-86.
16. Stephenson, A.L., et al., *Influence of nitrogen-limitation regime on the production by Chlorella vulgaris of lipids for biodiesel feedstocks.* Biofuels, 2009. **1**(1): p. 47-58.
17. Msanne, J., et al., *Metabolic and gene expression changes triggered by nitrogen deprivation in the photoautotrophically grown microalgae Chlamydomonas reinhardtii and Coccomyxa sp. C-169.* Phytochemistry, 2012. **75**: p. 50-9.

18. Griffiths, M. and S.L. Harrison, *Lipid productivity as a key characteristic for choosing algal species for biodiesel production.* Journal of Applied Phycology, 2009. **21**(5): p. 493-507.

19. Work, V.H., et al., *Increased lipid accumulation in the Chlamydomonas reinhardtii sta7-10 starchless isoamylase mutant and increased carbohydrate synthesis in complemented strains.* Eukaryot Cell, 2010. **9**(8): p. 1251-61.

20. Guarnieri, M.T., et al., *Examination of triacylglycerol biosynthetic pathways via de novo transcriptomic and proteomic analyses in an unsequenced microalga.* PLoS One, 2011. **6**(10): p. e25851.

21. Ohlrogge, J.B., J. Browse, and C.R. Somerville, *The genetics of plant lipids.* Biochim Biophys Acta, 1991. **1082**(1): p. 1-26.

22. Coleman, R.A. and D.P. Lee, *Enzymes of triacylglycerol synthesis and their regulation.* Prog Lipid Res, 2004. **43**(2): p. 134-76.

23. Czabany, T., et al., *Structural and biochemical properties of lipid particles from the yeast Saccharomyces cerevisiae.* J Biol Chem, 2008. **283**(25): p. 17065-74.

24. Beopoulos, A., T. Chardot, and J.M. Nicaud, *Yarrowia lipolytica: A model and a tool to understand the mechanisms implicated in lipid accumulation.* Biochimie, 2009. **91**(6): p. 692-6.

25. Fan, J., C. Andre, and C. Xu, *A chloroplast pathway for the de novo biosynthesis of triacylglycerol in Chlamydomonas reinhardtii.* FEBS Lett, 2011. **585**(12): p. 1985-91.

26. Nguyen, H.M., et al., *Proteomic profiling of oil bodies isolated from the unicellular green microalga Chlamydomonas reinhardtii: with focus on proteins involved in lipid metabolism.* Proteomics, 2011. **11**(21): p. 4266-73.

27. Gonzalez-Ballester, D., et al., *RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival.* Plant Cell, 2010. **22**(6): p. 2058-84.

28. Mus, F., et al., *Anaerobic acclimation in Chlamydomonas reinhardtii: anoxic gene expression, hydrogenase induction, and metabolic pathways.* J Biol Chem, 2007. **282**(35): p. 25475-86.

29. Dubini, A., et al., *Flexibility in anaerobic metabolism as revealed in a mutant of Chlamydomonas reinhardtii lacking hydrogenase activity.* J Biol Chem, 2009. **284**(11): p. 7201-13.

30. Miller, R., et al., *Changes in transcript abundance in Chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism.* Plant Physiol, 2010. **154**(4): p. 1737-52.

31. Moellering, E.R. and C. Benning, *RNA interference silencing of a major lipid droplet protein affects lipid droplet size in Chlamydomonas reinhardtii.* Eukaryot Cell, 2010. **9**(1): p. 97-106.

32. Rawsthorne, S., *Carbon flux and fatty acid synthesis in plants.* Prog Lipid Res, 2002. **41**(2): p. 182-96.

33. Rismani-Yazdi, H., et al., *Transcriptome sequencing and annotation of the microalgae Dunaliella tertiolecta: pathway description and gene discovery for production of next-generation biofuels.* BMC Genomics, 2011. **12**: p. 148.

34. Tardif, M., et al., *PredAlgo: a new subcellular localization prediction tool dedicated to green algae.* Mol Biol Evol, 2012. **29**(12): p. 3625-39.

35.   Ohlrogge, J. and J. Browse, *Lipid biosynthesis.* Plant Cell, 1995. **7**(7): p. 957-70.

36.   Goodson, C., et al., *Structural correlates of cytoplasmic and chloroplast lipid body synthesis in Chlamydomonas reinhardtii and stimulation of lipid body production with acetate boost.* Eukaryot Cell, 2011. **10**(12): p. 1592-606.

37.   Lippold, F., et al., *Fatty acid phytyl ester synthesis in chloroplasts of Arabidopsis.* Plant Cell, 2012. **24**(5): p. 2001-14.

38.   Ball, S.G., *The Intricate Pathway of Starch Biosynthesis and Degradation in the Monocellular Alga <i>Chlamydomonas reinhardtii</i>.* Australian Journal of Chemistry, 2002. **55**(2): p. 49-59.

39.   Fettke, J., et al., *Eukaryotic starch degradation: integration of plastidial and cytosolic pathways.* J Exp Bot, 2009. **60**(10): p. 2907-22.

40.   Li, Y., et al., *Inhibition of starch synthesis results in overproduction of lipids in Chlamydomonas reinhardtii.* Biotechnol Bioeng, 2010. **107**(2): p. 258-68.

41.   Bukau, B., J. Weissman, and A. Horwich, *Molecular chaperones and protein quality control.* Cell, 2006. **125**(3): p. 443-51.

42.   Rimaux, T., et al., *Expression of the arginine deiminase pathway genes in Lactobacillus sakei is strain dependent and is affected by the environmental pH.* Appl Environ Microbiol, 2012. **78**(14): p. 4874-83.

43.   Armbrust, E.V., et al., *The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism.* Science, 2004. **306**(5693): p. 79-86.

44.   Haines, R.J., L.C. Pendleton, and D.C. Eichler, *Argininosuccinate synthase: at the center of arginine metabolism.* Int J Biochem Mol Biol, 2011. **2**(1): p. 8-23.

45.   Wijffels, R.H. and M.J. Barbosa, *An outlook on microalgal biofuels.* Science, 2010. **329**(5993): p. 796-9.

46.   Finazzi, G., et al., *Nonphotochemical quenching of chlorophyll fluorescence in Chlamydomonas reinhardtii.* Biochemistry, 2006. **45**(5): p. 1490-8.

47.   Li-Beisson, Y., et al., *Acyl-lipid metabolism.* Arabidopsis Book, 2010. **8**: p. e0133.

48.   Boyle, N.R., et al., *Three acyltransferases and nitrogen-responsive regulator are implicated in nitrogen starvation-induced triacylglycerol accumulation in Chlamydomonas.* J Biol Chem, 2012. **287**(19): p. 15811-25.

49.   Zhang, M., et al., *DGAT1 and PDAT1 acyltransferases have overlapping functions in Arabidopsis triacylglycerol biosynthesis and are essential for normal pollen and seed development.* Plant Cell, 2009. **21**(12): p. 3885-901.

50.   Lu, Y. and T.D. Sharkey, *The importance of maltose in transitory starch breakdown.* Plant Cell Environ, 2006. **29**(3): p. 353-66.

51.   Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

52.   Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

53.   Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

54.   Swarbreck, D., et al., *The Arabidopsis Information Resource (TAIR): gene structure and function annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D1009-14.

55.   Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.

56.   Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.

57.   Bustin, S.A., et al., *The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.* Clin Chem, 2009. **55**(4): p. 611-22.

A
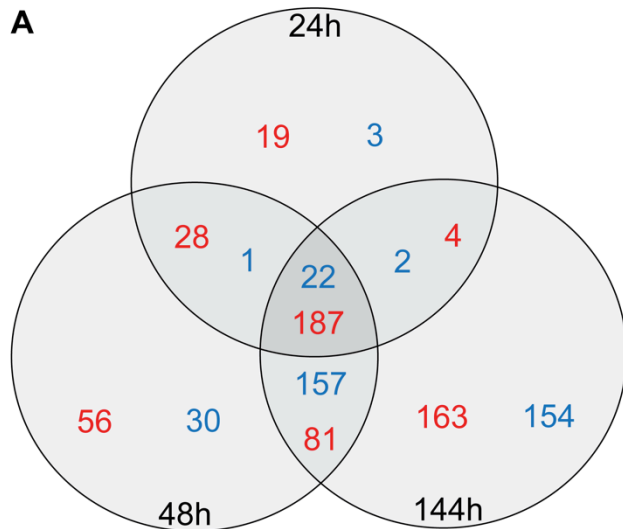
B

	Total	Lipid	Starch	Protein/AA
Upregulated	538	6	2	12
Downregulated	369	12	4	4

Figure 2.1: Summary of the differential expression data

A) Venn diagram showing the number of genes significantly upregulated (red) or downregulated (blue) between each time point. B) The number of differentially expressed genes for each of the pathways studied.
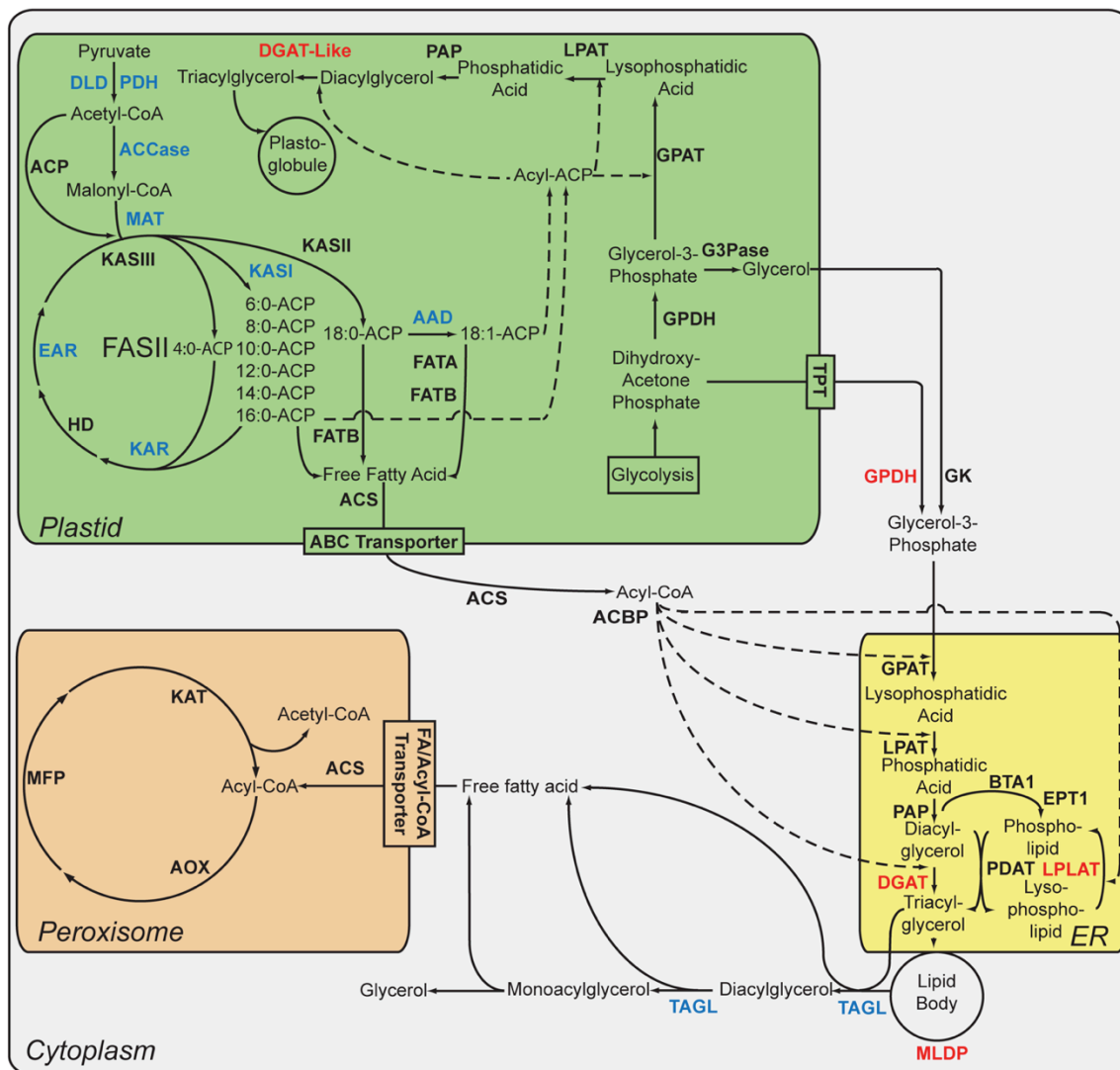
Figure 2.2: Pathways for fatty acid (FA) biosynthesis, TAG biosynthesis and degradation, and subcellular localization in *C. reinhardtii*

Expression of the corresponding genes in N-deprived cells relative to cells grown in nutrient replete medium is indicated by different colors: black, no significant change in expression; red, significantly upregulated transcripts; and blue, significantly downregulated transcripts. Abbreviations: DLD, dihydrolipoamide dehydrogenase; PDH, pyruvate dehydrogenase complex; ACCase, acetyl-CoA carboxylase; ACP, acyl-carrier protein; MAT, ACP-S-malonyl transferase; KAS, β-ketoacyl-ACP synthase; KAR, β-

ketoacyl-ACP reductase; HD, β-hydroxyacyl-ACP dehydrase; EAR, enoyl-ACP

reductase; AAD, Δ9 acyl-ACP desaturase; FAT, fatty acyl-ACP thioesterase; GPAT,

glycerol-3-phosphate acyltransferase; LPAT, lysophosphatidyl acyltransferase; PAP,

phosphatidic acid ahosphatase; DGAT-Like, diacylglycerol acyltransferase-like; GPDH,

glycerol-3-phosphate dehydrogenase; G3Pase, glycerol-3-phosphatase; ACS, acyl-CoA

synthetase; GK, glycerol kinase; DGAT, diacylglycerol acyltransferase; PDAT,

phospholipid:diacylglycerol acyltransferase; BTA, betaine lipid synthase; EPT,

ethanolamine phosphotransferase; LPLAT, lysophospholipid acyltransferase; ACBP,

acyl-CoA binding protein; MLDP, major lipid droplet protein; TAGL,

triacylglycerol/diacylglycerol lipase; AOX, acyl-CoA oxidase; MFP, multi-functional

protein (dienoyl-CoA reductase/enoyl-CoA dehydrogenase); KAT, ketoacyl-CoA

thiolase. See Table 2.1, Figure 2.3, and Figure A.2 for the transcripts whose expressions

were identified to be significantly differentially expressed.
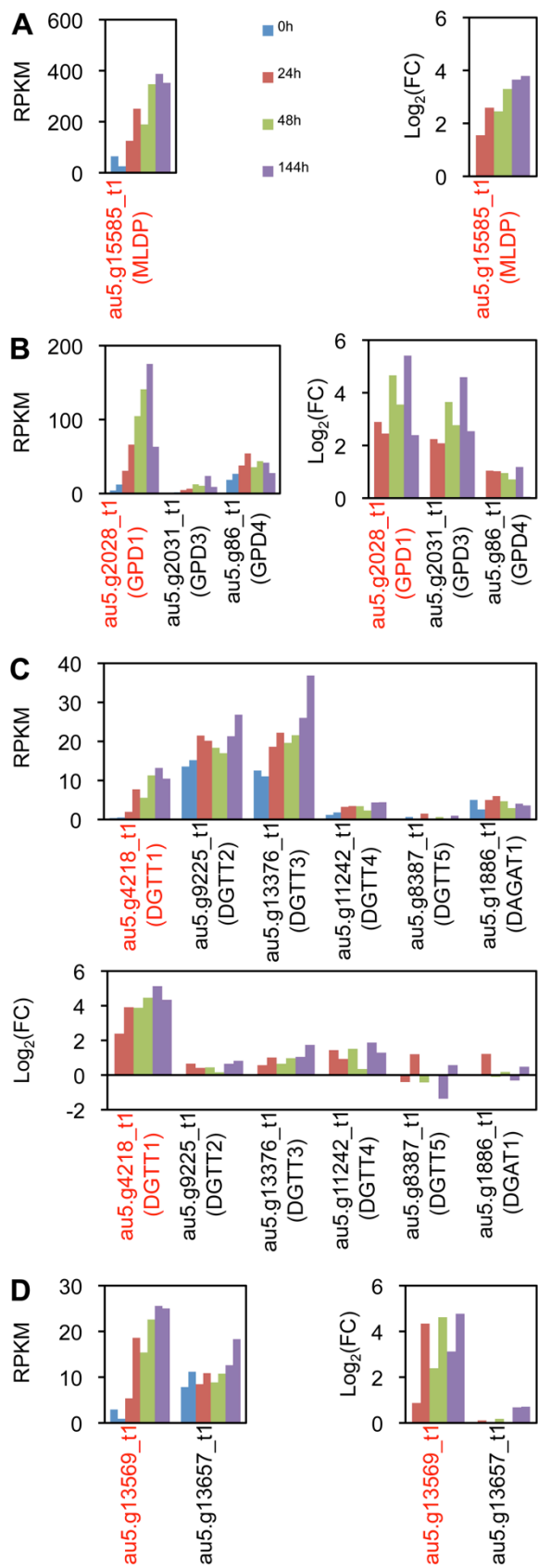
Figure 2.3: Expression analysis of significantly upregulated transcripts associated to FA and TAG metabolic pathways

The four enzyme families included (colored red in Figure 2.2) are: A. glycerol 3-phosphate dehydrogenase (*GPDH*), B. diacylglycerol acyltransferase (*DGAT*) C. lysophospholipid acyltransferase (*LPLAT*), and D. major lipid droplet protein (*MLDP*). Expression levels are presented as Reads per Kilobase per Million (RPKM) and *log₂*(Fold Change) for two independent experiments.
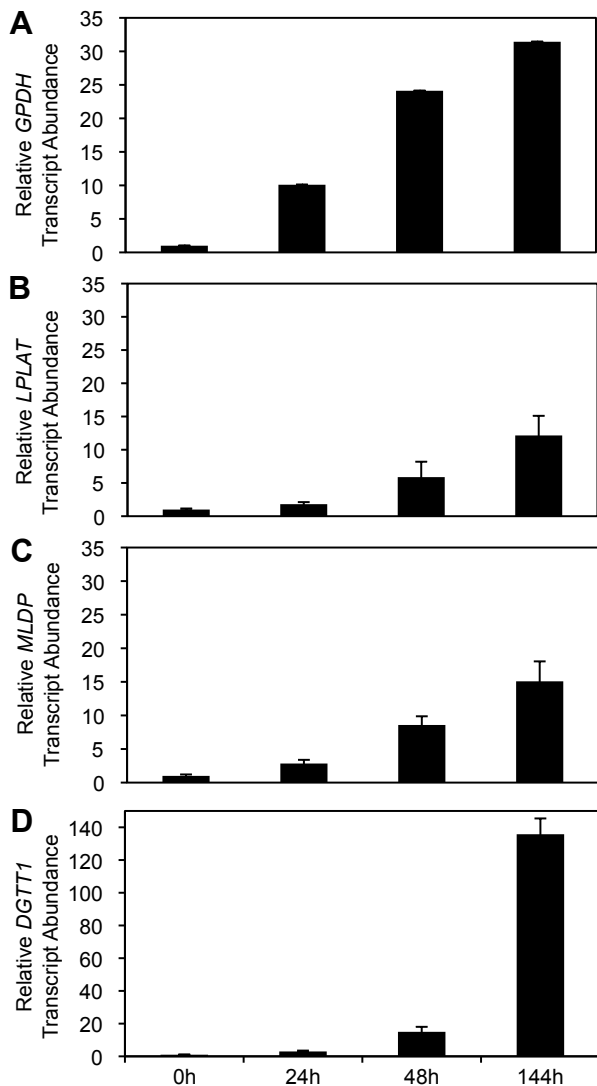
Figure 2.4: Real-time (qPCR) analysis of significantly upregulated TAG

biosynthesis genes identified by RNAseq

Relative abundance of A. Glycerol 3-Phosphate Dehydrogenase (*GPDH*), B.

Lysophospholipid Acyltransferase (*LPLAT*), C. Major Lipid Droplet Protein (*MLDP*),

and D. Diacylglycerol Acyltransferase (*DGTT1*) mRNAs were normalized to *Actin* in

wild-type *C. reinhardtii* cells grown in nutrient replete medium, or wild-type cells grown

in N-depleted medium for 24, 48, and 144 hours. Error bars represent standard deviations

from three independent biological replicates.
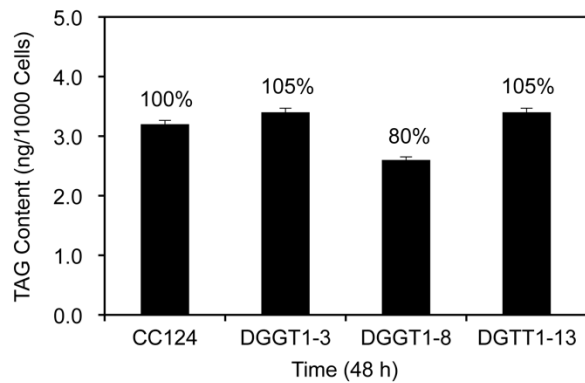
Figure 2.5: Triacylglycerol contents (nanograms per 1000 cells) determined by gas chromatography analysis after 48h nitrogen starvation

Values correspond to the average of three independent experiments.
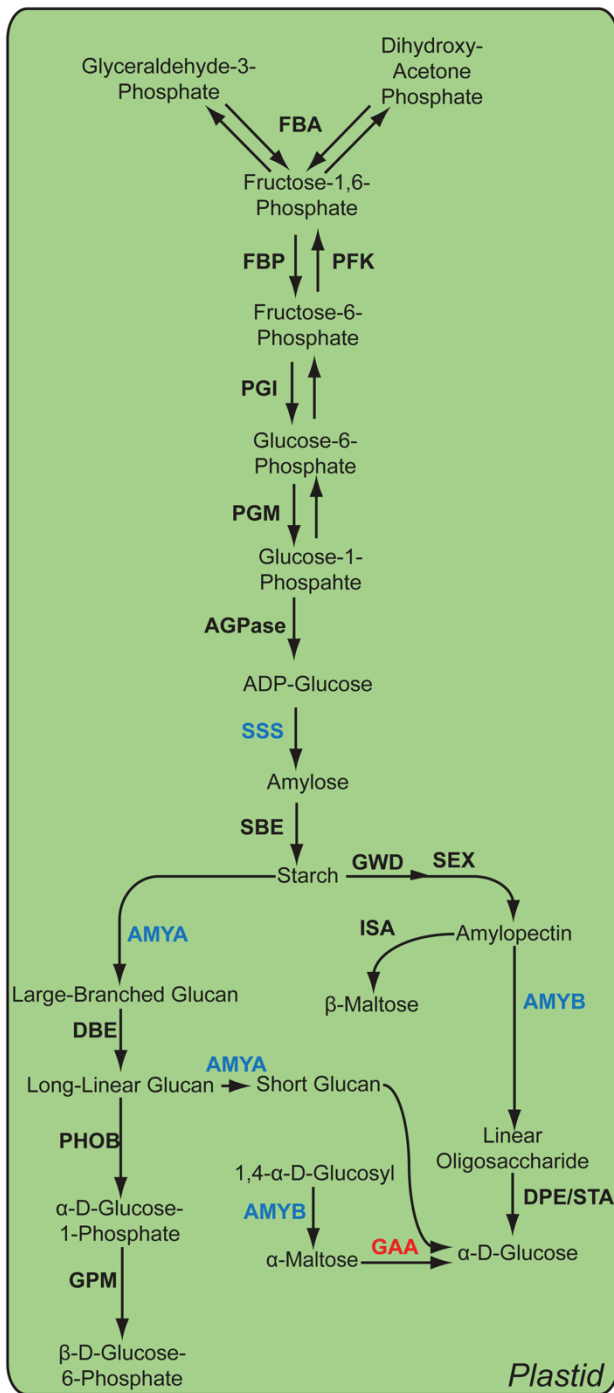
Figure 2.6: Pathways for starch biosynthesis and degradation, and subcellular localization in *C. reinhardtii*

Expression of the corresponding genes in N-deprived cells relative to cells grown in nutrient replete medium is indicated by different colors: black, no significant change in expression; red, significantly upregulated transcripts; and blue, significantly downregulated transcripts. Abbreviations: FBA, fructose bisphosphate aldolase; FBP, fructose-1,6-bisphosphatase; PFK, 6-phospho fructokinase; PGI, phosphoglucose isomerase; PGM, phosphoglucomutase; AGPase, ADP-glucose pyrophosphorylase; SSS, soluble starch synthase; SBE, starch branching enzyme; GWD, glucan water dikinase; AMYA, α-amylase; DBE, debranching enzyme; PHOB, starch phosphorylase; DPE/STA, 4-α-glucanotransferase; AMYB, β-amylase; SEX, 6-phosphogluco starch phosphatase; ISA, isoamylase; GAA, α-glucosidase. See Table 2.1 and Figure A.3 for the transcripts whose expressions were identified to be significantly differentially expressed.
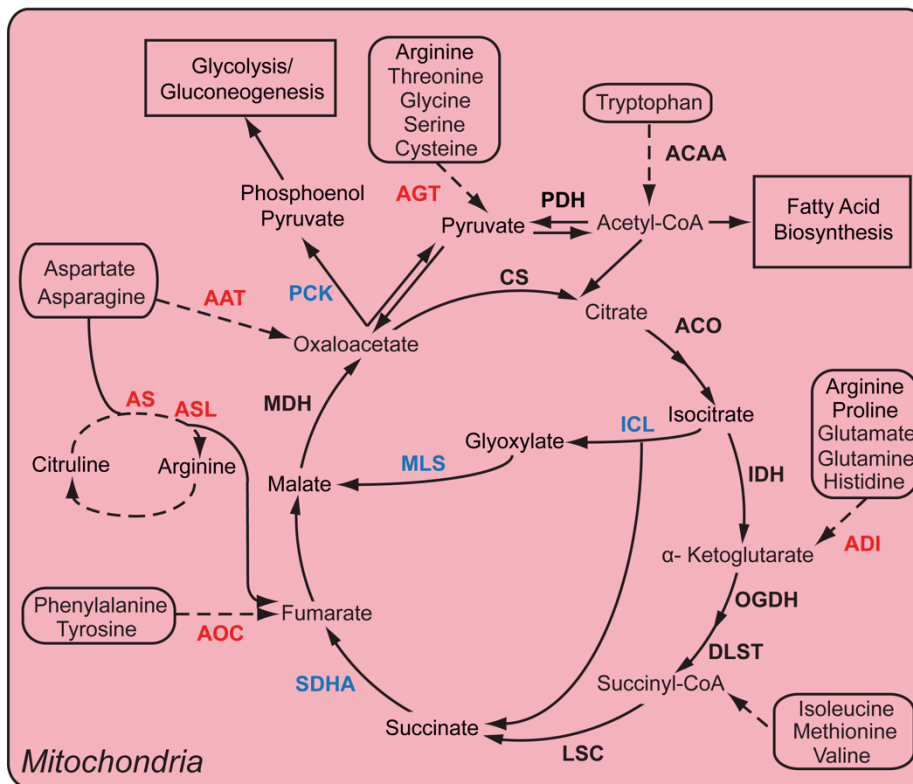
Figure 2.7: Overview of the amino acid degradation pathway and TCA cycle in *C. reinhardtii*

Expression of the corresponding genes in N-deprived cells relative to cells grown in nutrient replete medium is indicated by different colors: black, no significant change in expression; red, significantly upregulated transcripts; and blue, significantly downregulated transcripts. Abbreviations: ADI, agmatine deiminase; AOC, copper amine oxidase; AGT, alanine-glyoxylate transaminase; AS, argininosuccinate synthase; ASL, argininosuccinate lyase; SDHA, succinate dehydrogenase; AGT, alanine-glyoxylate transaminase; ACAA, acetyl-CoA acyltransferase; AAT, aspartate aminotransferase; CS, citrate synthase; ACO, aconitrate hydratase; IDH, isocitrate dehydrogenase; OGDH, 2-oxoglutarate dehydrogenase E1 component; DLST, 2-oxoglutarate dehydrogenase E2 component; LSC, succinyl-CoA synthase; MDH, malate dehydrogenase; ISL, isocitrate

lyase; MLS, malate synthase; PCK, phosphoenolpyruvate carboxylase. See Table 2.1,

Figure A.4 and for the transcripts whose expressions were identified to be significantly

differentially expressed.

Table 2.1: Summary of differentially expressed genes

| Enzyme	Symbol	Aug5 Model Name	Predicted Localization	RPKM			
				0h	24h	48h	144h
TAG Metabolism							
Pyruvate dehydrogenase	PDH	au5.g11028_t1	C	166.13	64.25	68.57	35.17
Dihydrolipoamide dehydrogenase	DLD	au5.g1301_t1	SP	66.04	19.57	23.25	10.94
Acetyl-CoA carboxylase	ACCase	au5.g1722_t1	O	111.05	30.67	34.35	14.06
		au5.g1722_t2	O	145.37	40.22	45.17	17.47
ACP-S-malonyl transferase	MAT	au5.g5494_t1	C	322.21	35.14	43.71	11.87
β-ketoacyl-ACP synthase	KASI	au5.g9840_t1	C	266.95	89.53	118.35	71.65
β-ketoacyl-ACP reductase	KAR	au5.g10590_t1	C	101.24	34.19	40.74	14.82
		au5.g2990_t1	M	117.83	27.30	16.18	9.28
Enoyl-ACP reductase	EAR	au5.g13296_t1	C	304.23	131.40	110.80	54.65
Δ9 acyl-ACP desaturase	AAD	au5.g7079_t1	C	79.59	57.83	47.96	21.03
Diacylglycerol acyltransferase	DGAT	au5.g14782_t1	O	1.31	3.97	3.71	5.91
		au5.g4218_t1	O	0.45	4.85	8.42	11.82
Glycerol 3-phosphate dehydrogenase	GPDH	au5.g2028_t1	C	8.10	48.36	122.76	119.18
		au5.g2031_t1	O	1.26	5.57	11.43	16.31
Lysophospholipid acyltransferase	LPLAT	au5.g13569_t1	O	1.93	11.98	18.99	25.30
Major lipid droplet protein	MLDP	au5.g15585_t1	O	45.03	188.07	268.19	370.06
Triacylglycerol/diacylglycerol lipase	TAGL	au5.g7029_t1	SP	36.63	11.16	5.80	12.69
		au5.g12306_t1	O	8.40	2.72	2.28	2.40
Starch Metabolism							
Soluble starch synthase	SSS	au5.g10852_t1	C	101.45	29.39	20.80	14.87
		au5.g7473_t1	C	114.61	123.57	67.93	8.65
		au5.g6335_t1	O	1.79	7.03	6.96	7.51
α-amylase	AMYA	au5.g15170_t1	O	15.52	4.83	1.53	1.91
β-amylase	AMYB	au5.g12805_t1	M	26.18	11.40	4.94	5.68
α-glucosidase	GAA	au5.g11038_t1	O	1.36	2.85	2.53	6.32
Protein Degradation							
GMP synthetase	GUA	au5.g13735_t1	O	2.18	98.78	82.18	97.05
Deiminase/IAA-amino acid hydrolase/aminoacylase	M20	au5.g4278_t1	SP	1.76	5.19	7.79	10.45
		au5.g4678_t1	SP	0.51	4.39	4.93	7.56
		au5.g12930_t1	O	3.73	11.63	12.79	18.76
Serine carboxypeptidase	S10	au5.g9131_t1	SP	13.00	41.17	39.87	51.68
Amino Acid Degradation							
Agmatine deiminase	ADI	au5.g1160_t1	M	1.31	46.90	63.32	50.03
Copper amine oxidase	AOC	au5.g987_t1	O	3.39	17.26	22.68	29.91
Alanine-glyoxylate transaminase	AGT	au5.g4992_t1	C	24.32	89.13	74.95	103.58
Aspartate aminotransferase	AAT	au5.g12556_t1	SP	12.76	45.49	36.43	47.36
Argininosuccinate synthase	AS	au5.g15792_t1	C	76.32	378.55	256.92	284.37
6-phosphogluconate dehydrogenase	PGD	au5.g3600_t1	O	27.59	100.18	88.41	95.06
Argininosuccinate lyase	ASL	au5.g1397_t1	O	10.50	47.52	33.22	35.08
Succinate dehydrogenase	SDHA	au5.g5444_t1	M	71.55	45.71	17.54	28.39
Isocitrate lyase	ICL	au5.g13056_t1	O	2451.26	1024.68	148.18	128.13
Malate synthase	MLS	au5.g2123_t1	O	882.50	392.10	83.90	92.90
Phosphoenolpyruvate carboxylase	PCK	au5.g9626_t1	C	479.04	252.41	29.47	19.18

Genes in the fatty acid biosynthesis, TAG biosynthesis and degradation, starch

biosynthesis and degradation, protein degradation, and amino acid degradation pathways

that are differentially expressed in nitrogen replete (0h) and nitrogen deplete (24h, 48h,

and 144h) growth conditions. The Reads per Kilobase per Million (RPKM) values

represent the mean value of both experiments. Abbreviations for localization: C,

chloroplast; M, mitochondria; SP, secretory pathway; ER, endoplasmic reticulum; O,

other.

CHAPTER 3- IDENTIFICATION OF AGO3 ASSOCIATED MIRNAS AND COMPUTATIONAL PREDICTION OF THEIR TARGETS IN THE GREEN ALGA *CHLAMYDOMONAS REINHARDTII*

Adam Voshall,[*] Eun-Jeong Kim,[*] Xinrong Ma,[*] Etsuko N. Moriyama,[§] and Heriberto Cerutti[§]

[*]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska 68588

[§]School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, Nebraska 68588

Running title: MicroRNAs and targets in Chlamydomonas

Key words: miRNA, argonaute, miRNA target, transcript cleavage, translation repression

Corresponding author: Heriberto Cerutti,

School of Biological Sciences and Center for Plant Science

Innovation,

University of Nebraska-Lincoln,

E211 Beadle Center,

Lincoln, Nebraska 68588-0666.

Phone: 402-472-0247

E-mail: hcerutti1@unl.edu

Abstract

The existence of small RNAs in *Chlamydomonas reinhardtii* has been established for several years but little is known about their role(s) in the regulation of endogenous transcripts and cellular processes. To define functional microRNAs (miRNAs) in Chlamydomonas, we characterized small RNAs associated with an argonaute protein, AGO3, by affinity purification and deep sequencing. Using a stringent set of criteria for canonical miRNA annotation, we identified 39 precursor miRNAs, which produce 46 unique, AGO3-associated miRNA sequences including 11 previously reported miRNAs and 35 novel ones. We also attempted to identify miRNA targets, based on the complementarity of miRNAs with candidate binding sites. Recent results with reporter constructs have indicated that Chlamydomonas miRNAs may regulate target genes through either transcript cleavage or translation repression. Thus, potential targets were divided into two categories depending on the extent of complementarity to a given miRNA, those likely to be regulated through cleavage and those likely to be regulated through translational repression. The search for cleavage targets identified 77 transcripts. However, only six of them showed an increase in mRNA levels in a mutant strain almost devoid of miRNAs. The search for translational repression targets, which used complementarity criteria more stringent than those experimentally required for a reduction in target protein levels, identified 506 transcripts. We also examined experimentally candidate targets regulated through cleavage or translational repression. Our results emphasize the difficulty of identifying genuine miRNAs and miRNA targets in *C. reinhardtii* and suggest that at least some miRNAs might regulate endogenous genes through translational repression. Formatted for submission to Genetics.

Introduction

RNA interference (RNAi) provides cells with a delicate means of gene expression

regulation through a variety of mechanisms [1-3]. MicroRNAs, components of the RNAi

machinery, consist of endogenously encoded 20-24 nucleotide (nt) RNA fragments [1].

In plants and some green algae, these fragments are predominantly derived from small

hairpin loops up to ~300 base pairs (bp) in length that are processed by RNase III

enzymes, the Dicer-like (DCL) family of proteins [3, 4]. In *Arabidopsis thaliana*, most

miRNAs are generated by DCL1 in two processing steps, from the initial transcript (the

primary miRNA or pri-miRNA) to a smaller hairpin miRNA precursor (pre-miRNA) and

then to a short duplex consisting of the miRNA and its complementary sequence (the

miRNA* or passenger strand) [3, 5]. The RNA duplex is then loaded into the RNA-

induced silencing complex (RISC), which contains an Argonaute (AGO) protein as a key

component and becomes active upon removal of the miRNA* strand. The active RISC

can repress expression of target mRNAs by either preventing translation or by cleaving

the transcript [2].

The extent of complementarity between miRNA and target mRNA has been

proposed to play a major role in determining whether silencing occurs through cleavage

or translational repression [2]. The first nucleotide at the 5' end of a miRNA is secured

within the Mid (middle) domain of Argonaute, making this nucleotide inaccessible to

bind with target mRNAs, although it is important for determining miRNA loading into

distinct AGOs (Figure 3.1) [6-11]. In this conformation, nucleotides 2-8 from the 5' end

of a miRNA (known as the seed region for its role in nucleating the pairing between an

AGO-associated miRNA and its mRNA target) face outward into the channel that will eventually contain the mRNA, while the remaining nucleotides are at first inaccessible and do not play an initial role in target selection (Figure 3.1) [12, 13]. In land plants, the complementarity of miRNAs and targets is often nearly perfect, resulting in cleavage of the target sequence between nucleotides 10 and 11 of the miRNA [14-17]. In animals, by contrast, miRNAs frequently display extensive mismatching with their targets and regulate target expression through translational repression and/or transcript destabilization independent of AGO-mediated cleavage [18-21]. There is a growing body of evidence indicating that plants can perform this type of regulation as well; however, it still appears to be much less common than regulation through direct cleavage [22-25].

In the green alga *Chlamydomonas reinhardtii*, miRNAs have been described relatively recently, based on deep sequencing of total cellular small RNA populations [4, 26-28]. However, the validity of many predicted Chlamydomonas miRNAs has been questioned of late because of their large hairpin precursor structures and the apparent imprecise processing of the small RNAs (sRNAs) [29, 30]. *C. reinhardtii* possesses a core miRNA/sRNA processing and effector machinery similar to that in higher plants [4]. This machinery includes three AGO proteins along with three DCL proteins. AGO1 and DCL1 appear to be primarily involved in the silencing of transposable elements whereas AGO3 seems to be most extensively involved in miRNA functions [4]. The quality control of miRNAs is at least partially directed by the MUT68 protein, which functions by placing untemplated nucleotides (generally uridyls) at the 3' end of the miRNA molecules, flagging them for degradation [28].

Little is known about miRNA/target interactions in green algae. Based on the

phylogenetic relationship to land plants, early studies suggested that miRNAs in *C. reinhardtii* would function primarily through cleavage, in a manner similar to those in higher land plants [26, 27]. However, very few of the circa 50 Chlamydomonas small RNAs deposited in miRBase have identifiable targets with near perfect complementarity. This lack of easily recognizable targets leaves open the possibility that miRNAs may function either largely or primarily through translational repression, in a manner more akin to those in animals. This possibility is strengthened by recent evidence, using artificial constructs, demonstrating that such regulation is possible in *C. reinhardtii* [31]. Thus, it remains unclear whether the majority of the described Chlamydomonas sRNAs are truly miRNAs, what their potential targets are, and by which mechanism(s) the may regulate endogenous gene expression.

In animals, the non-RISC-associated passenger strand is very labile and it has been suggested that the high stability of miRNAs reflects their RISC association, implying that sequencing of total cellular small RNA populations accurately reflects the RISC-associated sRNA populations [32]. This assumption has been recently challenged by studies suggesting that only a fraction of mature miRNAs is in fact AGO associated [33-35]. It remains uncertain whether this also applies to land plants and green algae but, to increase the likelihood of defining a set of functional, RISC-bound miRNAs in Chlamydomonas, we first identified AGO3-associated sRNAs by deep-sequencing. From these sequences, putative miRNAs, meeting canonical criteria, were selected and used to predict interacting targets by computational means. Lastly, the mechanism(s) of target repression was examined for a pair of small RNAs. Our results suggest the existence of a limited set of genuine miRNAs in *C. reinhardtii*, some of which may regulate target

transcripts preferentially by translation repression. However, this analysis is limited by the lack of a fully sequenced Chlamydomonas nuclear genome (see Discussion).

Materials and Methods

Transgenic strains, mutants, and culture conditions

Chlamydomonas cells were routinely grown in Tris-acetate-phosphate (TAP) medium [36]. Strain Maa7-IR44s, containing an inverted repeat transgene targeting the 3' UTR of the *MAA7* gene (encoding Tryptophan synthase β subunit), has been previously described [36]. Mut-20, deleted for *Tudor Staphylococcal Nuclease 1* (*TSN1*) and almost devoid of miRNAs, was obtained in an insertional mutagenesis screen designed to isolate mutants defective in RNAi-mediated translation repression [36]. For the isolation of AGO3-associated small RNAs, we fused the FLAG tag [37] to the N-terminal end of Chlamydomonas AGO3 (g16859). This construct was placed under the control of *psaD* regulatory sequences and transformed into Maa7-IR44s (Figure B.1). A similar construct was made fusing the FLAG tag to the coding sequence of the bleomycin gene and transformed into Maa7-IR44s.

Isolation of AGO3-associated small RNAs, library preparation, and sequencing

FLAG-tagged AGO3 was affinity purified from a cell lysate as previously described for a TAP-tagged protein [38]. RNAs associated with AGO3 were purified with TRI reagent (Molecular Research Center) and contaminant DNA was removed by DNase I treatment

(Ambion) (MA *et al.* 2013). Construction of cDNA libraries and Illumina sequencing were then carried out as previously reported (IBRAHIM *et al.* 2010).

Small RNA mapping and profiling

Sequenced reads were first mapped to version 9.1 of the phytozome *C. reinhardtii* genome [39], by using version 3.02 of Novoalign (www.novocraft.com) with the miRNA flag and with a score threshold of 15. Mapped reads were filtered to remove those showing alignments with gaps or mismatches as well as those that mapped to more than 5 locations in the genome. Both the total population of sequenced reads and the population of reads mapped to the genome were profiled based on length. Genome mapped reads were also classified as matching the chloroplast or the mitochondrial genomes, rRNAs, snRNAs, snoRNAs, tRNAs, other non-coding RNAs (ncRNAs), or transposable elements. The chloroplast and mitochondrial genomes were taken from Genbank (accession numbers BK000554 and NC_001638.1 respectively) along with the rRNAs, snRNAs, snoRNA, tRNAs, and ncRNAs sequences. Transposons sequences were taken from Repbase [40]. The sequence logo to show nucleotide composition bias by position was made using version 3.3 of Weblogo analyzing all unique reads 21nt in length [41].

Genomic clustering of sRNAs and miRNA identification

After removing those reads that mapped to known non-coding RNA categories, transposons, the chloroplast or the mitochondrial genomes, potential sRNA reads were processed following the steps illustrated in Figure 3.2A. The reads were first clustered by genomic location such that there was no more than 200 nt between adjacent reads,

regardless of strand. The reads on both strands in the same genomic location were placed in the same cluster. These clusters were analyzed further to differentiate pre-miRNA structures from other sRNA structures. The genomic sequence for each strand of the cluster was folded using version 2.1.5 of RNAfold from the Vienna RNA package [42]. Clusters with a genomic sequence < 50 nt or >1000 nt in length were excluded from folding, as they were deemed unlikely to code for canonical miRNAs [29]. Clusters containing sequence gaps (i.e., unsequenced genomic regions) were also excluded since the secondary structure folding of these regions cannot be unambiguously predicted. The secondary structures were parsed to determine if they fold into a hairpin. However, clusters whose highest expressed reads had more than four mismatches with the opposing arm of the hairpin were removed from further analyses, following the criteria set up by Tarver, Donoghue [29] for the annotation of plant miRNAs where complementary base pairing between the mature and star sequences is fairly precise with fewer than four mismatches and any asymmetric bulges involving only one or two nucleotides. The putative miRNA* sequence for the highest expressed read of each cluster was determined using the hairpin fold and the genomic sequence, with a two nucleotide offset to allow for the 3' overhang in the miRNA/miRNA* duplex. Clusters remaining after this filtering were manually curated based on the accuracy of processing of the 5' end of the predominant reads and the frequency of predominant reads present [29]. The miRNA* sequences for clusters with more than one predominant read were also determined manually based on the secondary structure.

MicroRNA target prediction

Figure 3.2B illustrates the process used for the prediction of miRNA targets. Potential

miRNA binding sites were determined by searching version 9.1 of the *C. reinhardtii*

transcriptome using version 2.1 of RNAhybrid [43]. This algorithm returns the most

energetically favorable pairing for each miRNA/target pair. The obtained hybrid pairs

were filtered based on the number and location of gaps, wobbles, and mismatches to limit

the predictions to only those that could regulate targets in a projected manner. For

cleavage targets, this search required perfect matching for nucleotides 2-8 (the seed

region, Figure 3.1B) and nucleotides 9-12 (the catalytic or center region, Figure 3.1B),

and no more than three GU wobbles and three mismatches or a gap of more than one

nucleotide in the remaining sequence. For translational repression targets, the constraints

for the catalytic region were relaxed to allow up to three mismatches or wobbles, with the

constraints for the seed-region and the 3' end pairing remaining the same. Additionally,

translational repression targets needed at least one mismatch or wobble in the catalytic

region to keep the two sets of predicted targets non-overlapping. MicroRNA encoding

transcripts, which may also be targeted for degradation by their encoded miRNAs, were

removed from the analysis.

Putative roles of the predicted targets were evaluated by using the annotations of

Chlamydomonas genes (if available) as well as conserved protein domains. Manual

annotations were taken from the Algal Functional Annotation site [44], and are based on

the protein annotations uploaded to Phytozome [45]. Conserved protein domains were

found by searching version 27.0 of the Pfam-A database [46] using as queries the protein

sequences of the predicted targets from version 9.1 of Phytozome with an e-value cutoff

of 1 using version 2.2.29 of BLASTp [47]. For proteins with multiple copies of the same

domain, each domain is only reported once.

RNA analyses and quantitative PCR assays

Total cell RNA was purified with TRI reagent, following the manufacturer's instructions.

For small RNA analyses, total RNA samples were resolved on 15% polyacrylamide/7M

urea gels and electroblotted to Hybond-XL membranes (GE Healthcare). Blots were

hybridized with ^{32}P-labeled DNA probes using the High Efficiency Hybridization System

as previously described (IBRAHIM *et al.* 2010). Specific miRNAs were detected by

hybridization with DNA oligonucleotides labeled at their 5' termini with [γ-^{32}P]ATP and

T4 Polynucleotide Kinase (New England Biolabs) [36]. Putative cleavage sites of the

Cre17.g697550 transcript were examined using a 5' RACE approach with the GeneRacer

kit (Life Technologies), as previously described (MOLNAR *et al.* 2007; YAMASAKI *et al.*

2013). First and nested PCR amplifications were performed using the following primers:

g697550(Race)-R1 (5'-CCTTGCACTTGAGGCACTGCACAA-3') and g697550(Race)-

R2 (5'-CGTGTGGGGGCGGGATGAT-3'). For quantitative PCR analyses, DNase I

treated RNA samples were used as template for first-strand cDNA synthesis using an

oligo(dT)$_{18}$ primer and SuperScript III reverse transcriptase (Life Technologies). Primer

pairs for the quantitative PCR amplifications were as follows: for *Cre17.g697550,*

g697550-F (5'-GAGAGGATCGCGGACAACC-3') and g697550-R (5'-

AGGACCGGTAGATGCTCTTGG-3'); and for *Cre16.g683650,* g683650-F (5'-

CAGTTTGAGCCCGACCTACG-3') and g683650-R (5'-CCACGCCGCACTCCAGC-

3'). DNA fragments were amplified and quantified with the RT^2 SYBR

Green/Fluorescein qPCR mastermix (Qiagen) using the iCycler Real Time PCR

Detection System (Bio-Rad).

Differential gene expression analysis

RNA samples from Maa7-IR44s and Mut-20, grown photoheterotrophically in TAP

medium, were purified with TRI reagent and treated with DNase I. Standard RNA-seq of

the samples was then performed using an Illumina Genome Analyzer II [48]. Sequencing

reads from both cell lines were mapped to the *C. reinhardtii* transcript models in version

9.1 of phytozome [39]. Reads were first trimmed to 36 nucleotides in length, to remove

low quality 3' end sequences before mapping. The alignment was performed using

Bowtie2 [49] with a seed length of 25, allowing 2 mismatches, and a maximum hit of 1.

Only reads that uniquely matched with a single transcript were counted. Raw gene counts

were determined by counting the number of reads aligned to each transcript. Transcript

abundance was examined as Reads per Kilobase per Million (RPKM), which normalizes

the read counts based on both transcript length and total number of reads, using the

following formula:

$$RPKM = \frac{10^9 C}{NL}$$

where C is the number of reads mapped to the transcript, N is the total number of mapped

reads in the library, and L is the length of the transcript in nucleotides [50]. To determine

changes in gene expression, transcript abundance for each transcript was compared

pairwise between Mut-20 and its parental strain Maa7-IR44s. Differences in gene

expression were examined as log_2(Fold Change), where Fold Change refers to the ratio of

RPKM values between the Mut-20 and the Maa7-IR44s (control) strains. Statistical

analysis to determine the significance of gene expression changes was performed using

version 1.16 of DESeq, for samples without replicates [51].

Immunoblot analysis

The Chlamydomonas Cre16.g683650 protein was immunodetected, following a standard

procedure (MA *et al.* 2013), by overnight incubation at 4 °C with a 1:10,000 dilution of a

rabbit antibody raised against a C-terminal peptide (GIKPSAHKRGGVRM) conjugated

to KLH (GenScript). A modification-insensitive polyclonal antibody (Abcam ab1791)

was used to detect histone H3.

Results

Identification of AGO3 associated small RNAs

A FLAG-tagged AGO3, under the control of the *psaD* promoter, was introduced into the

Maa7-IR44s strain [36] and its functional competence evaluated by examining its

subcellular localization (Figure B.1) and its association with the translationally repressed

MAA7 transcript (Figure B.2). Small RNAs associated with AGO3 were then isolated by

co-immunoprecipitation with the FLAG-tagged protein. Deep sequencing of the

corresponding small RNA library generated 20,589,616 reads (675,917 unique reads), of

which 12,482,999 (117,241 unique) mapped perfectly to the genome. Both the full library

as well as the genome matching reads revealed that 93% of the AGO3-associated sRNAs

are between 20-22nt in length (Figure 3.2A and 3.2B). Additionally, 98% of the sRNAs

have an U as the 5' nucleotide (Figure 3.2E), suggesting a strong preference by AGO3

for both length and starting nucleotide. Only 3% of the mapped reads matched known

noncoding RNAs or transposable elements (Figure 3.2C and 3.2D).

MicroRNA prediction

MicroRNAs were predicted based on the criteria outlined by Tarver *et al.* [29] (Figure

3.3A). To differentiate miRNAs from other small RNAs, all mapped reads were clustered

by genomic location such that within each cluster no reads were more than 200 nt apart,

regardless of strand, to allow identification of putative pre-miRNAs in each cluster.

Genomic sequences for each strand of each cluster were folded using RNAfold to

determine the secondary structure of putative pre-miRNAs [42]. In order to be classified

as a miRNA precursor, a cluster was required to fold into a hairpin, and have no more

than two predominant 5' processing sites. In addition, the main reads were required to

have no more than four mismatches in the opposing arm of the hairpin [29]. Using these

criteria, we identified 39 suitable pre-miRNAs, which produce 46 unique miRNA

sequences (Table 3.1). Of these 46 miRNAs, 28 also had their respective miRNA*

present in the library, while the remaining 18 did not (see Discussion). We found 11

previously reported *C. reinhardtii* miRNAs, and identified 35 novel ones. The 46

miRNAs accounted for 3,475,268 reads (roughly 30%) of the 12,482,999 genome-

mapped reads (Figure 3.2C and Table 3.1). The folding data and matching reads for all of

these clusters can be found in Figure B.3.

The expression of a subset of miRNAs was verified by northern blot analyses

(Figure 3.4A). Intriguingly, an abundant sequence in the small RNA library (based on

read frequency) did not have an identifiable precursor in the Chlamydomonas genome.

Yet, the corresponding small RNA (designated miR B) was easily detected in RNA blots

(Figure 3.4A) and appears to be a functional miRNA-like molecule (see below). Since the

current version of the *C. reinhardtii* genome contains sequence gaps, we hypothesize that

the precursor transcript encoding miR B might be located in an unsequenced region of the

genome. In support of this interpretation, for another small RNA (designated miR C), that

also is detectable in northern blots (Figure 3.4A) and appears to be functional (see

below), the putative pre-miRNA includes a sequence gap that precludes its unambiguous

folding into the expected stem-loop structure (Figure 3.4B).

MicroRNA target prediction

Target prediction for the identified 46 miRNA sequences relied on complementarity

between the miRNAs and putative binding sites on transcripts (Figure 3.3B). The binding

sites were uncovered using RNAhybrid to search against the Chlamydomonas

transcriptome in phytozome version 9.1 [43, 45]. Potential targets were divided into two

categories, depending on the extent of complementarity to a miRNA, those likely to be

regulated through cleavage and those likely to be regulated through translational

repression. Cleavage targets were set to require perfect matching between nucleotides 2-8

of the miRNA (the seed region; Figure 3.1B) and between nucleotides 9-12 (the catalytic

or center region, where AGO-mediated cleavage of the target occurs; Figure 3.1B). In

addition, they could have no more than three mismatches and three GU wobbles in the

remainder of the miRNA/target hybrid, or a gap of more than one nucleotide (Figure

3.3B).

Based on the above criteria, we identified 77 target transcripts but none

corresponded to a perfect miRNA:mRNA match. To provide some validation for these

findings, we examined transcript abundance of the putative targets in an RNAi defective

mutant, Mut-20, and in its parental strain, Maa7-IR44s. Mut-20 contains a deletion of the

gene coding for Tudor Staphylococcal Nuclease 1 (TSN1) (Figure B.4), a protein

previously implicated in RNA interference in metazoans, although its exact molecular

role remains undefined [52, 53]. Intriguingly, in *Trypanosoma brucei* a protein (TbRIF5)

containing tandem Staphylococcus nuclease (SNase) domains physically associates with

a Dicer-like protein [54], suggesting a possible involvement of proteins with SNase

motifs in small RNA biogenesis [55]. Notably, Chlamydomonas Mut-20 displays greatly

reduced levels of all tested small RNAs (Figure 3.4A; YAMASAKI *et al.* 2013) and, thus,

transcripts targeted for direct cleavage by RISC would be expected to show increased

accumulation in this mutant background. Yet, only six of the predicted targets showed at

least a two-fold increase in transcript abundance in Mut-20 relative to the wild type

control, and these changes were not statistically significant at an adjusted *p*-value <0.05.

Interestingly, only one of these six predicted targets has a functional annotation,

corresponding to a respiratory burst oxidase implicated in biotic stress responses.

Putative translational repression targets were set to require perfect matching in the

seed region and no more than three mismatches and three GU wobbles in the 3' end of

the hybrid duplex. However, unlike the pairing interaction for cleavage targets, they were

allowed to have one, but no more than three, mismatches or GU wobble pairs in the

catalytic region (Figure 3.3B). These criteria are more stringent than what has been

experimentally demonstrated as necessary for a reduction in target protein levels in *C.*

reinhardtii, based on tests with reporter constructs [31]. However, despite these stringent

requirements, we identified 506 transcripts potentially repressed at the translational level,

or more than 10 putative target sites per predicted miRNA on average. The great potential

for translation control, taken together with the limited number of putative cleavage

targets that showed increased transcript abundance in Mut-20, poses questions as to the

actual mechanism(s) of endogenous gene regulation by miRNAs in Chlamydomonas;

particularly since even sRNAs derived from inverted repeat transgenes and perfectly

complementary to a target transcript are capable of inhibiting protein synthesis without or

with only minimal mRNA destabilization [36].

sRNA-mediated endogenous gene regulation

To gain insight into how miRNAs and other sRNAs regulate endogenous genes in *C.*

reinhardtii, we examined mRNA abundance, and in one case also protein abundance, for

targets of two small RNAs in Mut-20 and in the Maa7-IR44s strain. We chose to test miR

B, even though it has no identifiable precursor RNA encoded in the Chlamydomonas

genome (see above), because it is the only small RNA for which we could

computationally find an unambiguous, perfectly complementary target transcript (Figure

3.5A), albeit coding for a protein of unknown function (Cre17.g697550). As expected for

RISC-mediated target cleavage, 5'RACE assays on RNA samples isolated from the strain

with a wild type background (Maa7-IR44s) detected *Cre17.g697550* transcripts

predominantly truncated in between the residues hybridizing to nucleotides 10 and 11 of

miR B (Figure 3.5A). Moreover, as anticipated for a mRNA regulated by miRNA-

directed cleavage, the *Cre17.g697550* transcript abundance increased considerably in Mut-20 (Figure 3.5B).

To examine whether Chlamydomonas sRNAs can regulate endogenous target transcripts by translation repression, we chose to test miR C because its predicted binding site in a protein kinase encoding mRNA (*Cre16.g683650*) displays several features similar to those characterized as involved in miRNA-mediated translation inhibition in animal systems [29]. The miR C binding site in *Cre16.g683650* is located in the 3' UTR, partly overlapping the stop codon, and it contains a mismatch with nucleotide 10 of the sRNA, which would preclude AGO-mediated cleavage (Figure 3.6A). As mentioned above, miR C maps next to a gap in the genome and, as a consequence, cannot be folded into a canonical hairpin with the miRNA sequence contained within the double stranded RNA stem, although extending the sequence out in either direction it does form a long hairpin structure (Figure 3.4B). As previously proposed for this kind of Chlamydomonas small RNAs mapping next to sequence gaps [26], miR C could correspond to a genuine miRNA or to a small interfering RNA (siRNA) processed from a very long double stranded RNA precursor. Interestingly, the transcript abundance of the miR C target, *Cre16.g683650*, was reduced in Mut-20 relative to the wild type (Figure 3.6B). In contrast, the protein abundance, detected by immunoblotting, was slightly increased in the mutant background (Figure 3.6C). These observations suggested that *Cre16.g683650* is indeed translationally repressed in wild type Chlamydomonas cells and that this repression is at least partly relieved in Mut-20, as expected for a sRNA/miRNA regulated transcript (see Discussion for details)

Discussion

miRNA populations in C. reinhardtii

The small RNAs associated with AGO3 show a strong bias for a length of ~21 nt and the presence of uracil as the 5'-nucleotide, both for unique as well as redundant reads. The same trends were reported in the original studies describing miRNAs in *C. reinhardtii*, which characterized total cellular small RNA populations [26, 27], but are much more pronounced in the AGO3-associated sRNAs. In particular, uracil was the most prevalent 5'-nucleotide in both of the previous reports, but it was present in less than 50% of the unique reads and a considerable number of small RNAs started with other nucleotides, most notably adenine [26, 27]. Since 80% of the AGO3-associated unique reads (98% of total mapped reads) begin with an uracil (Figure 3.2E), it seems likely that the more variable sRNAs sequenced in studies of total cellular small RNA populations represent those associated with all three Chlamydomonas AGO proteins, implying that AGO1 and/or AGO2 may have different preferences than AGO3 for the first nucleotide of a sRNA. Alternatively, as suggested for animal systems [35], only a fraction of mature sRNAs might be AGO associated in Chlamydomonas and total cellular small RNA populations may also include processed small RNAs that are not ordinarily incorporated into RISC. As a methodological drawback, the 5'-nucleotide bias had a significant impact on the identification of miRNAs* in libraries of AGO3-associated small RNAs, as those miRNAs* that begin with a nucleotide other than U are much less likely to be represented in the library. Because of this technical issue, the presence of the miRNA* sequence in the library was not used as a criteria for selecting canonical miRNAs.

By clustering all of the small RNAs associated with AGO3 based on their genomic locations, we were able to identify putative precursor RNA hairpins and classify the clusters as generating canonical miRNAs or other types of sRNAs, according to the characteristics of the hairpins and how the dominant reads appeared to be processed from them. Using the criteria set forth by Tarver *et al.* [29], we identified 35 novel miRNAs and confirmed 11 of the previously identified miRNAs, which together account for ~30% of the genome mapped small RNAs associated with AGO3 (Figure 3.2C). In principle, this would suggests that although AGO3 might play an important role in miRNA-mediated gene regulation in *C. reinhardtii*, the majority of the sRNAs that are associated with it correspond to other types of small RNAs, particularly those derived from very long hairpins that give rise to multiple products. However, a major caveat is that the Chlamydomonas nuclear genome is not fully sequenced and frequently small RNA sequences appear to lie next to sequence gaps (this work and MOLNAR *et al.* 2007). Thus, for many small RNAs it cannot be unambigously ascertained whether they might represent miRNAs or siRNAs (e.g., Figure 3.4B, miR C). Moreover, almost 40% of the high quality reads did not have a perfect match to the Chlamydomonas genome and were not analyzed further. Yet, at least some of these sequences may represent functional small RNAs, as suggested for miR B (Figure 3.5), and may be encoded in unsequenced regions of the genome. This technical problem currently prevents a comprehensive analysis of small RNA populations in *C. reinhardtii*.

Interestingly, the clusters that do meet the criteria for canonical miRNAs tend to have fewer mismatches between the miRNA and miRNA* than observed for plants or animals, which can have up to four or six mismatches within the duplex, respectively

[29]. In contrast, many of the miRNA encoding duplexes in *C. reinhardtii* show perfect complementary (Figure B.3). As a methodological difficulty, perfect complementarity in the stem of a hairpin encoding miRNA and miRNA* molecules results in both sequences mapping to both strands of the genome in a given cluster. However, mismatches in other parts of the hairpin stem often allow the identification of one strand with significant more reads mapping to it, suggesting that that strand corresponds to the pre-miRNA (Figure B.3). Based on these analyses, the 46 identified canonical miRNAs are encoded almost in equal proportion in intergenic regions of the Chlamydomonas nuclear genome or in introns or the 3'UTR of predicted (many hypothetical) protein genes (Table 3.1)

miRNA targets and mechanism(s) of endogenous gene regulation

Despite similarities in the biogenesis of miRNAs between *C. reinhardtii* and higher plants [4, 56], Chlamydomonas miRNAs do not appear to regulate targets primarily through the plant model of high complementarity between the miRNA and its target, resulting in transcript cleavage [17]. While we identified 77 target transcripts that had perfect complementarity to the corresponding miRNAs from nucleotides two to 12 (both the seed region and the region around the catalytic domain of AGOs; Figure 3.1B) and no more than three mismatches and three GU wobbles in the remaining region, only six of the predicted targets (7.8%) showed at least a two-fold increase in transcript level in the miRNA deficient Mut-20. The Chlamydomonas RNAi machinery does have the capability to operate by target transcript cleavage, as demonstrated with artificial miRNA constructs and reporters [26, 27, 31] and suggested by the analysis of the miR B predicted endogenous cleavage target *Cre17.g697550* (Figure 3.5). However, with the caveat that

some predictions may represent false positives (see below), most predicted endogenous targets do not appear to be subject to mRNA cleavage and degradation, since the anticipated increase in transcript abundance in Mut-20, which is almost devoid of sRNAs, was not observed.

When considering the possibility of miRNA-mediated translational repression in Chlamydomonas, even using complementarity criteria far more strict than necessary for reduction of protein levels in reporter constructs [31], we predicted on average more than 10 targets per miRNA (506 target transcripts for 46 miRNAs). However, it is unclear what fraction of these predicted targets might correspond to false positives. Again, the Chlamydomonas RNAi machinery does have the capability to cause translation repression of target transcripts, as demonstrated with inverted repeat transgenic constructs [36] and suggested by the analysis of the miR C predicted endogenous target, *Cre16.g683650* (Figure 3.6). The *Cre16.g683650* transcript abundance was reduced in Mut-20, suggesting that the mRNA may be stabilized by the presence of a functional RNAi machinery. We have previously demonstrated that transcripts translationally repressed by sRNA-mediated mechanisms remain associated with polyribosomes [36] which may perhaps protect against the action of ribonucleases. Interestingly, despite a decrease in *Cre16.g683650* mRNA amount to ~40% of wild type levels, the corresponding protein was ~20% more abundant in Mut-20 (Figure 3.6), indicating that *Cre16.g683650* is indeed translationally repressed in the wild type parental strain. However, if this case is representative of most endogenous miRNA targets, net changes in protein levels between mutant and wild type strains are fairly small making it very difficult the validation of predicted translation repression targets on a large scale.

Moreover, relaxing the miRNA:mRNA complementarity constraints for translational repression causes the number of predicted targets to rapidly increase, with nearly every gene in the Chlamydomonas transcriptome predicted to be a target by requiring only perfect pairing to the miRNA seed region (data not shown).

For both higher plants and animals, target prediction uses evolutionary conservation of the miRNA binding site to reduce the number of potential false positives [57-59]. However, the recent profiling of small RNAs and miRNAs in the green alga *Volvox carteri*, the closest relative to *C. reinhardtii* with a published genome, indicated that there is very little, if any, conservation of miRNAs between these two species [60]. Other features that are commonly used to reduce the number of potential false positives are the accessibility of the target site and the change of free energy for binding [61], the nucleotide composition of the miRNA and target analyzed with machine learning-base classifiers [62-64], and the presence of certain RNA-binding sites for protein modules overlapping the miRNA binding site [65-68]. Minimum free energy hybridization criteria incorporated into RNAhybrid [43] were used in our searches for targets whereas some of the other criteria would require a larger set of confirmed targets to implement machine learning based analyses and/or knowledge of RNA-binding sites for specific protein domains. Nonetheless, our results suggest, as recently described for reporter constructs [31], that miRNA regulation of endogenous transcript expression in Chlamydomonas may not involve the extensive complementarity that is typical of land plants [17] and may mainly operate, at least for certain targets, by translation repression. Yet, given the limitations outlined above, it remains uncertain what proportion of predicted Chlamydomonas targets is genuine.

Evolutionary role of miRNAs in green algae

In multicellular plants and animals, miRNAs play key roles in maintaining cell/tissue differentiation, and disruption in the production of miRNAs leads to abnormal growth, cancer, or is fatal to the organisms [69, 70]. However, the lack of miRNA conservation among examined green algae (as well as with any other organism; Nozawa, *et al.* 2012), the identification of few predicted targets with known function(s), and the modulatory nature of the regulation of some targets (~20% change in protein levels between wild type and an RNAi defective mutant) raise questions about the biological role(s) of miRNAs in *C. reinhardtii*. Furthermore, the miRNA deficient mutant, Mut-20, has no discernable phenotype when grown under laboratory conditions (data not shown), implying that miRNAs are not essential for cell survival. This is consistent with the observation that many unicellular eukaryotes, particularly those with small nuclear genomes, seem to have lost entirely the RNAi machinery or have retained only a basic set of RNAi components [71-74]. Whereas RNAi does play a role in transposon silencing in Chlamydomonas (CASAS-MOLLANO *et al.* 2008) and possibly in defense responses against (currently unidentified) viruses, defining the function of miRNAs will require additional work. For instance, miRNAs may only have a modulatory role in controlling gene expression under optimal laboratory conditions but they may become important regulators under certain stress conditions

The lack of miRNA conservation with *V. carteri* also raises questions about how these small RNAs are evolving in algae. The two main possibilities are that miRNAs are transient, with a very fast evolutionary turnover, or that all of the miRNAs evolved

independently in the two lineages after their divergence. MicroRNAs as accidental products may explain their limited functional role and justify a putative fast turnover rate in *C. reinhardtii*. As proposed in plants and animals [30, 69], hairpins may randomly form both in introns and in transcribed intergenic regions as the genome evolves. The RNAi machinery will process transcribed RNA hairpins generating endogenous small RNAs and potential miRNAs. However, these hairpins will evolve neutrally unless a processed small RNA confers a selective advantage or disadvantage to the cell. As a result, miRNA precursors could evolve, be processed for a certain amount of time, then accumulate sufficient mutations so that they are no longer processed into small RNAs, and eventually be lost from the genome, that is a pattern of birth-and-death evolution [30]. Despite *V. carteri* being the closest relative of *C. reinhardtii* with a sequenced genome, the two species diverged approximately 200 million years ago [25], possibly enough time for the complete divergence of fast evolving genome encoded miRNAs. Interestingly, while the miRNAs in higher plants that maintain tissue differentiation are highly conserved, those that are species specific often appear to follow this rapid turnover model of evolution [30, 69].

Conclusions

There is much work left to be done to understand what role(s) miRNAs play in *C. reinhardtii*, and why many small RNAs are produced in relatively high quantities yet play a small, seemingly non-essential role for the cells. With the caveat posed by the lack of a fully sequenced Chlamydomonas genome, our results suggest that most of the AGO3-associated small RNAs in *C. reinhardtii* are not canonical miRNAs and that regulation of

endogenous targets by miRNA-mediated cleavage is rare. While miR B, which does result in cleavage, almost completely abolishes expression of its target, many miRNAs predicted to operate by translational repression appear to either stabilize the mRNA target (as for miR C) or have no effect on transcript levels. Moreover, based on the analysis of the miR C target, these miRNAs seem to have only a modulatory effect on target protein levels. Indeed, *C. reinhardtii* miRNAs do not appear to regulate any gene essential for cell survival under laboratory conditions. It is tempting to speculate that the RNAi machinery initially evolved, in the common ancestor of eukaryotes, as a defense response against viruses and other genomic parasites [75] and that regulatory miRNAs evolved at a latter time mostly controlling newly arisen processes such as cell differentiation and development in multicellular organism. In unicellular eukaryotes, miRNAs arising as accidental products of random genome evolution may have no role in regulating ancestral, core cellular functions but may possibly be involved in more recently evolved responses to specific abiotic and/or biotic stresses.

Literature Cited

1.   Kim, V.N., *MicroRNA biogenesis: coordinated cropping and dicing.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 376-85.

2.   Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.

3.   Voinnet, O., *Origin, biogenesis, and activity of plant microRNAs.* Cell, 2009. **136**(4): p. 669-87.

4.   Casas-Mollano, J.A., et al., *Diversification of the core RNA interference machinery in Chlamydomonas reinhardtii and the role of DCL1 in transposon silencing.* Genetics, 2008. **179**(1): p. 69-81.

5.   Zhu, H., et al., *Bidirectional processing of pri-miRNAs with branched terminal loops by Arabidopsis Dicer-like1.* Nat Struct Mol Biol, 2013. **20**(9): p. 1106-15.

6.   Wang, Y., et al., *Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex.* Nature, 2008. **456**(7224): p. 921-6.

7.   Mi, S., et al., *Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide.* Cell, 2008. **133**(1): p. 116-27.

8.   Frank, F., N. Sonenberg, and B. Nagar, *Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2.* Nature, 2010. **465**(7299): p. 818-22.

9.   Frank, F., et al., *Arabidopsis Argonaute MID domains use their nucleotide specificity loop to sort small RNAs.* EMBO J, 2012. **31**(17): p. 3588-95.

10.   Zha, X., Q. Xia, and Y.A. Yuan, *Structural insights into small RNA sorting and mRNA target binding by Arabidopsis Argonaute Mid domains.* FEBS Lett, 2012. **586**(19): p. 3200-7.

11.   Endo, Y., H.O. Iwakawa, and Y. Tomari, *Arabidopsis ARGONAUTE7 selects miR390 through multiple checkpoints during RISC assembly.* EMBO Rep, 2013. **14**(7): p. 652-8.

12.   Wang, Y., et al., *Structure of the guide-strand-containing argonaute silencing complex.* Nature, 2008. **456**(7219): p. 209-13.

13.   Sheng, G., et al., *Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage.* Proc Natl Acad Sci U S A, 2014. **111**(2): p. 652-7.

14.   Rogers, K. and X. Chen, *Biogenesis, turnover, and mode of action of plant microRNAs.* Plant Cell, 2013. **25**(7): p. 2383-99.

15.   Zhou, M. and H. Luo, *MicroRNA-mediated gene regulation: potential applications for plant genetic engineering.* Plant Mol Biol, 2013. **83**(1-2): p. 59-75.

16.   Shen, Y., et al., *Combined small RNA and degradome sequencing reveals microRNA regulation during immature maize embryo dedifferentiation.* Biochem Biophys Res Commun, 2013. **441**(2): p. 425-30.

17.   Liu, Q., F. Wang, and M.J. Axtell, *Analysis of complementarity requirements for plant MicroRNA targeting using a Nicotiana benthamiana quantitative transient assay.* Plant Cell, 2014. **26**(2): p. 741-53.

18.   Carthew, R.W. and E.J. Sontheimer, *Origins and Mechanisms of miRNAs and siRNAs.* Cell, 2009. **136**(4): p. 642-55.

19. Djuranovic, S., A. Nahvi, and R. Green, *miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay.* Science, 2012. **336**(6078): p. 237-40.

20. Fujiwara, T. and T. Yada, *miRNA-target prediction based on transcriptional regulation.* BMC Genomics, 2013. **14 Suppl 2**: p. S3.

21. Zheng, H., et al., *Advances in the Techniques for the Prediction of microRNA Targets.* Int J Mol Sci, 2013. **14**(4): p. 8179-87.

22. Lanet, E., et al., *Biochemical evidence for translational repression by Arabidopsis microRNAs.* Plant Cell, 2009. **21**(6): p. 1762-8.

23. Beauclair, L., A. Yu, and N. Bouche, *microRNA-directed cleavage and translational repression of the copper chaperone for superoxide dismutase mRNA in Arabidopsis.* Plant J, 2010. **62**(3): p. 454-62.

24. Axtell, M.J., *Classification and comparison of small RNAs from plants.* Annu Rev Plant Biol, 2013. **64**: p. 137-59.

25. Chang, T.Y. and B.Y. Liao, *Flagellated algae protein evolution suggests the prevalence of lineage-specific rules governing evolutionary rates of eukaryotic proteins.* Genome Biol Evol, 2013. **5**(5): p. 913-22.

26. Molnar, A., et al., *miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii.* Nature, 2007. **447**(7148): p. 1126-9.

27. Zhao, T., et al., *A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii.* Genes Dev, 2007. **21**(10): p. 1190-203.

28. Ibrahim, F., et al., *Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase promotes their degradation in Chlamydomonas.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3906-11.

29. Tarver, J.E., P.C. Donoghue, and K.J. Peterson, *Do miRNAs have a deep evolutionary history?* Bioessays, 2012. **34**(10): p. 857-66.

30. Nozawa, M., S. Miura, and M. Nei, *Origins and evolution of microRNA genes in plant species.* Genome Biol Evol, 2012. **4**(3): p. 230-9.

31. Yamasaki, T., et al., *Complementarity to an miRNA seed region is sufficient to induce moderate repression of a target transcript in the unicellular green alga Chlamydomonas reinhardtii.* Plant J, 2013. **76**(6): p. 1045-56.

32. Winter, J. and S. Diederichs, *Argonaute proteins regulate microRNA stability: Increased microRNA abundance by Argonaute proteins is due to microRNA stabilization.* RNA Biol, 2011. **8**(6): p. 1149-57.

33. Janas, M.M., et al., *Alternative RISC assembly: binding and repression of microRNA-mRNA duplexes by human Ago proteins.* RNA, 2012. **18**(11): p. 2041-55.

34. Stalder, L., et al., *The rough endoplasmatic reticulum is a central nucleation site of siRNA-mediated RNA silencing.* EMBO J, 2013. **32**(8): p. 1115-27.

35. Flores, O., et al., *Differential RISC association of endogenous human microRNAs predicts their inhibitory potential.* Nucleic Acids Res, 2014. **42**(7): p. 4629-39.

36. Ma, X., et al., *Small interfering RNA-mediated translation repression alters ribosome sensitivity to inhibition by cycloheximide in Chlamydomonas reinhardtii.* Plant Cell, 2013. **25**(3): p. 985-98.

37. Einhauer, A. and A. Jungbauer, *The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins.* J Biochem Biophys Methods, 2001. **49**(1-3): p. 455-65.

38. van Dijk, K., et al., *Monomethyl histone H3 lysine 4 as an epigenetic mark for silenced euchromatin in Chlamydomonas.* Plant Cell, 2005. **17**(9): p. 2439-53.

39. Merchant, S.S., et al., *The Chlamydomonas genome reveals the evolution of key animal and plant functions.* Science, 2007. **318**(5848): p. 245-50.

40. Kapitonov, V.V. and J. Jurka, *A universal classification of eukaryotic transposable elements implemented in Repbase.* Nat Rev Genet, 2008. **9**(5): p. 411-2; author reply 414.

41. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.

42. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms Mol Biol, 2011. **6**: p. 26.

43. Kruger, J. and M. Rehmsmeier, *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.

44. Lopez, D., et al., *Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data.* BMC Bioinformatics, 2011. **12**: p. 282.

45. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1178-86.

46. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.

47. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

48. Brueggeman, A.J., et al., *Activation of the carbon concentrating mechanism by CO2 deprivation coincides with massive transcriptional restructuring in Chlamydomonas reinhardtii.* Plant Cell, 2012. **24**(5): p. 1860-75.

49. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

50. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

51. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

52. Gao, X., et al., *Tudor staphylococcal nuclease (Tudor-SN) participates in small ribonucleoprotein (snRNP) assembly via interacting with symmetrically dimethylated Sm proteins.* J Biol Chem, 2012. **287**(22): p. 18130-41.

53. Li, C.L., et al., *Structural and functional insights into human Tudor-SN, a key component linking RNA interference and editing.* Nucleic Acids Res, 2008. **36**(11): p. 3579-89.

54. Barnes, R.L., et al., *Comparative genomics reveals two novel RNAi factors in Trypanosoma brucei and provides insight into the core machinery.* PLoS Pathog, 2012. **8**(5): p. e1002678.

55. Burroughs, A.M., Y. Ando, and L. Aravind, *New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing.* Wiley Interdiscip Rev RNA, 2014. **5**(2): p. 141-81.

56.   Cerutti, H., et al., *RNA-mediated silencing in Algae: biological roles and tools for analysis of gene function.* Eukaryot Cell, 2011. **10**(9): p. 1164-72.

57.   Willmann, M.R. and R.S. Poethig, *Conservation and evolution of miRNA regulatory programs in plant development.* Curr Opin Plant Biol, 2007. **10**(5): p. 503-11.

58.   Thiebaut, F., et al., *Computational identification and analysis of novel sugarcane microRNAs.* BMC Genomics, 2012. **13**: p. 290.

59.   Allmer, J., *Computational and Bioinformatics Methods for MicroRNA Gene Prediction.* Methods Mol Biol, 2014. **1107**: p. 157-75.

60.   Li, J., Y. Wu, and Y. Qi, *MicroRNAs in a multicellular green alga Volvox carteri.* Sci China Life Sci, 2014. **57**(1): p. 36-45.

61.   Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.

62.   Yang, Y., Y.P. Wang, and K.B. Li, *MiRTif: a support vector machine-based microRNA target interaction filter.* BMC Bioinformatics, 2008. **9 Suppl 12**: p. S4.

63.   Mitra, R. and S. Bandyopadhyay, *MultiMiTar: a novel multi objective optimization based miRNA-target prediction method.* PLoS One, 2011. **6**(9): p. e24583.

64.   Kurubanjerdjit, N., et al., *Prediction of microRNA-regulated protein interaction pathways in Arabidopsis using machine learning algorithms.* Comput Biol Med, 2013. **43**(11): p. 1645-52.

65.   Didiano, D. and O. Hobert, *Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions.* Nat Struct Mol Biol, 2006. **13**(9): p. 849-51.

66.   Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

67.   Kedde, M., et al., *RNA-binding protein Dnd1 inhibits microRNA access to target mRNA.* Cell, 2007. **131**(7): p. 1273-86.

68.   Didiano, D. and O. Hobert, *Molecular architecture of a miRNA-regulated 3' UTR.* RNA, 2008. **14**(7): p. 1297-317.

69.   Fahlgren, N., et al., *MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana.* Plant Cell, 2010. **22**(4): p. 1074-89.

70.   Sheng, Y. and C. Previti, *Genomic features and computational identification of human microRNAs under long-range developmental regulation.* BMC Genomics, 2011. **12**: p. 270.

71.   Drinnenberg, I.A., G.R. Fink, and D.P. Bartel, *Compatibility with killer explains the rise of RNAi-deficient fungi.* Science, 2011. **333**(6049): p. 1592.

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

73.   Nakayashiki, H., N. Kadotani, and S. Mayama, *Evolution and diversification of RNA silencing proteins in fungi.* J Mol Evol, 2006. **63**(1): p. 127-35.

74.   Berezikov, E., *Evolution of microRNA diversity and regulation in animals.* Nat Rev Genet, 2011. **12**(12): p. 846-60.

75.   Cerutti, H. and J.A. Casas-Mollano, *On the origin and functions of RNA-mediated silencing: from protists to man.* Curr Genet, 2006. **50**(2): p. 81-99.

76.	Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
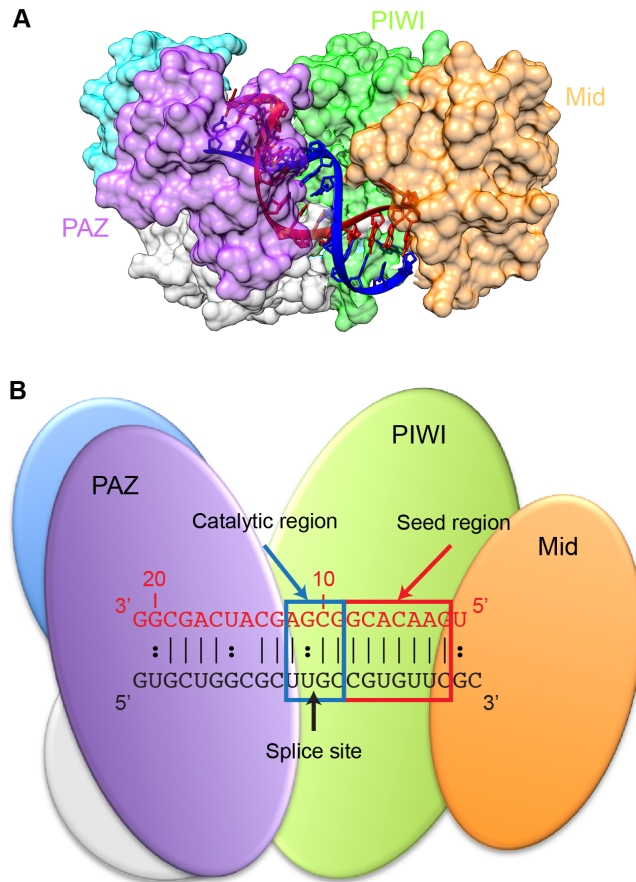
Figure 3.1: Structure of a canonical miRNA and its target transcript within an argonaute (AGO) protein

(A) Structure of the *Thermus thermophilus* RISC complex with an miRNA (in red) and a perfectly complementary target RNA (in blue), based on the 3D structure (PDB ID 4NCB, visualized using Chimera) [13, 76]. (B) Illustration of the duplex between the miRNA and its target RNA sequence. The miRNA is shown in red, and is oriented 3' to 5' with the 5' most nucleotide bound to the Mid domain of AGO. The target sequence is shown in black, with the cleavage site indicated between nucleotides 10 and 11 of the miRNA and located within the AGO PIWI domain. The catalytic and seed regions are indicated by surrounding blue and red boxes, respectively.

A

```
            Small RNAs

          Map to Genome

No ← ≤5 locations? → Yes
Discard

         Cluster small RNAs

        Fold genomic sequence
            for clusters

No ←  Form a hairpin? → Yes
Other sRNA

        Identify predominant
              reads

No ← ≤3 predominant reads? → Yes
Other sRNA

No ← <4 mismatches in main reads? → Yes
Other sRNA

              miRNA
```

B

```
              miRNAs

        Search transcriptome
          using RNAhybrid

        Analyze predicted
           binding sites

No ← Perfect match in seed region (nt 2-8)? → Yes
Discard

No ← ≤3 GU wobbles or mismatches in nt 9-12? → Yes
Discard

No ← ≤3 GU wobbles or mismatches in 3' end? → Yes
Discard

No ← 1-3 GU wobbles or mismatches in nt 9-12? → Yes
Cleavage

         Translational repression
```

Figure 3.2: Criteria and flowcharts for predicting canonical miRNAs and their targets

Criteria and flowcharts for predicting canonical miRNAs (A) and their targets (B).
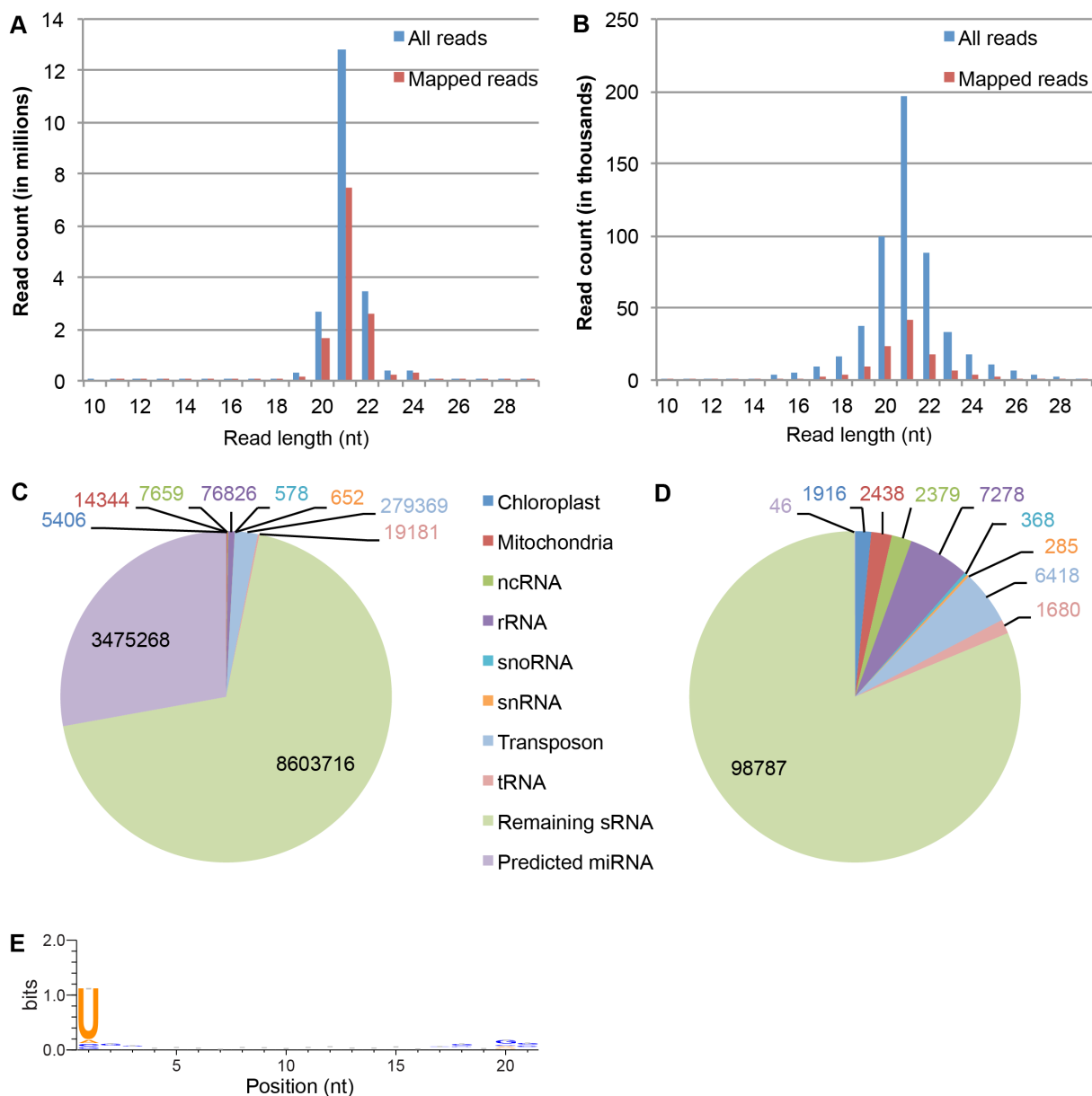
Figure 3.3: Classification of AGO3-associated Chlamydomonas small RNAs

(A) Size distribution of small RNAs associated with AGO3, displaying the frequencies for all reads in the library (blue) and for the reads that map to the genome (red). (B) Size distribution of unique small RNAs associated with AGO3. (C) Abundance (read counts) of all AGO3-associated small RNAs matching different categories of sequences. (D)

Abundance (read counts) of unique AGO3-associated small RNAs matching different

categories of sequences. (E) Nucleotide composition by position for all unique small

RNAs 21nt in length, represented in the sequence logo [41].
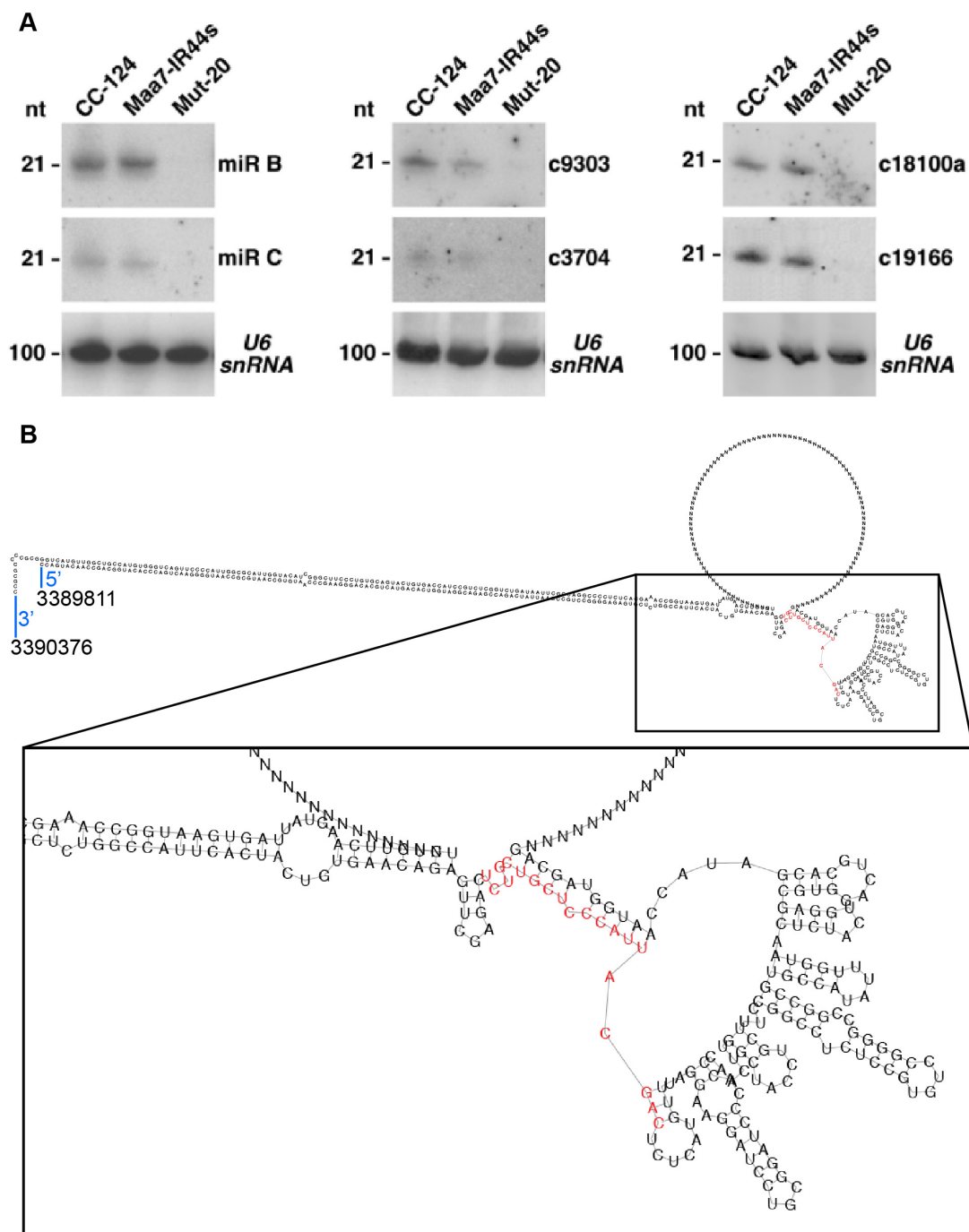
Figure 3.4: Detection of Chlamydomonas miRNAs by northern hybridization and predicted precursor structure for miR C

(A) Northern blot analyses of small RNAs isolated from the indicated strains and detected with probes specific for each miRNA. The same filters were reprobed with the

U6 small nuclear RNA sequence as a control for equivalent loading of the lanes. CC-124,

wild-type strain; Maa7-IR44s, CC-124 transformed with an IR transgene designed to

induce RNAi of *MAA7* (encoding tryptophan synthase β subunit); and Mut-20, *TSN1*

deletion mutant. (B) The putative secondary structure of the sequence surrounding miR C

was obtained with RNAfold. The miRNA sequence is shown in red. The pre-miRNA

structure is taken from nucleotides 3389811 to 3390377 of the minus strand of

Chromosome 4. This corresponds to 152 bases upstream and 393 bases downstream of

the miRNA sequence, and is based on boundaries of the small RNA clustering. The

entire pre-miRNA is contained within an intron of SRR18 (Cre04.g227500).
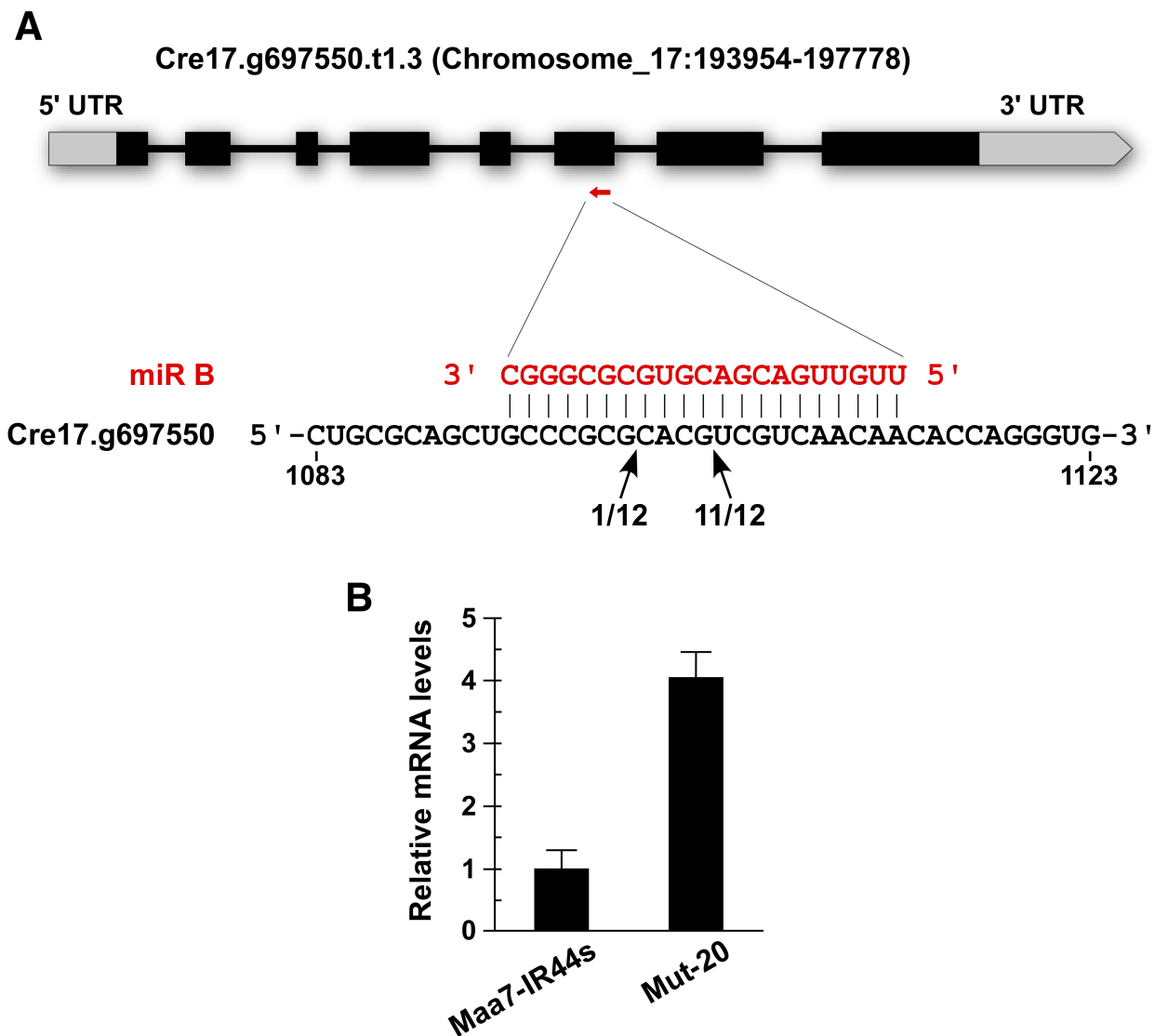
Figure 3.5: Analysis of the miR B target *Cre17.g697550*

(A) Diagram of the *Cre17.g697550* gene (encoding a predicted protein of unknown function) indicating the miR B-binding site within the coding sequence. Hybridization of miR C to the target site is shown below. The arrows and numbers indicate the position and frequency of 5'-termini of truncated target mRNAs examined by 5' RACE PCR. (B) Quantitative RT-PCR analysis of *Cre17.g697550* transcript levels in the indicated strains. Values are means +/- SD of three independent experiments.
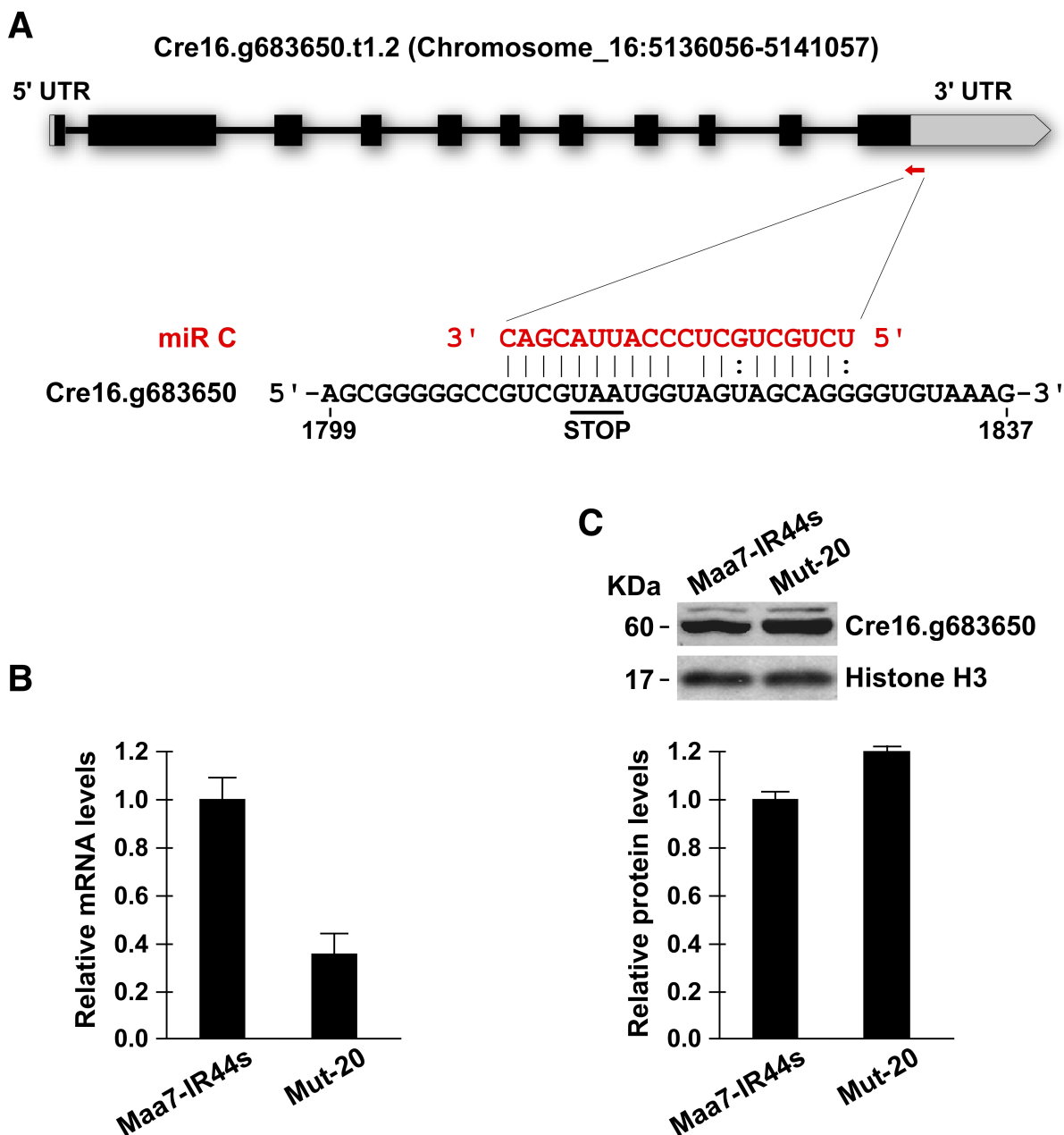
Figure 3.6: Analysis of the miR C target *Cre16.g683650*

(A) Diagram of the *Cre16.g683650* gene (encoding a predicted protein kinase) indicating

the miR C-binding site overlapping the stop codon. Hybridization of miR C to the target

site is shown below. The mismatch to nucleotide 10 of the miRNA would prevent

cleavage of the target mRNA. (B) Quantitative RT-PCR analysis of *Cre16.g683650*

transcript levels in the indicated strains. Values are means +/- SD of three independent

experiments. (C) Immunoblot analyses of Cre16.g683650 and histone H3 (loading

control) protein levels. The graph represents densitometric quantitation of the

Cre16.g683650 immunoblot signals (normalized to H3). Values are means +/- SD of

three independent experiments.

Table 3.1: AGO3-associated candidate miRNAs in *C. reinhardtii*

Cluster[a]	Alias[b]	miRNA	O[c]	Chr[d]	Start[e]	Location[f]	Gene[f]	PFam A[f]	Count[g]
1130	cr.02052	TTCGAAAAGTTGCACAGCGCC	+	1	5106348	Intron	Cre01.g035500	PI3K_C2, PI3Ka, PI3_PI4_kinase	631
2675a	cre-miR1174.1	TACCGGGCGTGGGGAGGGCAGG	-	10	3399959	3' UTR	Cre10.g444300		896822
2675b	cre-miR1174.2	TTACGGCTCCTTCTTATCGGC	+	10	3399983	Intergenic			647884
			-		3399872	3' UTR	Cre10.g444300		
3318		TCAATTGCTCTTTCTGATG	-	10	6282547	3' UTR	Cre10.g465000	Pkinase	3165
3704*		TGCACGTTGGTAAACAGTCCC	-	11	1442258	3' UTR	g11660		20465
3708		TAGTGCCAGCGCGCTCTCGGC	-	11	1448250	Intergenic			3004
3772*		TGGAAAATCGAGGCCGCTGGC	+	11	1713373	5' UTR	g11713		6521
			-		1713298	Intergenic			
3911*		TGGATGGTCGAAACCGCGTACA	-	11	2529518	3' UTR	Cre11.g476000		1595
3965*		TGGAGACTGTACGAGAGGCCA	-	11	2856089	3' UTR	Cre11.g477700		1237
4832a*		TTCGAGGTCGGCTGGAAGGTC	+	12	2444603	Intergenic			1847
4832b*		TGAACATAAAGCAAGACTTC	+	12	2444685	Intergenic			881
			-		2444818	5' UTR	g12428		
5094*		TCAATAACGGACCTGGAGGAC	+	12	3183096	Intergenic			1181
			-		3183213	3' UTR	g12598		
5858	cr.02116	TTCAGGTAGCGGGACCAGGTG	-	12	6402228	Intron	g13222	zf-CW, PWWP, Chromo, SNF2_N, Helicase_C	130801
6589*		TCTGACCGGTCGGTGCTGCCG	+	12	9542589	Intron	g13796		4758
7085a		TGACTCTCACTCCTACTCGGC	+	13	2001158	Intergenic			37202
			-		2001092	3' UTR	g14171		
7085b		TTCCACCTGACTGCTTGCTGA	-	13	2001069	3' UTR	g14171		12380
8319	cr.01824	TCCACGTTCTCGCCGCGCAGG	+	14	1950136	Intergenic			6989
8563a		TGCGGGGCCCGCAGGTAGCTGC	-	14	321886	Intron	Cre14.		704

ID	miRNA	Sequence	Strand	Chr	Position	Region	Gene	Protein	Distance
							g629200		
8563b	miR910	AGCAGCGTCGGGCTCGACCGC	-	14	3218785	Intron	Cre14.g629200		625
9303		TGGATGTTGCTTGCTGGATAC	+	16	185156	Intergenic			183618
			-		185089	3' UTR	Cre16.g694950	Pkinase	
9770a*	miR1157	TTGATGTCCGCTGGGAGCTGG	+	16	2231294	3' UTR	Cre16.g658782		7159
			-		2231401	Intergenic			
9770b*		TGCGAGCTGCGGAATGGCGAC	+	16	2231221	3' UTR	Cre16.g658782		1753
9770c*		TGCGTAGGACGTGGTAGGCGC	-	16	2231112	Intergenic			1908
11846		TGAGTGGTGGTCGACGGCAAT	+	17	3137574	3' UTR	Cre17.g721553		3013
12364		TCGGAGAAGCGGGTAGCTGA	+	17	5152874	Intergenic			122990
			-		5152806	5' UTR	g17790		
12551a		TAAACAGACAAGGCGACCGACA	-	17	6144115	3' UTR	g17937		18839
12551b		TCGCCTTGTCTGTTTATGTGG	+	17	6144121	Intergenic			9949
			-		6144182	3' UTR	g17937		
12615*		TGCAAAATCAAGACCGGGGGA	+	17	6674056	Intergenic			1243
			-		6674164	Intergenic			
12905	cr.02798	TGGCCGTACTACTATTGTC	-	2	739369	3' UTR	Cre02.g078500		5494
14705*		TCGGAAAGGAATCCAACGGCC	+	2	9101230	Intron	g2826		14785
			-		9101131	Intergenic			
14711*		TGAACACGGCGAGCATCAGCGG	+	2	9121266	3' UTR	g2830		1231
14712		TGCTTGCGCCCTCTAGCCGTC	+	2	9129575	3' UTR	g2831		8955
			-		9129507	Intergenic			
14719		TCGTGCTCGTCATCCCCTCG	+	2	9173370	3' UTR	g2836		8513
			-		9173305	Intergenic			
16411*		TGCGGTGAATGTGAATGATGG	+	3	6573968	3' UTR	Cre03.g195950		4117
			-		6573895	Intergenic			
17284*		TGACCACCCTGCAGCTGACGC	-	4	1804340	Intron	g4741	Exostosin	2415
17620	miR1153, cr.02052	TGGGCCATCGTATTACTATCAG	+	4	3100623	Intron	Cre04.g225700		377890
17755*		TAACTTAGTCGTCACAAGGCG	+	4	3694795	Intergenic			1588
18100a		TCGGTCAGCATTTCGTTTGG	+	5	935314	Intergenic			91516

18100b		TACCCGGAGTGGACGTCTTGC	+	5	935397	Intergenic			15493
19166		TTGGGCGGCGTTGTAAGATT	+	6	2201623	Intergenic			86172
			-		2201668	Intron	g5972	WD40	
19538	miR1162, cr.02386	TGTTGTAGTAGTTTAGCCCTGC	+	6	3067434	Intron	Cre06. g274550	Pkinase, Pkinase_Tyr	307549
20399		TAGAGCTCGAAGAACTTGGGA	+	6	6776173	Intergenic			362932
20462	miR1152, cr.01969	TAAGAAGGTGCGCTGTCTTGA	-	6	7063794	Intron	g6871	UBA_e1_thiolCys , UBACT, ThiF, UAE_UbL	1701
21642		TCGACGCGGTGATGGGCCTGG	+	7	2713524	CDS	Cre07. g331114		2702
22587		TGGCTTTCGTCGGTCCTAGG	+	8	121861	Intergenic			26418
			-		121923	3' UTR	g8351		
25074		TTGGACGCGGACCCGGCGCAG	+	9	6518859	Intron	g10105	TENA_THI-4	26633
			-		6519021	Intergenic			

[a]Asterisks in the cluster number indicate that no miRNA* was identified in the small RNA library. Letters in the cluster number indicate the existence of more than one predominant product.

[b] Names for previously identified miRNAs are taken from Molnar, *et al.* (2007), Zhao, *et al.* (2007), and Yamasaki, *et al.* (2013)

[c]Strand orientation of the mapped reads. In cases where the miRNA is processed from a region of the hairpin where the two arms are perfectly complementary, the miRNA will map to both strands within the same locus. In these cases, the coordinates and region are shown for both strands (see Figure B.3).

[d]Chromosome where the cluster is located.

[e]Start location for the miRNA (5' nucleotide), relative to the + strand. For reads that map to the – strand, this represents the position of the miRNA 3' nucleotide.

[f]If the miRNA maps to a gene, the location of the miRNA within the gene is shown along with information about the gene itself including conserved Pfam A domains.

[g]Reads counts for each miRNA are shown in absolute numbers and are not normalized by library size.

CHAPTER 4- Changes in miRNA populations between heterotrophic, photoautotrophic, and nitrogen deplete conditions for the green alga *Chlamydomonas reinhardtii*

Introduction

While the biological role of microRNAs (miRNAs) in unicellular organisms such as *Chlamydamonas reinhardtii* remains unclear, one possibility is that they are involved in the cell's response to environmental stress similar to higher plants [1-8]. In these stress conditions, such as nutrient deprivation (including nitrogen), a subset of the overall miRNA population is differentially or uniquely expressed [6-9]. These condition specific miRNAs often have very few, if any, targets within the metabolic pathways involved in the stress response. They instead generally target transcription factors, RNA and DNA binding proteins, enzymes involved in post-translational modification of proteins, and protein kinases, making the overall impact of the differential expression difficult to determine. For miRNAs differentially expressed during nitrogen starvation in *Arabidopsis thaliana*, only a handful of experimentally validated targets across all of the differentially expressed miRNAs identified were metabolic enzymes, two of which (a fucosyltransferase and haloacid dehalogenase-like hydrolase) are involved in carbohydrate metabolism and the rest, interestingly, appear to be involved in copper homeostasis [7].

An analysis of putative miRNAs differentially expressed during sulfur starvation in *C. reinhardtii* similarly had very few metabolic enzymes predicted as targets [3]. However, given the difficulty in predicting functional targets for miRNAs in *C.*

reinhardtii makes the reliability of the predictions questionable without further experimental analysis (See Chapter 3). Indeed, several of the predicted targets that are essential for the cell's acclimation to sulfur deprivation are upregulated when the miRNAs predicted to target them are also upregulated [3, 10]. This expression pattern makes it unlikely that these targets are regulated by cleavage as predicted, though these genes could still be targeted by translational repression, especially under the model that these miRNAs modulate final protein levels rather than silencing their targets.

Studies examining miRNA populations between species have found that new miRNAs can arise spontaneously, culminating in a plasticity in the overall miRNA populations [11-15]. In plants, this can be seen in the number of miRNAs that are either species-specific or that are contained in only a few very closely related species [11, 12]. Within Arabidopsis, there is evidence that these sequences can arise from random sequences that form small or partial hairpins transcripts [13]. The spontaneous production of novel miRNAs from random sequences could account for a large number of unique miRNAs sequences between *A. thaliana* and *A. lyrata*, roughly 13% of the total miRNA sequences for each species, despite their relatively recent speciation [14]. As these novel miRNAs arise in plants, they are initially very lowly expressed, while miRNAs conserved between species generally show higher expression levels [16].

Here, in order to identify a possible role of miRNAs in *C. reinhardtii*, we investigate the changes in overall population of miRNAs and their targets of *C. reinhardtii* cells grown in heterotrophic conditions, in photoautotrophic conditions with a source of nitrogen, and in photoautotrophic conditions without a source of nitrogen to determine how miRNAs are involved in the response to nitrogen starvation. Our results

suggest that, as in Arabidopsis, very few of the predicted targets are metabolic enzymes. The low expression of condition-specific miRNAs is similar to newly evolved miRNAs identified in higher plants, while the miRNAs expressed across all conditions tended have a much higher expression.

Materials and Methods

Transgenic strains, mutants, and culture conditions

Chlamydomonas cells were routinely grown photoheterotrophically in Tris-acetate-phosphate (TAP) medium [17] or photoautotrophically in nitrogen-replete high salt (HS+N) medium (Sueoka, 1960). Cells from the nitrogen-replete HS medium were collected by centrifugation and resuspended at a density of 5 x 10^5 cells mL^{-1} in regular HS media or in the same media lacking nitrogen (HS-N). Strain Maa7-IR44s, containing an inverted repeat transgene targeting the 3' UTR of the *MAA7* gene (encoding Tryptophan synthase β subunit), has been previously described [17]. Mut-20, deleted for *Tudor Staphylococcal Nuclease 1* (*TSN1*) and almost devoid of miRNAs, was obtained in an insertional mutagenesis screen designed to isolate mutants defective in RNAi-mediated translation repression [17]. For the isolation of AGO3-associated small RNAs, we fused the FLAG tag [18] to the N-terminal end of Chlamydomonas AGO3 (g16859). This construct was placed under the control of *psaD* regulatory sequences and transformed into Maa7-IR44s (Figure B.1). A similar construct was made fusing the FLAG tag to the coding sequence of the bleomycin gene and transformed into Maa7-IR44s.

Isolation of AGO3-associated small RNAs, library preparation, and sequencing

FLAG-tagged AGO3 was affinity purified from a cell lysate as previously described for a TAP-tagged protein [19]. RNAs associated with AGO3 were purified with TRI reagent (Molecular Research Center) and contaminant DNA was removed by DNase I treatment (Ambion) (MA *et al.* 2013). Construction of cDNA libraries and Illumina sequencing were then carried out as previously reported (IBRAHIM *et al.* 2010).

Small RNA mapping and profiling

Sequenced reads were first mapped to version 10 of the phytozome *C. reinhardtii* genome [20], by using version 3.02 of Novoalign (www.novocraft.com) with the miRNA flag and with a score threshold of 15. Mapped reads were filtered to remove those showing alignments with gaps or mismatches as well as those that mapped to more than five locations in the genome. The expression level in counts per pillion (CPM) for each read was determined by the formula:

$$CPM = \frac{10^6 C}{N}$$

where *C* is the number of reads in the library and *N* is the total number of mapped reads in the library. Genome mapped reads were also classified as matching the chloroplast or the mitochondrial genomes, ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs,) other non-coding RNAs (ncRNAs), or transposable elements. The *C. reinhardtii* chloroplast and mitochondrial genomes were taken from Genbank (accession numbers BK000554 and NC_001638.1, respectively) along with the rRNAs, snRNAs, snoRNA, tRNAs, and ncRNAs sequences. Transposons sequences were taken from Repbase [21].

Genomic clustering of sRNAs and miRNA identification

After removing those reads that mapped to known non-coding RNA categories,

transposons, the chloroplast or the mitochondrial genomes, potential small RNA (sRNA)

reads were processed following the steps illustrated in Figure 3.2A. The reads were first

clustered by genomic location such that there was no more than 200 nt between adjacent

reads, regardless of strand. The reads on both strands in the same genomic location were

placed in the same cluster. These clusters were analyzed further to differentiate pre-

miRNA structures from other sRNA structures. The genomic sequence for each strand of

the cluster was folded using version 2.1.5 of RNAfold from the Vienna RNA package

[22]. Clusters with a genomic sequence < 50 nt or >1000 nt in length were excluded from

folding, as they were deemed unlikely to code for canonical miRNAs [23]. Clusters

containing sequence gaps (*i.e.*, unsequenced genomic regions) were also excluded since

the secondary structures of these regions cannot be unambiguously predicted. The

secondary structures were parsed to determine if they fold into a hairpin. However,

clusters whose highest expressed reads had more than four mismatches with the opposing

arm of the hairpin were removed from further analyses, following the criteria proposed

by Tarver, Donoghue [23] for the annotation of plant miRNAs where complementary

base pairing between the mature miRNA and miRNA* sequences is fairly precise with

fewer than four mismatches and any asymmetric bulges involving only one or two

nucleotides. The putative miRNA* sequence for the highest expressed read of each

cluster was determined using the hairpin fold and the genomic sequence, with a two

nucleotide offset to allow for the 3'-overhang in the miRNA/miRNA* duplex. The

remaining clusters were filtered using a classifier based on a support vector machine

(SVM) [24]. The descriptors used for the SVM classifier are: the accuracy of processing

of the 5'-end of the predominant reads, the frequency of predominant reads present, size

of the pre-miRNA hairpin structure, and the extent of complementarity between the two

arms of the hairpins, all of which are based in the criteria for canonical miRNAs in

Tarvers, *et al*. [23]. The accuracy of the 5' processing is determined by the percentage of

reads that overlap with the potential miRNA sequences that start at the same 5'

nucleotide as the potential miRNA sequence. The classifier was trained on the manually

curated list of 46 positive and 154 negative miRNAs generated in Chapter 3. The

classifier was generated using version 3.18 of libsvm [24] with a radial basis function

kernel, with the parameters $c = 8.0$ and $g = 0.03125$, which were determined by a grid

search over the parameters. This classifier was verified with a 4-fold cross validation and

had an overall accuracy rate of 87%, with a sensitivity of 0.734, a specificity of 0.932,

and a MCC of 0.690. These predictions were used as a pre-filter for classification and

both the positive and negative predictions were further manually curated to ensure they

remain consistent with the predictions made in Chapter 3.

MicroRNA target prediction

The process used for the prediction of miRNA targets is illustrated Figure 3.2B. Potential

miRNA-binding sites were determined by searching version 10 of the *C. reinhardtii*

transcriptome using version 2.1 of RNAhybrid [25]. This algorithm returns the most

energetically favorable pairing for each miRNA/target pair. The obtained hybrid pairs

were filtered based on the number and location of gaps, wobbles, and mismatches to limit

the predictions to only those that could regulate targets in a projected manner. For

cleavage targets, this search required perfect matching for nucleotides 2-8 (the seed

region, Figure 3.1B) and nucleotides 9-12 (the catalytic or center region, Figure 3.1B),

and no more than three GU wobbles and three mismatches or a gap of more than one

nucleotide in the remaining sequence. For translational-repression targets, the constraints

for the catalytic region were relaxed to allow up to three mismatches or wobbles, with the

constraints for the seed-region and the 3'-end pairing remaining the same. Additionally,

translational-repression targets needed at least one mismatch or wobble in the catalytic

region to keep the two sets of predicted targets non-overlapping. MicroRNA-encoding

transcripts, which may also be targeted for degradation by their encoded miRNAs, were

removed from the analysis.

Putative roles of the predicted targets were evaluated by using the annotations of

Chlamydomonas genes (if available) as well as conserved protein domains. Manual

annotations were taken from the Algal Functional Annotation Tool [26] which are based

on the protein annotations uploaded to Phytozome [27].

Total RNA isolation

Total RNA was isolated with TriReagent (Molecular Research Center, Inc.) in

accordance with the manufacturer's instructions from *C. reinhardtii* grown under N-

replete and N-deplete conditions. Agarose gel electrophoresis (1.5%) was used to monitor

the quality of the extracted RNA by checking the integrity of the RNA bands. The

purified total RNA concentration was measured using a Thermo Scientific NanoDrop

2000c spectrophotometer. Total RNA isolated from these cells was used in

transcriptomic, real-time quantitative polymerase chain reaction, and semi-quantitative polymerase chain reaction experiments.

Transcriptome analysis

Transcriptomic changes in *C. reinhardtii* grown photoautotrophically under N-replete and N-deplete conditions were measured using RNAseq. Experiments were performed independently twice and libraries were sequenced with the Illumina GAIIx analyzer. One library was sequenced at the Joint Genomic Institute (JGI), the other library was sequenced on the Univeristy of Nebraska-Lincoln campus. The Illumina reads from both experiments were mapped to the Augustus v5.0 transcript models for *C. reinhardtii* (available from http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html). The reads were first trimmed to 36 nucleotides in length to remove low quality 3′-ends before mapping. The alignment was performed using Burrows-Wheeler Aligner (BWA; v0.5.7) [28] with a seed length of 25, allowing 2 mismatches. A perl script (A. Voshall, unpublished) was used to ensure that only reads that uniquely matched with a single transcript were counted. Raw gene counts were determined by counting the number of reads aligned to each transcript. To examine transcript abundance, gene expression was analyzed as Reads per Kilobase per Million (RPKM), which normalizes the read counts based on both transcript length and total number of reads, using the following formula:

$$RPKM = \frac{10^9 C}{NL}$$

where C is the number of reads mapped to the transcript, N is the total number of mapped reads in the library, and L is the length of the transcript in nucleotides (Mortazavi *et al.,*

2008). To determine the change in expression, each of the three time points during the N-deplete condition were compared pairwise with the control (N-replete) condition for each experiment. Differences in gene expression were examined as log_2(Fold Change), where "Fold Change" refers to the ratio of RPKM values, at each time point, between treatment and control. Statistical analysis of the data was performed using the DESeq package (version 1.18) [29]. The q-values (adjusted p-values from DESeq) for each time point compared to time 0 were calculated using both replicates. All transcripts that have at least one time point with a q-value ≤ 0.05 and at least a 3-fold change in expression (a log_2(Fold Change) of 1.585) are considered significant. The Augustus v5 transcript IDs were converted to the current Phytozome 10 transcript IDs using the transcript name conversion file on the Phytozome 10 website (http://genome.jgi.doe.gov/ pages/dynamicOrganism Download.jsf?organism=PhytozomeV10) along with the gene name, description, and PANTHR annotations. MapMan annotations for the genes targeted miRNA in HS+N and HS-N conditions were taken from version 1.0 of the MapMan Chlamydomonas annotations (http://mapman.gabipd.org/web/guest/mapmanstore), which are based on the Phytozome 9 transcript models [30]. The first two categories of the bin numbers were used to determine the functional category for the target genes.

Arabidopsis miRNA and target prediction

The miRNA prediction and expression information for *A. thaliana* and *A. lyrata* were taken from Ma *et al.* [31]. The miRNA and gene expression levels for each biological replicate (given in raw read counts) were converted to CPM using the formula given

above using the total library size reported by the authors. The average expression for *A. lyrata*, which consisted of three libraries, was the average of the CPMs calculated for each miRNA across all three libraries. Species-specific miRNAs were determined by comparing the miRNA names present between the datasets for *A. thaliana* and *A. lyrata* and cross-referenced with the in-text results. Targets for the species-specific miRNAs were predicted using version 1.6 of Target Finder [32], searching against version 10 of the TAIR *A. thaliana* transcriptome (taken from Phytozome 10) and the Phytozome 10 version of the *A. lyrata* transcriptome [27]. All of the predictions were taken as-is from each reference, with no manual curation or filtering.

Comparison of miRNA expression and number of predicted targets

All of the statistical analyses were performed in R using standard libraries. The histograms of miRNA expression were generated by binning each miRNA into expression categories of 500 CPM based on their average CPM. miRNAs that had no identifiable targets or only targeted the transcript that they are derived from were excluded from all further analyses as these miRNAs would not have an evolutionary constraint on their expression level. A cutoff of 500 CPM was used to differentiate highly expressed ($\geq$500 CPM) and lowly expressed (<500 CPM) miRNAs, which was chosen due to the the distributions of expressions and target number observed in *C. reinhardtii*. The numbers of targets between highly expressed and lowly expressed miRNAs were compared using the Welch's *t*-tests as well as Wilcoxon Rank Sum (MWW), and Cohen's *d* was used to determine the effect size for the two groups.

Results

Changes in miRNA population between conditions

Using the criteria established in Chapter 3, we identified 102 miRNA sequences across

three growth conditions: TAP media, HS+N, and HS-N (Figure 4.1, Table 4.1). These

predictions include 45 of the 46 miRNAs identified in the photoheterotrophic media

(TAP) used in Chapter 3 as well as 57 additional miRNA sequences. The sequence that

was present in the Chapter 3 but removed here was misclassified in Chapter 3 due to

human error in determining the precision of the 5' processing and did not meet the

criteria specified. Nine additional miRNAs were identified in TAP that were not

identified in Chapter 3, which meet all of the criteria specified but were misclassified in

Chapter 3 due to human error in identifying predominant reads in the clusters. In total,

54 of the 102 miRNAs identified here are present in TAP. Of the 48 additional miRNAs

identified in this analysis with photoautotrophic (HS) media, two are present in miRBase

(cre-miR918 and cre-miR9897) and the remaining 46 are novel predictions. The majority

(60 of the 102) miRNAs are shared between at least two conditions, with roughly a

quarter (29 of 102) common between all three conditions. While Chlamydomonas grown

photoheterotrophically in TAP media had the most condition specific miRNAs (19 of

102), there are more many more miRNAs (48 of 102) that are specific to photoheterophic

conditions, of which 25 are common to Chlamydomonas grown in photoheterotrophic

conditions regardless of the presence of a nitrogen source than specific to TAP. As

expected, there are no miRNAs found in both TAP (which contains a nitrogen source)

and HS-N media that are not also found in HS+N media.

While the predicted miRNA population varies between conditions, only one predicted miRNA (miR_t42) is completely missing from libraries for the other conditions (Table 4.1). For the remaining 101 miRNA, the pre-miRNA structure is still transcribed and processed, but no longer meets the criteria to be classified as a miRNA. For most of the sequences, this occurs because the 5' processing of the predominant reads no longer satisfies the requirements, which becomes more common as the read count for the predominant reads decreases (data not shown). In the case miR_t42, the pre-miRNA is also transcribed in all conditions, with the remaining predominant reads still present. However, it only meets the requirements for miRNA classification within TAP media.

The expression for miRNAs also varies between conditions, including clusters that meet the criteria for miRNA classification in all three conditions (Table 4.1). Most of these variations are within an approximately 2-fold change between conditions. Although some miRNAs show differences in expression levels, *e.g.*, 3-fold change with miR1153, without more replicates, the results are inconclusive.

Changes in predicted miRNA targets between conditions

Changing miRNA populations in each condition result in the corresponding changes in the predicted targets for both cleavage and translational repression for those conditions (Figure 4.2A and 4.2B). Differences in the number of targets predicted for each miRNA lead to large variations in the number of condition specific targets identified, ranging from an average of 1.6 cleavage targets and 11 translational repression targets per miRNA in TAP specific miRNAs (30 targets and 209 targets respectively for 19 miRNA) to an average of 5.2 cleavage targets and 12.4 translational repression targets for HS+N

specific miRNAs (73 targets and 174 targets, respectively, for 14 miRNAs) (Table 4.1).

The average number of targets for miRNAs exclusive to HS+N and HS-N media is

skewed heavily by miR_t59, which alone accounts for roughly half of the predicted

targets for both modes of regulation (48 of 90 cleavage targets and 242 of 444

translational repression targets). Excluding this one miRNA, the number targets per

miRNA is comparable to that of the TAP-specific and miRNAs common to TAP and

HS+N, with an average of 1.8 cleavage targets and 8.4 translational repression targets per

miRNA. Three of the predicted miRNAs (one TAP-specific, and two common to HS+N

and HS-N) had no predicted targets for either mode of regulation.

As shown in Figure 4.3A, the vast majority of the predicted miRNAs have low

levels of expression (lower than 500 CPM), and these miRNAs tend to have a higher

number of predicted targets than those with higher expression (Figure 4.4 and Table 4.1).

This trend is most visible in the extreme cases. While the most highly expressed miRNAs

have no more than 4 cleavage targets and 9 translational repression targets, those that

have extremely low expression were predicted to have up to 48 cleavage targets and 242

targets for a single miRNA. When the miRNA population was divided into two

categories: low expression miRNAs (average expression less than 500 CPM) and high

expression (average expression greater or equal to 500 CPM), the average number of

targets per miRNA was significantly different (Figure 4.6, blue boxes). The lowly

expressed miRNAs have an average of 5.11 (+- 8.76; n=47) cleavage targets per miRNA,

whereas highly expressed miRNAs average only 1.78 (+- 1.56; n=18) cleavage targets

per miRNA ($p = 0.0154$ with a two-tailed Welch's t-test, $p = 0.0156$ for MWW, Cohen's

$d = 0.538$). This trend is also present for translational repression targets, but with smaller

differences between highly expressed and lowly expressed miRNAs that are not significantly different. The lowly expressed miRNAs have an average of 15.31 (+- 31.12; n=70) translational repression targets per miRNA, whereas highly expressed miRNAs average only 7.56 (+- 31.12; n=27) translational repression targets per miRNA ($p = 0.053$ with a two-tailed Welch's t-test, $p = 0.267$ for MWW, Cohen's $d = 0.290$).

Comparison with newly evolved, species-specific miRNAs in Arabidopsis

Because the low expression and high target count for many of the miRNAs identified in *C. reinhardtii* resemble the descriptions of newly evolved miRNAs in higher plants [13, 14, 31-33], we performed the same analyses on miRNAs specific to *Arabidopsis thaliana* and *A. lyrata*. Similar to what we observed with the *C. reinhardtii* miRNAs, the vast majority of the species-specific miRNAs in both Arabidopsis species are lowly expressed (< 500 CPM) (Figures 4.3B and 4.3C). They also show the same relationship between the expression level of the miRNAs and the number of predicted targets (Figures 4.5A and 4.6). The difference is most prominent in *A. thaliana*, with an average of 9.25 (+- 8.948; n=40) cleavage targets per lowly expressed miRNA and only 2.75 (+- 0.957; n=4) targets per highly expressed miRNA ($p < 0.0001$ with two-tailed Welch's t-test, $p = 0.075$ MWW, Cohen's $d = 1.318$). The trend is still present in *A. lyrata*, with an average of 4.73 (+- 4.981; n=34) targets per lowly expressed miRNA and 2.80 (+- 1.095; n=5) targets per highly expressed miRNA. However, because only a few of the *A. lyrata* specific miRNAs with an average expression greater that 500 CPM have identifiable targets and the difference between the two groups of miRNAs are smaller, the difference is not significant ($p = 0.058$ with a two-tailed Welch's t-test, $p = 0.915$ for MWW, Cohen's $d =$

1.008). The relationship between the number of predicted targets and the expression of the miRNA for species-specific miRNAs in Arabidopsis is in sharp contrast the relationship for miRNAs conserved between the two species, for which many highly expressed miRNAs have ~20 predicted targets (Figures 4.5B and 4.5C).

miRNA targets related to nitrogen starvation

To determine what role miRNAs play in the response to nitrogen starvation, we examined the predicted targets of miRNAs that present in either HS+N or HS-N excluding all of the miRNAs that are found in both or neither condition. The HS+N miRNAs include those that are HS+N-specific and those that are found in both TAP and HS+N. The HS-N miRNAs are only those that are HS-N exclusive, as there are no miRNAs found in both TAP and HS-N, but not HS+N (Figure 4.1). These combinations result in 84 cleavage targets and 215 translational repression targets potentially regulated by miRNAs HS+N, but not HS-N, and conversely 31 cleavage targets and 101 translational repression targets potentially regulated in HS-N, but not HS+N (Figure 4.2 and Table 4.1). The vast majority of these targets in both conditions have no known function (Figures 4.6 and 4.7). Very few of the targets that do have known functions are metabolic enzymes. Instead they have a wide variety of roles within the cells, including transcription factors, DNA and RNA binding proteins, posttranslational modification of proteins, and transporters.

Of the 84 cleavage targets and 215 translational repression targets in HS+N, five and ten, respectively, are significantly differentially expressed during nitrogen starvation (Table 4.2). However, none of these differentially expressed targets showed at least a 2-fold change in expression in the miRNA-deficient Mut20 cells, and several of them are

down-regulated during nitrogen starvation, despite the absence of the miRNA targeting them in HS-N conditions. Similarly, of the 31 cleavage targets and 101 translational repression targets in HS-N, only three cleavage targets and three translational repression targets are significantly differentially expressed during nitrogen starvation. As for HS+N, a number of these targets are up-regulated during nitrogen starvation, despite the differential presence of the miRNAs targeting them. However, one of the cleavage targets, FAP281, is down-regulated during nitrogen starvation, and shows the 3-fold up-regulation in Mut20 expected for a cleavage target. The remaining differentially expressed targets again have no known function. Those that do have predicted functions perform a variety of non-metabolic functions, with only three of the 21 targets predicted to have metabolic activity.

Discussion

The changes in miRNA population between the three conditions investigated in this study, cells grown phototrophically in TAP media, photoautotrophically in HS media with a nitrogen source (HS+N), and photoautotrophically in HS media without a nitrogen source (HS-N), sheds light onto how gene regulation by miRNAs change in response, as well as the relationship between the expression level of miRNAs and the constraints on the number of potential targets for that miRNA. When comparing the predicted targets for both cleavage and translational repression for the miRNAs that are differentially present between HS+N and HS-N conditions, the target functions closely resemble those of Arabidopsis during nitrogen starvation [7]. For both organisms, very few of the transcripts targeted are metabolic enzymes, with most the targets instead comprising a

wide range of non-metabolic roles, including transcription factors, DNA and RNA binding proteins, post-translational protein modification, transporters, and ion channels. While most of these targets do not show an increase in expression in the Mut20 cells that lack miRNA what would be expected for true cleavage targets, they could still represent translational repression targets or the expression in the mutant could be altered by other compensatory mechanisms. These targets could allow each of the miRNA to have a far-reaching impact on abundance, stability, and function of proteins and metabolic pathways far beyond the scope of the individual miRNA targets. While these miRNAs do not appear to play a large role in directly regulating the metabolic pathways surrounding lipid production, it remains possible that these pathways are still impacted by the downstream effects of miRNA gene silencing.

In addition to the targets that are similar between Chlamydomonas and Arabidopsis in nitrogen starvation, two of the 21 differentially expressed targets (DHC13 and FAP281) are related to flagella in Chlamydomonas, including the only predicted cleavage target that shows the expected two-fold up-regulation in the Mut20 cells that lack miRNA. While not related to the accumulation of lipids seen during nitrogen starvation, these targets are still interesting as nitrogen starvation triggers gamete formation in Chlamydomonas, which entails a remodeling of the flagella to accommodate mating [34, 35]. DHC13 (also known as ODA11) is a flagellar outer arm dynein heavy chain alpha protein, which provides a major structural component for the outer dynein arm and plays a role in flagellar motility [36, 37]. According to Phytozome 10, FAP281 is a coiled-coil protein found in the flagellar proteasome, and the transcript is up-regulated during regeneration, however its exact role is not known [38, 39]. The impact

of this regulation remains unclear, as the miRNA-deficient Mut20 cells do not appear to have any swim-speed or mating phenotypes, despite lacking this regulatory mechanism (data not shown).

The expression pattern for the miRNAs that are condition specific or present in only two out of the three conditions is also intriguing. Especially for miRNAs that are specific to HS+N or present in HS+N and TAP, but not HS-N, the overall expression of the miRNAs is much lower than for the miRNAs that are present in all three conditions, with an average expression level $1/20^{th}$ that of the miRNAs present in all three conditions (Table 4.1). The expression pattern for these miRNAs resembles those found in newly evolved miRNAs in higher plants, where as new miRNAs emerge, they are first expressed at a very low level and may or may not be functional (see Figures 4.3B, 4.3C, 4.5A, and 4.6), whereas miRNAs that become conserved between species often exhibit a much higher expression level, regardless of the number of targets (Figures 4.5B, 4.5C) [13-16, 31-33]. This possibility is further strengthened by the relationship between the number of predicted targets (both cleavage targets and translational repression targets) and the average expression for the miRNAs across all of the conditions that it is predicted in (Figure 4.4). As the average expression for the miRNA sequence increases, there is a marked decrease in the number of potential targets, especially above 500 CPM. Conversely, the miRNAs that have many targets (more 7 cleavage targets or than 30 translational repression targets) have among the lowest expression levels. In particular, increasing expression appears to provide an evolutionary constraint on the number of genes targeted for each miRNA. If the miRNAs are not expressed at a high enough level to be functional or have a significant impact, there are no constraints on the number or

type of genes they are complementary to, and thus are predicted to target. Similarly, the expression levels of those miRNAs with many targets or targets that would be detrimental to target would also be constrained to prevent them from becoming functional. For example, miR_t39, which is encoded within three separate pre-miRNA structures in HS+N and HS-N media and has a relatively high expression level with an average expression of 2499 CPM, has no identifiable targets. It indicates that this miRNA is not constrained even though its expression level is not low. Presumably this miRNA would continue to evolve neutrally until either it accumulates mutations and indels until degenerated and lost or the mature miRNA becomes sufficiently complementary to a target transcript for it to bind, at which point the constraints apparent for the miRNAs with identifiable targets would also apply to this miRNA.

There are many possible mechanisms for novel miRNAs to evolve in both plants and animals. In plants, where it is common for not only the binding site of the miRNA to be highly complementary to the mature miRNA, but also for the region surrounding the binding site to be complementary to part of the pre-miRNA, one possibility is a partial inverted duplication of part of the target sequence [40, 41]. Such inverted duplication can occur as the result of the genomic rearrangements and segmental duplications the evolution of plants [42]. Another similar possibility, which appears possible in both plants and animals, is the insertion of transposable elements creating an inverted repeat [43-45]. In both of these cases, in addition to formation of inverted repeats in the genome, the sequences must also contain sufficient promoter regions to be transcribed, giving rise to a number of potential hairpin sequences in the genomes that are either too weakly transcribed to be functional, or not transcribed at all, which has a significant impact on

the prediction non-conserved miRNAs that evolve in such a manner. One final possibility, which is more likely in animals with their low degree of complementarity between the miRNAs and their targets than in plants, is that the miRNA machinery could process any transcript that forms a hairpin structure [46, 47]. In this model, a huge variety of small RNAs could be processed and loaded into the RISC complex, regardless of whether they are able to target a transcript. As it appears to be the case here, when these new sequences do arise, the level of expression for each is also subjected to selective pressure, allowing tight control of the regulatory roles these miRNA play within the cells [15, 16].

This analysis also highlights the importance of the growth conditions used for performing miRNA analysis. While both TAP (which contains acetate and provides photoheterotrophic conditions for the cells) and HS (which does not contain a carbon source and provides photoautotrophic conditions) are considered standard growth conditions, roughly 30% of the total miRNAs identified are specific to one or the other conditions, and three times as many miRNAs shared between HS+N and HS-N conditions than between TAP and HS+N (Figure 4.1, Table 1). Coupled with the observation that only one of the miRNAs are completely absent in any condition, but rather that the processing of the pre-miRNAs causes those miRNAs to meet the criteria in some conditions but not in others, this can cause miRNAs identified in one experiment to no longer appear to be a canonical miRNA in a second experiment despite still being present in the library if any changes are made to the growth conditions. To what extent this variation occurs within the same condition remains an open question that further replication of the Ago3 cross-linking experiment will be necessary to answer.

Literature Cited

1. Zhao, T., et al., *A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii.* Genes Dev, 2007. **21**(10): p. 1190-203.
2. Molnar, A., et al., *miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii.* Nature, 2007. **447**(7148): p. 1126-9.
3. Shu, L. and Z. Hu, *Characterization and differential expression of microRNAs elicited by sulfur deprivation in Chlamydomonas reinhardtii.* BMC Genomics, 2012. **13**: p. 108.
4. He, H., L. He, and M. Gu, *Role of microRNAs in aluminum stress in plants.* Plant Cell Rep, 2014.
5. Gupta, O.P., et al., *MicroRNA mediated regulation of metal toxicity in plants: present status and future perspectives.* Plant Mol Biol, 2014. **84**(1-2): p. 1-18.
6. Khraiwesh, B., J.K. Zhu, and J. Zhu, *Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants.* Biochim Biophys Acta, 2012. **1819**(2): p. 137-48.
7. Liang, G., H. He, and D. Yu, *Identification of nitrogen starvation-responsive microRNAs in Arabidopsis thaliana.* PLoS One, 2012. **7**(11): p. e48951.
8. Shukla, L.I., V. Chinnusamy, and R. Sunkar, *The role of microRNAs and other endogenous small RNAs in plant stress responses.* Biochim Biophys Acta, 2008. **1779**(11): p. 743-8.
9. Ding, Y.F. and C. Zhu, *The role of microRNAs in copper and cadmium homeostasis.* Biochem Biophys Res Commun, 2009. **386**(1): p. 6-10.
10. Gonzalez-Ballester, D., et al., *RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival.* Plant Cell, 2010. **22**(6): p. 2058-84.
11. Loh, Y.H., S.V. Yi, and J.T. Streelman, *Evolution of microRNAs and the diversification of species.* Genome Biol Evol, 2011. **3**: p. 55-65.
12. Barakat, A., et al., *Conservation and divergence of microRNAs in Populus.* BMC Genomics, 2007. **8**: p. 481.
13. Felippes, F.F., et al., *Evolution of Arabidopsis thaliana microRNAs from random sequences.* RNA, 2008. **14**(12): p. 2455-9.
14. Fahlgren, N., et al., *MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana.* Plant Cell, 2010. **22**(4): p. 1074-89.
15. Takuno, S. and H. Innan, *Selection fine-tunes the expression of microRNA target genes in Arabidopsis thaliana.* Mol Biol Evol, 2011. **28**(9): p. 2429-34.
16. Montes, R.A., et al., *Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs.* Nat Commun, 2014. **5**: p. 3722.
17. Ma, X., et al., *Small interfering RNA-mediated translation repression alters ribosome sensitivity to inhibition by cycloheximide in Chlamydomonas reinhardtii.* Plant Cell, 2013. **25**(3): p. 985-98.
18. Einhauer, A. and A. Jungbauer, *The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins.* J Biochem Biophys Methods, 2001. **49**(1-3): p. 455-65.

19. van Dijk, K., et al., *Monomethyl histone H3 lysine 4 as an epigenetic mark for silenced euchromatin in Chlamydomonas.* Plant Cell, 2005. **17**(9): p. 2439-53.

20. Merchant, S.S., et al., *The Chlamydomonas genome reveals the evolution of key animal and plant functions.* Science, 2007. **318**(5848): p. 245-50.

21. Kapitonov, V.V. and J. Jurka, *A universal classification of eukaryotic transposable elements implemented in Repbase.* Nat Rev Genet, 2008. **9**(5): p. 411-2; author reply 414.

22. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms Mol Biol, 2011. **6**: p. 26.

23. Tarver, J.E., P.C. Donoghue, and K.J. Peterson, *Do miRNAs have a deep evolutionary history?* Bioessays, 2012. **34**(10): p. 857-66.

24. Chang, C.C. and C.J. Lin, *LIBSVM: A Library for Support Vector Machines.* Acm Transactions on Intelligent Systems and Technology, 2011. **2**(3).

25. Kruger, J. and M. Rehmsmeier, *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.

26. Lopez, D., et al., *Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data.* BMC Bioinformatics, 2011. **12**: p. 282.

27. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1178-86.

28. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

29. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

30. Thimm, O., et al., *MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.* Plant J, 2004. **37**(6): p. 914-39.

31. Ma, Z., C. Coruh, and M.J. Axtell, *Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus.* Plant Cell, 2010. **22**(4): p. 1090-103.

32. Fahlgren, N., et al., *High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes.* PLoS One, 2007. **2**(2): p. e219.

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

34. Sager, R. and S. Granick, *Nutritional control of sexuality in Chlamydomonas reinhardi.* J Gen Physiol, 1954. **37**(6): p. 729-42.

35. Martin, N.C. and U.W. Goodenough, *Gametic differentiation in Chlamydomonas reinhardtii. I. Production of gametes and their fine structure.* J Cell Biol, 1975. **67**(3): p. 587-605.

36. Ishikawa, T., H. Sakakibara, and K. Oiwa, *The architecture of outer dynein arms in situ.* J Mol Biol, 2007. **368**(5): p. 1249-58.

37. Alper, J.D., M. Tovar, and J. Howard, *Displacement-weighted velocity analysis of gliding assays reveals that Chlamydomonas axonemal dynein preferentially moves conspecific microtubules.* Biophys J, 2013. **104**(9): p. 1989-98.

38.   Stolc, V., et al., *Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3703-7.

39.   Pazour, G.J., et al., *Proteomic analysis of a eukaryotic cilium.* J Cell Biol, 2005. **170**(1): p. 103-13.

40.   Allen, E., et al., *Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana.* Nat Genet, 2004. **36**(12): p. 1282-90.

41.   Rajagopalan, R., et al., *A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.* Genes Dev, 2006. **20**(24): p. 3407-25.

42.   Li, A. and L. Mao, *Evolution of plant microRNA gene families.* Cell Res, 2007. **17**(3): p. 212-8.

43.   Piriyapongsa, J. and I.K. Jordan, *A family of human microRNA genes from miniature inverted-repeat transposable elements.* PLoS One, 2007. **2**(2): p. e203.

44.   Piriyapongsa, J., L. Marino-Ramirez, and I.K. Jordan, *Origin and evolution of human microRNAs from transposable elements.* Genetics, 2007. **176**(2): p. 1323-37.

45.   Piriyapongsa, J. and I.K. Jordan, *Dual coding of siRNAs and miRNAs by plant transposable elements.* RNA, 2008. **14**(5): p. 814-21.

46.   Svoboda, P. and A. Di Cara, *Hairpin RNA: a secondary structure of primary importance.* Cell Mol Life Sci, 2006. **63**(7-8): p. 901-8.

47.   Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs.* Nat Genet, 2005. **37**(7): p. 766-70.
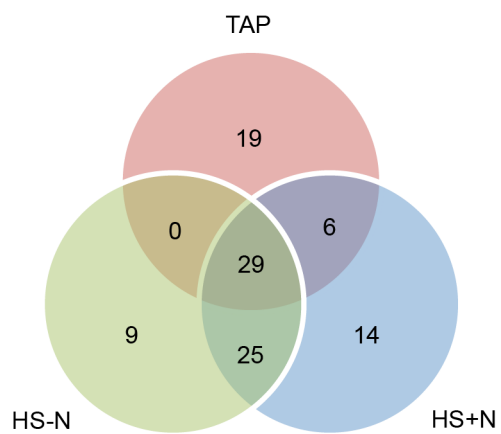
Figure 4.1: Comparison of predicted miRNAs TAP, HS+N, and HS-N growth media
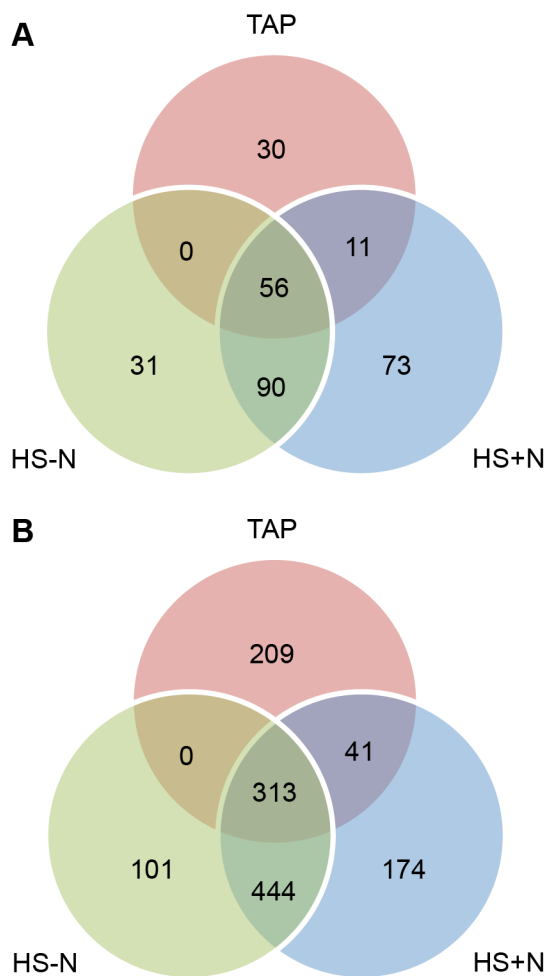
Figure 4.2 Comparison of predicted miRNA targets in TAP, HS+N, and HS-N growth media

Venn diagrams show the overlap of the number of predicted miRNA cleavage targets (A) and translational repression targets (B) across each of the three growth conditions.
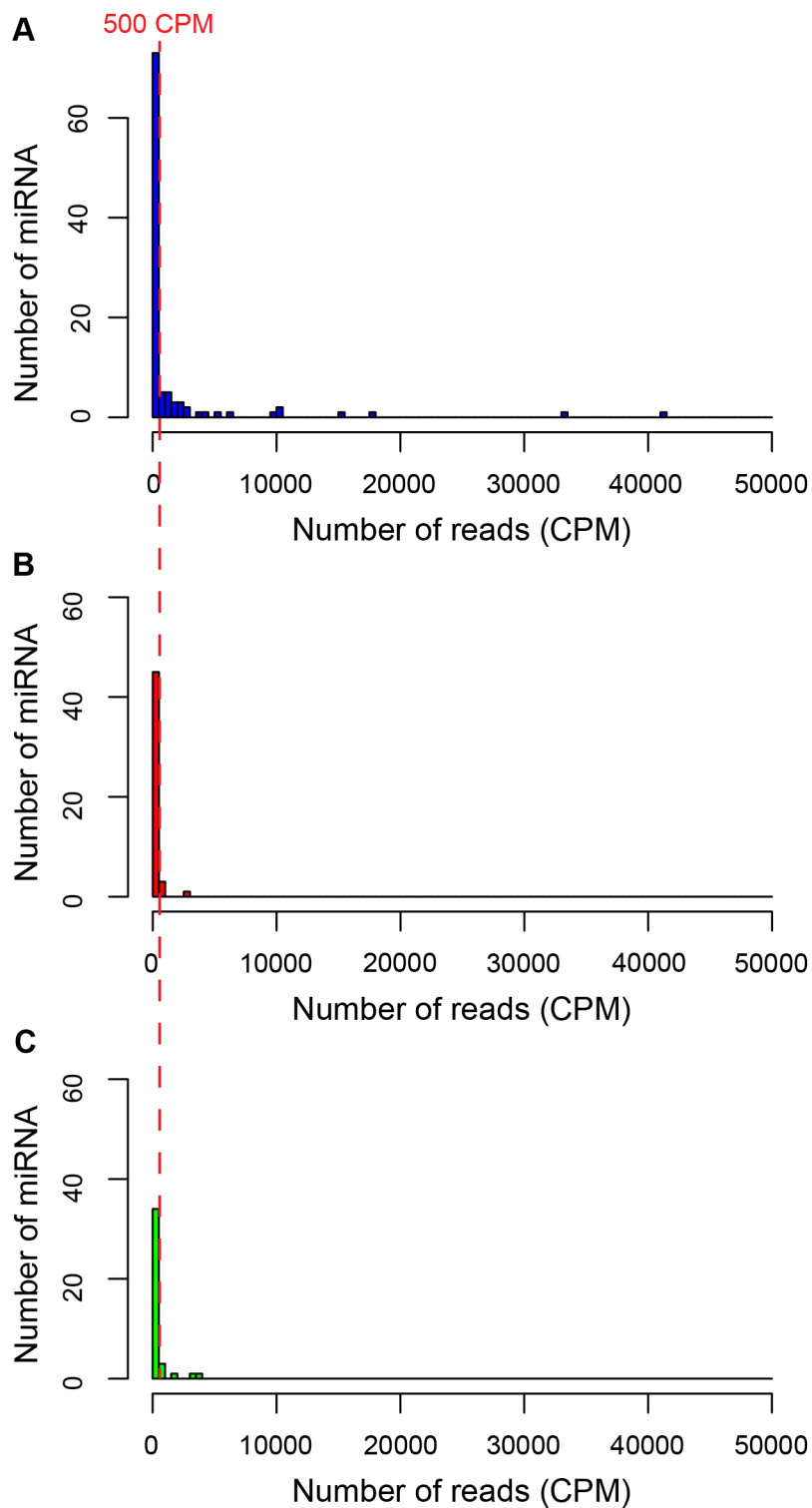
Figure 4.3 Distributions of miRNA expression levels for *C. reinhardtii* and species specific miRNAs in *A. thaliana and A. lyrata*

The dashed line at 500 CPM represents the cutoff between lowly expressed and highly

expressed miRNA across all three species.
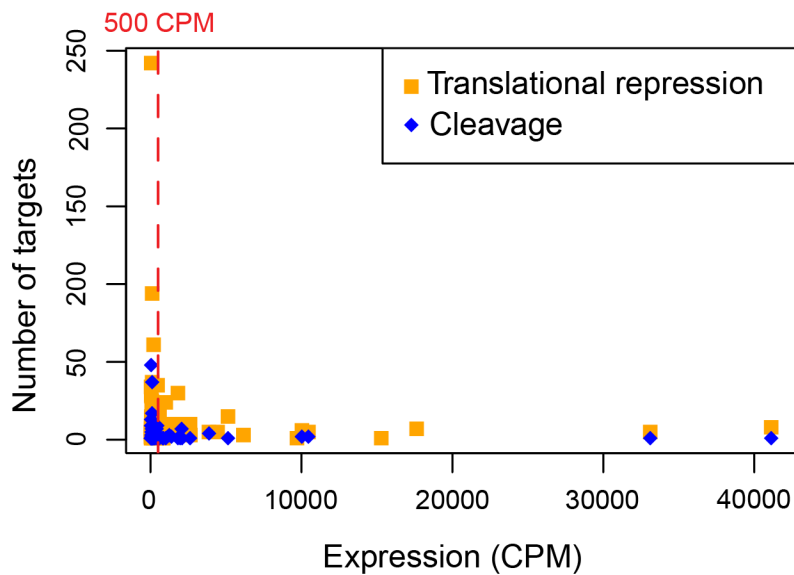
Figure 4.4 Relationship between miRNA expression level and number of targets in

C. reinhardtii

Scatter plot shows the relationship between the expression level of an miRNA and the

number of targets (blue diamonds for cleavage and orange squares for translational

repression) predicted for that miRNA in *C. reinhardtii*. The dashed line at 500 CPM

represents the cutoff between lowly expressed and highly expressed miRNA.
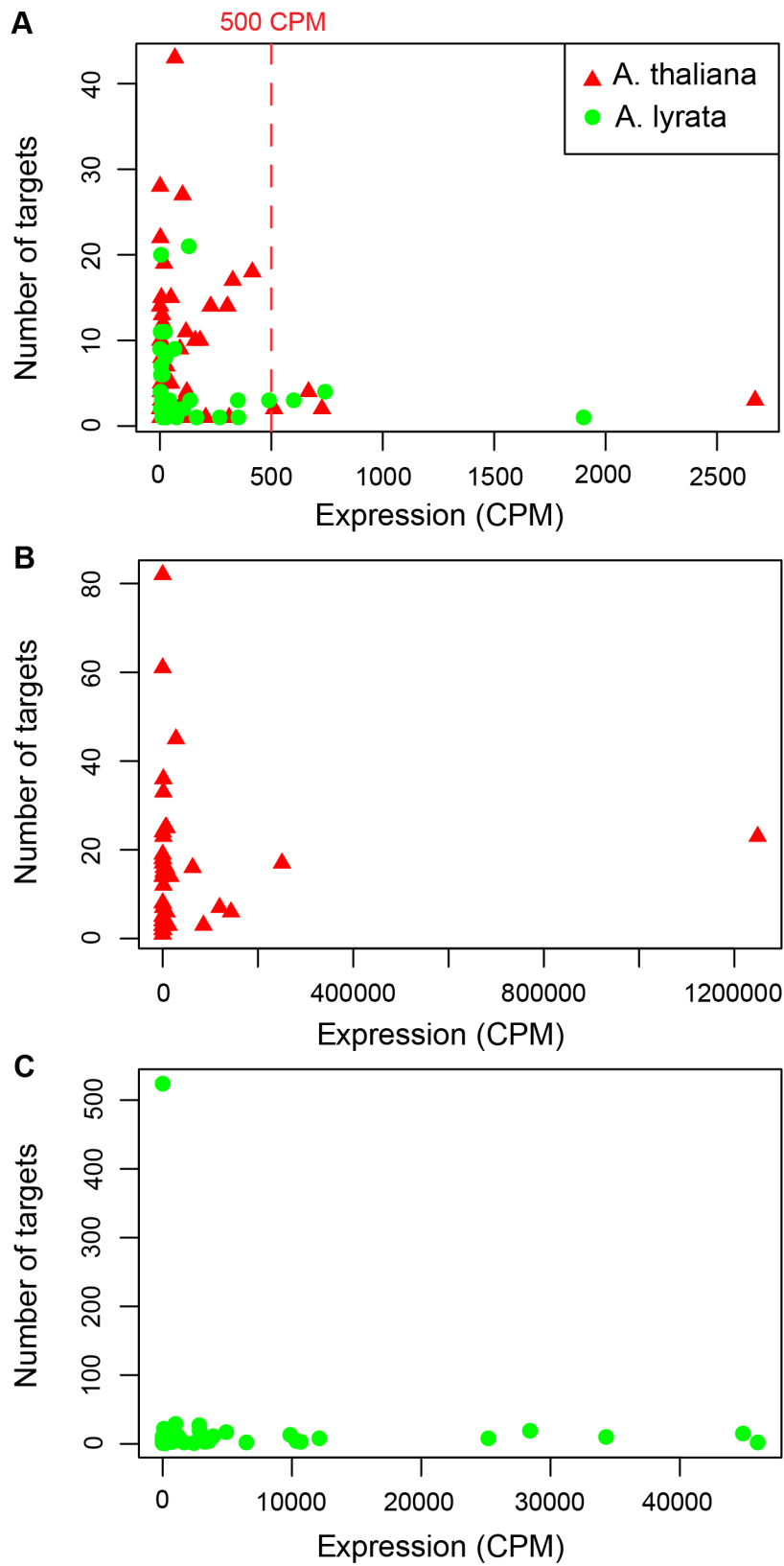
Figure 4.5 Relationship between miRNA expression level and number of targets in

A. thaliana **and** *A. lyrata*

Scatter plots shows the relationship between the expression level of an miRNA and the number of targets predicted for that miRNA for (A) species specific miRNA in *A. thaliana* (red) and *A. lyrata* (green), (B) conserved miRNAs in *A. thaliana* and (C) conserved miRNAs in *A. lyrata*. The dashed line at 500 CPM represents the cutoff between lowly expressed and highly expressed miRNA.
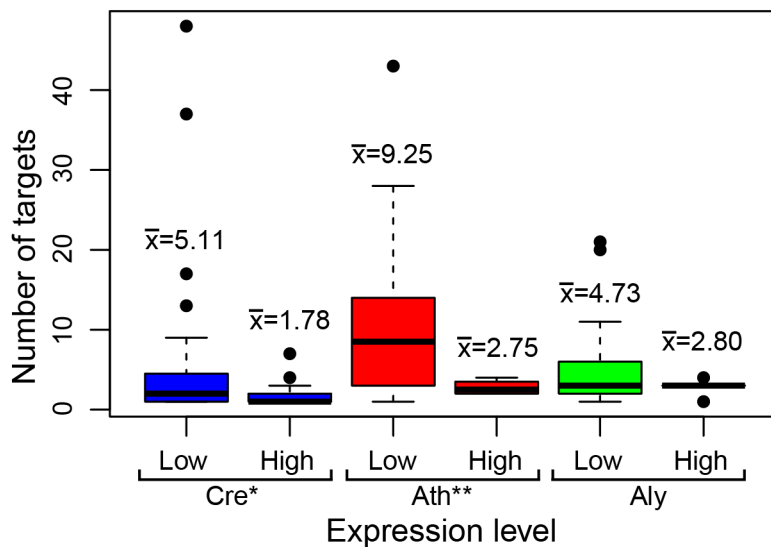
Figure 4.6 Comparison of the number of targets between lowly expressed and highly expressed miRNAs

Boxplots show the number of cleavage targets for miRNAs with low expression (<500 CPM) and high expression (≥500 CPM) in *C. reinhardtii* (blue), *A. thaliana* (red), and *A. lyrata* (green). The mean numbers of targets ($\overline{x}$) for each category are shown above the boxes. The target numbers are compared between the two expression level groups: *p < 0.05 and **p < 0.0001 (Two-tailed Welch's *t*-test).
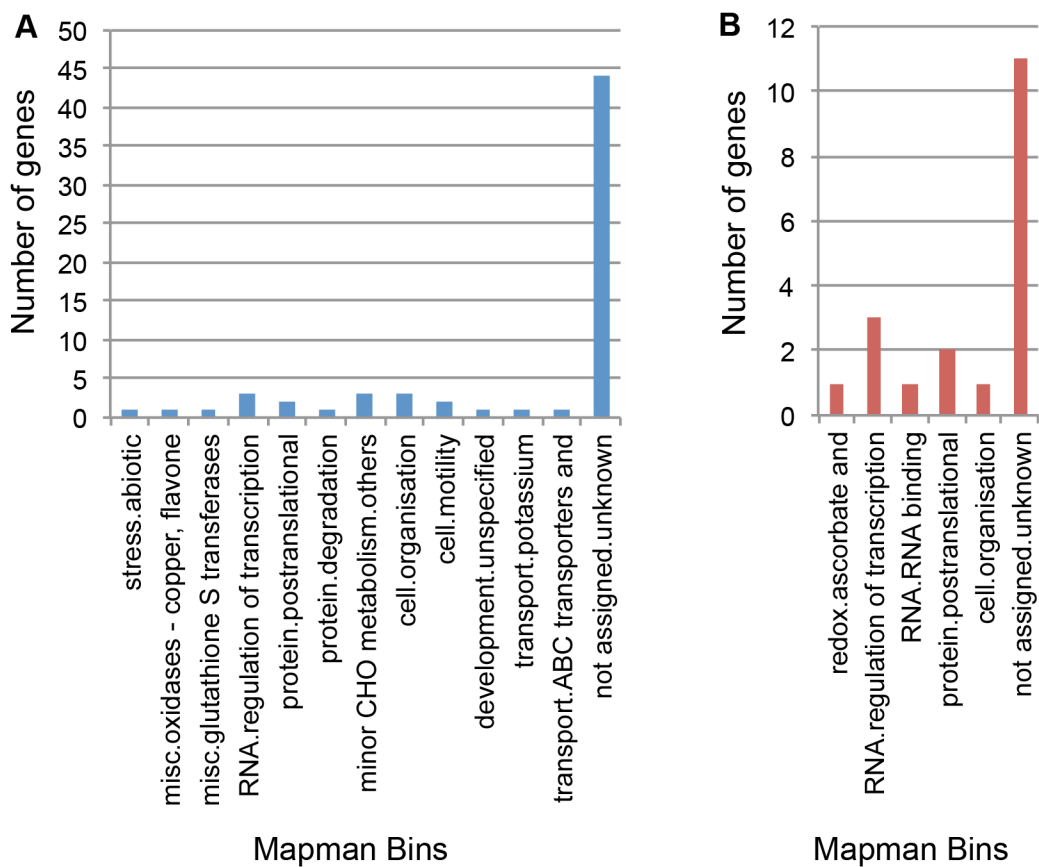
Figure 4.6: Mapman bins for cleavage targets differentially regulated in HS+N (A) and HS-N (B) media.

Figure A: Number of genes vs Mapman Bins (blue bars). X-axis categories: lipid metabolism.exotics (steroids, squalene etc), amino acid metabolism.synthesis, amino acid metabolism.degradation, secondary metabolism.unspecified, Co-factor and vitamine metabolism.ubiquinone, nucleotide metabolism.degradation, Biodegradation of Xenobiotics.lactoylglutathione, misc.cytochrome P450, RNA.processing, RNA.regulation of transcription, RNA.RNA binding, DNA.synthesis/chromatin structure, DNA.repair, protein.aa activation, protein.synthesis, protein.targeting, protein.postranslational modification, protein.degradation, protein.folding, signalling.receptor kinases, signalling.calcium, signalling.phosphinositides, signalling.G-proteins, cell.organisation, cell.division, cell.cycle, cell.motility, development.unspecified, transport.metal, transport.ABC transporters and multidrug, transport.metabolite transporters at the, transport.misc, not assigned.unknown

Figure B: Number of genes vs Mapman Bins (red bars). X-axis categories: PS.lightreaction, secondary metabolism.isoprenoids, hormone metabolism.auxin, tetrapyrrole synthesis.chlorophyll b synthase, major CHO metabolism.synthesis, major CHO metabolism.degradation, nucleotide metabolism.synthesis, misc.acid and other phosphatases, misc.short chain dehydrogenase/reductase (SDR), RNA.regulation of transcription, RNA.RNA binding, DNA.synthesis/chromatin structure, DNA.repair, protein.synthesis, protein.targeting, protein.postranslational modification, protein.degradation, protein.folding, signalling.calcium, cell.organisation, cell.division, transport.misc, not assigned.unknown, glycolysis.cytosolic branch

Figure 4.7: : Mapman bins for translational repression targets differentially regulated between HS+N (A) and HS-N media (B).

Table 4.1: List of predicted miRNAs

Read	miRNA	Alias[a]	TAP[b]	TAP CPM[c]	HS+N	HS+N CPM	HS-N	HS-N CPM	Cleavage targets[d]	TR targets
TTCGAAAAGTTGCACAGCGCC	miR_t94	--	1130	31	--	(7)	--	(13)	0	2
TACCGGGCGTGGGGAGGGCAGG	miR_t10	miR9897-5p	2675	43557	1857	45759	3871	34052	1	9
TTACGGCTCCTTCTTATCGGC	miR_t90	miR9897-3p	2675	31467	1857	39880	3871	27972	2	5
TGCTTGGATCCGTTTGGATGG	miR_t69	--	--	(12)	--	(41)	4648	58	1	15
TCGGAAGGGAGAACGGTGGCC	miR_t39	--	--	(1041)	2251, 3680, 5595	3273	4654, 7413, 11395	1724	0	0
TCGGAAGGGAGAACGGTGGTC	miR_t40	--	--	(988)	2251, 3680, 5595	3409	4654, 7413, 11395	1842	1	10
TGGACAGCGAACAACTTTGAGC	miR_t71	--	--	(429)	--	(46)	4654, 11395	1387	2	10
TCAATTGCTCTTTCTGATG	miR_t30	--	3318	154	2358	90	4837	272	0	5
TCAGATAGAGCAATTGATG	miR_t31	--	--	(15)	2358	58	4837	101	0	6
TGCACGTTGGTAAACAGTCCC	miR_t55	--	3704	994	2690	671	--	(594)	1	5
TGCAGGGGGACTTAGTGGAGA	miR_t56	--	3704	228	2690	107	--	(103)	1	6
TAGTGCCAGCGCGCTCTCGGC	miR_t24	--	3708	146	2699	409	5482	662	7	15
TGGAAAATCGAGGCCGCTGGC	miR_t70	--	3772	317	2729	2457	5570	556	0	0
TGGATGGTCGAAACCGCGTACA	miR_t75	--	3911	77	--	(15)	--	(21)	0	3
TGGAGACTGTACGAGAGGCCA	miR_t73	--	3965	60	2890	81	5847	45	0	3
TGGAGACTGAACGAGAGGC	miR_t72	--	--	(15)	2890	26	--	(8)	1	5
TTTCGAGGTACGAGGACCGGT	miR_t101	--	--	(19)	3163	93	--	(59)	1	3
TACGTCGGCGCCAGCTTCACC	miR_t13	--	--	(9)	3542	40	--	(2)	14	5
TTCGAGGTCGGCTGGAAGGTC	miR_t95	--	4832	90	3548	816	--	(353)	4	18
TCAATAACGGACCTGGAGGAC	miR_t29	--	5094	57	--	(13)	--	(23)	0	1
TGGCATGGCTGGAGGAGCAGG	miR_t78	--	--	(64)	3952	282	8069	131	2	14

TTCAGGTAGCGGGACCAGGTG	miR_t91	miR1157	5858	6353	4247	1515	8689	3762	4	5
TCTGACCGGTCGGTGCTGCCG	miR_t46	--	6589	231	--	(126)	--	(39)	1	9
TCGCGGTACAGGCTCTTAGGC	miR_t37	--	--	(666)	4707	1544	9668	2572	7	10
TAGCATCGTACCGTAATTGT	miR_t18	--	--	(22)	5047	125	10379	212	0	3
TAGCATCGTACCGTTATTGT	miR_t19	--	--	(5)	5047	27	10379	48	1	9
TGACTCTCACTCCTACTCGGC	miR_t50	--	7085	1807	5048	3534	10380	2599	1	3
TTCCACCTGACTGCTTGCTGA	miR_t93	--	7085	601	--	(39)	--	(109)	1	16
TCCACGTTCTCGCCGCGCAGG	miR_t32	--	8319	339	5973	62	--	(84)	4	3
TCGAGATGCCCCTGCAGTACC	miR_t34	--	8319	38	5973	33	--	(30)	0	7
AGCAGCGTCGGGCTCGACCGC	miR_t2	miR910	8563	30	6153	39	12470	54	0	29
TGCGGGGCCCGCAGGTAGCTGC	miR_t64	--	8563	34	--	(2)	--	(1)	9	34
TGCGGCCTGGTTTATTGGAGA	miR_t63	--	--	(1625)	6204	143	12551	462	1	18
TGGATGTTGCTTGCTGGATAC	miR_t76	--	9303	8918	6666	13144	13325	7059	1	1
TAAGCCGGGGTTGGTAGCAAC	miR_t8	--	9303	1535	--	(339)	--	(182)	0	4
TTGATGTCCGCTGGGAGCTGG	miR_t98	--	9770	348	--	(83)	--	(162)	1	22
TGCGTAGGACGTGGTAGGCGC	miR_t66	--	9770	93	--	(16)	--	(32)	0	6
TGCGAGCTGCGGAATGGCGAC	miR_t57	--	9770	85	--	(227)	--	(165)	7	25
TGCACGCTGTGACTGTCTAGC	miR_t54	--	--	(103)	--	(97)	15652	221	4	15
TGGCTGGTGAATTTGTAGTAG	miR_t80	--	--	(351)	8193	163	16325	389	2	14
TATGACTGACAGCAGGGTGCC	miR_t27	--	--	(32)	8193	121	16325	317	1	4
TGCGCGAGGATGAGATGCTGG	miR_t61	--	--	(50)	8314	117	16579	38	3	21
TGAGTGGTGGTCGACGGCAAT	miR_t51	--	11846	146	8411	72	16786	76	0	2
TCGGAGAAGCGGGTAGCTGA	miR_t41	--	12364	5973	8753	5831	17472	3599	2	15
GAGAACATCAAGGAGCTGCGG	miR_t3	--	--	(39)	8762	78	17485	50	0	2
TAAACAGACAAGGCGACCGACA	miR_t5	--	12551	915	8879	1038	17767	478	2	1
TCGCCTTGTCTGTTTATGTGG	miR_t36	--	12551	483	8879	36	17767	101	8	61
TAAACAGACAAGGCGACCGA	miR_t4	--	12551	76	8879	62	17767	51	0	0
TGCAAAATCAAGACCGGGGGA	miR_t52	--	12615	60	--	(16)	--	(6)	0	3

TGGCCGTACTACTATTGTC	miR_t79	--	12905	267	9173	428	18300	695	1	35
TAGGAACAAGGACGGAGGCGC	miR_t21	--	--	(26)	9395	132	--	(99)	1	3
TGTGATGGGTGTCAAGGGCC	miR_t88	--	--	(10)	9395	55	--	(24)	1	14
TTGACACTCATCACACTGGC	miR_t97	--	--	(38)	9395	30	--	(42)	1	4
TCGATGTTGTTGCGCTGCGGC	miR_t35	--	--	(429)	9510, 9511	964	19015	1520	3	5
TCGGTGTTGTTGCGCTGCGGC	miR_t44	--	--	(114)	9510	363	19015	472	1	13
TGCTGCGGCGTGCTGCTGGACC	miR_t67	--	--	(340)	--	(4)	19015	105	17	37
TGCGCAAGAAGCTGGACCTCG	miR_t58	--	--	(73)	9900	86	19822	40	3	13
TCGGAAAGGAATCCAACGGCC	miR_t38	--	14705	718	10470	866	20969	976	1	1
TGAACACGGCGAGCATCAGCGG	miR_t48	--	14711	60	--	(9)	--	(2)	9	28
TGCTTGCGCCCTCTAGCCGTC	miR_t68	--	14712	435	10475	265	20977	229	4	18
TAGAGGGCGCAAGCACGCA	miR_t17	--	--	(39)	10475	110	20977	60	0	2
TCGTGCTCGTCATCCCCTCG	miR_t45	--	14719	413	10491	116	21002	145	3	16
TAGCCGTCAGGAGCTGTGAGC	miR_t20	--	--	(15)	--	(15)	21002	73	2	10
TGGAGGGCGCAAGCACGCAACG	miR_t74	--	14719	47	10491	24	--	(13)	1	2
TAGTCAGAACAGAGCACTAGC	miR_t23	--	--	(10)	10497	136	21011	440	0	0
TGCGGTGAATGTGAATGATGG	miR_t65	--	16411	200	11646	202	23405	270	2	17
TGACCACCCTGCAGCTGACGC	miR_t49	--	17284	117	--	(10)	--	(3)	1	9
TGGTGATGCTCTGAGTTGGGC	miR_t86	--	--	(36)	12071	192	24260	170	1	15
TACCTCCACCACACTACCCACC	miR_t11	--	--	(320)	12111	445	--	(815)	9	19
TAGTGTGGTGGAGGTAGCAGG	miR_t25	--	--	(560)	12111	418	--	(117)	4	8
TGGTCGTAGTGAGCGAGTGAAG	miR_t85	--	--	(100)	12340	126	--	(139)	0	2
TGGGCCATCGTATTACTATCAG	miR_t83	miR1153-5p	17620	18353	12524	5658	25083	6066	2	6
TAACTTAGTCGTCACAAGGCG	miR_t6	--	17755	77	--	(2)	--	(51)	0	0
TGCGCAGCGGCATCATCTGGA	miR_t59	--	--	(7)	12646	39	25305	41	48	242
TCGGTCAGCATTTCGTTTGG	miR_t43	--	18100	4445	--	(427)	--	(622)	0	5
TCGGTCAGCATCTCGATTGGC	miR_t42	--	18100	1010	--	0	--	0	1	24
TACCCGGAGTGGACGTCTTGC	miR_t9	--	18100	752	--	(2816)	--	(1663)	0	6

TACCTGAAGCGGACATCTTGC	miR_t12	miR918	18100	225	--	(484)	--	(279)	0	4
TGCGGAGTGTGAAAGGGCTGC	miR_t62	--	--	(56)	--	(50)	25611	90	2	4
TTGAACAACTCTCCCACGCTG	miR_t96	--	--	(44)	12945	362	--	(50)	1	3
TTTGGCTGTCCCTTGCATGAGC	miR_t102	--	--	(38)	--	(8)	25938	126	2	8
TGAAAAGTGTGGAATGGGCGG	miR_t47	--	--	(35)	13213	151	26384	89	2	2
TTCCACACTTTTCACGCCTCT	miR_t92	--	--	(523)	13213	107	26384	118	2	3
TTGGGCGGCGTTGTAAGATT	miR_t100	--	19166	4185	13545	634	27033	623	1	30
TGTTGTAGTAGTTTAGCCCTGC	miR_t89	miR1162-3p	19538	14937	13708	8622	27350	7821	2	5
TAGAGCTCGAAGAACTTGGGA	miR_t15	--	20399	17627	--	(1553)	--	(7832)	0	7
TAGAGCTCGAAGAACTTGGGAAGG	miR_t16	--	--	(12984)	--	(303)	28483	15292	0	1
TGGGAAGGAGGACGGTCACGC	miR_t82	--	--	(69)	14240	119	28483	66	2	11
TAGAGCTCGAAGAACTTGGG	miR_t14	--	--	(754)	14240	2089	--	(2009)	1	3
TAAGAAGGTGCGCTGTCTTGA	miR_t7	miR1152	20462	83	14276	89	28573	47	5	4
TGGGCTGTAATGACGGGACG	miR_t84	--	--	(817)	14374	1866	28800	2063	2	6
TGTCAATGACAGTCCTGCAG	miR_t87	--	--	(33)	14687	42	--	(25)	0	10
TATCGGGCGGCAAGCAAGGCGC	miR_t26	--	--	(9)	15016	53	30117	95	1	4
TATTGACGCGCTGCACCAAGC	miR_t28	--	--	(4)	--	(23)	30117	33	1	1
TCGACGCGGTGATGGGCCTGG	miR_t33	--	21642	131	15037	170	30154	310	3	12
TGCAAGGGCTGCGAGCAGGAG	miR_t53	--	--	(34)	15230	262	30575	72	7	15
TAGGACCGACGAAAGCCACT	miR_t22	--	22587	416	15726	105	31500	198	2	2
TGGCTTTCGTCGGTCCTAGG	miR_t81	--	22587	1283	15726	90	31500	120	2	1
AAAAGAAGGTCGCCGAGCTCG	miR_t1	--	--	(39)	17342	192	--	(30)	2	1
TGGATTGATCCCAGCCAGGCG	miR_t77	--	25047	3912	17556	8357	34925	6200	0	3
TTGGACGCGGACCCGGCGCAG	miR_t99	--	25074	1294	17579	923	34946	871	0	9
TGCGCAGCTGCTCCTCCTTCT	miR_t60	--	--	(97)	17579	107	--	(69)	37	94

[a]Alias names are taken from miRBase, Tarvers *et al.*, Shu and Hu, and Nozawa *et al.*

[b]Cluster numbers for each growth condition are given if the read meets the criteria for miRNAs in that condition

[c]Numbers in parenthesis represent the expression of the reads (in CPM) for conditions where the read does not meet the criteria for

miRNAs

[d]Number of predicted targets

Table 4.2: HS+N and HS-N miRNA targets differentially expressed between conditions

Target	miRNA	Direction of differential expression[a]	Gene name	Mut20 log$_2$(FC)	Panther predicted function	Mode of regulation
HS+N Differentially expressed targets						
Cre01.g012750.t1.1	miR_t14	Up	--	0.34	VOLTAGE AND LIGAND GATED POTASSIUM CHANNEL	Cleavage
Cre03.g145127.t1.1	miR_t60	Down	DHC13	-0.80	DYNEIN HEAVY CHAIN FAMILY PROTEIN	Cleavage
Cre04.g216826.t3.1	miR_t60	Up	--	--	--	Cleavage
Cre05.g244300.t1.1	miR_t60	Up	--	0.31	--	Cleavage
Cre12.g488500.t1.2	miR_t60	Down	ARC6	-0.57	--	Cleavage
Cre01.g003376.t1.1	miR_t60	Up	--	0.67	DNAJ HOMOLOG SUBFAMILY C MEMBER	Translational Repression
Cre02.g086650.t1.2	miR_t60	Down	--	-0.33	STRUCTURAL MAINTENANCE OF CHROMOSOMES SMC FAMILY MEMBER, STRUCTURAL MAINTENANCE OF CHROMOSOMES PROTEIN 2	Translational Repression
Cre03.g143987.t1.1	miR_t60	Down	--	-1.45	--	Translational Repression

Cre03.g199450.t1.2	miR_t60	Down	MINE2	-0.35	--	Translational Repression
Cre04.g224500.t1.1	miR_t88	Down	--	-1.40	ATP-BINDING CASSETTE TRANSPORTER	Translational Repression
Cre04.g225750.t2.1	miR_t11	Up	--	-0.15	-	Translational Repression
Cre06.g289400.t2.1	miR_t87	Down	--	0.40	3',5'-CYCLIC PHOSPHODIESTERASE PDE-3-RELATED,CYCLIC NUCLEOTIDE PHOSPHODIESTERASE	Translational Repression
Cre10.g441700.t1.1	miR_t60	Down	--	0.34	--	Translational Repression
Cre02.g089650.t1.1	miR_t95	Down	--	-0.99	--	Translational Repression
Cre10.g432600.t1.1	miR_t56	Down	--	-0.52	ARYLSULFATASE	Translational Repression

HS-N Differentially expressed targets

Cre05.g242301.t1.1	miR_t102	Up	--	0.86	--	Cleavage
Cre16.g668700.t1.2	miR_t67	Down	--	-0.13	--	Cleavage
Cre18.g749747.t1.1	miR_t67	Down	FAP281	1.80	--	Cleavage
Cre01.g007901.t3.1	miR_t54	Up	--	--	MULTIDRUG RESISTANCE PROTEIN	Translational Repression
Cre07.g338000.t1.2	miR_t71	Down	MCM2	1.25	DNA REPLICATION LICENSING FACTOR	Translational Repression

					MCM2,DNA REPLICATION LICENSING FACTOR	
Cre12.g551700.t1.2	miR_t67	Up	--	-0.48	--	Translational Repression
Cre12.g555803.t1.1	miR_t67	Down	--	-0.37	NADH DEHYDROGENASE-RELATED,APOPTOSIS-INDUCING FACTOR 2	Translational Repression

[a]Direction of the change in expression during nitrogen starvation.

CHAPTER 5- Conclusions

The overall conclusion from our studies in Chlamydomonas was that miRNAs do not play a direct role in the transcriptomic response to nitrogen starvation seen in Chapter 2. None of the differentially expressed genes identified in the metabolic pathways relating to TAG biosynthesis and degradation, starch biosynthesis, or amino acid metabolism were directly targeted by miRNAs that are differentially expressed between HS+N and HS-N conditions. Instead, these miRNAs target many genes that are capable of propagating the effects of the regulation across many pathways, albeit in currently unpredictable ways. The exact impacts of these miRNAs are difficult to determine, however, as the vast majority of the predicted targets have no known function, including nearly all of the targets that are differentially expressed during nitrogen starvation. Coupled with the difficulty in verifying predicted targets, especially those predicted to regulate genes through translational repression, much work remains to be done to create a clear picture of what functional role miRNAs play in this and other unicellular organisms.

The results presented here do have implications for the evolution of miRNAs in Chlamydomonas, especially when comparing the expression of miRNAs between different growth conditions. These results reflect what has been seen in other organisms, especially higher plants, for the evolution of novel miRNAs within a species. As these miRNAs first evolve, they often show very weak expression and may or may not have any functional targets. As they transition from being evolutionarily transient to being conserved between species, the expression is generally higher. This trend is reflected in the expression levels of miRNAs between conditions, with those expressed in all three

conditions generally showing several fold higher read counts than those found in only one or two of the conditions. Additionally, the number of targets predicted for an individual miRNA is negatively correlated with the average expression for that miRNA, suggesting that, again similar to newly evolved miRNA in higher plants, selective pressure over which genes are regulated increases as expression of the miRNA increases. The presence of an miRNA with new predicted targets, but a relatively high level of expression, also supports the hypothesis that many of these miRNAs are newly evolved and have not yet (and may never) acquire a definite regulatory role.

The differential expression of miRNAs in different growth conditions lays the foundation for these cells to be able to utilize the miRNA machinery to adapt to changing environmental conditions. Furthermore, Chlamydomonas is unique in that is a model organism which expresses miRNAs, but in which RNAi deficient mutants are not lethal. This combination makes Chlamydomonas a strong candidate for studying the evolution of miRNAs, and establishing the functional role that this mode of regulation may have played for early eukaryotes prior to the evolution of multicellularity.

185

Appendix A- Supplemental figures for Chapter 2
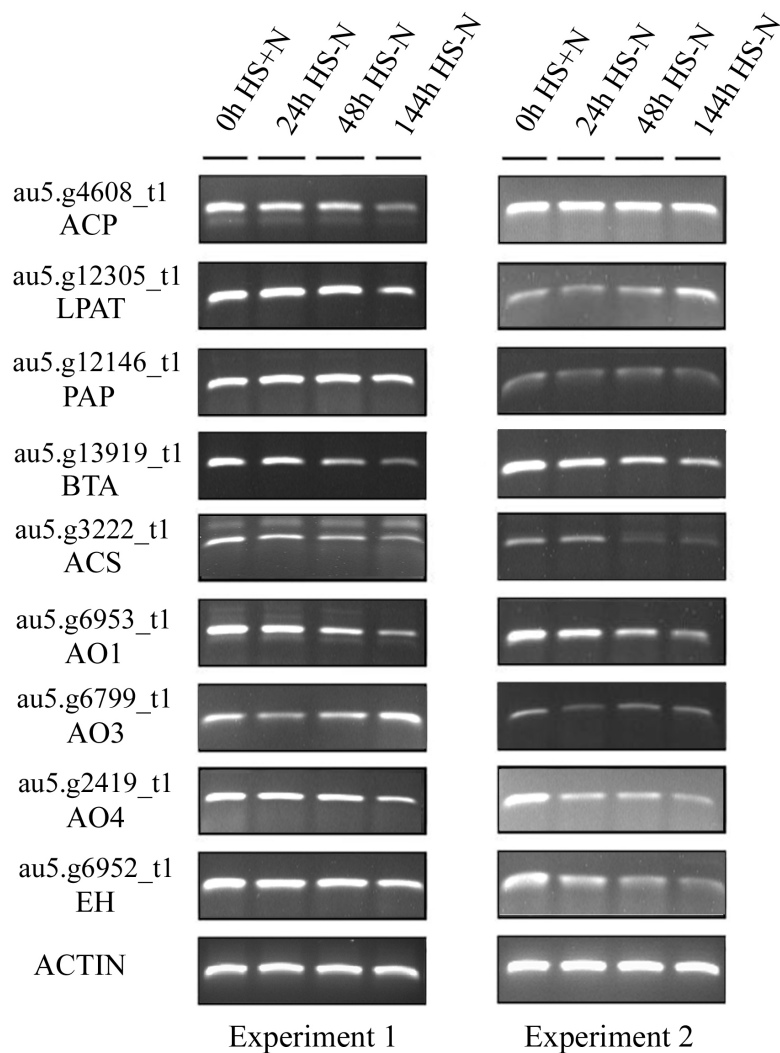
Figure A.1: Expression of TAG metabolism genes in *C. reinhardtii* during nitrogen deprivation under photoautotrophic conditions
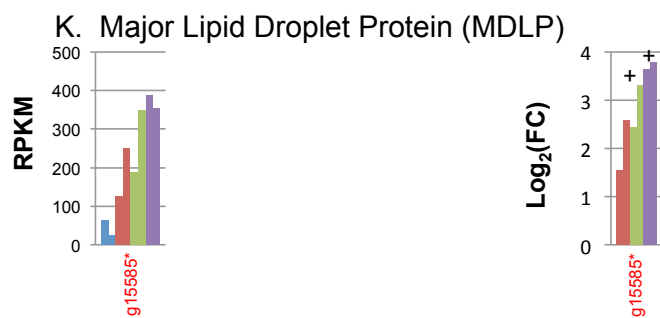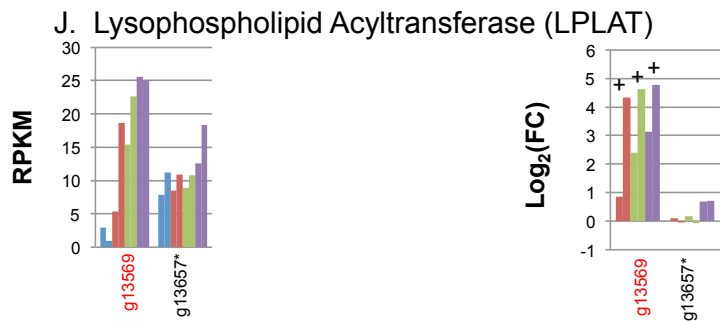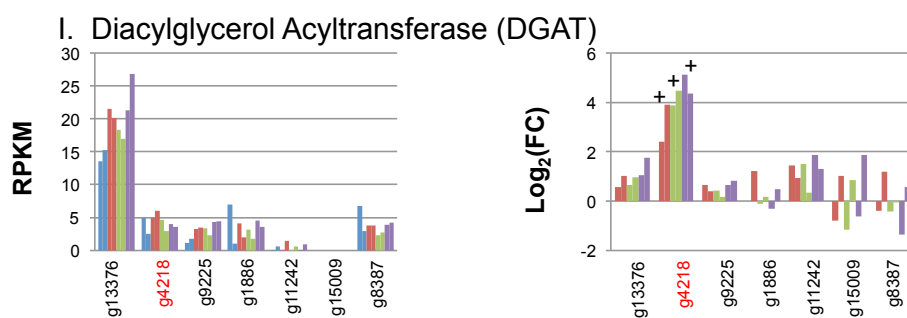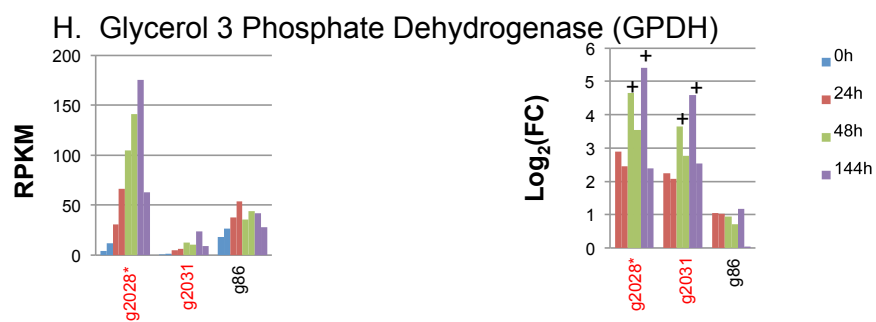
Cells were grown for the indicated times in HS medium with (+N) or without (-N) nitrogen. Transcript abundance corresponding to specific genes showing a marked difference with the mRNA level between control time points of both experiments was analyzed by semi-quantitative RT-PCR with Actin as control. Examined genes included

186

those encoding for Acyl-Carrier Protein (*ACP*), lysophosphatidyl acyltransferase (*LPAT*), phosphatidic acid phosphatase (*PAP*), betaine lipid synthase (*BTA*), acyl-CoA synthase (*ACS*), acyl-CoA oxidase (*AO*), and enoyl-CoA reductase (*EH*).
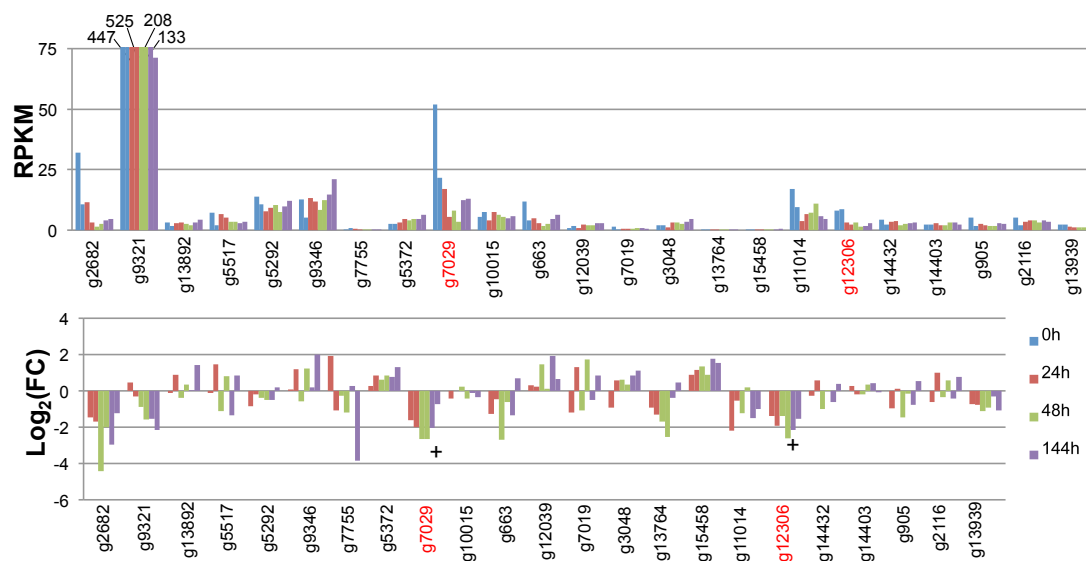
**Supplemental Figure 2: Expression analysis for significant FA and TAG enzymes**

A. Pyruvate Dehydrogenase (PDH)

B. Acetyl-CoA Carboxylase (ACCase)

C. β-ketoacyl-ACP Synthase (KASI)

D. β -Ketoacyl-ACP Reductase (KAR)

E. Enoyl-ACP Reductase (EAR)

F. Δ9 Acyl-ACP Desaturase (AAD)

G. DGAT-Like

H. Glycerol 3 Phosphate Dehydrogenase (GPDH)

I. Diacylglycerol Acyltransferase (DGAT)

J. Lysophospholipid Acyltransferase (LPLAT)

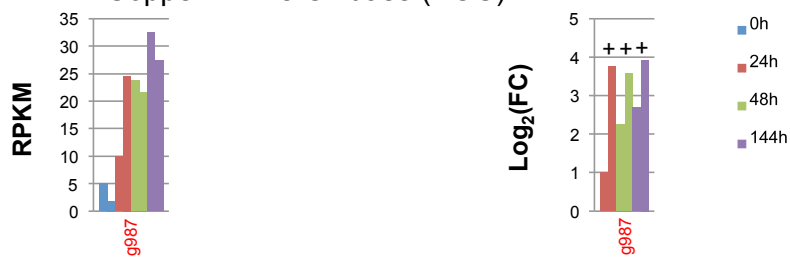K. Major Lipid Droplet Protein (MDLP)
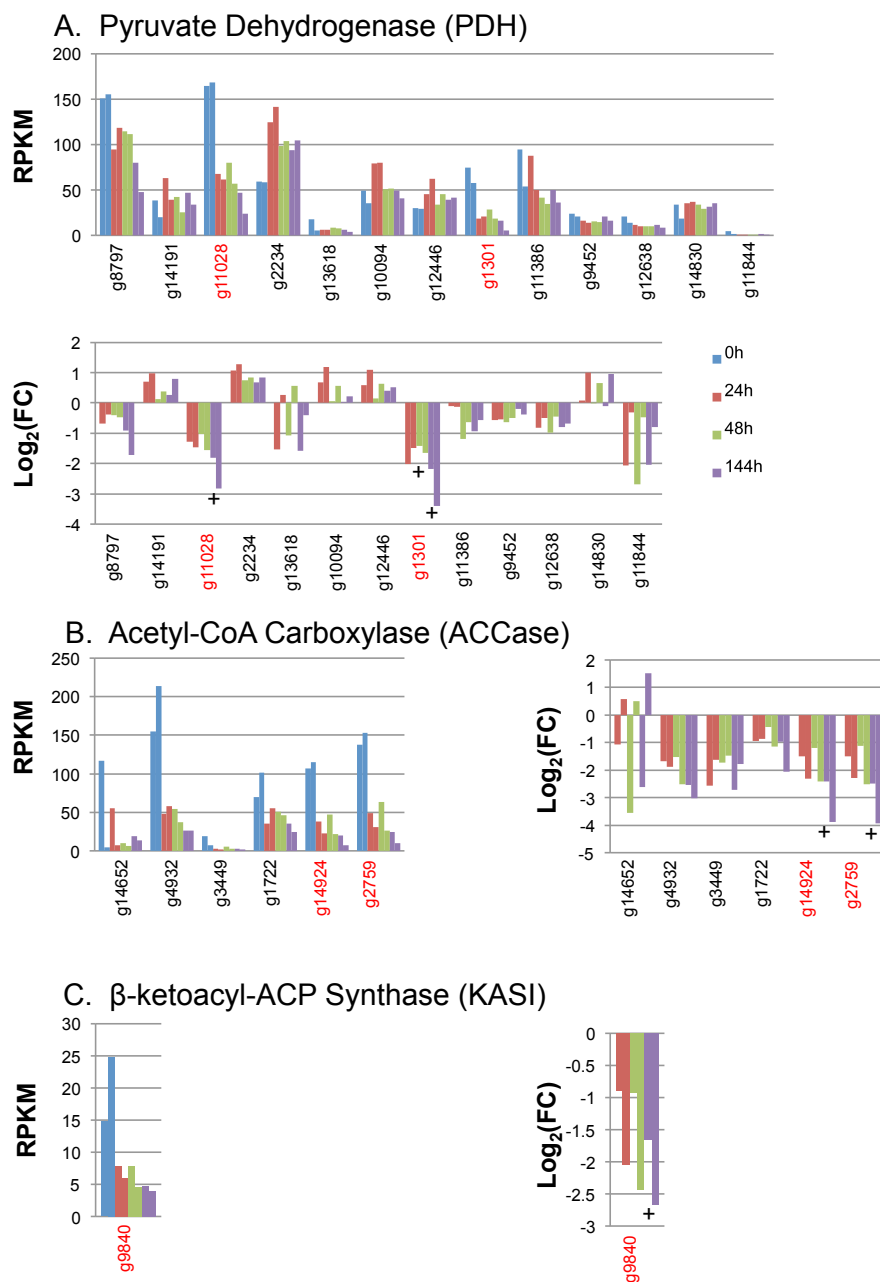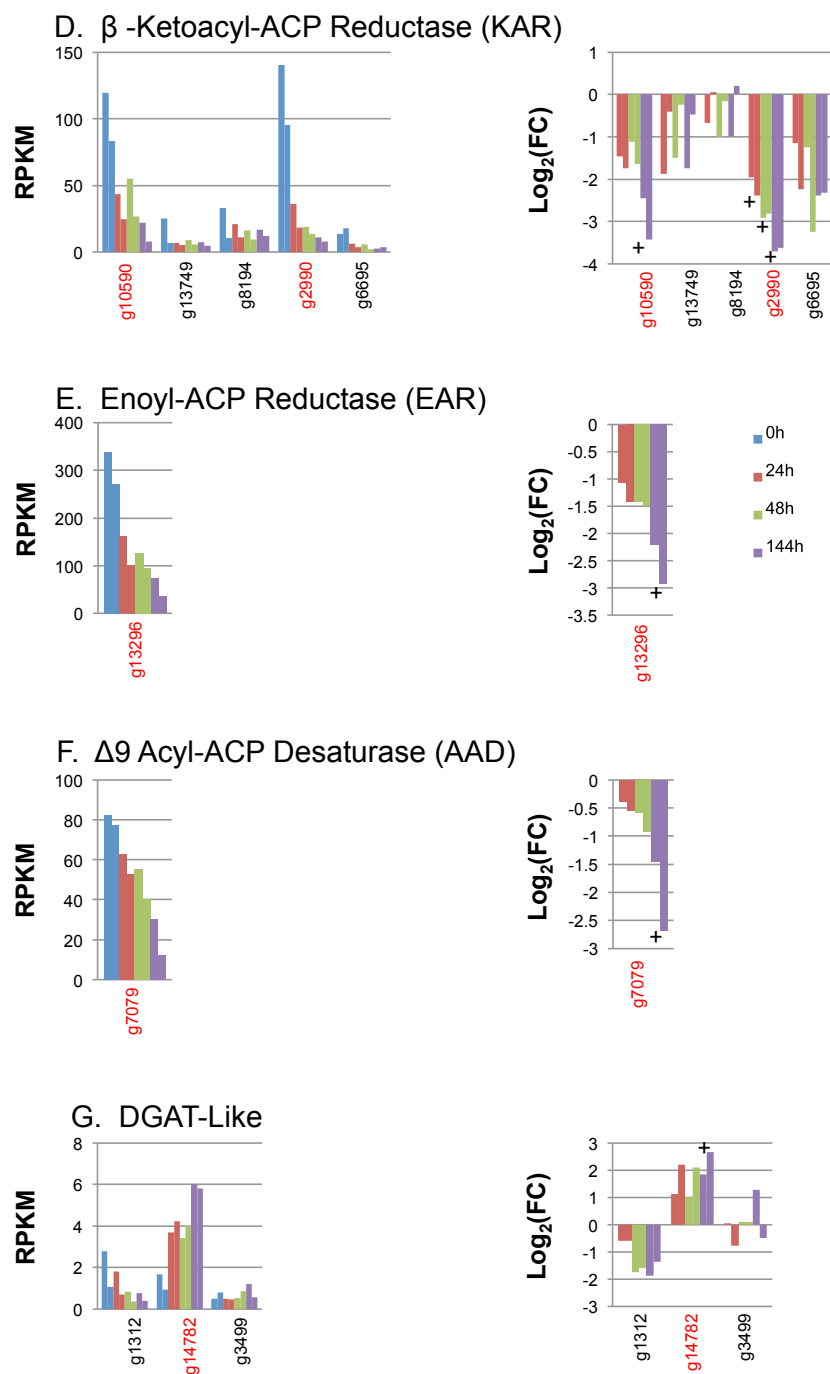
L. TAG/DAG Lipase (TAGL)

Figure A.2: Expression analysis of significantly differentially expressed transcripts associated to FA and TAG metabolic pathways

Names in red are transcripts with at least one significant time point. Plus signs over bars indicate time points that are significant.

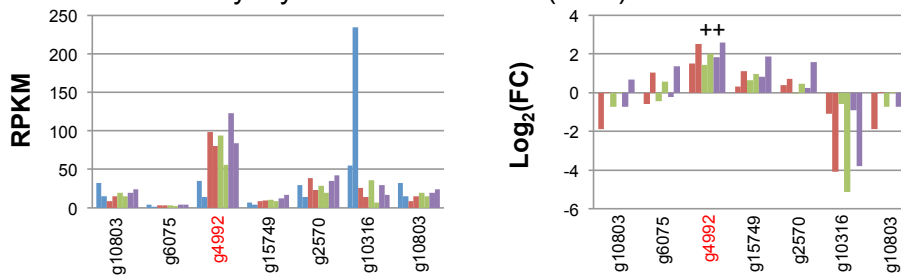**Supplemental Figure 3: Expression analysis for significant starch enzymes**

A. Starch Synthase (SSS)

B. Alpha Amylase (AMYA)

C. Beta Amylase (AMYB)

D. Alpha Glucosidase (GAA)
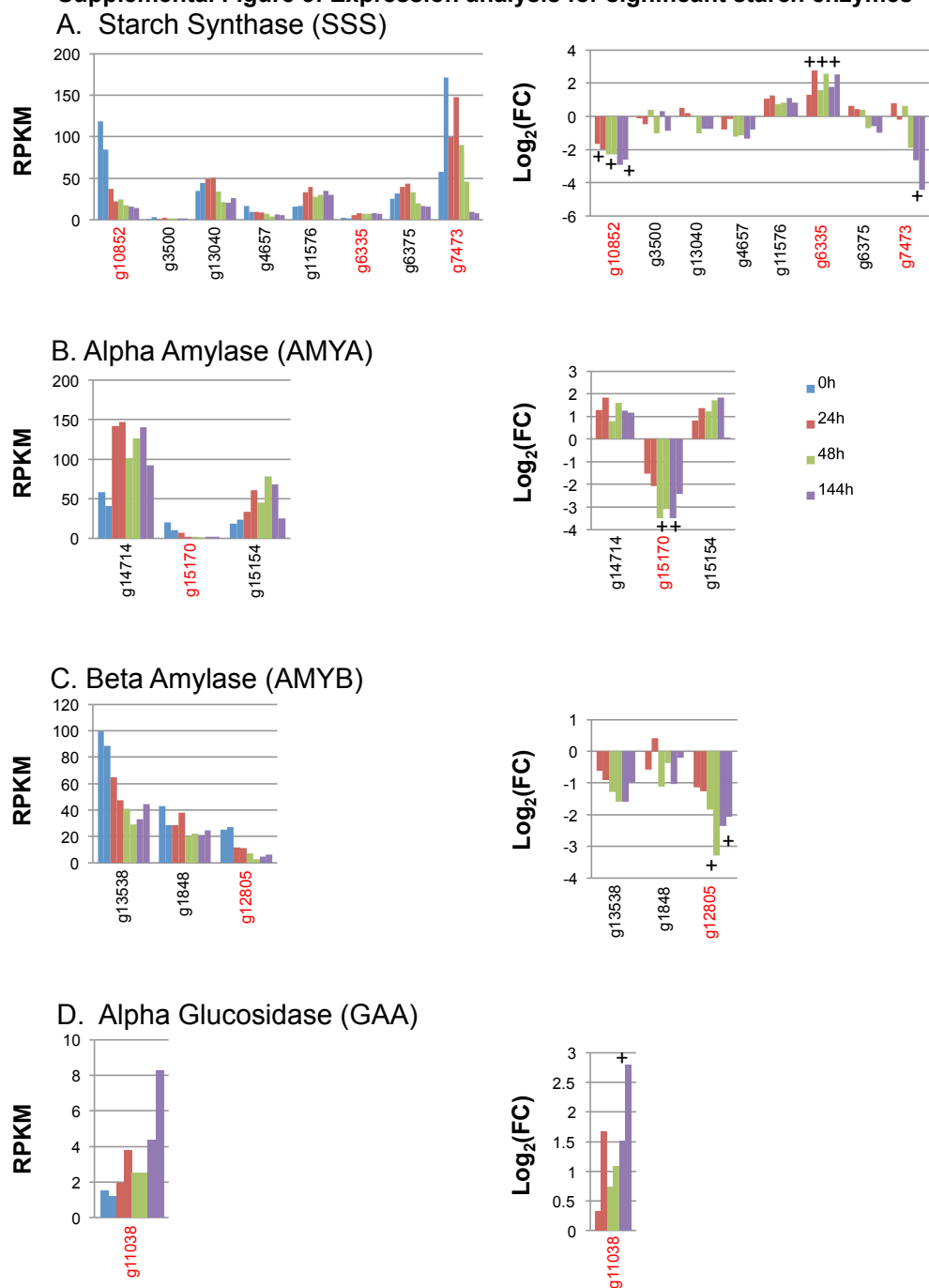
Figure A.3: Expression analysis of significantly differentially expressed transcripts associated to starch metabolic pathways

Names in red are transcripts with at least one significant time point. Plus signs over bars indicate time points that are significant.

Supplemental Figure 4: Expression analysis for significant protein and amino acid enzymes

A. GMP Synthetase (GUA)

B. Deiminase/IAA-amino acid Hydrolase/Aminoacylase (M20)

C. Serine Carboxypeptidase (S10)

D. Agmatine Deiminase (ADI)
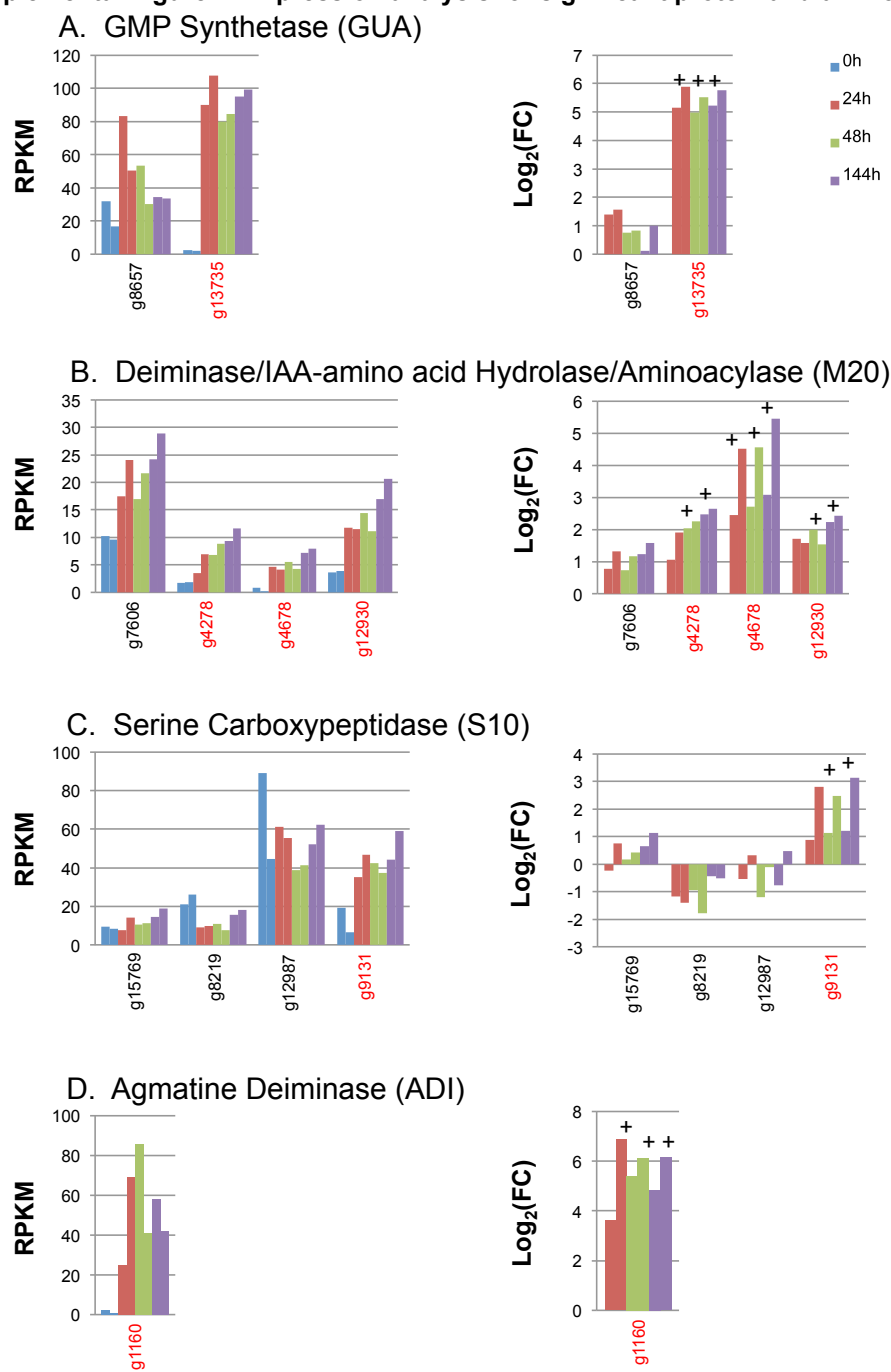
E. Copper Amine Oxidase (AOC)

F. Dihydrolipoamide Dehydrogenase (DLD)

G. Alanine-Glyoxylate Transaminase (AGT)

H. Aspartate Aminotransferase (AAT)

I. Argininosuccinate Synthase (AS)

J. 6-Phosphogluconate Dehydrogenase (PGD)

K. Argininosuccinate Lyase (ASL)

L. Succinate Dehydrogenase (SDHA)
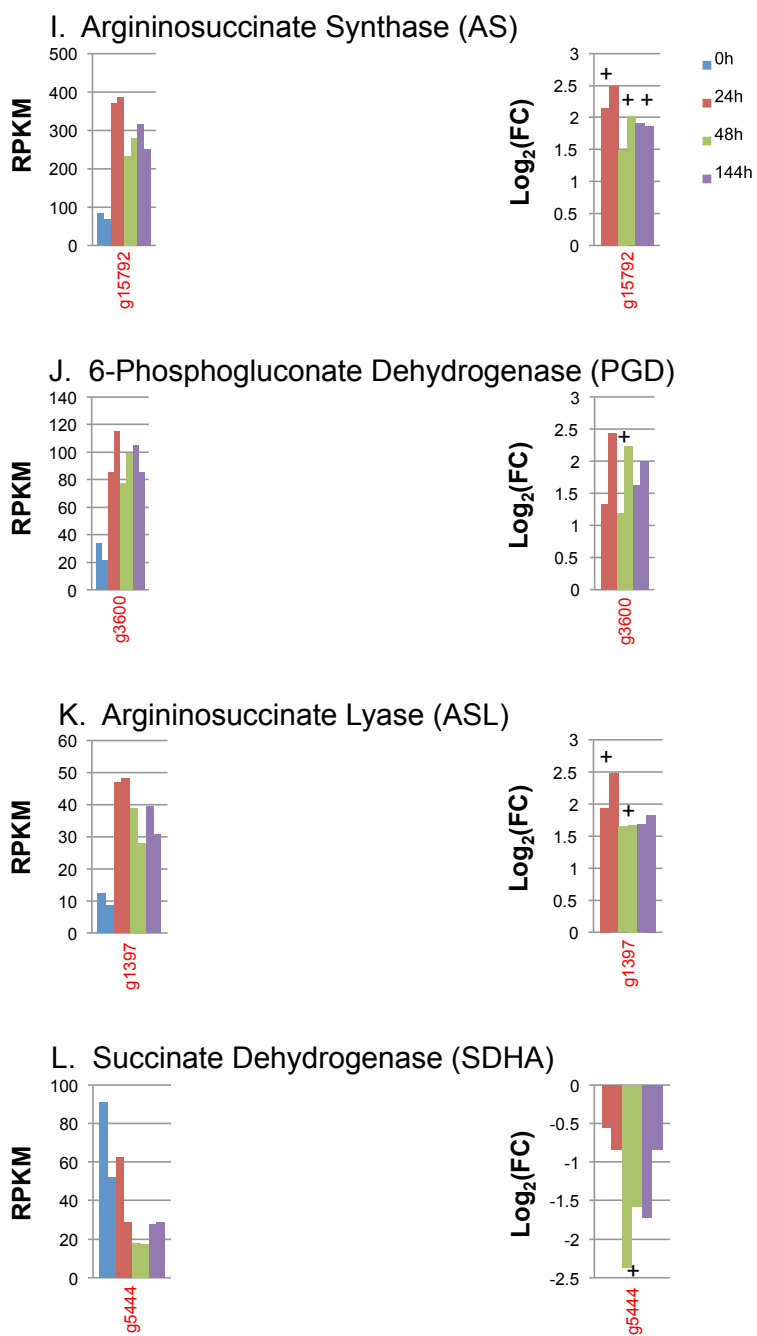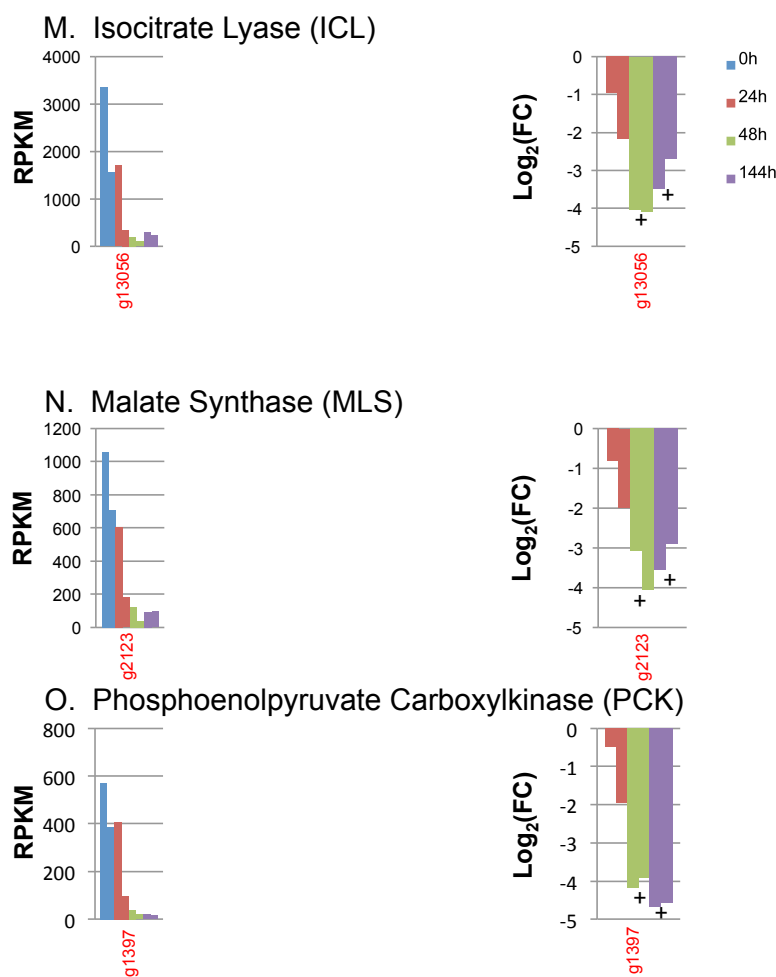
Figure A.4: Expression analysis of significantly differentially expressed transcripts associated to protein and amino acid catabolic pathways

Names in red are transcripts with at least one significant time point. Plus signs over bars indicate time points that are significant.

Appendix B– Supplemental figures for Chapter 3
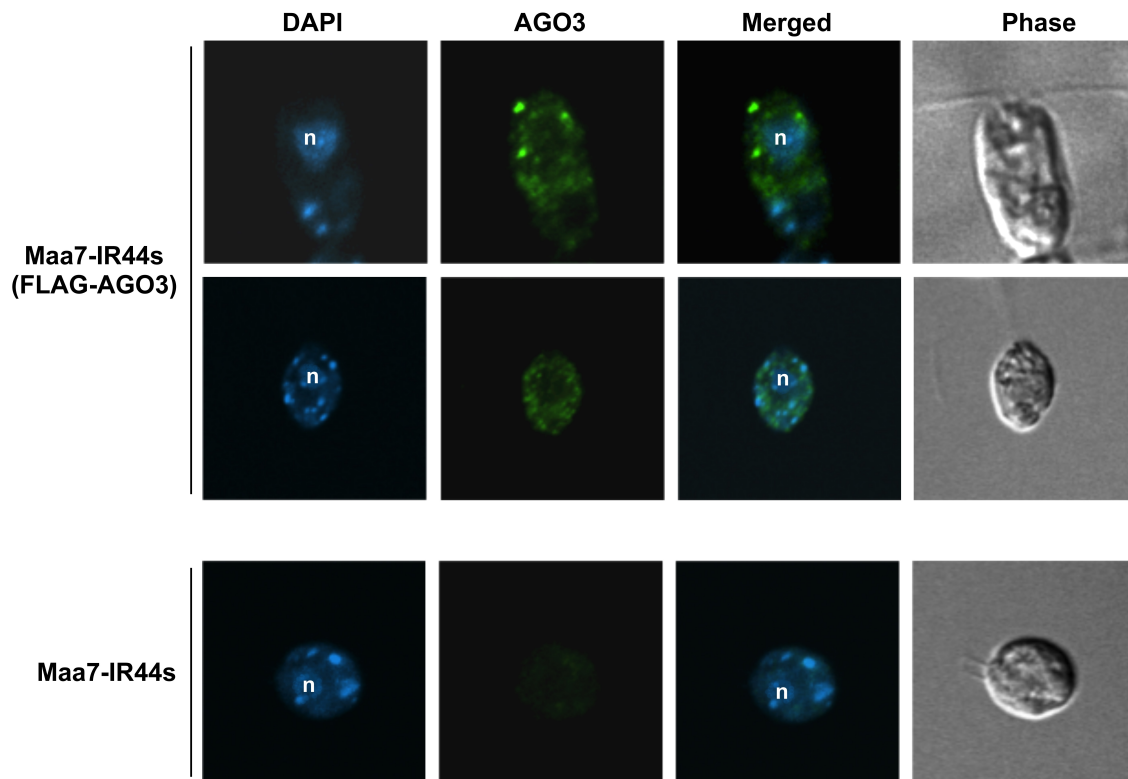
Figure B.1: Immunofluorescence localization of FLAG-tagged AGO3 in *C. reinhardtii.*

The panels indicate DAPI staining of nuclear and organellar DNA (DAPI, blue), immunolocalization of epitope-tagged AGO3, detected with an antibody conjugated to Alexa Fluor 488 (AGO3, green), the merged images (Merged), and the phase contrast images of the cells (Phase). Representative images are shown with the location of the nucleus indicated by n. AGO3 displayed predominantly cytoplasmic localization with a punctuated appearance. Maa7-IR44s-(FLAG-AGO3), transgenic strain expressing a FLAG-tagged AGO3 under the control of the *psaD* promoter. Maa7-IR44s, parental, non-transgenic strain (used as a negative control).

A

Input   IP: anti-FLAG

B
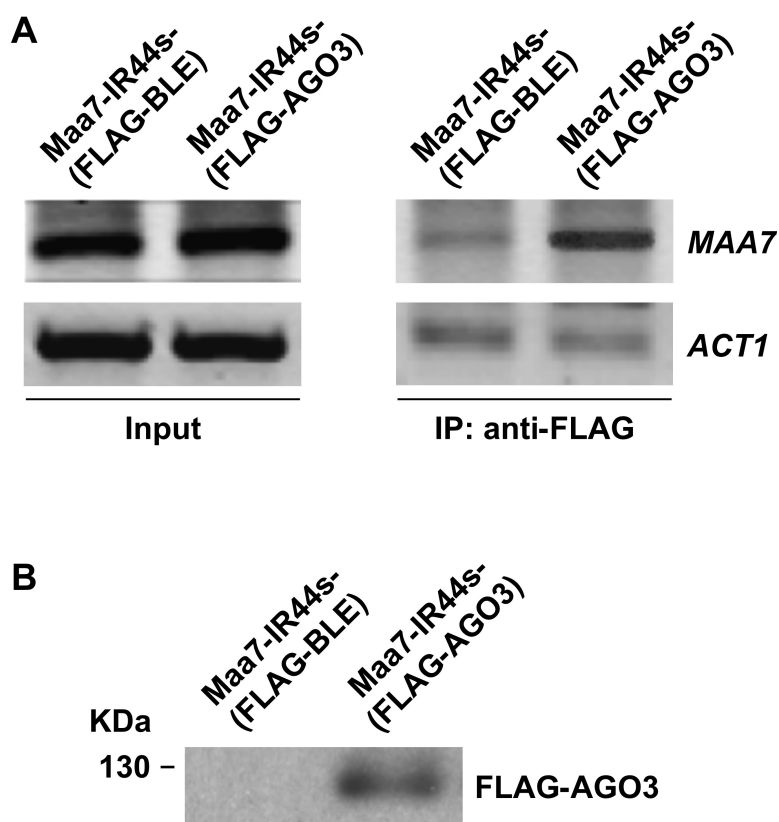
Figure B.2: FLAG-tagged AGO3 associates with the *MAA7* transcript, which is translationally repressed by small RNAs in the Chlamydomonas strain Maa7-IR44s

[1]. (A) Immunoprecipitations were performed with an anti-FLAG antibody on the indicated strains. Total RNA associated with the immunoprecipitated proteins was purified and used as template in RT-PCR reactions to detect the *MAA7* and *ACT1* transcripts (right panels). Total RNA isolated from the cell lysates (prior to immunoprecipitation) was used as template in RT-PCR reactions to determine input transcript levels (left panels). The *MAA7* transcript co-immunoprecipitated with FLAG-tagged AGO3, at levels well above those detected as background in immunoprecipitations with a FLAG-tagged BLE protein. This suggested that the introduced FLAG-tagged AGO3 is functional and associates with an sRNA-repressed

mRNA. In contrast, the *ACTIN1* (*ACT1*) transcript, not targeted for regulation by small RNAs, was only detected at background levels in the immunoprecipitates. Maa7-IR44s-(FLAG-AGO3), transgenic strain expressing a FLAG-tagged AGO3 under the control of the *psaD* promoter. Maa7-IR44s-(FLAG-BLE), transgenic strain expressing a FLAG-tagged BLE protein (conferring resistance to bleomycin). (B) Immunoblot analysis of the FLAG-AGO3 protein, detected with an anti-AGO3 antibody [2], in the indicated strains. Similar loading of the lanes was verified by Coomassie-blue staining of an equivalent gel (not shown).
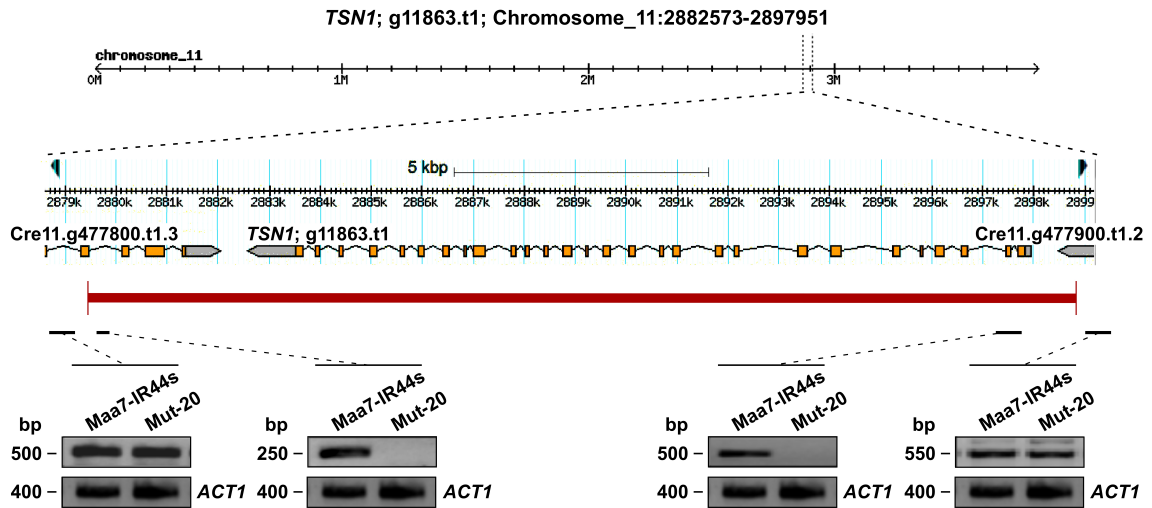
Figure Apx B.3: Mut-20 contains a deletion encompassing the *Tudor Staphylococcal Nuclease 1* (*TSN1*) gene

Diagram of the *TSN1* (*g11863*) gene and surrounding chromosomal regions. The red horizontal bar indicates the extent of the deletion in Mut-20. PCR analyses of genomic DNA in Mut-20 and in the parental strain, Maa7-IR44s, are shown below the diagram.

Literature Cited

1.	Ma, X., et al., *Small interfering RNA-mediated translation repression alters ribosome sensitivity to inhibition by cycloheximide in Chlamydomonas reinhardtii.* Plant Cell, 2013. **25**(3): p. 985-98.

2.	Ibrahim, F., et al., *Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase promotes their degradation in Chlamydomonas.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3906-11.