The Pennsylvania State University

The Graduate School

Graduate Program in Plant Biology

IDENTIFICATION OF SMALL RNA PRODUCING GENES IN THE MOSS *PHYSCOMITRELLA PATENS*

A Dissertation in

Plant Biology

by

Ceyda Coruh

© 2014 Ceyda Coruh

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2014

The dissertation of Ceyda Coruh was reviewed and approved* by the following:

Michael J. Axtell
Associate Professor of Biology
Dissertation Advisor
Chair of Committee

Claude dePamphilis
Professor of Biology

Sarah M. Assmann
Waller Professor of Biology

Anton Nekrutenko
Associate Professor of Biochemistry and Molecular Biology

Teh-hui Kao
Distinguished Professor of Biochemistry and Molecular Biology
Chair, Intercollege Graduate Degree Program in Plant Biology

*Signatures are on file in the Graduate School

ABSTRACT

In plants, a significant fraction of the genome is responsible for making regulatory small RNAs. These ubiquitous, endogenous small RNAs are currently categorized into two groups: microRNAs (miRNAs) and small interfering RNAs (siRNAs). They are produced by Dicer-Like (DCL) proteins and utilized by Argonaute (AGO) proteins to guide repressive regulation of target mRNAs and/or chromatin selected on the basis of small RNA-target complementarity at the transcriptional or post-transcriptional levels. 21 nt miRNAs and 24 nt heterochromatic siRNAs are the two major types of small RNAs found in angiosperms (flowering plants).

The small RNA populations in angiosperms are dominated by 24 nt heterochromatic siRNAs which derive from intergenic, repetitive regions and mediate DNA methylation and repressive histone modifications to targeted loci in angiosperms. However, the existence and extent of heterochromatic siRNAs in other land plant lineages has been less clear. The failure to identify 24 nt heterochromatic siRNA accumulation by initial small RNA-seq attempts from several other species including gymnosperms (Dolgosheina et al. 2008), and the lycophyte *Selaginella* (Banks et al. 2011) has raised the question whether the heterochromatic siRNA pathway is angiosperm-specific. Previous work in *Physcomitrella* provides evidence that supports the hypothesis that the heterochromatic siRNA pathway is an ancestral trait that was present in the last common ancestor of bryophytes and all other subsequently diverged lineages of plants (Cho et al. 2008). However, comprehensive annotation of small RNA genes in the basal lineage *Physcomitrella* is still lacking and an investigation of small RNA populations in this model organism would shed more light on the evolution of regulatory small RNA pathways in land plants.

With the advent of next-generation sequencing, small RNA-seq has become a good resource for producing enormous volumes of data on plant miRNA and siRNA expression. Therefore, we produced extensive small RNA-seq data (more than 10^8 mapped reads) to annotate small RNA genes in ten-day-old protonemata from wild-type *Physcomitrella*. ShortStack is a recently developed tool to analyze small RNA-seq data with respect to a reference genome and to provide a comprehensive annotation of *de novo* discovered small RNA genes. Utilizing ShortStack, we identified 16,024 distinct *DCL*-dependent small RNA producing loci and classified them into five different groupings based on the RNA secondary structure evaluation and the predominant small RNA size. These *Physcomitrella* small RNA producing loci is now available in our developing web server (plantsmallrnagenes.psu.edu).

In order to investigate the features of heterochromatic siRNAs, we revisited the *Physcomitrella* genome to find functional orthologs of the heterochromatic siRNA genes. We identified candidate proteins that could potentially be involved in the accumulation of heterochromatic siRNAs and created mutants to perform genetic analysis. With the power of consistent biological replicates, differential expression analysis on small RNA-seq data revealed that the accumulation of siRNAs from 23-24 nt siRNA loci depends upon *Physcomitrella* homologs of *DICER-LIKE 3* (*DCL3*), *RNA-DEPENDENT RNA POLYMERASE 2* (*RDR2*), and the largest sub-unit of *DNA-DEPENDENT RNA POLYMERASE IV* (Pol IV), with the largest sub-unit of a Pol V homolog contributing to expression at a smaller subset of the loci. These data lead us to conclude that *Physcomitrella* utilizes a heterochromatic siRNA pathway fundamentally similar to that of flowering plants. In contrast to angiosperms, we identified a *Physcomitrella*-specific *MINIMAL DICER-LIKE* (*mDCL*) gene, which lacks the N-terminal helicase domain typical of DCL proteins, but contains the 'catalytic core' (the PAZ domain and the twin RNAseIII domains) of the DCL proteins. We showed that *Physcomitrella* heterochromatic siRNAs are not solely composed of 24 nt siRNAs as seen in angiosperms, but rather contain equal mixtures of 23 and 24 nt siRNAs. Interestingly, *Physcomitrella*-specific mDCL is found to be specifically required for 23 nt siRNA accumulation from these loci. Overall, our data lead us to conclude that heterochromatic siRNAs, and their biogenesis pathways, are largely but not completely identical between angiosperms and basal land plants, as represented by the bryophyte, *Physcomitrella patens.*

Significant effort has been made in small RNA gene annotation, but this progress has been unevenly distributed, with *MIRNA* loci in particular receiving a disproportionate share of the attention. We believe that further efforts at comprehensive and consistent reference annotations of all types of small RNA producing genes, and improvements in the dissemination of such annotations, will greatly enhance the future of plant genomics. Our developing web server (plantsmallrnagenes.psu.edu), which currently hosts small RNA gene annotations of just two species, *Amborella trichopoda* and *Physcomitrella patens,* is intended to serve this purpose. In particular, we look forward to the day when researchers seeking to study small RNAs will be liberated from the need to "re-invent the wheel" by generating their own *de novo* annotations of small RNA-producing genes with each analysis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

***** Large datasets provided in separate files online.

PREFACE

Chapter 1

Chapter 1 includes a published work (Coruh et al. 2014) which is reproduced here with minor modifications.

Chapter 2

Chapter 2 is currently under review at *Genome Biology*.

Authors' Contributions:

Ceyda Coruh and Michael Axtell analyzed small RNA-seq data.
Sung Hyun (Joseph) Cho generated the *Physcomitrella* mutant lines and performed phenotypic analyses.
Saima Shahid performed the miRNA analysis.
Qikun Liu and Sung Hyun (Joseph) Cho prepared small RNA-seq libraries.
Andrzej Wierzbicki generously provided the *Physcomitrella* Pol IV and Pol V largest sub-unit sequences.

ACKNOWLEDGEMENTS

I am very grateful to my dissertation advisor, Dr. Michael J. Axtell, for his continuous guidance, encouragement, and patience during my graduate study. He always encouraged me when things went well, and gave me an alternative perspective when they fell apart. I would like to express my appreciation to my committee, Drs. Sarah Assmann, Claude dePamphilis and Anton Nekrutenko for their support and inspiring discussions of my research.

I thank all former and current members of the Axtell lab for creating a wonderful, warm environment. My special thanks goes to Joseph Cho who introduced me to how to work with the moss *Physcomitrella*. I'm also very grateful to the 'cave' people, namely, Charles Addo-Quaye, Zhaorong Ma, and Saima Shahid, for their inspiring discussions and help on all of my bioinformatics-related questions. I thank Jo Ann Snyder who trained me to handle radioactive materials. I feel so lucky to have gotten to know Qikun Liu, Feng Wang, Cathy Lin, Charles Page, and Seth Polydore, as they are all great friends and colleagues.

I owe many thanks to my dear friends Elif Balin and Esra Kurum for their great friendship and support over the years. I also deeply appreciate my State College "family" with whom I enjoyed State College the most: Esra, Ali, Sinem, Ozhan, Betul, Mahir and Hakan. I thank Nedim and Deniz for their continuous encouragement during my dissertation writing. I also thank Sedat who helped me in writing my initial scripts.

I am deeply grateful to my wonderful parents who provided me a joyful childhood, taught me how to cope with hard times, and tried their best to give me a great education. Words cannot express how much I appreciate the sacrifices they have made in order to assure a better life-story for me. I deeply appreciate my brother for all the joy, fun, support, and sometimes trouble =) that he brings to my life. Last but not least, I would also like to express my deep gratitude to my host family, Marilyn and Don Keat, simply, Ma and Pa, for making me part of their family. They always make me feel I'm home and inspire me to reach out..

Oddly enough, I can't resist thanking State College as a whole, as my intuition tells this Istanbul woman, I will miss this wonderful, cute little town.

"Nothing in life is to be feared. It is only to be understood."
Marie Curie

[1]Chapter 1

Introduction

1.1 *Physcomitrella patens* as a model organism

Embryophytes (land plants) originated from a group of green algae, charophytes, about 475 million years ago (Ma) and can be categorized into two groups: bryophytes (non-vascular land plants) and tracheophytes (vascular land plants). Three bryophyte lineages; the marchantiophyta (liverworts), the anthocerotophyta (hornworts), and the bryophyta (mosses), comprise the earliest lineages of land plants (Kenrick and Crane 1997; Troitsky et al. 2007). Seedless vascular plants form a group whose extant descendents can be categorized into three divisions: the pterophyta (ferns), the lycophyta (lycophytes) and the sphenophyta (horsetails) (Fig. 1.1). On the other hand, more recently diverged seed plants can be categorized into two groups: gymnosperms and angiosperms (flowering plants) (Nickrent et al. 2000). Angiosperms comprise the most diverse, geographically widespread and economically important group of plants. Mosses, positioned as one of the early diverging lineages of embryophytes, provide an ideal resource for comparative studies to illuminate evolutionary changes in land plants. Comparative genomics involving mosses is a powerful tool to better understand how plants conquered land, acquired genes important for tolerating terrestrial stresses (e.g. water deficiency), lost genes associated with aquatic environments and developed hormone signaling pathyways for coordinating multicellular growth.

[1] Chapter 1.6 includes a published paper from *Current Opinion in Plant Biology* (Coruh et al. 2014) and is reproduced here with minor modifications.

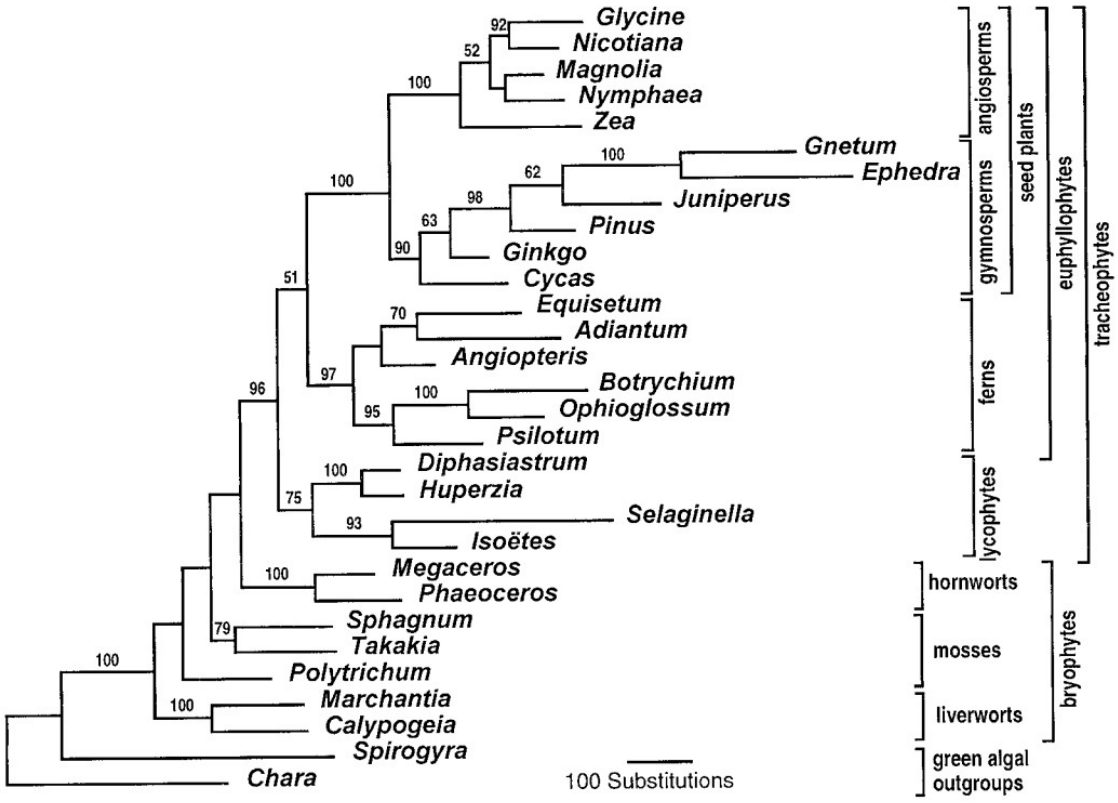

Figure 1.1: Phylogeny of land plants.

Land plant relationships derived from an MP13Ti analysis of the four-gene data set (chloroplast *rbcl* and small-subunit rDNA from all three plant genomes). A single tree (length = 7,074) was obtained with a consistency index (excluding uninformative sites) of 0.3900, a retention index of 0.4756, and a rescaled consistency index of 0.2332. Values above the branches are bootstrap percentages derived from 100 replications (Nickrent et al. 2000).

Alternation of generations occurs in plants and certain groups of algae, and refers to the alternation of multicellular haploid generation (gametophyte) and multicellular diploid generation (sporophyte). Bryophytes show alternation of generations in which the haploid gametophyte stage (producing gametes by mitosis) dominates over the diploid sporophyte stage (producing spores by meiosis). In contrast, all vascular plants are sporophyte dominant and have smaller, short-lived gametophytes. In mosses and other bryophytes, the sporophyte is dependent on the gametophyte as it grows out of the archegonium, the female gametangium that produces eggs (Cove 2005). On the

other hand, the sporophytic stage is the dominant life cycle in both seedless and seed-bearing vascular plants, with an independent gametophyte in seedless vascular plants (e.g. ferns) and sporophyte-dependent reduced gametophytes in seed-bearing plants (e.g. angiosperms) (Campbell et al., 1999).

The sporangium is the structure where diploid sporophyte produces haploid spores via meiosis. Plants can be divided into two groups based on the types of the spores: homosporous and heterosporous plants. Bryophytes and most seedless vascular plants, including ferns and horsetails, are homosporous plants, in which sporophyte produces a single type of spore. Each spore develops into a bisexual gametophyte having both archegonia (female sex organ producing eggs) and antheridia (male sex organ producing sperms) (Campbell et al., 1999). In contrast, heterosporous plants (seed-bearing plants and some seed-free lycophytes such as *Selaginella*) have two types of sporangia; megasporangia and microsporangia, producing megaspores and microspores, respectively. The megasporangium, protected by the ovary wall, contains megaspore mother cells (megasporocytes) which undergo meiosis to produce megaspores. Likewise, the microsporangium, within the anther, contains microspore mother cells (microsporocytes) which undergo meiosis to produce microspores. The resulting megaspores and microspores develop into female and male gametophytes to produce eggs and sperms, respectively (Ambrose and Purugganan 2012).

There are some experimental advantages to working with mosses such as *Physcomitrella patens*. Given its dominant haploid gametophyte phase, *P. patens* is one of the earliest land plants and given its dominant haploid gametophyte phase, loss-of-function mutations are simpler to phenotype than in species with a dominant diploid sporophytic stage (Cove 2005). Not only their simplicity of genetic studies but also their amenability to *in vitro* tissue culture techniques makes mosses a useful model organism. The dominant haploid gametophyte is comprised of protonemata and gametophores, upon which the gametes are produced (Fig. 1.2). Because protoplasts isolated from the filamentous young protonemal tissue in *P. patens*, they provide an ideal genetic material with relatively simpler mutant isolation and genetic analysis. Unlike most other plants, successful transformation of *P. patens* can be achieved through homologous recombination if the transforming DNA contains significant homology to the target locus (Schaefer and Zrÿd 1997). The transformed protoplasts can then regenerate directly into

protonemal tissue in a manner similar to spore germination, thus providing abundant haploid gametophytic tissue for genetic analysis (Cove 2005; Cove et al. 2006).

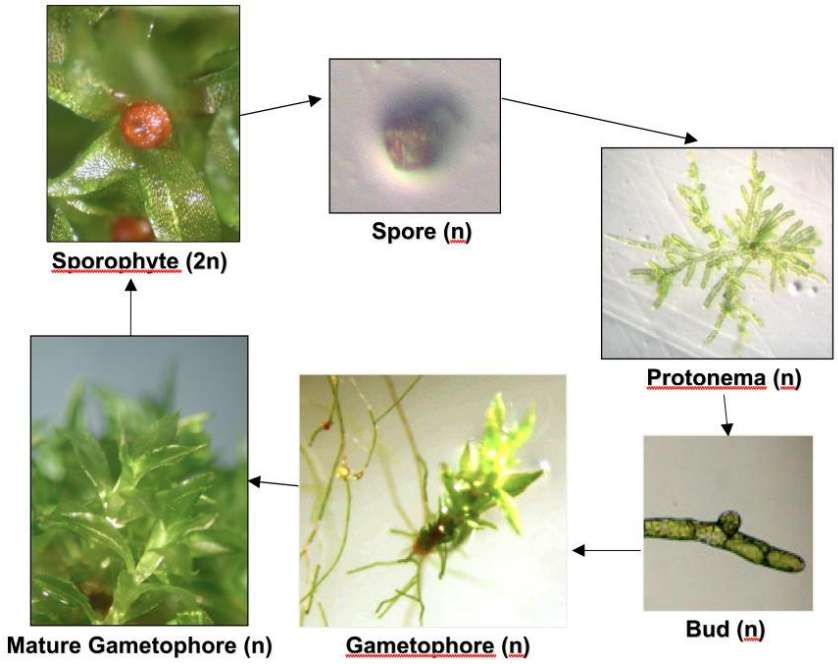

Figure 1.2: **Development and life cycle of the moss *Physcomitrella patens*.** Photos by Sung Hyun Cho.

1.2 Overview of small RNAs

Although unnoticed until the beginning of the 21st century, endogenous small RNAs and their associated silencing pathways are ubiquitous in eukaryotes. They are evolutionarily conserved since the last common ancestor of plants and animals (Jones-Rhoades et al. 2006). Endogenous small RNAs are diverse, non-coding and play important regulatory roles. There are two main distinctions between different types of small RNAs in terms of their origin of biogenesis: some are derived from single-stranded precursors with an imperfect stem-loop structure, known as the “hairpin” structure, and thereby are referred as hairpinRNAs (hpRNAs), whereas others are processed from double-stranded RNA (dsRNA) precursors (Axtell 2013a). Endogenous small RNAs,

whose functions are relatively well defined and can be categorized into three groups: microRNAs (miRNAs), short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). These non-coding small RNAs have distinct sizes with miRNAs typically being 21-22 nucleotides (nts), siRNAs typically being 21-24 nts and piRNAs typically 26-31 nt in length (Röther and Meister 2011; Axtell 2013a).

Although miRNAs are deeply conserved in both plant and animal kingdoms, there are significant differences between the biogenesis and scope of miRNA-mediated target regulation. The first discovered miRNA, *lin-4* RNA, controls the timing of larval development in the nematode *C. elegans* via pairing with the 3'UTR of *lin-14* mRNA and thus repressing its translation (Lee et al. 1993; Wightman et al. 1993). Subsequent studies revealed the presence of a large class of regulatory ~22 nt noncoding miRNAs which have been found in various organisms, from viruses to metazoans and plants (Jones-Rhoades et al. 2006).

In animals, long primary transcripts (pri-miRNAs) are cleaved by a ribonuclease III enzyme called Drosha to produce stem-loop structures (pre-miRNAs) in the nucleus, and pre-miRNAs are then processed into mature miRNAs by another ribonuclease III enzyme called Dicer in the cytoplasm (Lee et al. 2002, 2003). However, plants have no clear ortholog to Drosha. Instead, both pri-miRNAs and pre-miRNAs are processed by the same enzyme, Dicer-like 1 (DCL1), a Dicer homolog that affects mature miRNA levels in *Arabidopsis* and is found in the nucleus (Kurihara and Watanabe 2004; Park et al. 2002). In animals, miRNA-regulated processes generally include the control of cell proliferation, cell death, and timing in development (Abrahante et al. 2003; Brennecke et al. 2003; Johnston and Hobert 2003; Lin et al. 2003); whereas in plants, miRNAs have been found to play critical roles in regulating leaf morphology (Chen et al. 2002; Chen 2004), flower development (Palatnik et al. 2003), stress responses (Jones-Rhoades and Bartel 2004; Navarro et al. 2006), and nutrient homeostasis (Fujii et al. 2005; Chiou et al. 2006).

siRNAs were first identified in plants (Hamilton and Baulcombe 1999) and are more abundant than miRNAs in flowering plants (Jones-Rhoades et al. 2006). Deep sequencing of small RNAs has revealed that many miRNAs are conserved between relatively close organisms, whereas most endogenous siRNAs are very diverse (Llave et al. 2002a; Sunkar and Zhu 2004; Sunkar et al. 2005; Lu et al. 2006; Rajagopalan et al. 2006). siRNA-mediated silencing confers viral and bacterial resistance (Voinnet 2001;

Waterhouse et al. 2001; Katiyar-Agarwal et al. 2006), protects the genome from mobile DNA elements (Tabara et al. 1999; Wu-Scharf et al. 2000), and acclimates plants to abiotic stress (Borsani et al. 2005). Although silencing pathways utilizing small RNAs have much in common, there are some fundamental distinctions between the three classes of small RNAs, particularly in regard to their origin of biogenesis, evolutionary conservation and their targets (Bartel and Bartel 2003). The most striking difference between miRNAs and siRNAs is related to their origin of biogenesis: while siRNAs are processed from long, dsRNA duplexes formed by intermolecular hybridization of complementary RNA strands (Elbashir et al. 2001), miRNA precursors are single RNA molecules that fold back to form an imperfect stem-loop ("hairpin") structure (Lagos-Quintana et al. 2001; Lee and Ambros 2001).

Piwi subfamily proteins have only been identified in animals, thus, piRNAs have been exclusively observed in animals. In contrast to miRNAs and siRNAs, which are processed from the helical regions of RNA precursors by Dicer or DCLs, Dicer-independent piRNAs are presumably produced from long, non-helical RNA precursors (Brennecke et al. 2007; Gunawardane et al. 2007). piRNAs were found to be 2'-O-methylated at their 3'ends in *Drosophila* and mouse (Houwing et al. 2007; Kirino and Mourelatos 2007a; Ohara et al. 2007; Kirino and Mourelatos 2007b; Ohara et al. 2007; Horwich et al. 2007; Saito et al. 2007). Unlike AGO-binding miRNAs and siRNAs, piRNAs specifically interact with Piwi subfamily proteins that are enriched in the germline of many animals (Girard et al. 2006; Aravin et al. 2006; Lau et al. 2006; Watanabe et al. 2006; Grivna et al. 2006). Repeat-associated small RNAs derived from transposons were shown to interact with Piwi subfamily proteins in *Drosophila* (Vagin et al. 2006; Saito et al. 2006). Piwi/piRNA complexes were shown to be essential in germline maintenance by silencing transposons in *Drosophila*, mouse and zebrafish (Aravin et al. 2006; Watanabe et al. 2006; Carmell et al. 2007; Houwing et al. 2007). Repression of transposition in the animal germline is presumably mediated by both epigenetic suppression and transposon mRNA slicing.

1.3 Plant miRNAs

MiRNA pathway is deeply conserved within both the animal and plant lineages; however, the fact that not a single miRNA is found to be common between plants and animals suggests that plant miRNAs and animal miRNAs have evolved independently. Although several miRNA families were shown to be highly conserved between basal land plants and angiosperms; the majority of miRNAs have lineage-specific, distinct small RNA sequences, yet evolved to play common biological functions in plants (Axtell et al. 2007; Axtell and Bowman 2008). Initially identified in *Arabidopsis thaliana* and *Oryza sativa* (rice), plant miRNAs are generally 21 nt long and mostly depend on *DCL1* for their biogenesis (Axtell 2013a). Deeply conserved *MIRNA* families have relatively high expression levels and contain multiple paralogous loci (Rajagopalan et al. 2006; Axtell et al. 2007; Chávez Montes et al. 2014). The majority of these conserved miRNAs target transcription factors, while the targets of non-conserved miRNAs have more diverse functions (Fahlgren et al. 2007; Howell et al. 2007).

MIRNA genes are transcribed by RNA polymerase II (Pol II) mostly from intergenic regions of the genome, and the resulting pri-miRNAs are stabilized by the addition of a 5'cap and a 3'polyadenine tail (Cai et al. 2004; Lee et al. 2004; Xie et al. 2005). In plants, pri-miRNAs are processed into stem-loop hairpin-like precursors (pre-miRNAs) and further processed predominantly by DCL1 into short dsRNA consisting of mature miRNA guide and passenger (miRNA*) strands with 2-nt 3' overhangs. DCL1 is part of a family of four DCL proteins (Margis et al. 2006). Each DCL protein produces distinct sizes of small RNAs: DCL1 and DCL4 typically generate 21 nt long small RNAs, DCL2 generates 22 nts, and DCL3 generates 24 nt long small RNAs (Xie et al. 2004; Akbergenov et al. 2006; Deleris et al. 2006; Cuperus et al. 2010). Most plant miRNAs are 21 nt long and most *MIRNA* genes depend on DCL1, the double stranded RNA-binding protein HYPONASTIC LEAVES 1 (HYL1), the C2H2-zinc finger protein SERRATE (SE), the RNA-binding protein DAWDLE (DDL), and the nuclear cap-binding complex (CBC) for miRNA biogenesis (Kurihara and Watanabe 2004; Voinnet 2009). As opposed to the relatively recently evolved miRNAs, *DCL1*-dependent miRNA biogenesis seems to be specialized for older miRNAs. For instance, biogenesis of two of the younger miRNAs, AtMIR822 and AtMIR839, depend on *DCL4*, instead of *DCL1*, possibly because secondary structures of their precursors have not evolved sufficiently to be

preferentially recognized by DCL1 (Rajagopalan et al. 2006). The 3' nts of the DCL1-generated miRNA/miRNA* duplexes are then 2'-*O*-methylated by the methyltransferase HEN1 to be protected from exonuclease degradation and exported to the cytoplasm by the Exportin 5 ortholog HASTY (HST) (Park et al. 2002; Yu et al. 2005; Park et al. 2005).

Most miRNAs are loaded into a member of the AGO1-clade of AGO proteins for downstream target repression (Vaucheret 2008). Sufficient pairing between the AGO-loaded miRNA and the target mRNA was shown to mediate slicing of the target mRNA by the endonucleolytic cleavage of the associated AGO protein (Llave et al. 2002b; Dunoyer et al. 2004; Carbonell et al. 2012). Later, some studies revealed the loss of miRNA target proteins despite the lack of an apparent decrease in the target mRNAs, a phenomenon previously observed in animals (Aukerman and Sakai 2003; Chen 2004). Subsequent studies have further confirmed that translational repression is a widespread mechanism by which plant miRNAs regulate their targets (Gandikota et al. 2007; Brodersen et al. 2008; Yang et al. 2012). Altogether, the current data indicate that miRNA-mediated target gene repression in plants involves both AGO-catalyzed mRNA destabilization and/or translation inhibition (Gandikota et al. 2007; Brodersen et al. 2008; Carbonell et al. 2012).

1.4 Heterochromatic siRNAs

1.4.1 Biogenesis and function of heterochromatic siRNAs

Heterochromatic siRNAs are derived from intergenic and/or repetitive regions of the genome and are associated with 5-methyl cytosine (5-mC), particularly at asymmetric CHH sites (where H = A, T, or C), and histone H3 lysine 9 (H3K9) methylation marks (Law and Jacobsen 2010; Law et al. 2013; Zhang et al. 2013a). Heterochromatic siRNAs are characterized by their distinct sizes, which have been typically described as 24 nt long (Axtell 2013a). Proteins required for the production and accumulation of heterochromatic siRNAs include one of the non-canonical, plant-specific DNA-dependent RNA polymerases, Pol IV, an RNA-DEPENDENT RNA POLYMERASE 2 (RDR2), and DICER-LIKE 3 (DCL3) (Xie et al. 2004; Herr et al. 2005; Kanno et al. 2005; Onodera et al. 2005; Pontier et al. 2005; Wierzbicki et al. 2008). Precursors of the heterochromatic siRNAs are transcribed by Pol IV in the nucleus (Wierzbicki 2012).

Recent studies suggest that histone modification and siRNA-guided DNA methylation form a positive feedback loop to reinforce transcriptional silencing (Fig. 1.3). In this model, the presence of H3K9 methyl marks lead SAWADEE HOMEODOMAIN HOMOLOG 1/DNA-BINDING TRANSCRIPTION FACTOR 1 (SHH1/DTF1) to recruit Pol IV to the loci where precursors of heterochromatic siRNAs are transcribed (Law et al. 2013; Zhang et al. 2013a). Those single-stranded RNAs are converted into dsRNAs by RDR2, which are then processed by DCL3 to generate 24 nt siRNAs (Xie et al. 2004; Daxinger et al. 2009).

Once they are cleaved by DCL3, one strand of the siRNA duplex is loaded into one of the AGO4-clade AGOs (AGO4, AGO6 and AGO9 in *Arabidopsis*) for processing (Havecker et al. 2010). Unlike miRNAs, slicing activity of the AGO4 has been shown to be required for the loading of siRNAs into associated AGO in the cytoplasm (Ye et al. 2012). Another plant-specific DNA-dependent RNA polymerase, Pol V, was shown to produce transcripts that serve as scaffolds for the binding of AGO4-loaded heterochromatic siRNAs at the vicinity of these siRNA production (Wierzbicki et al. 2008). Sufficient complementarity between the AGO4-bound heterochromatic siRNAs and nascent Pol V transcripts leads to the further recruitment of other chromatin modifying enzymes, such as histone-modifying enzymes and *de novo* cytosine methyltransferase DRM2, to trigger the *de novo* deposition of repressive chromatin modifications including CHH methylation and H3K9 histone dimethylation at the vicinity of Pol V occupancy (Wierzbicki et al. 2012; Zhong et al. 2012). A very recent study suggests that methyl-DNA binding SUVH2/SUVH9 proteins recruit Pol V to the pre-existing DNA methylation marks to induce subsequent transcription which acts as a self-reinforcing loop to maintain transcriptional repression via DNA methylation (Johnson et al. 2014). This epigenetic modification guided by AGO4-associated heterochromatic siRNAs is also recognized as RNA-directed DNA methylation (RdDM) in plants (Havecker et al. 2010).

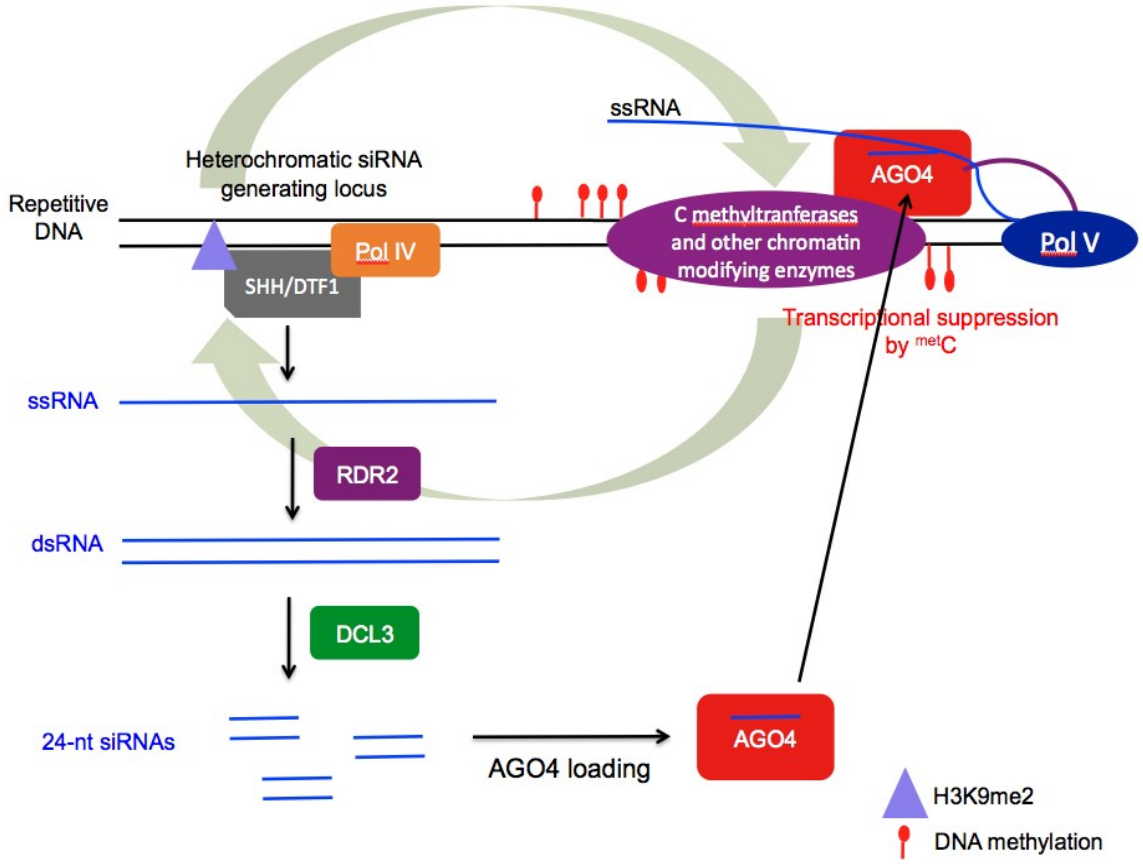

Figure 1.3: **Current model for biogenesis and function of heterochromatic siRNAs in plants.**

Left side of the figure illustrates how heterochromatic siRNAs are produced. Right side of the figure represents their downstream targeting once they are bound to AGO. Purple triangle indicates the pre-existing H3K9 dimethylation mark, while the red lollipop represents cytosine methylation. Arrows indicate the positive feedback between 5-mC and histone methylation. Abbreviations: Pol IV, RNA polymerase IV; SHH/DTF1, SAWADEE HOMEODOMAIN HOMOLOG 1/DNA-BINDING TRANSCRIPTION FACTOR 1; RDR2, RNA-dependent RNA polymerase 2; DCL3, DICER-LIKE 3; AGO4, ARGONAUTE 4; Pol V, RNA polymerase V; ssRNA, single-stranded RNA; dsRNA, double-stranded RNA.

1.4.2 Conservation of heterochromatic siRNA pathway

Research on plant species other than *Arabidopsis thaliana* has revealed conservation of heterochromatic siRNAs and the components involved in their biogenesis and downstream functions. For instance, accumulation of 24 nt heterochromatic siRNAs in young maize ears is dramatically reduced in the absence of *MOP1*, *RDR2* ortholog in maize (Nobuta et al. 2008). Also, *OsDCL3a* and *OsRDR2*, *DCL3* and *RDR2* orthologs in rice, are responsible for the accumulation of 24 nt small RNAs, which constitute the majority of the small RNA population in wild-type rice, and deposit DNA methylation at their target loci (Wu et al. 2010). Deep sequencing of small RNAs and phylogenetic analyses of the moss *P. patens* reveal the presence of the ortholog for angiosperm DCL3 accompanied with 22-24 nt siRNA production at repetitive and intergenic regions of the genome (Cho et al. 2008). This suggests that the heterochromatic siRNA pathway evolved prior to the radiation of the seed-bearing plants (Axtell et al. 2007). However, the failure to identify a *DCL3* ortholog in the conifer, *Pinus contorta,* coupled with the apparent lack of 24 nt small RNAs suggests that 24 nt siRNAs, which are involved in heterochromatin silencing in angiosperms, might have been lost in the conifer lineage (Dolgosheina et al. 2008). When compared to *Oryza sativa s*mall RNA profiles, the absence of 24 nt siRNAs in *P. contorta* was coupled with the relatively higher diversity of 21 nt small RNAs, some of which bear features of siRNAs (Morin et al. 2008).

Despite the genome size variation among gymnosperms (Burleigh et al. 2012), the fact that *Pinus* species have larger genomes containing appreciable amounts of retrotransposons (Morse et al. 2009) suggests that the pine genome should somehow be protected from transposable elements (TEs), as 24 nt siRNA–mediated silencing has evolved for this purpose in angiosperms. The compensatory mechanism in pines could be explained by the functional replacement of the heterochromatic siRNAs by 21 nt small RNAs (Morin et al. 2008). The fact that the deeper-branching *Selaginella* lacks the accumulation of 24 nt small RNAs despite the presence of full siRNA pathway machinery suggests that heterochromatic siRNA silencing pathway could be temporally and/or spatially controlled (Banks et al. 2011). In contrast, a very recent study revealed the accumulation of 24 nt small RNAs in the fern *Marsilea* (Chávez Montes et al. 2014). Taken together, heterochromatic siRNA-induced genome silencing appears to be an ancestral pathway within the land plants, as the deeper-branching *Physcomitrella*

possesses DCL3-generated 24 nt siRNA accumulation at certain regions of its genome, yet available data from species other than flowering plants suggest that heterochromatin silencing might involve different sizes of small RNAs and/or tissue-specificity (Axtell 2013a). For instance, gymnosperms show a great variation in their small RNA size accumulations, with the obvious presence of 24 nts in *Cycas*, but the apparent lack of 24 nts in *Picea*. (Chávez Montes et al. 2014). Interestingly, even though the heterochromatic siRNA pathway appears to be conserved across different species, individual heterochromatic siRNA loci are quite distinct, as opposed to the individual miRNA loci, which can be conserved among multiple species.

1.4.3 Role of RNA-directed DNA methylation in heterochromatin assembly

In eukaryotes, small RNA-mediated gene silencing is a widespread phenomenon involved in viral resistance, gene regulation, and genome maintenance. Small RNAs can induce both transcriptional gene silencing (TGS) by deploying repressive epigenetic modifications, such as DNA methylation and histone methylation, and post-transcriptional gene silencing (PTGS) by means of transcript degradation or translation inhibition (Matzke and Mosher 2014). The first evidence that links small RNAs with DNA methylation came from plants infected with viroids where dsRNAs were shown to trigger RNA-interference (RNAi) (Wassenegger et al. 1994; Mette et al. 2000). Later, it was found that proteins involved in RNAi are also required for the RdDM pathway which is composed of a number of different proteins (Matzke and Birchler 2005). Subsequent studies of other organisms, such as fission yeast, demonstrated that epigenetic regulation of heterochromatin is best characterized by DNA methylation and covalent histone modifications, both of which involve RNAi (Volpe et al. 2002). In plants, maintenance of the genome integrity is primarily executed by repressing heterochromatin, which is enriched in repetitive DNA elements. Suppression of the heterochromatic genome is exerted through TGS, with cytosine methylation and histone H3 methylation on lysine-9 dimethylation (H3K9me2) being the best characterized mechanisms (Martienssen and Colot 2001; Zilberman et al. 2003). Available data show that cytosine methylation is a eukaryotic gene-silencing mechanism which protects the genome from transposable elements (TEs) and regulates expression of genes whose promoters contain repetitive elements (Lippman and Martienssen 2004; Matzke and Mosher 2014).

Silencing of repetitive DNA through RdDM was described in *Arabidopsis* where 24 nt siRNAs, processed from long dsRNA molecules by DCL3 processing, are loaded into AGO4 in order to induce subsequent DNA methylation at the target loci (Baulcombe 2004; Zilberman et al. 2004; Wierzbicki et al. 2008). In addition to RNAi proteins, RdDM requires DNA methyltransferases, namely, chromomethylase 3 (CMT3), and domains rearranged methyltransferase 1 and 2 (DRM1/DRM2) (Lindroth et al. 2001; Cao and Jacobsen 2002, 3; Huettel et al. 2007). Two plant-specific Pol II-related RNA-dependent RNA polymerases, Pol IV and Pol V; the chromatin-remodeling protein DEFECTIVE IN RNA-DIRECTED DNA METHYLATION (DRD1); and DEFECTIVE IN MERISTEM SILENCING (DMS3) comprise the other components of the RdDM pathway (Zhang et al. 2006; Pikaard et al. 2008; Matzke et al. 2009; Kanno et al. 2010). Methylation of an endogenous *FWA* locus at its promoter containing tandem repeats, and other repeat-containing loci, such as *MEA-ISR* and *SUP* were shown to be dependent on RdDM pathway components (Soppe et al. 2000; Cao and Jacobsen 2002). Accumulation of repeat-associated siRNAs at the *FWA* promoter coupled with the presence of an asymmetric CHH methylation were shown to be *dcl3*-, *rdr2*-, and *ago4*-dependent (Chan et al. 2004; Lippman et al. 2004). Similarly, the *SINE* element containing locus *AtSN1* accumulates siRNAs in an RdDM pathway-dependent manner (Zilberman et al. 2003; Xie et al. 2004). The spread of DNA methylation, which is constrained by the small RNA-Pol V-generated transcript sequence homology, is a distinct feature of the RdDM pathway (Fig. 1.3). Overall, RdDM provides a sequence-specific silencing mechanism to maintain genome integrity by repressing heterochromatin; therefore, investigating the sequence requirements for heterochromatic siRNAs to repress transposons and repeats is of important interest for future experiments.

1.5 *DICER* and *Dicer-Like* (*DCL*) genes

1.5.1 Functional domain organization of Dicer genes

Dicer is the primary RNA recognition and processing protein in the RNAi machinery. It anchors dsRNAs and cuts it into small RNA duplexes that act as sequence-specific regulators after incorporation into associated AGOs (Ketting et al. 2001; Hammond et al. 2001). Dicer proteins are generally composed of an N-terminal

DExD ATPase/RNA helicase (Bass 2000), a central dsRNA binding domain DUF 283 (Dlakić 2006), a Piwi/Argonaute/Zwille (PAZ) domain (Matsuda et al. 2000), two catalytic RNase III domains, and a C-terminal dsRNA binding domain (dsRBD) (Tabara et al. 1999, 2002) (Fig. 1.4). The PAZ domain anchors the 3'-end of the dsRNA which is then cleaved by the two RNase III domains, RNase IIIa and RNase IIIb, forming an intramolecular heterodimer (Ketting et al. 2001; Yan et al. 2003; Ma et al. 2004; Zhang et al. 2004; MacRae and Doudna 2007). Cleavage by RNAse III domains generates a duplex with 2-nt 3'-overhangs, both of whose strands posses a 5'-monophosphate and 3'-hydroxyl. The distance between the PAZ domain and the active sites of the RNase III domains determines the length of the small RNA product. Therefore, this "catalytic core" is sufficient for Dicer to act as a molecular ruler for some organisms (MacRae et al. 2006, 2007). However, some animal Dicers possess an N-terminal extension of the PAZ domain which has been shown to be critical for precise processing of miRNAs (Park et al. 2011). The absence of the PAZ domain in ciliate, fungal and algal Dicers points out possible Dicer-interacting proteins which might lead Dicer to its template (Margis et al. 2006). The fact that R2D2 interacts only with PAZ-less Dicer-2 in *Drosophila* but not with PAZ-containing Dicer-1 suggests that adaptor molecules might positively regulate Dicer-2 dependent siRNA production as R2D2-Dicer-2 complex, but not the PAZ-less Dicer2 alone, binds to siRNAs (Liu et al. 2003, 2; Carmell and Hannon 2004).

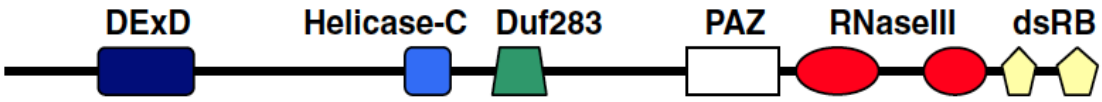

Figure 1.4: **Domains typically found in DCL or DCR proteins.** Adapted from (Margis et al. 2006). Common domains of DCL or DCR proteins are illustrated from the N-terminal to the C-terminal region of the protein: N-terminal DExD Helicase domain, dsRNA binding Duf283 domain, dsRNA-anchoring PAZ domain, two catalytic RNase III domains, and the C-terminal dsRNA binding domain.

The helicase domain is suggested to facilitate Dicer's movement along its dsRNA substrate and to be involved in the processing of siRNAs, but not miRNAs (Welker et al. 2011). The preference of *Drosophila* Dicer1, which does not have a functional helicase domain, for miRNA biogenesis suggests that the helicase domain could be used as a surveillance system for recognizing the ends of viral RNA and transposable elements, and Dicer2 might have evolved for this antiviral function as it contains the helicase

domain (Welker et al. 2011; Mukherjee et al. 2013). It is important to point out that Dicers in some unicellular eukaryotes do not necessarily include all of these functional domains. For instance, the lack of the dsRBD and the helicase domain in *Schizosaccharomyces pombe* (fission yeast), *Giardia lamblia* (protozoan parasite), and *Tetrahymena thermophila* (ciliate protozoa) suggests that Dicer dsRBD and Dicer helicase domain might have co-evolved (Mochizuki and Gorovsky 2005; Margis et al. 2006; MacRae and Doudna 2007).

1.5.2 Evolution of eukaryote Dicer genes

The evolution and diversification of the *DICER* gene family has been investigated using the available complete and near-complete genome sequences of various eukaryotic organisms (Mukherjee et al. 2013). Phylogenetic analysis supports independent expansions of the ancient Dicer protein in animals, plants, and fungi; yet Dicer paralogs in animals and plants appear to have a monophyletic origin (Fig. 1.5; Bernstein et al. 2001; Mukherjee et al. 2013). Based on the homology-based gene identification analyses, current data suggest an early eukaryotic origin of Dicer with evidence supporting its presence in animals, plants, fungi and many protozoan lineages, but not in bacteria and Archaea (Cerutti and Casas-Mollano 2006). The *DICER* gene family emerged early in eukaryotes and independently diverged in plants and animals. This is also reflected in the changing number of Dicer family members in different lineages: insects and fungi have two *Dicer-Like* genes, while many animals, including humans, have only one Dicer gene (Hammond 2005; Margis et al. 2006).

Most model plants, on the other hand, contain four DCL enzymes, which are suggested to have originated very early in plant evolution, and rapidly diversified before the divergence of moss from higher plants (Mukherjee et al. 2013). It appears that the number of *DCL*s has increased throughout the evolution of plants, in contrast to the decrease observed during the course of animal evolution. Current data show that plant *DCL*s are involved not only in regulating development but also in forming a defense system against viruses and transposons (Margis et al. 2006). In contrast, mammals have only one Dicer to process miRNAs and miRNAs were shown to predominantly decrease target mRNA levels (Guo et al. 2010). It has been suggested that unlike plants, there was no need for a Dicer-dependent defense mechanism in mammals as they have evolved immune systems to protect themselves against invaders (Margis et al. 2006).

Similarly, it appears that antiviral Dicer2 in *Drosophila* was lost in lineages with alternative antiviral defense mechanisms (Mukherjee et al. 2013).

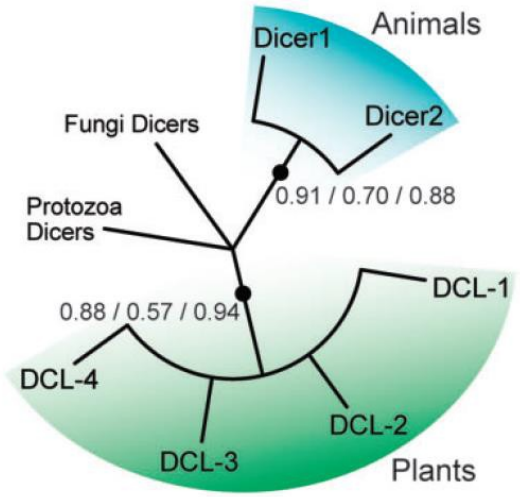

Figure 1.5: **Phylogenetic analysis of Dicers in different kingdoms.** Adapted from (Mukherjee et al. 2013). Support for monophyletic expansions of Dicer paralogs in animals and plants is plotted. Support is given as SH-like aLRT scores/maximum-likelihood bootstrap proportions/Bayesian posterior probabilities.

1.5.3 *DCL* genes in plants

Unlike animals, plant Dicer genes are more numerous (typically four *DCL*s 1-4) suggesting a functional diversification for each Dicer gene. However, it has been shown that paralogs of *DCL*s involved in antiviral defense might partially compensate for each other, leaving a potential for more research on how *DCL*s diverged throughout plant evolution (Gasciolli et al. 2005). Based on the phylogenetic analysis using available genome sequences, it was suggested that a *DCL* gene underwent a rapid four-way duplication early in plant evolution, before or around the divergence of moss from higher plants (Mukherjee et al. 2013). Although estimating the precise timing of DCL2/4 duplication is enigmatic, current data supports a very early emergence of four of the *DCL*s coincident with the origin of multicellularity.

Plant DCL1s, DCL3s and DCL4s contain a second dsRBD, a unique feature absent in non-plant Dicers. Double-stranded RNA binding motif (drsm) domains typically guide the hand-off of the template RNA from Dicer to an AGO protein, so differences in

these C-terminal dsRBDs might play critical roles for determining downstream partners of Dicers in the RNAi pathway (Marques et al. 2010). The Dicer PAZ domain has a unique feature conserved between plants and animals, and binds the RNA-ends mainly through electrostatic interactions by bearing a positively charged pocket (Ma et al. 2004; Wang et al. 2009). However, DCL4 PAZ RNA-binding pocket appears to be variable even between plant species such as *Arabidopsis* and rice, and primarily positively charged particularly in monocots (Mukherjee et al. 2013). This variation in the RNA-binding properties of DCL4 could reflect an ongoing change in response to the long-term evolutionary arms race with viral factors as *DCL4* seems to be specialized for antiviral immunity (Bouche et al. 2006; Deleris et al. 2006; Mukherjee et al. 2013).

The Dicer gene family in *Arabidopsis thaliana* has four members two of which, *DCL1* and *DCL3*, have relatively better understood functions (Schauer et al. 2002). *DCL1* is required for miRNA biogenesis (Papp et al. 2003; Finnegan et al. 2003; Xie et al. 2004), *DCL3* mainly generates 24 nt siRNAs corresponding to retroelements and transposons which maintains heterochromatin silencing (Hamilton et al. 2002; Xie et al. 2004). These four types of *DCL*s were found to be present in other flowering plants, such as poplar and rice. Poplar contains single orthologs of *AtDCL1*, *AtDCL3*, and *AtDCL2*, and two orthologs for *AtDCL2*. Rice has single orthologs of *AtDCL1* and *AtDCL4* with a pair of orthologs of *AtDCL2* and *AtDCL3* (Margis et al. 2006). *DCL2* paralogs in both poplar and rice appear to be quite similar to each other with 85% and 99% amino acid sequence similarity, respectively. *DCL3* paralogs in rice, *OsDCL3A* and *OsDCL3B*, are more divergent with only 57% similarity at the amino acid level (Margis et al. 2006). Phylogenetic analysis suggests that the duplication event that created paralogs of *DCL3* in rice occurred prior to the common ancestor of barley and rice (Margis et al. 2006).

1.6 Annotation of small RNA-producing genes in plants

1.6.1 Complications in miRNA annotation

*MIRNA*s (the loci which produce mature miRNAs) have received much attention and are thus the best annotated type of small RNA genes in plants. *MIRNA* annotations are disseminated by miRBase (Kozomara and Griffiths-Jones 2011). Currently, miRBase

(release 20) houses annotations of hundreds of *MIRNA* genes from 72 plant species. Community accepted standards specific for the features of plant *MIRNA*s guide miRBase submissions (Meyers et al. 2008). The basic premise of miRBase is that a hairpin RNA transcribed from the *MIRNA* locus is processed to ultimately yield a single functional mature miRNA; the minimal miRBase entry consists simply of a hairpin and a single linked mature miRNA sequence. However, the reality of miRNA expression is now known to be much more complex.

Related *MIRNA* hairpins often produce mature miRNAs that vary in length, sequence, or both. This variation can result from expression of multiple paralogous *MIRNAs* that differ slightly in sequence, creating several slightly different mature miRNAs. Another, very common type of miRNA variation is the result of differentially processed and/or truncated RNAs from the same hairpin (Fig. 1.6A). To illustrate how common such variation is, we aligned small RNA-seq data from wild-type *Arabidopsis* flowers and leaves (NCBI GEO GSM738731 and GSM738727; (Liu et al. 2012a)) to the *Arabidopsis* nuclear genome, and compared the alignments to annotations from miRBase 20. Precision^{ann} values (the fraction of all alignments to a hairpin corresponding to the miRBase-annotated mature miRNA) were often very poor (Fig. 1.6B). The distribution of precision^{max} values (the fraction of all alignments to a hairpin corresponding to the most abundantly observed small RNA) values was better, but nonetheless showed that it is very rare for an annotated *MIRNA* hairpin to produce just one discrete RNA (Fig. 1.6C). In our analysis the most abundant RNA was NOT annotated as the mature miRNA for the majority of *Arabidopsis MIRNA* loci (Fig. 1.6D). According to our current understanding, only AGO-loaded small RNAs are functional. There is no guarantee that all RNAs observed via small RNA-seq are AGO-bound. We therefore aligned a set of small RNAs that co-immunoprecipitated with a major *Arabidopsis* AGO protein, AGO1 (NCBI GEO GSM989351; (Carbonell et al. 2012)), and performed a similar analysis. Based on the known preferences of AGO1 for RNA binding, this analysis was limited to *MIRNA* loci whose annotated mature miRNAs were 21 nts with a 5'-U. The distributions of precision values improved (Fig. 1.6E-F), as did the concordance between miRBase annotations and the observed most abundant RNAs (Fig. 1.6G). Nonetheless, extensive heterogeneity in miRNA accumulation was still apparent for nearly all known *MIRNA* loci. Two conclusions emerge from this simple analysis. First: there are large discrepancies between empirical data and miRBase in

terms of annotation of the mature miRNA. Second: even putting aside potential errors in annotation of mature miRNAs, nearly all known *MIRNA* hairpins produce more than a single product.

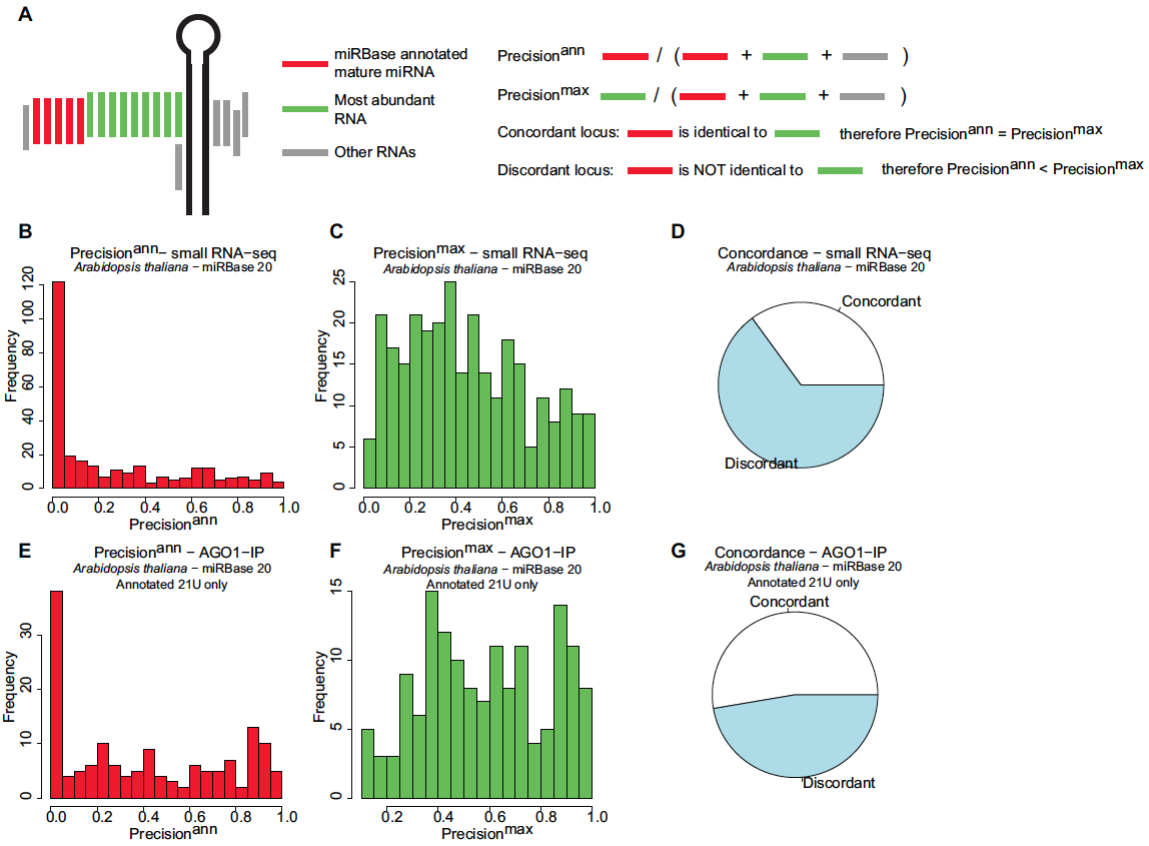

Figure 1.6: ***MIRNA* hairpins produce more than one product.** Adapted from (Coruh et al. 2014). **(A)** Schematic of a typical *MIRNA* locus with aligned reads from small RNA-seq, and explanation of terms and calculations. **(B)** Distribution of precision^{ann} values from *Arabidopsis MIRNA* loci with respect to miRBase 20. Based on genome alignment of a small RNA-seq dataset comprised of NCBI GEO GSM738731 and GSM738727. **(C)** As in B, except for precision^{max} values. **(D)** Frequency of concordance between miRBase 20 annotations of the mature miRNA, and the observed most abundant RNA for the small RNA-seq data. **(E)** As in B, except using small RNAs from an AGO1-IP experiment (NCBI GEO GSM989351), and restricting the analysis to *MIRNA* loci annotated with a mature miRNA 21 nt in length with a 5' U. **(F)** As in E, except for precision^{max} values. **(G)** Frequency of concordance as in D, except for AGO1-IP data and restricting the analysis to *MIRNA* loci annotated with a mature miRNA 21 nt in length with a 5' U.

One type of alternative RNA that arises from *MIRNA* hairpins are miRNA*s. In the canonical viewpoint of miRNA biogenesis, the miRNA* is defined as the strand of the

initial miRNA/miRNA* duplex that is discarded at the time of AGO-loading. However, there is ample evidence demonstrating that miRNA*s can also be AGO-loaded and functional. Many miRNA*s are enriched in AGO1 immunoprecipitates (Manavella et al. 2012), others associate with AGO2 (Zhang et al. 2011), and several have known functions (Zhang et al. 2011; Manavella et al. 2013). Positional variants outside of the annotated miRNA/miRNA* pair are also prominent features of plant *MIRNA* hairpin processing and they are known to have functional consequences (Vaucheret 2009). A very extensive study by Jeong et al. demonstrated that heterogeneity in *MIRNA* processing is quite common in *Arabidopsis*, and that in many cases there is compelling evidence for the functional relevance of these processing variants (Jeong et al. 2013).

Additional complexity in miRNA annotation arises due to various modifications of mature miRNAs that occur after dicing. HEN1 is a methyltransferase that catalyzes 2'-O-methylation of the 3'-most nucleotide of plant miRNAs and siRNAs (Yu et al. 2005). In *hen1* mutants, miRNAs display extensive 3'-truncations coupled with addition of non-templated nts (predominantly U) at the 3' end (Zhai et al. 2013). The truncated and tailed variants occur after the miRNAs are loaded onto the AGO1 protein, implying that these modifications could potentially affect the target specificity of the miRNAs. Importantly, 3'-truncation and 3' non-templated tailing also occur for some miRNAs in the wild-type background (Zhai et al. 2013), implying that this may be a mechanism used in normal conditions to modulate miRNA target specificity or mechanism of action.

1.6.2 *MIRNA* superfamilies

Another challenge in miRNA annotation is to accurately describe the evolutionary relationships between *MIRNA* loci. *MIRNA* loci are commonly grouped into families (which are assigned the same number) based on high levels of sequence similarity. However, the existence of *MIRNA* superfamilies, whose members have evidence of common descent and functions despite extensive sequence diversification, complicates this system. In one extreme example, both *Physcomitrella patens* (a moss) and flowering plants express miRNAs (miR904 and miR168, respectively) that target *AGO1* mRNAs, but the mature miRNAs have no detectable sequence similarity (Axtell et al. 2007). Whether this situation arose because of convergent evolution or extensive sequence diversification of a single ancestral miRNA is not clear. The miR482/2118 superfamily of miRNAs comprise a sequence-diverse set of mature miRNAs that are present in many

plant species, and frequently function to target nucleotide binding site-leucine-rich repeat (*NB-LRR*) innate immune receptor mRNAs (Zhai et al. 2011; Shivaprasad et al. 2012; Li et al. 2012), as well as other RNAs (Johnson et al. 2009). A second set of plant *MIRNA* superfamilies is comprised of the miR390, miR4376, and miR7122 superfamilies (Xia et al. 2013). Members of the miR390 superfamily are highly conserved in most plant species, but miR4376 and miR7122 superfamilies have highly diverse mature miRNAs in various species. Careful sequence analysis provides compelling evidence that the miR390, miR4376, and miR7122 superfamilies are all related by common descent (Xia et al. 2013). Curiously, all of these described superfamilies serve as initiators of secondary siRNA biogenesis. The observation of superfamilies whose members have diverged to the edge of reliable alignments suggests that many other evolutionary relationships between superficially unrelated *MIRNA*s may exist.

1.6.3 The annotation gap

miRBase is the main source for *MIRNA* annotations for all organisms. However, it is critical to emphasize just how minor the contribution of miRNAs is to the total small RNA expression profile of plants. To illustrate this, we compared *Arabidopsis* small RNA-seq alignments to miRBase annotations. As a counter-point, we also compared aligned polyA+ RNA-seq data to the TAIR10 mRNA annotations. The small RNA-seq dataset was from flowers and leaves as used in Figure 1.6 (Liu et al. 2012a). The RNA-seq dataset was also derived from flowers and leaves, and comprised 101 nt single-end reads from polyA-enriched samples (Liu et al. 2012b). To minimize contamination with breakdown products of abundant RNAs, rRNA, tRNA, snRNA, and snoRNA regions of the reference genome sequence were masked prior to alignment, and only alignments of 20-24 nt reads were retained for the small RNA-seq data. The RNA-seq data were aligned using a spliced aligner (TopHat; (Trapnell et al. 2009)) and randomly down-sampled to achieve an approximately equal number of alignments compared to the small RNA-seq data (32.5E6 and 35.7E6 alignments for the small RNA-seq and RNA-seq data, respectively). For the purposes of illustration, we considered a genomic position active if it had a coverage >= 0.1 reads per million, which equated to a depth of four or more alignments for both datasets. Based on this analysis, roughly 34 million and 12 million nucleotides of the *Arabidopsis* genome expressed significant polyA+ and small RNA, respectively (Fig. 1.7A). There was very little overlap between the two, indicating

that expression of long polyA+ RNA and 20-24 nt RNAs is usually mutually exclusive in these tissues. Annotated *MIRNA* loci account for only a tiny fraction of the genome that actively produces 20-24 nt RNAs (Fig. 1.7B, left). In contrast, nearly all of the polyA+ RNA-seq is explained by existing gene annotations (TAIR10; Fig. 1.7B, right). In terms of abundance, small RNAs aligned to annotated *MIRNA* hairpins were in the minority; however, nearly all of the polyA+ RNA-seq alignments fell within annotated genes (Fig. 1.7C). We do not believe this analysis implies a vast amount of un-annotated *MIRNA* loci. Instead, it highlights the fact that the majority of expressed plant small RNAs are NOT miRNAs, and that these in total account for roughly 10% of the *Arabidopsis* genome. Clearly, there is large 'annotation gap' between the empirical knowledge of small RNA expression and the annotations of small RNAs provided by miRBase alone.

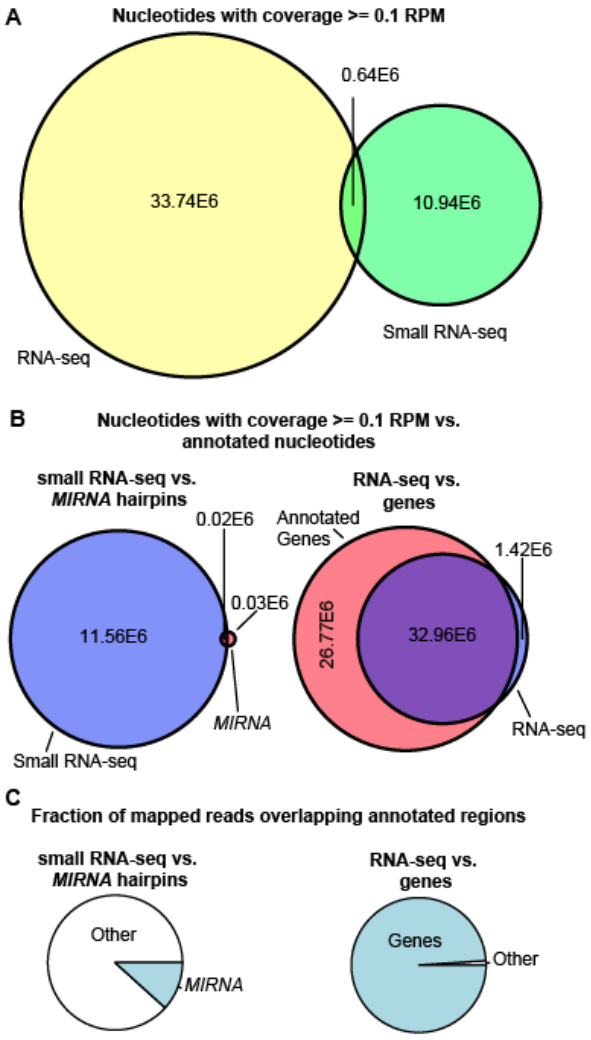

Figure 1.7: **The annotation gap: comparison of observed expression data to annotations for small RNAs (NCBI GEO GSM738731 and GSM738727) and polyA+ RNAs (NCBI GEO GSM946222 and GSM946223) in *Arabidopsis*.** Adapted from (Coruh et al. 2014). **(A)** Area-proportional Venn diagram showing the extent (number of nts) of significant (defined as a coverage of >= 0.1 read per million) polyA+ RNA (RNA-seq) and small RNA-seq expression in the *Arabidopsis* genome. **(B)** Area-proportional Venn diagrams illustrating the overlap between areas of significant small RNA-seq or RNA-seq expression and annotated regions in *Arabidopsis* (left: small RNA-seq vs. miRBase, right: RNA-seq vs. TAIR10 genes including introns). **(C)** Pie charts illustrating the proportion of aligned small RNA-seq reads overlapping *MIRNA* annotation (left), or the proportion of RNA-seq reads overlapping TAIR10 gene annotations including introns (right) for *Arabidopsis*.

1.6.4 Other hairpin-derived RNAs

Long inverted-repeat containing hairpin RNAs (hpRNAs) have long been used to manipulate plant mRNA expression levels (Chuang and Meyerowitz 2000; Wesley et al. 2001). Small RNAs derived from artificial hpRNA constructs are processed in a manner similar to the processing of viral RNAs and drive silencing of endogenous and exogenous genes as well as trigger long distance signals in *Arabidopsis* (Fusaro et al. 2006). Genome-wide scans find substantial correlations between small RNA accumulation and hairpins that do not qualify as miRNAs (Henderson et al. 2006; Axtell 2013b), implying that endogenous hpRNAs may be widespread. Only a few endogenous hpRNA loci have been characterized in depth. These include the *IR71* and *IR2039* loci in *Arabidopsis* (Dunoyer et al. 2010) and the *Mu killer* locus in maize (Slotkin et al. 2005). Systematic annotation of endogenous hpRNA loci has not yet been reported, and there are not yet clear community-accepted standards for discerning hpRNA loci. Nonetheless, the presence of endogenous hpRNA loci in different plant species such as the *IR71* and *IR2039* in *Arabidopsis*, and the *Mu killer* locus in maize suggests that there may be a great number of such genes.

1.6.5 Secondary, phased siRNAs

Secondary siRNAs are characterized by a distinct small RNA biogenesis pathway that requires the slicing of a primary transcript by a specific miRNA or other secondary siRNAs. The cleaved transcript is converted into a dsRNA by an RNA-dependent RNA polymerase and then processed by a DCL protein into siRNAs (Fei et al. 2013). Because the location of the initial cut is specified by an upstream small RNA cleavage, dicing of the dsRNA with a defined start point generates siRNAs in a “phased” pattern. Most annotated secondary siRNAs have been found using several similar algorithms based upon this characteristic phased pattern (Chen et al. 2007; Howell et al. 2007). However, in contrast to *MIRNA* loci, there is as yet no centralized database or registry devoted to this class of small RNA loci.

The classic examples of phased secondary siRNA loci are several families of non protein-coding RNAs termed *TRANS ACTING siRNA* (*TAS*) loci. Some phased siRNAs can repress target mRNAs in *trans*, hence the term *trans*-acting siRNAs (tasiRNAs). The extensively-studied *TAS3a/b/c* family is targeted at two sites by miR390 and produces conserved tasiRNAs that target *Auxin Response Factor* (*ARF*) mRNAs involved in

developmental timing and leaf polarity (Nogueira et al. 2007). *TAS3* is a particularly well-conserved *TAS* locus, and even has homologs in the moss *Physcomitrella patens* (Axtell et al. 2006). A linkage between miR390-controlled *TAS3* loci and a novel miR156-controlled *TAS* family, *TAS6*, has been identified in *Physcomitrella* (Arif et al. 2012; Cho et al. 2012). *TAS6a* and *TAS3a* are present on the same primary transcript, which has four miRNA target sites. *TAS6a* and *TAS3a* regions have two targets sites for miR156 and miR390, respectively. Inhibition of miR156 and over-expression of miR390 both delayed gametophore development, and resulted in the increased production of miR390-triggered tasiRNAs (Cho et al. 2012). These data demonstrate that *TAS* transcripts can serve as integration points that sense and respond to the accumulation of multiple miRNAs.

Protein-coding genes also can spawn secondary, phased siRNAs. Phased siRNAs from diverse sets of protein-coding genes have been observed in multiple plant species (reviewed by (Fei et al. 2013)). Assuming that some of the induced secondary siRNAs can act as tasiRNAs to target other members of large gene families, secondary phased siRNA production from protein-coding mRNAs may serve as a mechanism to achieve coordinate post-transcriptional repression for many transcripts at once. One example of special interest is members of the miR482/2118 superfamily, which target *NB-LRR* disease resistance mRNAs. In *Medicago truncatula*, miR2118, miR2109, and miR1507 cause large amounts of phased secondary siRNAs from at least 71 *NB-LRR* mRNAs (Zhai et al. 2011). High accumulation of these three miRNAs is seen across the Fabaceae (Zhai et al. 2011). In tobacco, miR6019 and miR6020 target the *N* resistance gene and cause extensive production of secondary phased siRNAs (Li et al. 2012). In tomato, sequence diverse members of the miR482 family also target large numbers of *NB-LRR* mRNAs, which in turn produce phased siRNAs (Shivaprasad et al. 2012). Importantly, both viral and bacterial infections of tomato correlate with decreased miR482 accumulation and increased *NB-LRR* accumulation (Shivaprasad et al. 2012). This suggests that pathogen-induced suppression of miRNA levels could serve to enhance *NB-LRR* expression, perhaps priming plant defense responses. This has the potential to be a wide-spread mechanism, as *NB-LRR* mRNAs are potent sources of phased siRNAs in many plant species, including the conifer *Picea abies* (Kallman et al. 2013).

1.6.6 Annotation of heterochromatic siRNAs

Heterochromatic siRNAs are the major components of the small RNA populations in most tissues of most plant species examined to date. Most angiosperm genomes have thousands of loci that produce heterochromatic siRNAs. They are the specificity determinants that guide the process of RNA-directed DNA methylation (RdDM), likely via the targeting of nascent long non-coding RNAs produced by a specialized DNA-dependent RNA polymerase, Pol V (reviewed by (Wierzbicki 2012)).

Less attention has been paid to systematic annotation of individual heterochromatic siRNA loci, and there is no miRBase-type registry or database for these types of genes. Several groups have, however, reported the results of in-house computational approaches that defined heterochromatic siRNA loci on the basis of simple clustering methods coupled with analysis of heterochromatic siRNA mutants (Mosher et al. 2008; Cho et al. 2008; Lee et al. 2012). Two recent studies have described the role of the SHH1/DTF1 DNA-binding protein in guiding the formation of *Arabidopsis* heterochromatic siRNAs and in the process defined sets of heterochromatic siRNA loci. Law et al. (Law et al. 2013) defined ~12,500 heterochromatic siRNA loci by clustering of uniquely mapping 24 nt siRNAs. Similarly, Zhang et al. (Zhang et al. 2013a) defined 4,187 loci comprised mainly of 24 nt RNAs that were strongly down-regulated in *dtf1* mutants. Both studies showed that *SHH1/DTF1* is a major regulator of heterochromatic siRNA levels. Importantly, SHH1/DTF1 is suggested to recruit Pol IV, which transcribes the precursors of heterochromatic siRNAs (Wierzbicki 2012), to loci based upon the presence of H3K9 methylation marks (Law et al. 2013; Zhang et al. 2013a). These data suggest that prior deposition of repressive histone modifications is a pre-requisite for heterochromatic siRNA biogenesis.

Several lines of evidence indicate that heterochromatic siRNA gene annotation should not depend on a rigid siRNA size requirement of 24 nts. *Arabidopsis* transposable elements that normally produce 24 nt heterochromatic siRNAs instead begin to produce appreciable amounts of 21-22 nt siRNAs in the dedifferentiated cell suspension cultures (Tanurdzic et al. 2008) and pollen (Slotkin et al. 2009). Nuthikattu et al. (Nuthikattu et al. 2013) demonstrated that, upon global erasure of DNA methylation in the *Arabidopsis ddm1* mutant, 15 families of transposable elements begin to produce very high amounts of 21-22 nt siRNAs. These are dependent upon *RDR6*, which had previously been associated with secondary and phased siRNAs, but not heterochromatic

siRNAs. The *RDR6*-dependent 21-22 nt siRNAs were capable of directing RdDM, making them *bona fide* heterochromatic siRNAs (Nuthikattu et al. 2013). Similarly, Marí-Ordóñez et al. (Marí-Ordóñez et al. 2013) also demonstrated that an epigenetically re-activated transposon, *EVD*, initially is targeted by 21-22 nt siRNAs. Over multiple generations of inbreeding, *EVD* is eventually silenced by RdDM. Interestingly, over the course of several generations, *EVD*-derived siRNAs transitioned from *RDR6*-dependent 21-22 nt siRNAs to Pol IV-dependent 24 nt siRNAs. Together, these studies suggest a model in which active transposable elements are first targeted by the secondary siRNA pathway, which makes 21-22 nt siRNAs that can cause both transcriptional and post-transcriptional silencing. Later, there is a gradual handoff to the 24 nt, Pol IV / Pol V heterochromatic siRNA pathway as the transcriptional silencing of the element becomes firmly entrenched. This implies that the prevalence of 24 nt heterochromatic siRNAs across many plant genomes represents a final 'maintenance' state for transposons and retroviruses that invaded long ago. There is evidence indicating that 21-22 nt 'initiation' state heterochromatic siRNA loci also exist in wild-type plants. Genome-wide analysis of DNA methylation in *Arabidopsis rdr6* mutants identified 138 loci with *RDR6*-dependent DNA methylation, most of which were associated with 21-22 nt siRNAs and distinct from the DNA methylation caused by the canonical heterochromatic siRNA pathway (Stroud et al. 2013). In maize, which has a huge load of very active transposons, there are large numbers of 22 nt small RNAs that are not dependent on the canonical 24 nt heterochromatic siRNA pathway (Nobuta et al. 2008).

1.6.7 Resources for creating and disseminating annotations

A great number of programs geared specifically to *MIRNA* locus annotation exist, with several that are specialized for the unique features of plant *MIRNA*s (Yang and Li 2011; Xie et al. 2012; Qian et al. 2012). Several related algorithms designed to detect the unique phasing signature of phased siRNA loci have also been described (Chen et al. 2007; Howell et al. 2007; De Paoli et al. 2009). General purpose clustering methods that define loci of small RNA production based on small RNA-seq alignments also are available (MacLean et al. 2010; Pantano et al. 2011; Hardcastle et al. 2012; Chen et al. 2012a). The UEA sRNA workbench (Stocks et al. 2012) contains several stand-alone programs that individually address *MIRNA* annotation, general small RNA cluster identification, and phased siRNA locus annotation. Our program, ShortStack (Axtell

2013b; Shahid and Axtell 2014), generates annotations of *MIRNA* loci, other hpRNA loci, phased siRNA loci, and all other types of small RNA loci. Recent versions of ShortStack have added the capability to handle read-trimming and alignment of data (Shahid and Axtell 2014), making it an integrated solution to generate small RNA gene annotations from raw small RNA-seq data.

Several web-based resources exist to disseminate plant small RNA gene annotations and related small RNA-seq alignment data (Table 1.1). As discussed above, miRBase (Kozomara and Griffiths-Jones 2014) is the central repository and arbitrator for *MIRNA* loci from all species. The Meyers Lab maintains one of the most extensive small RNA web servers, primarily focused on plant species (Nakano 2006). At present, 15 plant species are represented, each with easily queried databases of aligned small RNA-seq data, and custom-built genome browsers. Other web servers focus on small RNA-seq alignments and annotations for specified species (Backman et al. 2008; Johnson et al. 2007). To the best of our knowledge, the current web servers are primarily focused on providing and visualizing small RNA-seq alignment data, as opposed to the curation and dissemination of stable reference annotations (with the exception of *MIRNA*s from miRBase). To address this, we are developing a web server (plantsmallrnagenes.psu.edu) whose focus goes beyond delivery and visualization of alignment data by adding comprehensive reference annotations for small RNA-producing loci. As of this writing, the site hosts annotations for just two species (*Amborella trichopoda* and *Physcomitrella patens*), but a major expansion is planned over the next year.

Table 1.1. Selected websites that disseminate plant small RNA alignments and/or annotations.

Site Name	URL	Species currently present	Comments	Citation
miRBase	http://www.mirbase.org/	72 plants (as of version 20)	Disseminates *MIRNA* hairpin and mature miRNA annotations for all species, including plants.	(Kozomara and Griffiths-Jones 2011)
University of Delaware SBS databases	http://mpss.udel.edu/	15 plant species	Small RNA-seq, RNA-seq, PARE/degradome, and other high-throughput datasets with search functions and a custom-built genome browser for each species	(Nakano 2006)
ASRP	http://asrp.danforthcenter.org/	*Arabidopsis thaliana*	Disseminates small RNA-seq datasets and features a genome-browser.	(Backman et al. 2008)
CSRDB	http://sundarlab.ucdavis.edu/smrnas/	Maize and rice	Queryable small RNA-seq data along with target predictions and genome browsers	(Johnson et al. 2007)
The plant small RNA genes web server at Penn State	http://plantsmallrnagenes.psu.edu/	*Physcomitrella patens* and *Amborella trichopoda*	Disseminates global reference annotations of small RNA producing genes (all types), along with full datasets and genome browsers.	(Coruh et al. 2014)

1.7 Objectives

The primary objective of my research was to create reference annotations for the small RNA-producing genes in the deep-branching moss, *Physcomitrella patens*. We now know that regulatory small RNAs account for a significant fraction of the genome in plants. However, genome-wide, comprehensive annotation of small RNA genes has not been documented for the basal plants such as *Physcomitrella*. Therefore, I aimed to utilize small RNA sequencing (small RNA-seq) data analysis to characterize different types of small RNAs in wild-type *Physcomitrella.* In recent years, small RNA-seq, enabled by cheap and fast high throughput parallel DNA sequencing technologies, has become a powerful tool to generate enormous amount of data on plant small RNAs. Substantial effort has been put into improving next generation sequencing technologies, which has resulted in increased coverage (thereby high sensitivity) in sequenced libraries. Despite the tremendous increase in data volume, however, there are not yet community-accepted standards for categorizing different small RNA-producing loci and most of the current criteria rely on the previously characterized types of small RNAs. In

order to identify endogenous small RNA-producing genes reliably, I analyzed wild-type small RNA-seq reads using ShortStack (Axtell 2013b; Shahid and Axtell) and provided a comprehensive annotation and quantification of small RNA genes in *Physcomitrella*.

The secondary objective of my research was to test whether the heterochromatic siRNA pathway is conserved in the moss *Physcomitrella*. Abundance of heterochromatic siRNAs has been extensively shown in angiosperms but there are not sufficient data showing their existence in other land plants. It has been shown that the conifer genome does not encode a *DCL3* paralog (Dolgosheina et al. 2008), which is the major component for heterochromatic siRNA accumulation in flowering plants. However, we have previously demonstrated that the *Physcomitrella DCL3* homolog is required for the accumulation of 22, 23, and 24 nt RNAs from a handful of siRNA 'hotspots' (Cho et al. 2008). Based on our previous results, I hypothesized that heterochromatic siRNAs are expressed in the basal land plant *Physcomitrella*. To test this hypothesis, I used extensive small RNA-seq analysis in wild-type *Physcomitrella* in comparison with several *Physcomitrella* mutants potentially defective in the RNAi pathway.

Chapter 2 provides a comprehensive reference annotation of small RNA genes in wild-type *Physcomitrella*. It also compares the profiles of small RNAs in wild-type plants and RNAi-defective mutants in order to find which candidate proteins are involved in the accumulation of certain types of small RNAs. I aimed to generate a better understanding of these small RNA-producing genes by applying a differential expression analysis and a co-occupancy analysis of these loci against various genomic features. Proteins involved in the heterochromatic siRNA biogenesis are thereby identified in the basal plant *Physcomitrella*.

In Chapter 3, I summarize the conclusions of Chapter 2 and briefly discuss the challenges and key goals of annotating small RNA genes in plants reliably. I also discuss the prospects of small RNA gene annotation.

[2]Chapter 2

Comprehensive annotation of *Physcomitrella patens* small RNA loci reveals 23nt heterochromatic siRNAs dependent on a minimal Dicer-Like gene

2.1 Summary

Many eukaryotic small RNAs serve as sequence-specific negative regulators of target mRNAs and/or chromatin. They are involved in a variety of biological processes including viral resistance, gene regulation, and genome maintenance. In angiosperms (flowering plants), the two most abundant endogenous small RNA populations are usually 21 nt microRNAs (miRNAs) and 24 nt heterochromatic short interfering RNAs (siRNAs). Heterochromatic siRNAs derive from repetitive regions and direct DNA methylation and repressive histone modifications to targeted loci. Despite their prevalence in angiosperms, the existence and extent of heterochromatic siRNAs in other land plant lineages has been unclear. Analysis of extensive small RNA-sequencing (small RNA-seq) data from the moss *Physcomitrella patens* identified over 14,000 loci that produce mostly 23-24 nt siRNAs. These loci tend to overlap intergenic regions, transposons, and regions of dense 5-mC, while avoiding genes (here we consider genes excluding promoters). Accumulation of siRNAs from these loci depends upon *Physcomitrella* homologs of *DICER-LIKE 3* (*DCL3*), *RNA-DEPENDENT RNA POLYMERASE 2* (*RDR2*), and the largest sub-unit of *DNA-DEPENDENT RNA POLYMERASE IV* (Pol IV), with the largest sub-unit of a Pol V homolog contributing to expression at a smaller subset of the loci. A *MINIMAL DICER-LIKE* (*mDCL*) gene, which

[2] The work presented in Chapter 2 has been submitted for review at *Genome Biology*.
Authors: Ceyda Coruh*, Sung Hyun Cho*, Saima Shahid, Qikun Liu, Andrzej Wierzbicki, Michael J. Axtell. (* indicates co-first authors)
Authors' Contributions: Ceyda Coruh and Michael Axtell analyzed small RNA-seq data. Sung Hyun (Joseph) Cho generated the *Physcomitrella* mutant lines and characterized their phenotypes. Saima Shahid performed the miRNA analysis. Qikun Liu and Sung Hyun (Joseph) Cho prepared the small RNA- seq libraries. Andrzej Wierzbicki generously provided the *Physcomitrella* Pol IV and Pol V largest sub-unit sequences.

lacks the N-terminal helicase domain typical of DCL proteins, is specifically required for 23 nt siRNA accumulation from these loci. We conclude that heterochromatic siRNAs, and their biogenesis pathways, are largely identical between angiosperms and *Physcomitrella patens*, with the notable exception of the *Physcomitrella*-specific use of *mDCL* to produce 23 nt siRNAs.

2.2 Introduction

Small non-coding RNAs regulate gene expression to control growth, development, differentiation, genome integrity, and stress response mechanisms in eukaryotic organisms. There are two main categories of endogenous small RNAs in plants: microRNAs (miRNAs) and short-interfering RNAs (siRNAs). Although the silencing pathways utilizing small RNAs have much in common, there are some fundamental distinctions between the two classes of small RNAs, particularly in regard to their biogenesis, evolutionary conservation, targets, and modes of action (Axtell 2013a). Most importantly, miRNAs and siRNAs differ in their precursors: while siRNA precursors are the products of intermolecular hybridization of two complementary RNA strands forming double-stranded RNA (dsRNA) duplexes, miRNAs are derived from single RNA molecules that fold back to form self-complementary “hairpin” RNAs. Endogenous siRNAs are the dominant small RNA type in many plant species, while miRNAs have received more attention, particularly in regard to annotations of specific loci.

Heterochromatin, which contains repetitive sequences and transposable elements, is silenced by conserved epigenetic modifications of histones and DNA. Epigenetic silencing is believed to prevent abnormal chromosomal rearrangements, and activation of transposons which can cause mutations if they are integrated into genes (Lippman and Martienssen 2004). In flowering plants, siRNAs are known to induce DNA methylation at targeted genomic regions (Matzke and Birchler 2005). Repressive histone modifications and siRNA- directed DNA methylation form a positive feedback loop to reinforce transcriptional silencing. This pathway is particularly well understood in *Arabidopsis thaliana*, where the presence of H3K9 methylation leads the SAWADEE HOMEODOMAIN HOMOLOG 1 / DNA-BINDING TRANSCRIPTION FACTOR 1 (SHH/DTF1) protein to recruit an alternative DNA-dependent RNA polymerase (Pol IV)

to chromatin (Law et al. 2013; Zhang et al. 2013a). Pol IV transcribes precursors of heterochromatic siRNAs, which are promptly converted into dsRNAs by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2), and then processed by DICER-LIKE 3 (DCL3) to generate 24 nt siRNAs (Xie et al. 2004; Daxinger et al. 2009). The 24 nt siRNAs are then bound to ARGONAUTE 4 (AGO4) or another AGO4-clade AGO protein and seek nascent transcripts produced by another alternative DNA-dependent RNA polymerase, Pol V (Wierzbicki et al. 2008, 2009). Binding of an AGO4-siRNA complex to Pol V nascent transcripts is thought to recruit DNA- and histone-methyltransferases to the vicinity of the target chromatin.

24 nt heterochromatic siRNAs dominate endogenous small RNA populations in most tissues of most angiosperms, but their presence in other land plants has been less clear. Early small RNA-seq efforts from the mosses *Physcomitrella patens* (Arazi et al. 2005) and *Polytrichum juniperinum* (Axtell and Bartel 2005), several gymnosperm species (Dolgosheina et al. 2008), and the lycophyte *Selaginella moellendorffii* (Banks et al. 2011) all found a conspicuous absence of endogenous 24 nt RNAs. It has also been suggested that conifers lack homologs of *DCL3* (Dolgosheina et al. 2008). However, there are several hints suggesting that the heterochromatic siRNA pathway may indeed be present outside of angiosperms. Significant amounts of 24 nt RNAs have been observed in gymnosperms in a highly tissue-specific manner (Nystedt et al. 2013; Zhang et al. 2013b). The *Selaginella* genome contains *DCL3*, *RDR2*, *AGO4*, and Pol IV / V largest sub-unit homologs (Banks et al. 2011), suggesting that the absence of 24 nt RNAs in initial small RNA-seq libraries may be due to tissue-restricted expression. Finally, our previous analysis demonstrated that the *Physcomitrella DCL3* homolog is required for the accumulation of 22, 23, and 24 nt RNAs from a handful of siRNA 'hotspots' (Cho et al. 2008). Nonetheless, conclusive description of a *bona fide* heterochromatic siRNA system in plants basal to the angiosperms has yet to be described. In this study, we used extensive small RNA-seq analysis in wild-type and several *Physcomitrella* mutants (two *RDR*s, Pol IV, Pol V, two canonical *DCL*s, and a minimal *DCL* gene) to rigorously test the hypothesis that heterochromatic siRNAs are expressed in this basal land plant.

2.3 Results

2.3.1 Most *DCL*-derived small RNA loci produce mixtures of 23-24 nt small RNAs in *Physcomitrella*.

Several previous studies have annotated *Physcomitrella* miRNAs and endogenous siRNAs using small RNA-seq (Arazi et al. 2005; Axtell et al. 2006; Fattash et al. 2007; Cho et al. 2008; Arif et al. 2012). However, these previous small RNA-seq efforts have all had quite low sequencing depth by current standards (less than 5 x 105 mapped reads per library in all cases). Therefore, to create a more comprehensive annotation of *Physcomitrella* small RNA genes, we obtained ten small RNA-seq libraries (from six biological replicates; four samples were run twice) from ten-day old wild-type protonemata totaling more than 10^8 mapped reads (Table 2.1). The majority of the small RNAs aligned to the nuclear genome, while a substantial minority aligned to the plastid genome (Fig. 2.1A). We identified 39,975 small RNA-producing loci using ShortStack (Axtell 2013b) (Table 2.1). For each locus, the fraction of included reads between 20-24 nts in length was calculated, and a cutoff of 0.8 was used to discriminate non-*DCL*-derived loci from *DCL*-derived loci (Fig. 2.1B). Two loci that were clearly *MIRNA*s, Ppv2-0_Cluster_10211 and Ppv2-0_Cluster_27602 (Table 2.4) with at least 60% of their small RNAs being 21 nt in length, barely missed this cutoff and were rescued and retained as *DCL* loci. Roughly half of the non-*DCL*-derived loci originated from mRNAs (Fig. 2.1C). In terms of small RNA abundance, however, the bulk of the non-*DCL* reads originated from the plastid genome or nuclear rRNA genes (Fig. 2.1D). In contrast, nearly all of the loci and the abundance from *DCL* loci came from un-annotated regions of the nuclear genome. We conclude that most of the non-*DCL* loci likely represent fragments of abundant RNAs arising from the plastid genome or nuclear rRNA genes and unrelated to the *DCL*/*AGO* system of regulatory small RNAs. Therefore, we focused on the 16,024 *DCL* loci for the remainder of the study.

Table 2.1: *Physcomitrella patens* small RNA-seq libraries

Library	Genotype	Strain	Number of Mapped Reads (x 10^6)	GEO GSE	GEO GSM	Libraries Re-sequenced
1	Wild-type	Gransden 2004	15.0	GSE44900	GSM1093595	-
2	Wild-type	Gransden 2004	17.9	GSE44900	GSM1194292	Re-run of Library 1
3	Wild-type	Gransden 2004	15.8	GSE44900	GSM1093596	-
4	Wild-type	Gransden 2004	18.6	GSE44900	GSM1194293	Re-run of Library 3
5	Wild-type	Gransden 2004	11.4	GSE44900	GSM1194296	-
6	Wild-type	Gransden 2004	13.7	GSE44900	GSM1194297	-
7	Wild-type	Gransden 2009	9.8	GSE44900	GSM1093597	-
8	Wild-type	Gransden 2009	11.6	GSE44900	GSM1194294	Re-run of Library 7
9	Wild-type	Gransden 2009	12.5	GSE44900	GSM1093598	-
10	Wild-type	Gransden 2009	14.6	GSE44900	GSM1194295	Re-run of Library 9
11	*rdr2-1*	Gransden 2004	10.9	GSE51419	GSM1245155	-
12	*rdr2-1*	Gransden 2004	13.0	GSE51419	GSM1245157	Re-run of Library 11
13	*rdr2-1*	Gransden 2004	14.8	GSE51419	GSM1245156	-
14	*rdr2-1*	Gransden 2004	17.1	GSE51419	GSM1245158	Re-run of Library 13
15	*rdr2-2*	Gransden 2004	17.6	GSE51419	GSM1245159	-
16	*rdr2-2*	Gransden 2004	19.3	GSE51419	GSM1245160	-
17	*rdr6-19*	Gransden 2004	13.6	GSE51419	GSM1245161	-
18	*rdr6-19*	Gransden 2004	11.7	GSE51419	GSM1245162	-
19	*rdr6-35*	Gransden 2004	14.2	GSE51419	GSM1245163	-
20	*rdr6-35*	Gransden 2004	16.4	GSE51419	GSM1245164	-
21	*dcl3-5*	Gransden 2004	12.2	GSE51419	GSM1245131	-
22	*dcl3-5*	Gransden 2004	13.6	GSE51419	GSM1245132	-
23	*dcl3-10*	Gransden 2004	24.7	GSE51419	GSM1245133	-
24	*dcl3-10*	Gransden 2004	13.3	GSE51419	GSM1245134	-
25	*dcl4-1*	Gransden 2004	9.1	GSE51419	GSM1245135	-
26	*dcl4-1*	Gransden 2004	16.9	GSE51419	GSM1245136	-
27	*mdcl-77*	Gransden 2009	13.0	GSE51419	GSM1245137	-
28	*mdcl-77*	Gransden 2009	15.4	GSE51419	GSM1245141	Re-run of Library 27
29	*mdcl-77*	Gransden 2009	12.4	GSE51419	GSM1245138	-
30	*mdcl-77*	Gransden 2009	14.5	GSE51419	GSM1245142	Re-run of Library 29
31	*mdcl-107*	Gransden 2009	20.1	GSE51419	GSM1245139	-
32	*mdcl-107*	Gransden 2009	23.5	GSE51419	GSM1245143	Re-run of Library 31
33	*mdcl-107*	Gransden 2009	12.7	GSE51419	GSM1245140	-
34	*mdcl-107*	Gransden 2009	15.0	GSE51419	GSM1245144	Re-run of Library 33
35	*nrpe1a_128*	Gransden 2004	15.1	GSE51419	GSM1245145	-
36	*nrpe1a_128*	Gransden 2004	17.6	GSE51419	GSM1245149	Re-run of Library 35
37	*nrpe1a_128*	Gransden 2004	13.5	GSE51419	GSM1245146	-
38	*nrpe1a_128*	Gransden 2004	15.9	GSE51419	GSM1245150	Re-run of Library 37
39	*nrpd1_12*	Gransden 2004	14.7	GSE51419	GSM1245153	-
40	*nrpd1_12*	Gransden 2004	16.9	GSE51419	GSM1245154	-

All annotated small RNA loci were classified into three categories: *MIRNA* loci, non-*MIRNA* hairpin-RNA loci (HP), and siRNA loci. Most *DCL*-derived small RNA loci were classified as siRNA loci (Fig. 2.1E). We refer to the predominant size of RNA produced by a locus as the "DicerCall". The majority were siRNA loci with DicerCalls of 23 or 24 (Fig. 2.1E). However, when analyzed by total abundance, 21 nt RNAs dominated *MIRNA*, HP, and siRNA loci (Fig. 2.1F). Thus, we conclude that a relatively small number of *MIRNA*, HP, and siRNA loci produce large amounts of 21 nt RNAs, while a much larger set of mostly siRNA loci produce more modest amounts of 23 and 24 nt RNAs.

DicerCall is a somewhat crude indicator, and could mask cases where loci actually tend to produce mixtures of different sized RNAs. For *MIRNA*s, the DicerCall generally reflected a strong majority of RNAs of the corresponding DicerCall size (Fig. 2.1G). However, HP and siRNA loci with DicerCalls of 23 or 24 in fact often produce mixtures of 23 and 24 nt RNAs (Fig. 2.1H, I). For further analyses, we classified five different groupings of *DCL* loci (Fig. 2.1J), listed here in order from most-to-least numerous: 23-24 nt siRNA loci, 20-22 nt siRNA loci, 23-24 nt HP loci, 20-22 nt HP loci, and *MIRNA* loci (Fig. 2.1J).

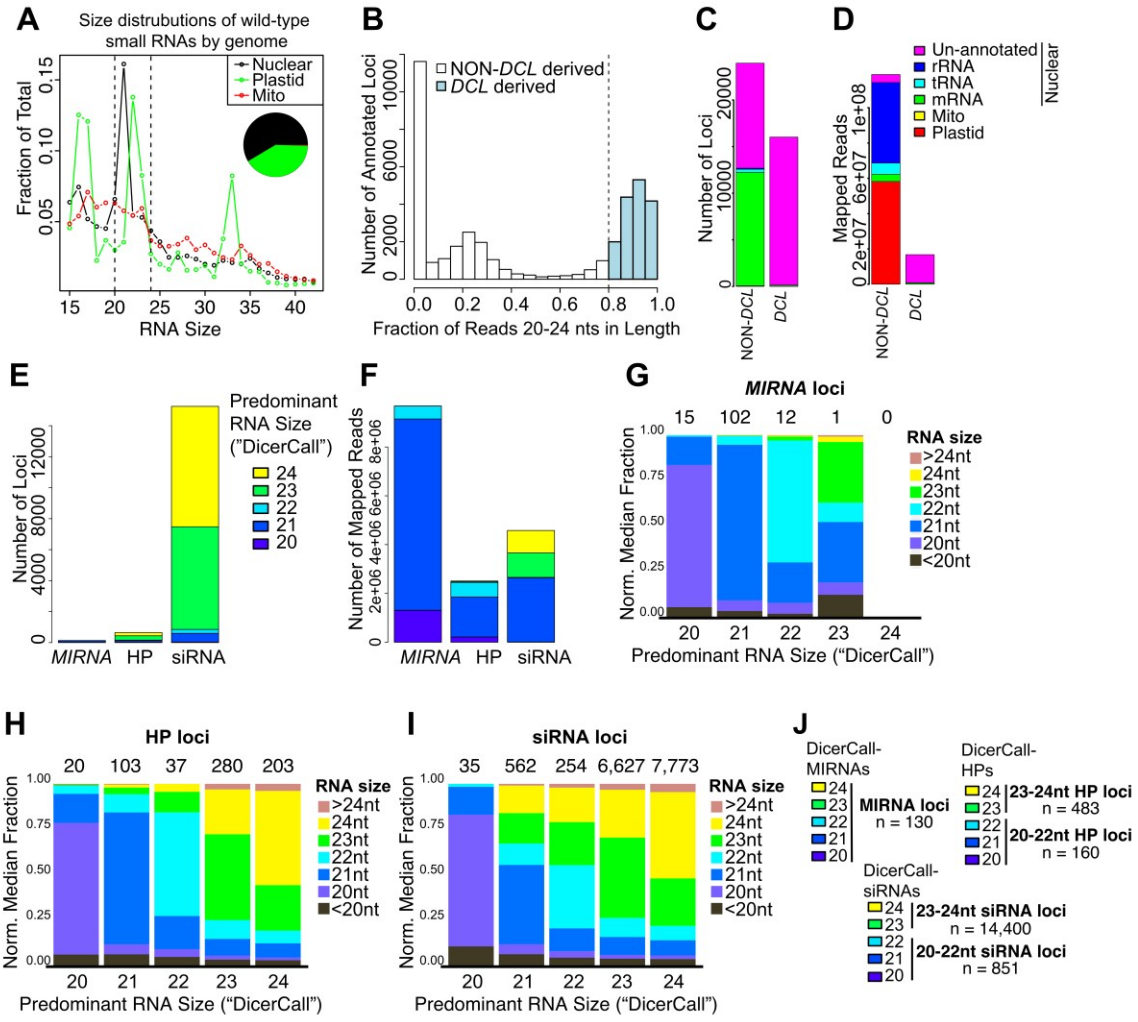

Figure 2.1: **Properties of *Physcomitrella patens* small RNA genes.**

(*A*) Size distributions of wild-type small RNAs aligned to the indicated genomes. Vertical dotted lines indicate the *DCL* size range of 20-24 nts. **(*B*)** Histogram showing the fraction

of reads between 20-24 nts within annotated small RNA loci. Dotted line at 0.8 indicates the cutoff for the 'DicerCall'; loci with a fraction of < 0.8 were given a DicerCall of "N", while those >= 0.8 were given a DicerCall equal to the most abundant small RNA size in the locus. **(*C*)** Tallies of small RNA loci by different genomic regions. **(*D*)** Abundance of small RNA reads by different genomic regions. **(*E*)** Classification of *DCL*-derived small RNA loci, counted either by number of loci or (*F*) by total small RNA abundance. **(*G*)** Small RNA size distributions within each class of DicerCall at *MIRNA* loci. **(*H*)** Same as in *G* except for HP loci. **(*I*)** Same as in *G* except for siRNA loci. **(*J*)** Definitions for five categories of *Physcomitrella DCL*-derived small RNA-producing loci.

2.3.2 *Physcomitrella* 23-24 nt siRNA loci are associated with repeats, transposons, and regions with dense 5-methyl cytosine.

We performed co-occupancy analysis of the five groupings of *DCL*-derived small RNA loci against various genomic features. Two broad patterns are apparent. At one extreme, *MIRNA*s, and to a lesser extent 20-22 nt hpRNA loci, avoid regions with dense 5-mC, transposons, and repeats, but can occasionally overlap with genes (Fig. 2.2A, C). At the other extreme, 23-24 nt siRNA and 23-24 nt HP loci are enriched for overlaps with 5-mC-dense regions of all contexts, transposons, and repeats, and are severely depleted in overlaps with genes (Fig. 2.2A, C). 20-22 nt siRNA loci are intermediate between these two patterns. Consistent with a previous analysis (Zemach et al. 2010), *Physcomitrella* 5-mC regions are almost entirely confined to intergenic regions, enriched for association with repeats and transposons, and avoid genes (Fig. 2.2B, D). We conclude that *Physcomitrella* 23-24 nt siRNA and HP loci are heterochromatic siRNAs, with grossly similar genomic arrangements as the heterochromatic siRNAs of higher plants.

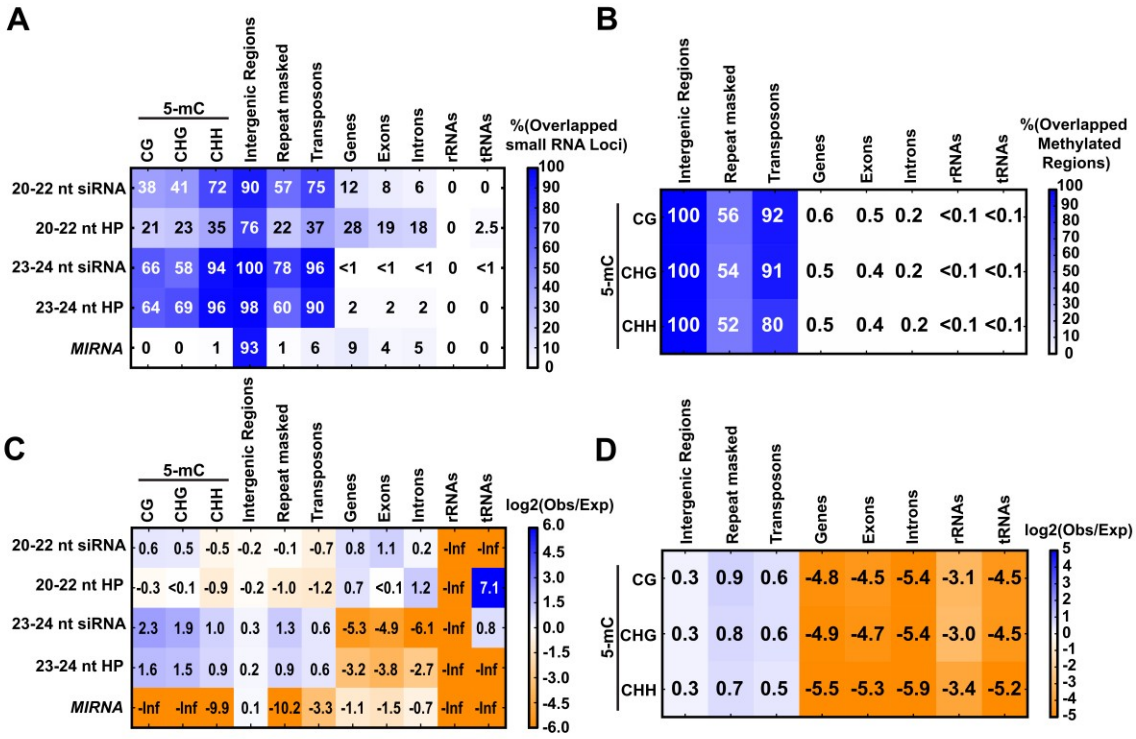

Figure 2.2: Genomic features of *Physcomitrella* small RNA-producing loci.

(*A*) Percentages of small RNA-producing loci that overlap various genomic features (% Overlap = (# small RNA loci overlapping with one or more of the indicated genomic feature / # total small RNA loci) * 100). **(*B*)** Same as in *A* except for regions of dense DNA methylation relative to different genomic features. **(*C*)** Heatmap showing log2 (observed overlapped bases / expected overlapped bases) for each of the pair-wise comparisons shown. Cell values are shown in bold text. **(*D*)** As in *C* except for regions of dense DNA methylation.

2.3.3 Improved *Physcomitrella MIRNA* annotations

Our entirely *de novo* annotation of *MIRNA*s found 130 loci, of which 107 were already annotated in miRBase release 20 (Kozomara and Griffiths-Jones 2014) (Fig. 2.3A, Table 2.3 (Table2.3_Pp_MIRNA_loci_v1.6.txt) and Table 2.4). We compared our mature miRNAs from novel loci with all mature miRNAs present in miRBase release 20. Two of the novel loci were found to be paralogs of known families (miR1027 and

miR1065), but the remaining 21 new loci do not belong to any previously known plant miRNA families (Fig. 2.3A).

Table 2.4: Summary of ShortStack-annotated miRNAs

Locus	Name	miRNA	miRNA mappings	miRNA-star	miRNA-star mappings	Total mappings
scaffold_1:3540225-3540393	Ppv2-0_MIR537b	UUGAGGUGUUUCUACAGGCU, UUGAGGUGUUUCUACAGGCUA	644, 7746	GACUGUAGAAACACCUGAAGU, ACUGUAGAAACACCUGAAGU	2069, 51	21867
scaffold_1:3555310-3555478	Ppv2-0_MIR537c	UUGAGGUGUUUCUACAGGCUA, UUGAGGUGUUUCUACAGGCU	792, 7630	GACUGUAGAAACACCUGAAGU, ACUGUAGAAACACCUGAAGU	2108, 66	21918
scaffold_2:3223712-3223838	Ppv2-0_Cluster_685	UCAUGUGCCUGUUGUAGUUCC	37	AACUUCAAAGGCUCAUGAGA	10	361
scaffold_3:2089924-2090047	Ppv2-0_MIR156c	UGACAGAAGAGAGUGAGCAC	43087	GCUCACUCUCUUCAUGUCACG	1130	52983
scaffold_3:3110232-3110386	Ppv2-0_MIR2084	AAUCCAUCGAAGCAGGCGCUU	432	GCGCCUGCAUUGUUGGAUUGU	93	1340
scaffold_5:35728-35801	Ppv2-0_MIR1033d	UGACGGGUCGUGAUGGGCACU, UGACGGGUCGUGAUGGGCACUC	354, 418	AUGUCCAUGACGAUCUGUCAAC, UGUCCAUGACGAUCUGUCAAC	267, 1	1316
scaffold_5:3307166-3307340	Ppv2-0_MIR902j	AGAAGGAUCUGCAACAUAGA, AGAAGGAUCUGCAACAUAGAC	1697, 2234	AUAUGUUGCAGAUUCUUCAUU, UAUGUUGCAGAUUCUUCAUU	87, 11	4569
scaffold_8:1132567-1132675	Ppv2-0_MIR1049	UCUCUCUUAGCCAAACAGUCU	3703	ACCUGUUGGCGAAAGAGAGUCG	234	4658
scaffold_8:1592328-1592485	Ppv2-0_Cluster_2776	UUUGCAUUGCACACUAUAAUG	31	UUGUGGUGUGCAAUGUAAGUU	7	41
scaffold_11:1661361-1661470	Ppv2-0_MIR902e	ACGAAGGUCUGCAUCAUAGU	65652	UAUGAUGCAGAUUCUUCAUCU	135	71392
scaffold_12:1858517-1858631	Ppv2-0_MIR156a	UGACAGAAGAGAGUGAGCAC	43605	GCUCACUCUCUUCCUGUCGCA	34	51260
scaffold_13:1365519-1365663	Ppv2-0_MIR1033a	UGACGGGUCGUGAUGGGCACUC	407	AUGUCCAUCACAACCUGCCAAC	268	1507
scaffold_14:198801-198944	Ppv2-0_MIR477e	AGAAGCCUUUGUGGGAGAGGG	6671	CUCUCCCUCAAAGGCUUCCAA	991	15996
scaffold_14:2066036-2066119	Ppv2-0_MIR166f	UCGGACCAGGCUUCAUUCCCC	32254	GGAAUGCUGCAUGGUCUGAUG	255	38174
scaffold_15:2756708-2756840	Ppv2-0_Cluster_4718	CUCAUGAGUGAUGGUAUGUGC	21	ACAUAUCAUCACUCAAGAGAG	4	43
scaffold_16:536012-536103	Ppv2-0_Cluster_4787	GUUGGGAUAAAAAGAACAAUA	27	UUGUUCUUAGUAUCCCAACUG	9	130
scaffold_17:2119355-2119544	Ppv2-0_MIR902g	ACGAAGGUCUGCAUCAUAGU	65928	UAUGAUGCAGAUUCUUCAUCU	117	71660
scaffold_19:8682-8967	Ppv2-0_Cluster_5476	UAUUAUGUCUAUACACUCACC, AUUAUGUCUAUACACUCACCC	207, 117	GUGUGUGUAUAGAUAUAAUGGG, UGUGUGUAUAGAUAUAAUGGG	1, 14	386
scaffold_19:89684-89892	Ppv2-0_MIR319a	CUUGGACUGAAGGGAGCUCC	171978	AGCUCCGUUUCGGUCCAAUAG	12718	254422
scaffold_19:293821-294034	Ppv2-0_MIR171b	UGAGCCGCGCCAAUAUCACAU	8794	GUGUUAUUGGGCCCGCUCAAU	4	10429
scaffold_19:562783-562931	Ppv2-0_MIR902h	UUAUGAUGUAGAUUCUUCAUC	4189	AGAAGGGUCUACAUCAUAAAC	1335	6201
scaffold_19:828533-828660	Ppv2-0_MIR533d	CUCACAGUCUGCACAGCUCUC	30402	GAGCUGUUCAGACUCUGAGAG	2345	39164
scaffold_19:1175559-1175703	Ppv2-0_MIR160c	CGCCUGGCUCCCCUGCAUGCCA	81	GCGUGUGGGGGUCAGACAGG	52	237
scaffold_19:1176196-1176349	Ppv2-0_MIR160b	CGCCUGGCUCCCUGUAUGCCA	297	GCGUUCAGGGAGUCAAGCAGA	97	780
scaffold_19:2735033-2735152	Ppv2-0_MIR538b	CGGACAUAGCCUUCAUGCAUG	55179	UUGCAUGGAGUCUAUGUCUGGA	7382	87118
scaffold_21:64659-64801	Ppv2-0_MIR166g	UCGGACCAGGCUUCAUUCCCC	32156	GGAAUGCCCCCUGGCCCGAAG	16	37129
scaffold_25:2584327-2584455	Ppv2-0_Cluster_7156	GAAAUGUAGAAGCAGGUUUUG	23	UAACCUGGUUCUACAUUCAUG	9	63
scaffold_26:1562334-1562447	Ppv2-0_MIR166b	UCGGACCAGGCUUCAUUCCCC	32256	GGAAUGCCGCAUGGACCGAAG	585	37918
scaffold_27:2077215-2077338	Ppv2-0_MIR477h	GUUGGAUGCCUUAGUGGGAGA	935	UCCCUCAAAGGCUUCCAACUA	506	2658
scaffold_29:989057-989147	Ppv2-0_MIR1033b	UGCUCAACACGAUCUGUCAAC, UGACGGGUCGUGAUGGGCACUC	472, 430	AUGCUCAACACGAUCUGUCAAC, UGACGGGUCGUGAUGGGCACU	83, 355	2028
scaffold_30:1354118-1354281	Ppv2-0_MIR166a	UCGGACCAGGCUUCAUUCCCC	32370	GGAAUGCCCCCUGGCCCGAAG	22	37262
scaffold_31:1002602-1002725	Ppv2-0_MIR533b	CUCACAGUCUGUACAGCUCUC	152	GAGCUGUCCAGGCUGUGAGGG	78	394
scaffold_34:1432904-1433088	Ppv2-0_MIR1028b	CGGCAUUGUGGACCUAAGACC	158863	UCUUAGAUCUACAAUGCCACC	76478	243803
scaffold_36:478448-478590	Ppv2-0_MIR166e	UCGGACCAGGCUUCAUUCCCC	32400	GGAAUGCCGCAUGGACCGAAG	609	38533
scaffold_41:581030-581127	Ppv2-0_Cluster_10211	UGAGUAGAUUCUUUGUAUAGA	19	UAUACAUAGAAUCUACUCUGU	4	35
scaffold_43:1696878-1697030	Ppv2-0_MIR1024a	UCUGGUUGGAUUGUAGGCCUC	71915	CGCCCUGCAAUUCAAGCAGACU	292	159086
scaffold_44:180008-180118	Ppv2-0_MIR535d	UGACAACGAGAGAGAGCACGC	132928	GUGCCCUCUCCCGUCGUCACC	60371	302946
scaffold_45:1376158-1376296	Ppv2-0_Cluster_11079	UGUGGUUGGAAUUGUUAGAGC	28	UCUAAUAAUUCCAACUACAGG	2	59
scaffold_45:1704524-1704683	Ppv2-0_MIR902d	AUGAAGGUCUGCAUCGUAGC	933	UAUGAUGCAGAUUCUUCAUCU	134	1359
scaffold_46:1288883-1288966	Ppv2-0_Cluster_11259	CUUUGUUGGAACUGGGGAGU	41	UCCCCGGUUCCAACAAGCUC	39	188
scaffold_49:1228150-1228303	Ppv2-0_MIR1062	UCCCUCACGUGGUGUUUGCAGC	68	UGCAAACACCCGUAUGGGACG	9	100
scaffold_51:599238-599388	Ppv2-0_MIR898a	UUGCUGUGCACUACUUAGUACA	88246	UGCUAGGCAGUGCACAGCGAUA	7	101776
scaffold_51:950572-950744	Ppv2-0_MIR1023c	AGGGAAUCGGAAAGAGUGUAC	64	CCACUCUCUCCGUUUCCCUUC	1	72
scaffold_52:1225394-1225592	Ppv2-0_MIR537a	GACUGUAGAAACACCUGAAGC, UUGAGGUGUUUCUACAGGCU	10595, 7758	ACUGUAGAAACACCUGAAGC, UUGAGGUGUUUCUACAGGCUA	39, 7916	31012
scaffold_55:1836358-1836576	Ppv2-0_Cluster_12833	CUACGGGGUCGUUUUGCUCGU	60	CAGCAGAUCGACCUCGUAGCA	26	226
scaffold_56:756196-756284	Ppv2-0_MIR535b	UGACAACGAGAGAGAGCACGC	132376	GUGCCCUCUCCCGCUGUCGCC	6562	148244
scaffold_58:126289-126469	Ppv2-0_MIR1031a	AAGCUUCACAGAACAUUGACU	24572	UCAAUGCUCUCUGGAGCUUCU	829	26934
scaffold_59:1334038-1334181	Ppv2-0_MIR1023a	AGAGAAUUGGAGAGAGUGCAU	30122	ACACUCUCUCCAUUUCUCUAC	3589	40290
scaffold_64:1475293-1475396	Ppv2-0_MIR166l	UCCGGACCAGGCUUCAUUCCC, CCGGACCAGGCUUCAUUCCCC	16560, 13833	GGAAUGACGUGUGGCCCGUAG, GAAUGACGUGUGGCCCGUAGC	44, 454	41032
scaffold_66:898021-898134	Ppv2-0_Cluster_14458	CACUCGAAGUCACUGUACCUG	927	GGUACAGAGGCUUCGAGUGGG	2	2810
scaffold_67:1713097-1713245	Ppv2-0_Cluster_14654	GGCAGAAACGGGCAGAGCUCC	238	AGCUCUGCCCGUGUCUGCCCG	2	375
scaffold_71:1665539-1665796	Ppv2-0_MIR171a	UGAGCCGCGCCAAUAUCACAU	8891	GUGAUAUUGGUGGCGCUCAAU	360	11479
scaffold_72:21038-21325	Ppv2-0_MIR1069	CUUAUCAUUGGAUUGAGCACC	102	UGCUCACUGUAAUGAUAAAGC	3	341
scaffold_74:647191-647360	Ppv2-0_MIR902k	ACGAAGGAUCUGCAAUAUAAA	1402	UAUGUUGCAGAUUCUUCAUUU	20	3073
scaffold_77:341294-341451	Ppv2-0_MIR1216	UUGAUGGUGAUGCGCUUGUAUC, UGAUGGUGAUGCGCUUGUAUC	3567, 4662	UACAAGCCCAUCACCCUCAACG, UACAAGCCCAUCACCCUCAAC	505, 244	9643
scaffold_79:1036196-1036295	Ppv2-0_MIR1222d	UGCUGGUGAACUCCUUCAUGC, UGCUGGUGAACUCCUUCAUG	5879, 5671	UUUAAGGGGUUCACUGGUAUA, UUAAGGGGUUCACUGGUAUA	695, 20	18089
scaffold_80:53814-54102	Ppv2-0_Cluster_16488	UCUGUCUGCCGUCUCUCAGCU	43	GUGAGGGAGAGUGGACAGAGC	3	158
scaffold_80:1382873-1383005	Ppv2-0_MIR1052	CCGCAAUCAGUAAAAGGGAGU	3844	UUCCUUUUAUUGAUUGUGGUA	2610	13379
scaffold_80:1498991-1499142	Ppv2-0_MIR390b	AAGCUCAGGAGGGAUAGCGCC	9773	CGCUAUCCAUUCUGAGCUUUG	506	12069
scaffold_83:1094111-1094181	Ppv2-0_MIR1222c	UCUGAAGGAGUUCAUUGGUAC	66	ACCAGUGCGUUCCUCCAAACC	23	135
scaffold_83:1142088-1142269	Ppv2-0_Cluster_17003	UAUAGAUUGACUGGGGUAACU	828	UUACCCCAGUCAAACUAUUCC	24	1455
scaffold_86:401355-401510	Ppv2-0_Cluster_17386	UGCGAGAAGUCAGAGACAGUC	293	CAGUCUCUGACUUCUCGCAUG	47	489
scaffold_89:1180768-1180881	Ppv2-0_MIR533a	CUCACAGUCUGCACAGCUCUC	30664	GAGCUGGCCAGGCUGUGAGGG	2313	47045
scaffold_94:882754-882972	Ppv2-0_MIR1036	UGUGGAGUCCGUAAUUAGCUG, GUGGAGUCCGUAAUUAGCUGC	1597, 1625	AGCUAAUUAAGGAUUCUACAC, GCUAAUUAAGGAUUCUACACA	129, 11	3743
scaffold_94:1388997-1389108	Ppv2-0_MIR1027a	UUUCUAUCUUCUCUUCCAAUC,	1851,	AUUGGAAGAGCAGAUCGAAAA,	52,	4978

		UUCUAUCUUCUCUUCCAAUCU	1294	UUGGAAGAGCAGAUCGAAAAA	313	
scaffold_94:1430752-1430863	Ppv2-0_MIR1027b	UUUCUAUCUUCUCUUCCAAUC, UUCUAUCUUCUCUUCCAAUCU	1925, 1398	AUUGGAAGAGCAGAUCGAAAA, UUGGAAGAGCAGAUCGAAAAA	54, 322	5194
scaffold_96:1006502-1006657	Ppv2-0_MIR1050	UGACCACCUUGAUUCCGGCCU	24717	GCUGAUAUCAGGGUGGUCACA	376	25960
scaffold_99:433973-434143	Ppv2-0_MIR319e	CUUGGACUGAAGGGAGCUCCC	1221695	GAGCUCCUUCCGGUUCAAUAG	38	1438733
scaffold_100:50329-50546	Ppv2-0_MIR533c	GAGCUGUUCAGACUAUGAGAA	206590	CUCACAGUCUGCACAGCUCUC	30261	275813
scaffold_100:542192-542372	Ppv2-0_MIR1217	AAUUUGAAGCAUGAUGUCAAG	2862	UGGUAUCAUGUUGCAAAUGGC	1344	5128
scaffold_101:1152005-1152185	Ppv2-0_MIR535c	UGACAACGAGAGAGAGCACGC	132698	GUGCCCUUUCCCGUUGUCGCC	7683	149132
scaffold_106:956218-956410	Ppv2-0_MIR1221	UGGCCCAUGCACAGUAUCUACG	4721	UGGAUGGUGUGCAGGGUCAAA	572	6765
scaffold_107:19112-19238	Ppv2-0_MIR1212	CGUGGGACAGCAUAGAAUGCG, CGUGGGACAGCAUAGAAUGC	36955, 23626	CAUCCUCUGCUGUGCCCACAUG, AUCCUCUGCUGUGCCCACAUG	1813, 13	68953
scaffold_112:1004893-1005105	Ppv2-0_MIR1054	GGAGUAGAAGGGAGGGCUUAACU, UAAACCCUCUCUCUAUUCCUG	955, 1070	GGAGUAGAAGGGAGGGCUUAAC, GUAAACCCUCUCUCUAUUCCUG	170, 1	3395
scaffold_117:1077909-1077998	Ppv2-0_Cluster_21057	UCAAAAUUUCAAGAUAGUAUG	16	UGCUACCUUGAAAUUUUGUUG	2	27
scaffold_119:1056015-1056219	Ppv2-0_MIR390a	AAGCUCAGGAGGGAUAGCGCC	9931	CGUUAUCCAUUCUGAGCUUUG	1903	18731
scaffold_126:342504-342653	Ppv2-0_MIR898b	UUGCUGUGCACUACUUAGUAC	10668	GCUGAGUAGUGCACAGCAAUA	1	34583
scaffold_128:563507-563616	Ppv2-0_MIR1045	AAGUGCUGGCUUUUUGACGUU, CUGGCUUUUUGACGUUUUGCU	461, 303	CAAGACGUCAAACAACCAGCA, CGUCAAACAACCAGCACUUCG	15, 18	953
scaffold_139:409391-409574	Ppv2-0_MIR1079	CGUGUUGAGGCUUGUUAAAG	75	UAAGCAAGCUUCAACAAGCU	5	152
scaffold_141:592759-592855	Ppv2-0_MIR1214	UACUAUGAGAAUCUCGCGGCC	245000	CUGCGAUUUUCUCAUAGCAGA	1	287350
scaffold_150:632148-632294	Ppv2-0_Cluster_24121	UGCCAAAAUCAAUCAACUCGAC	50	CAAGUUGAUUGAUUUGGCACA	8	141
scaffold_161:204425-204629	Ppv2-0_MIR1220a	UUGCACUCCUCUAUCUCCCUC	1162	GGAAGAUAGAGGAGUUCAAGA	43	2089
scaffold_165:985495-985631	Ppv2-0_Cluster_25395	UCCAUACCACUGCUAGAGACU	120	UCUCUAGCAGUGAUAUGGUGG	1	293
scaffold_167:174995-175231	Ppv2-0_MIR160h	CGCCUGGCUCCUUGUAUGCCA	133	GCACUCAGGGAGUCAAGCAGG	51	340
scaffold_167:958055-958142	Ppv2-0_MIR156b	UGACAGAAGAGAGUGAGCAC	44078	GCUCACUCUCUUCAUGUCGCG	58	48126
scaffold_170:61531-61657	Ppv2-0_Cluster_25701	UCAAAGAUCUCCGUCUGCCCU	91316	AGCAGGUGGCGAUCUUUGAAG	3933	151446
scaffold_173:168098-168342	Ppv2-0_MIR1215	UCAUUGCAAAACUGUAUACGA, GUAUACAGUUUCGCUGUGGAAC	13388, 10797	CUCAUUGCAAAACUGUAUACGA, GUAUACAGUUUCGCUGUGGAA	20, 5	36440
scaffold_179:820129-820230	Ppv2-0_MIR902c	ACGAAGGUCUGCAUCAUAGU	65444	UAUGAUGCAGAUUCUUCAUCU	132	71200
scaffold_184:577127-577210	Ppv2-0_MIR535a	UGACAACGAGAGAGAGCACGC	132868	GUGCCCUCUACCGUUGUCGCC	13932	155731
scaffold_197:428698-428808	Ppv2-0_Cluster_27602	UCGGCAGCGUUAUUCUUUGAC	18	CAAAGAGCAACGUUGUCGAAU	2	35
scaffold_202:211291-211472	Ppv2-0_MIR1034	AUUACUUUGGCAGCGCUGUGC, UUACUUUGGCAGCGCUGUGCU	395, 290	CACUACGCUGCCAAAGUCAUG, ACUACGCUGCCAAAGUCAUGA	51, 5	954
scaffold_208:523161-523318	Ppv2-0_MIR1032	AGGUGACUGCCUGGAAUUGGG	1971	CAAUUCCAGGCAAGCCCCUGU	32	7510
scaffold_217:384222-384354	Ppv2-0_MIR1035	CGUUUUGUGAGCUAAGAAGGU	436	CGUCUUAGCCCACAAAACGAA	36	720
scaffold_219:310476-310605	Ppv2-0_MIR1222a	UGCUAGUAAACUCCUUCAUGC	45529	UUGAAGGAGUUCAUUGGUAUA	1803	59790
scaffold_220:156609-156824	Ppv2-0_MIR1067	ACAUACUGAAGUUUGAUGCCA	40	GCAUCAAACGUCAGCAUGUGU	26	89
scaffold_234:21891-22056	Ppv2-0_MIR477f	GUCGGAAGCCUUCGAGGGAGA	2986	UCCCUCAAAGGCUUCCAACAA	60	4164
scaffold_234:55723-55899	Ppv2-0_MIR319b	CUUGGACUGAAGGGAGCUCC	172482	AGCUCUUUUCAGUCCAGUAG	134	251792
scaffold_234:189677-189961	Ppv2-0_MIR904a	CCCCUAAAUCUUGGCAAGACC	330396	UCUUGUCAAUGUUUAGGGGCA	44853	433010
scaffold_234:416698-416810	Ppv2-0_MIR1026b	UGAGAAAGACUUGAGAGGACA	28469	UUCUCUCUAGUCUUUCUUGGA	2336	32827
scaffold_243:593008-593194	Ppv2-0_MIR1028c	UGGCAUUGUAGGUUUAAGAGC	130752	UCUUAGAUCUACAAUGCCUCU	8436	146722
scaffold_245:121321-121538	Ppv2-0_MIR1076	UCCAAGCACUUAUCGCACCCUG, CCAAGCACUUAUCGCACCCUG	7584, 5716	AGGUGCGAUAAUUGCUCGAAG, AGGUGCGAUAAUUGCUCGAAGA	20, 3	17342
scaffold_248:120187-120296	Ppv2-0_MIR1048	UAGAACAUGAGUGUAGACGAC	371	UGUCUACACUCAUGUUCUAGA	6	507
scaffold_256:157999-158087	Ppv2-0_MIR1033e	UGACGGGUCGUGAUGGGCACUC	440	AUGUCCAGCACAACCUGUCAAC	6	1081
scaffold_257:30590-30731	Ppv2-0_MIR1023b	AGAGAAUUGAAGAGAGUGCAU	640	ACACUCUCUCCAUUUCUCUGC	353	1206
scaffold_264:204261-204402	Ppv2-0_MIR390c	GAGCUCAGGAGGGAUAGCGCC, CGCUGUCCAUUCUGAGCAUU	23617, 58222	AGCUCAGGAGGGAUAGCGCC, CGCUGUCCAUUCUGAGCAUUG	339, 10927	103830
scaffold_266:191728-191909	Ppv2-0_MIR319d	CUUGGACUGAAGGGAGCUCCC	1222276	GAGCUUUCUUCGGUCCAAUAG	16893	2222478
scaffold_266:312421-312594	Ppv2-0_MIR904b	CCCCUAAAUGUUGGCAAGACC	217975	UCUUGUCAAUGUUUAGGGGCA	44611	320400
scaffold_275:298232-298433	Ppv2-0_MIR477g	GUUGGAAGCCUUCGUGGGAGA	7454	UCCCUCAAAGGCUUCCAACAA	58	11943
scaffold_281:80504-80600	Ppv2-0_Cluster_32741	UCUGAUAUCAACUCCAGCUCU	124	AGUUCGAGUUGGUAUCAAAUA	11	174
scaffold_287:109482-109573	Ppv2-0_MIR1042	UGUCUAGUCUCUCCACGGCCCG	165	GGCUGUGCAGGAACUAGAUAGG	12	291
scaffold_287:109856-109954	Ppv2-0_MIR1043	UUGCGCGUGAAUUUGAAGGCU	2457	CCUUCAACUUCAUGCGUGCAA	5	4727
scaffold_291:313490-313651	Ppv2-0_MIR1033c	UGCCCAACACGAUCUGUCAAC	2549	UGACGGGUCGUGAUGGGCACU	352	4341
scaffold_309:256461-256548	Ppv2-0_Cluster_34103	CCACUCGUUCAUUGUGAAACU	1078	UUUCACAUCGAACGAGUGUCU	731	3357
scaffold_313:116809-116997	Ppv2-0_MIR1073	UGAAUGAUAUUAACGUCCACG	27	UGGGCGUUAAUACAUUUCAUA	3	48
scaffold_313:118181-118386	Ppv2-0_MIR408b	UGCACUGCCUCUUCCCUGGCU	209	CCAGCGUGAGGCAAUGCAUG	188	509
scaffold_325:243086-243186	Ppv2-0_MIR1039	GGGGAGACGGGCUCAAGGAUC	17007	UCUUUGGGUCUUUCUCUCCUG	717	21075
scaffold_336:345183-345323	Ppv2-0_MIR1024b	UCUGGUUGGAUUGUAGGCCUC	71162	CGCCUUGCAAUUCAAGCAGACU	3453	161726
scaffold_345:14285-14472	Ppv2-0_MIR2082	UGUGUGUUCCGCUUCUUCUUU	6038	AGAAGAAGCGGCUUACGCAAU	77	8882
scaffold_369:340080-340267	Ppv2-0_MIR538c	CUAGAGUAUCAUGGAGGCCCG	407049	AGUCUCCAUGCUUCUCUGACG	14419	516703
scaffold_381:111313-111462	Ppv2-0_MIR1029	UCUCUCUCAACCAACCAUACU, CUCUCUCAACCAACCAUACUG	1284, 1190	GUAUUUUUGGUUUGAGAGAAAG, UAUUUUUGGUUUGAGAGAAAGG	159, 121	4174
scaffold_391:252756-252878	Ppv2-0_MIR1055	UUAGGGGUGUUUUCCAGUGACU	116	CCACUGGAAACACCCCUGAAU	4	154
scaffold_422:146945-147131	Ppv2-0_MIR160e	CCAUACAGGGAGUCAGACAGA	24832	UGCCUGGCUCCCUGUAUGCCA	265	30735
scaffold_427:191987-192179	Ppv2-0_MIR529a	CGAAGAGAGAGAGCACAGCCC	3190	GCUGUGCUCUCUCUCUUCUGG	24	4329
scaffold_433:104785-104919	Ppv2-0_MIR160g	CCAUGCAGGGAGUCAGACAGA	75	UGCCUGGCUCCUUGUAUGCCA	2	113
scaffold_439:50265-50549	Ppv2-0_MIR534a	AUGCAACUUGUGUGGACAGACU	255885	UAUGUCCAUUGCAGUUGCAUAC	4550	327002
scaffold_448:172817-172935	Ppv2-0_MIR1223c	AAGGUGUGCGACUCUAUAAUC	16422	UUGUAGAGUCAUACACCUCUA	208	19570
scaffold_448:173163-173288	Ppv2-0_MIR1223a	UUGUAGAGUCAUACACCUCCA	3073	AGGGUGUGUGACUCUAUAAUC	709	5669
scaffold_511:22738-22849	Ppv2-0_Cluster_38636	UUUCUAUCUUCUCUUCCAAUC, UUCUAUCUUCUCUUCCAAUCU	1961, 1279	AUUGGAAGAGCAGAUCGAAAA, UUGGAAGAGCAGAUCGAAAAA	53, 306	4984
scaffold_536:16047-16144	Ppv2-0_MIR902f	ACGAAGAUCUGCAUCAUAAC	9608	UAUGAUGCAGAUUCUUCAUCU	126	10631
scaffold_551:45746-45873	Ppv2-0_MIR529c	CGAAGAGAGAGAGCACAGCCC	3182	GCUGUGCUCUCUCUUUUCCAG	229	4693

miRBase release 20 lists 229 *Physcomitrella MIRNA* loci, of which 105 are annotated as high-confidence based upon older small RNA-seq datasets (Kozomara and Griffiths-Jones 2014). Our deeper dataset coupled with improved *MIRNA* annotation methods allowed us to further assess these prior annotations. Most *Physcomitrella*

miRBase loci (217 out of 229) were discovered as small RNA producing loci in our analysis (Table 2.5). Only 109 of the prior miRBase annotations satisfied the strict structure and expression criteria we imposed to designate a *MIRNA* locus with at least 80% of its small RNAs falling within the *DCL* size range of 20-24 nts (Fig. 2.3B). Interestingly, the overlap between those 109 and the loci noted as "high confidence" loci in miRBase 20 (Kozomara and Griffiths-Jones 2014) was not very high. Only 56 of the 105 miRBase 20 high confidence loci were accepted by our analysis (Fig. 2.3B, Table 2.5). We attribute this to the much greater sequencing depth, and consequent increased specificity, that our new small RNA-seq data allowed.

Table 2.5: All miRBase loci and overlapping ShortStack loci

miRBase miRNA (release 20)	miRNA family	miRBase high confidence miRNA	miRBase locus (scaffold:start-end strand)	Overlapping ShortStack locus name	ShortStack locus (scaffold:start-end strand)	ShortStack annotation	ShortStack DicerCall
ppt-MIR156c	156	High confidence	scaffold_3:2089939-2090033 +	Ppv2-0_MIR156c	scaffold_3:2089924-2090047 +	MIRNA	20
ppt-MIR156a	156	High confidence	scaffold_12:1858517-1858631 -	Ppv2-0_MIR156a	scaffold_12:1858517-1858631 -	MIRNA	20
ppt-MIR156b	156	-	scaffold_167:958010-958229 +	Ppv2-0_MIR156b	scaffold_167:958055-958142 +	MIRNA	20
ppt-MIR160c	160	High confidence	scaffold_19:1175581-1175680 -	Ppv2-0_MIR160c	scaffold_19:1175559-1175703 -	MIRNA	21
ppt-MIR160b	160	-	scaffold_19:1176220-1176322 -	Ppv2-0_MIR160b	scaffold_19:1176196-1176349 -	MIRNA	21
ppt-MIR160f	160	High confidence	scaffold_29:139485-139622 +	Ppv2-0_Cluster_7798	scaffold_29:139508-139598 +	Non-HP	N
ppt-MIR160a	160	High confidence	scaffold_104:1047225-1047323 -	Ppv2-0_Cluster_19658	scaffold_104:1047222-1047324 -	HP	21
ppt-MIR160i	160	-	scaffold_167:174013-174151 -	Ppv2-0_Cluster_25472	scaffold_167:174039-174143 -	Non-HP	N
ppt-MIR160d	160	High confidence	scaffold_167:174750-174849 -	Ppv2-0_Cluster_25473	scaffold_167:174757-174860 -	Non-HP	N
ppt-MIR160h	160	-	scaffold_167:175044-175179 -	Ppv2-0_MIR160h	scaffold_167:174995-175231 -	MIRNA	21
ppt-MIR160e	160	High confidence	scaffold_422:146965-147102 -	Ppv2-0_MIR160e	scaffold_422:146945-147131 -	MIRNA	21
ppt-MIR160g	160	High confidence	scaffold_433:104783-104920 +	Ppv2-0_MIR160g	scaffold_433:104785-104919 +	MIRNA	21
ppt-MIR166c	166	-	scaffold_14:2065312-2065434 +	Ppv2-0_Cluster_4425	scaffold_14:2065341-2065592 +	Non-HP	21
ppt-MIR166d	166	-	scaffold_14:2065490-2065617 +	Ppv2-0_Cluster_4425	scaffold_14:2065341-2065592 +	Non-HP	21
ppt-MIR166f	166	High confidence	scaffold_14:2066016-2066138 +	Ppv2-0_MIR166f	scaffold_14:2066036-2066119 +	MIRNA	21
ppt-MIR166g	166	-	scaffold_21:64657-64802 +	Ppv2-0_MIR166g	scaffold_21:64659-64801 +	MIRNA	21
ppt-MIR166b	166	High confidence	scaffold_26:1562339-1562441 +	Ppv2-0_MIR166b	scaffold_26:1562334-1562447 +	MIRNA	21
ppt-MIR166a	166	-	scaffold_30:1354148-1354249 -	Ppv2-0_MIR166a	scaffold_30:1354118-1354281 -	MIRNA	21
ppt-MIR166e	166	High confidence	scaffold_36:478447-478592 -	Ppv2-0_MIR166e	scaffold_36:478448-478590 -	MIRNA	21
ppt-MIR166l	166	-	scaffold_64:1475278-1475410 +	Ppv2-0_MIR166l	scaffold_64:1475293-1475396 +	MIRNA	21
ppt-MIR166i	166	-	scaffold_127:459114-459233 +	Ppv2-0_Cluster_22016	scaffold_127:459143-459380 +	Non-HP	21
ppt-MIR166j	166	-	scaffold_127:459279-459409 +	Ppv2-0_Cluster_22016	scaffold_127:459143-459380 +	Non-HP	21
ppt-MIR166m	166	-	scaffold_487:54440-54569 -	Ppv2-0_Cluster_38487	scaffold_487:54468-54491 -	Non-HP	N
ppt-MIR166h	166	-	scaffold_487:54795-54920 -	Ppv2-0_Cluster_38488	scaffold_487:54821-55074 -	Non-HP	21
ppt-MIR166k	166	-	scaffold_487:54980-55102 -	Ppv2-0_Cluster_38488	scaffold_487:54821-55074 -	Non-HP	21
ppt-MIR167	167	-	N/A	N/A	N/A	N/A	N/A
ppt-MIR171b	171	-	scaffold_19:293884-293975 -	Ppv2-0_MIR171b	scaffold_19:293821-294034 -	MIRNA	21
ppt-MIR171a	171	High confidence	scaffold_71:1665613-1665703 +	Ppv2-0_MIR171a	scaffold_71:1665539-1665796 +	MIRNA	21
ppt-MIR319a	319	-	scaffold_19:89704-89872 +	Ppv2-0_MIR319a	scaffold_19:89684-89892 +	MIRNA	20
ppt-MIR319c	319	-	scaffold_29:1728494-1728682 -	Ppv2-0_Cluster_7886	scaffold_29:1728435-1728653 -	HP	21
ppt-MIR319e	319	-	scaffold_99:433971-434144 +	Ppv2-0_MIR319e	scaffold_99:433973-434143 +	MIRNA	21
ppt-MIR319b	319	High confidence	scaffold_234:55730-55889 -	Ppv2-0_MIR319b	scaffold_234:55723-55899 -	MIRNA	20
ppt-MIR319d	319	-	scaffold_266:191734-191903 +	Ppv2-0_MIR319d	scaffold_266:191728-191909 +	MIRNA	21
ppt-MIR390b	390	-	scaffold_80:1498969-1499143 -	Ppv2-0_MIR390b	scaffold_80:1498991-1499142 -	MIRNA	21
ppt-MIR390a	390	-	scaffold_119:1056046-1056182 -	Ppv2-0_MIR390a	scaffold_119:1056015-1056219 -	MIRNA	21
ppt-MIR390c	390	High confidence	scaffold_264:204264-204398 -	Ppv2-0_MIR390c	scaffold_264:204261-204402 -	MIRNA	20
ppt-MIR395	395	-	scaffold_261:428635-428812 -	N/A	N/A	N/A	N/A
ppt-MIR408b	408	-	scaffold_313:118222-118365 -	Ppv2-0_MIR408b	scaffold_313:118181-118386 -	MIRNA	21
ppt-MIR408a	408	-	scaffold_333:250024-250171 -	N/A	N/A	N/A	N/A
ppt-MIR414	414	-	scaffold_123:538836-539011 -	Ppv2-0_Cluster_21627	scaffold_123:538843-538890 +	Non-HP	N
ppt-MIR419	419	-	scaffold_177:53968-54049 -	Ppv2-0_Cluster_26157	scaffold_177:53958-53993 -	Non-HP	N
ppt-MIR477e	477	-	scaffold_14:198811-198937 -	Ppv2-0_MIR477e	scaffold_14:198801-198944 -	MIRNA	21
ppt-MIR477h	477	-	scaffold_27:2077199-2077353 +	Ppv2-0_MIR477h	scaffold_27:2077215-2077338 +	MIRNA	21
ppt-MIR477c	477	-	scaffold_50:2006194-2006278 +	Ppv2-0_Cluster_11994	scaffold_50:2006188-2006286 +	HP	22
ppt-MIR477d	477	-	scaffold_165:751611-751764 +	Ppv2-0_Cluster_25379	scaffold_165:751640-751734 +	Non-HP	N
ppt-MIR477b	477	-	scaffold_216:329066-329235 -	Ppv2-0_Cluster_28943	scaffold_216:329097-329203 -	Non-HP	N
ppt-MIR477f	477	-	scaffold_234:21912-22038 -	Ppv2-0_MIR477f	scaffold_234:21891-22056 -	MIRNA	21
ppt-MIR477a	477	High confidence	scaffold_242:477499-477583 +	Ppv2-0_Cluster_30458	scaffold_242:477503-477583 +	Non-HP	N
ppt-MIR477g	477	High confidence	scaffold_275:298272-298399 +	Ppv2-0_MIR477g	scaffold_275:298232-298433 +	MIRNA	21
ppt-MIR529f	529	High confidence	scaffold_18:1793125-1793246 +	Ppv2-0_Cluster_5415	scaffold_18:1793154-1793382	Non-HP	21
ppt-MIR529g	529	High confidence	scaffold_18:1793338-1793486 +	Ppv2-0_Cluster_5415	scaffold_18:1793154-1793382	Non-HP	21
ppt-MIR529b	529	-	scaffold_40:841404-841552 -	Ppv2-0_Cluster_10038	scaffold_40:841428-841745 -	Non-HP	21
ppt-MIR529d	529	-	scaffold_40:841646-841781 -	Ppv2-0_Cluster_10038	scaffold_40:841428-841745 -	Non-HP	21
ppt-MIR529a	529	High confidence	scaffold_427:191894-192042 -	Ppv2-0_MIR529a	scaffold_427:191987-192179 -	MIRNA	21
ppt-MIR529e	529	High confidence	scaffold_427:192122-192243 -	Ppv2-0_MIR529a	scaffold_427:191987-192179 -	MIRNA	21
ppt-MIR529c	529	-	scaffold_551:45735-45885 -	Ppv2-0_MIR529c	scaffold_551:45746-45873 -	MIRNA	21

ppt-MIR533d	533	-	scaffold_19:828528-828665 +	Ppv2-0_MIR533d	scaffold_19:828533-828660 +	MIRNA	21
ppt-MIR533b	533	-	scaffold_31:1002600-1002724 -	Ppv2-0_MIR533b	scaffold_31:1002602-1002725 -	MIRNA	21
ppt-MIR533a	533	High confidence	scaffold_89:1180775-1180876 -	Ppv2-0_MIR533a	scaffold_89:1180768-1180881 -	MIRNA	21
ppt-MIR533c	533	High confidence	scaffold_100:50354-50524 -	Ppv2-0_MIR533c	scaffold_100:50329-50546 -	MIRNA	21
ppt-MIR533e	533	High confidence	scaffold_224:413758-413910 +	Ppv2-0_Cluster_29429	scaffold_224:413791-413880 +	Non-HP	22
ppt-MIR534b	534	High confidence	scaffold_23:1118731-1119009 -	Ppv2-0_Cluster_6579	scaffold_23:1118843-1119136 -	HP	22
ppt-MIR534a	534	-	scaffold_439:50310-50509 -	Ppv2-0_MIR534a	scaffold_439:50265-50549 -	MIRNA	22
ppt-MIR535d	535	High confidence	scaffold_44:180020-180104 +	Ppv2-0_MIR535d	scaffold_44:180008-180118 +	MIRNA	21
ppt-MIR535b	535	High confidence	scaffold_56:756200-756280 -	Ppv2-0_MIR535b	scaffold_56:756196-756284 -	MIRNA	21
ppt-MIR535c	535	High confidence	scaffold_101:1152052-1152132 +	Ppv2-0_MIR535c	scaffold_101:1152005-1152185 +	MIRNA	21
ppt-MIR535a	535	High confidence	scaffold_184:577100-577234 -	Ppv2-0_MIR535a	scaffold_184:577127-577210 -	MIRNA	21
ppt-MIR536b	536	-	scaffold_16:354484-354705 -	Ppv2-0_Cluster_4771	scaffold_16:354403-354590 -	HP	22
ppt-MIR536c	536	-	scaffold_25:2619793-2620014 -	Ppv2-0_Cluster_7160	scaffold_25:2619723-2619986 -	Non-HP	N
ppt-MIR536f	536	-	scaffold_56:1822559-1822734 +	Ppv2-0_Cluster_12998	scaffold_56:1822684-1822707 +	Non-HP	N
ppt-MIR536a	536	-	scaffold_56:1827125-1827323 -	Ppv2-0_Cluster_13001	scaffold_56:1827130-1827187 -	HP	22
ppt-MIR536d	536	-	scaffold_73:909101-909543 +	Ppv2-0_Cluster_15539	scaffold_73:909129-909151 +	Non-HP	21
ppt-MIR536d	536	-	scaffold_73:909101-909543 +	Ppv2-0_Cluster_15540	scaffold_73:909407-909517 +	HP	22
ppt-MIR536e	536	-	scaffold_82:417955-418418 +	Ppv2-0_Cluster_16783	scaffold_82:417970-418050 +	HP	21
ppt-MIR536e	536	-	scaffold_82:417955-418418 +	Ppv2-0_Cluster_16784	scaffold_82:418375-418541 +	HP	22
ppt-MIR537b	537	-	scaffold_1:3540236-3540384 +	Ppv2-0_MIR537b	scaffold_1:3540225-3540393 +	MIRNA	21
ppt-MIR537c	537	-	scaffold_1:3555321-3555469 +	Ppv2-0_MIR537c	scaffold_1:3555310-3555478 +	MIRNA	21
ppt-MIR537d	537	High confidence	scaffold_2:4065022-4065199 -	Ppv2-0_Cluster_765	scaffold_2:4065040-4065176 -	Non-HP	N
ppt-MIR537a	537	High confidence	scaffold_52:1225399-1225588 +	Ppv2-0_MIR537a	scaffold_52:1225394-1225592 +	MIRNA	21
ppt-MIR538b	538	High confidence	scaffold_19:2735034-2735151 -	Ppv2-0_MIR538b	scaffold_19:2735033-2735152 -	MIRNA	21
ppt-MIR538a	538	-	scaffold_234:166194-166372 -	Ppv2-0_Cluster_29985	scaffold_234:166203-166361 -	Non-HP	21
ppt-MIR538c	538	High confidence	scaffold_369:340096-340255 -	Ppv2-0_MIR538c	scaffold_369:340080-340267 -	MIRNA	21
ppt-MIR893	893	-	scaffold_348:182950-183169 +	Ppv2-0_Cluster_35524	scaffold_348:182924-183218 +	HP	21
ppt-MIR894	894	-	scaffold_10:1562414-1562633 -	Ppv2-0_Cluster_3260	scaffold_10:1562563-1562584 -	Non-HP	N
ppt-MIR895	895	-	scaffold_433:159731-159952 +	N/A	N/A	N/A	N/A
ppt-MIR897	897	-	scaffold_19:1445467-1445689 +	Ppv2-0_Cluster_5630	scaffold_19:1445509-1445588 +	HP	24
ppt-MIR898a	898	-	scaffold_51:599206-599425 +	Ppv2-0_MIR898a	scaffold_51:599238-599388 +	MIRNA	22
ppt-MIR898b	898	-	scaffold_126:342471-342690 +	Ppv2-0_MIR898b	scaffold_126:342504-342653 +	MIRNA	22
ppt-MIR899	899	-	scaffold_459:1646-1867 +	Ppv2-0_Cluster_38130	scaffold_459:1390-2025 +	Non-HP	21
ppt-MIR900	900	-	scaffold_374:104328-104547 -	Ppv2-0_Cluster_36303	scaffold_374:104372-104448	Non-HP	21
ppt-MIR901	901	-	scaffold_182:652806-653026 -	Ppv2-0_Cluster_26593	scaffold_182:642869-655801	Non-HP	23
ppt-MIR902b	902	-	scaffold_2:4263201-4263423 -	Ppv2-0_Cluster_782	scaffold_2:4263250-4263331 -	Non-HP	N
ppt-MIR902j	902	High confidence	scaffold_5:3307166-3307338 +	Ppv2-0_MIR902j	scaffold_5:3307166-3307340 +	MIRNA	21
ppt-MIR902l	902	-	scaffold_6:1347479-1347612 +	Ppv2-0_Cluster_2155	scaffold_6:1347511-1347583 +	Non-HP	N
ppt-MIR902e	902	High confidence	scaffold_11:1661349-1661481 +	Ppv2-0_MIR902e	scaffold_11:1661361-1661470 +	MIRNA	20
ppt-MIR902g	902	High confidence	scaffold_17:2119374-2119514 +	Ppv2-0_MIR902g	scaffold_17:2119355-2119544 +	MIRNA	20
ppt-MIR902h	902	High confidence	scaffold_19:562795-562922 -	Ppv2-0_MIR902h	scaffold_19:562783-562931 -	MIRNA	21
ppt-MIR902d	902	High confidence	scaffold_45:1704540-1704668 -	Ppv2-0_MIR902d	scaffold_45:1704524-1704683 -	MIRNA	20
ppt-MIR902a	902	-	scaffold_45:1758930-1759152 +	Ppv2-0_Cluster_11124	scaffold_45:1759034-1759100 +	Non-HP	N
ppt-MIR902k	902	High confidence	scaffold_74:647206-647347 -	Ppv2-0_MIR902k	scaffold_74:647191-647360 -	MIRNA	21
ppt-MIR902c	902	High confidence	scaffold_179:820112-820245 +	Ppv2-0_MIR902c	scaffold_179:820129-820230 +	MIRNA	20
ppt-MIR902i	902	High confidence	scaffold_488:34764-34883 -	Ppv2-0_Cluster_38492	scaffold_488:34739-34908 -	HP	21
ppt-MIR902f	902	-	scaffold_536:16034-16156 +	Ppv2-0_MIR902f	scaffold_536:16047-16144 +	MIRNA	20
ppt-MIR903	903		scaffold_3506:8104-8315 +	Ppv2-0_Cluster_39779	scaffold_3506:7625-9357 +	Non-HP	N
ppt-MIR904a	904	-	scaffold_234:189702-189921 -	Ppv2-0_MIR904a	scaffold_234:189677-189961 -	MIRNA	21
ppt-MIR904b	904	-	scaffold_266:312404-312623 +	Ppv2-0_MIR904b	scaffold_266:312421-312594 +	MIRNA	21
ppt-MIR1023c	1023	High confidence	scaffold_51:950574-950741 -	Ppv2-0_MIR1023c	scaffold_51:950572-950744 -	MIRNA	21
ppt-MIR1023e	1023	High confidence	scaffold_51:960497-960616 +	N/A	N/A	N/A	N/A
ppt-MIR1023a	1023	High confidence	scaffold_59:1334025-1334194 +	Ppv2-0_MIR1023a	scaffold_59:1334038-1334181 +	MIRNA	21
ppt-MIR1023d	1023	-	scaffold_126:664283-664423 -	N/A	N/A	N/A	N/A
ppt-MIR1023b	1023	High confidence	scaffold_257:30575-30746 -	Ppv2-0_MIR1023b	scaffold_257:30590-30731 -	MIRNA	21
ppt-MIR1024a	1024	High confidence	scaffold_43:1696881-1697042 -	Ppv2-0_MIR1024a	scaffold_43:1696878-1697030 -	MIRNA	20
ppt-MIR1024b	1024	High confidence	scaffold_336:345174-345297 -	Ppv2-0_MIR1024b	scaffold_336:345183-345323 -	MIRNA	20
ppt-MIR1025	1025	High confidence	scaffold_52:1805710-1805919 -	Ppv2-0_Cluster_12305	scaffold_52:1805712-1805916 -	HP	20
ppt-MIR1026b	1026	High confidence	scaffold_234:416678-416829 +	Ppv2-0_MIR1026b	scaffold_234:416698-416810 +	MIRNA	21
ppt-MIR1026a	1026	-	scaffold_275:115212-115382 -	Ppv2-0_Cluster_32436	scaffold_275:115212-115383 -	HP	21
ppt-MIR1027a	1027	-	scaffold_94:1388982-1389123 -	Ppv2-0_MIR1027a	scaffold_94:1388997-1389108 -	MIRNA	21
ppt-MIR1027b	1027	-	scaffold_94:1430737-1430878 -	Ppv2-0_MIR1027b	scaffold_94:1430752-1430863 -	MIRNA	21
ppt-MIR1028b	1028	High confidence	scaffold_34:1432898-1433091 +	Ppv2-0_MIR1028b	scaffold_34:1432904-1433088 +	MIRNA	21
ppt-MIR1028a	1028	High confidence	scaffold_167:878809-879047 +	Ppv2-0_Cluster_25536	scaffold_167:878838-878855 +	Non-HP	N
ppt-MIR1028c	1028	High confidence	scaffold_243:593002-593198 -	Ppv2-0_MIR1028c	scaffold_243:593008-593194 -	MIRNA	21
ppt-MIR1029	1029	-	scaffold_381:111310-111465 +	Ppv2-0_MIR1029	scaffold_381:111313-111462 +	MIRNA	21
ppt-MIR1030d	1030	-	scaffold_25:514775-514992 +	Ppv2-0_Cluster_6993	scaffold_25:514804-515269 +	Non-HP	21
ppt-MIR1030e	1030	High confidence	scaffold_25:515080-515287 +	Ppv2-0_Cluster_6993	scaffold_25:514804-515269 +	Non-HP	21
ppt-MIR1030a	1030	High confidence	scaffold_110:19833-20047 +	Ppv2-0_Cluster_20199	scaffold_110:19616-20018 +	Non-HP	21
ppt-MIR1030b	1030	High confidence	scaffold_217:672403-672572 +	Ppv2-0_Cluster_29010	scaffold_217:672412-672686 +	Non-HP	21
ppt-MIR1030j	1030	-	scaffold_245:435034-435255 -	Ppv2-0_Cluster_30632	scaffold_245:435204-435504 -	Non-HP	21
ppt-MIR1030g	1030	High confidence	scaffold_245:435319-435526 -	Ppv2-0_Cluster_30632	scaffold_245:435204-435504 -	Non-HP	21
ppt-MIR1030h	1030	High confidence	scaffold_305:497882-498052 -	Ppv2-0_Cluster_33966	scaffold_305:498018-498347 -	Non-HP	21
ppt-MIR1030c	1030	High confidence	scaffold_305:498157-498374 -	Ppv2-0_Cluster_33966	scaffold_305:498018-498347 -	Non-HP	21
ppt-MIR1030i	1030	-	scaffold_435:112340-112569 -	Ppv2-0_Cluster_37820	scaffold_435:112517-112820 -	Non-HP	21
ppt-MIR1030f	1030	High confidence	scaffold_435:112638-112833 -	Ppv2-0_Cluster_37820	scaffold_435:112517-112820 -	Non-HP	21
ppt-MIR1031a	1031	-	scaffold_58:126275-126482 +	Ppv2-0_MIR1031a	scaffold_58:126289-126469 +	MIRNA	21
ppt-MIR1031b	1031	-	scaffold_79:1240135-1240332 -	Ppv2-0_Cluster_16456	scaffold_79:1240128-1240188 -	HP	21
ppt-MIR1032	1032	-	scaffold_208:523177-523303 -	Ppv2-0_MIR1032	scaffold_208:523161-523318 -	MIRNA	21
ppt-MIR1033d	1033	-	scaffold_5:35705-35823 +	Ppv2-0_MIR1033d	scaffold_5:35728-35801 +	MIRNA	22
ppt-MIR1033a	1033	-	scaffold_13:1365532-1365650 -	Ppv2-0_MIR1033a	scaffold_13:1365519-1365663 -	MIRNA	22
ppt-MIR1033b	1033	-	scaffold_29:989044-989162 -	Ppv2-0_MIR1033b	scaffold_29:989057-989147 -	MIRNA	21
ppt-MIR1033e	1033	-	scaffold_256:157985-158103 -	Ppv2-0_MIR1033e	scaffold_256:157999-158087 -	MIRNA	21
ppt-MIR1033c	1033	-	scaffold_291:313507-313625 -	Ppv2-0_MIR1033c	scaffold_291:313490-313651 -	MIRNA	21
ppt-MIR1034	1034	High confidence	scaffold_202:211306-211450 -	Ppv2-0_MIR1034	scaffold_202:211291-211472 -	MIRNA	21
ppt-MIR1035	1035	High confidence	scaffold_217:384203-384372 +	Ppv2-0_MIR1035	scaffold_217:384222-384354 +	MIRNA	21
ppt-MIR1036	1036	High confidence	scaffold_94:882776-882952 -	Ppv2-0_MIR1036	scaffold_94:882754-882972 -	MIRNA	21
ppt-MIR1037	1037	-	scaffold_100:446962-447089 +	Ppv2-0_Cluster_19168	scaffold_100:446988-447067	Non-HP	N
ppt-MIR1038	1038	High confidence	scaffold_8:1028777-1028896 +	N/A	N/A	N/A	N/A
ppt-MIR1039	1039	-	scaffold_325:243071-243200 +	Ppv2-0_MIR1039	scaffold_325:243086-243186 +	MIRNA	21

ppt-MIR1040	1040	High confidence	scaffold_62:1568544-1568685 +	Ppv2-0_Cluster_13871	scaffold_62:1568570-1568658 +	Non-HP	N
ppt-MIR1041	1041	-	scaffold_8:62755-62896 +	Ppv2-0_Cluster_2657	scaffold_8:62784-62808	Non-HP	21
ppt-MIR1042	1042	-	scaffold_287:109461-109593 +	Ppv2-0_MIR1042	scaffold_287:109482-109573 +	MIRNA	22
ppt-MIR1043	1043	-	scaffold_287:109841-109969 +	Ppv2-0_MIR1043	scaffold_287:109856-109954 +	MIRNA	21
ppt-MIR1044	1044	High confidence	scaffold_30:1040926-1041057 +	Ppv2-0_Cluster_8007	scaffold_30:1040966-1041020 +	Non-HP	N
ppt-MIR1045	1045	-	scaffold_128:563490-563634 -	Ppv2-0_MIR1045	scaffold_128:563507-563616 -	MIRNA	21
ppt-MIR1046	1046	-	scaffold_301:448437-448576 +	N/A	N/A	N/A	N/A
ppt-MIR1047	1047	High confidence	scaffold_30:2152683-2152827 +	Ppv2-0_Cluster_8096	scaffold_30:2152717-2152797 +	Non-HP	N
ppt-MIR1048	1048	High confidence	scaffold_248:120170-120314 -	Ppv2-0_MIR1048	scaffold_248:120187-120296 -	MIRNA	21
ppt-MIR1049	1049	-	scaffold_8:1132552-1132690 -	Ppv2-0_MIR1049	scaffold_8:1132567-1132675 -	MIRNA	21
ppt-MIR1050	1050	High confidence	scaffold_96:1006513-1006663 +	Ppv2-0_MIR1050	scaffold_96:1006502-1006657 +	MIRNA	21
ppt-MIR1051	1051	-	scaffold_56:1824620-1824816 +	Ppv2-0_Cluster_13000	scaffold_56:1824605-1824831 +	HP	21
ppt-MIR1052	1052	High confidence	scaffold_80:1382853-1383025 +	Ppv2-0_MIR1052	scaffold_80:1382873-1383005 +	MIRNA	21
ppt-MIR1053	1053	-	scaffold_38:1275612-1275732 +	Ppv2-0_Cluster_9673	scaffold_38:1275653-1275832 +	Non-HP	N
ppt-MIR1054	1054	-	scaffold_112:1004878-1005122 -	Ppv2-0_MIR1054	scaffold_112:1004893-1005105 -	MIRNA	23
ppt-MIR1055	1055	-	scaffold_391:252741-252894 -	Ppv2-0_MIR1055	scaffold_391:252756-252878 -	MIRNA	22
ppt-MIR1056	1056	-	scaffold_193:686926-687070 +	Ppv2-0_Cluster_27375	scaffold_193:686955-687049 +	Non-HP	21
ppt-MIR1057	1057	-	scaffold_22:898467-898582 -	Ppv2-0_Cluster_6307	scaffold_22:898526-898549 -	Non-HP	N
ppt-MIR1058	1058	High confidence	scaffold_71:1697955-1698096 +	Ppv2-0_Cluster_15290	scaffold_71:1697985-1698075	Non-HP	N
ppt-MIR1059	1059	High confidence	scaffold_242:277353-277567 +	Ppv2-0_Cluster_30438	scaffold_242:277369-277562 +	Non-HP	N
ppt-MIR1060	1060	High confidence	scaffold_58:1496367-1496511 -	Ppv2-0_Cluster_13253	scaffold_58:1496460-1496482 -	Non-HP	N
ppt-MIR1061	1061	-	scaffold_64:284784-284915 +	Ppv2-0_Cluster_14109	scaffold_64:284813-284881 +	Non-HP	N
ppt-MIR1062	1062	High confidence	scaffold_49:1228135-1228318 -	Ppv2-0_MIR1062	scaffold_49:1228150-1228303 -	MIRNA	22
ppt-MIR1063g	1063	High confidence	scaffold_20:864059-864369 -	Ppv2-0_Cluster_5830	scaffold_20:864036-864340 -	Non-HP	20
ppt-MIR1063f	1063	High confidence	scaffold_63:917568-917838 -	Ppv2-0_Cluster_13978	scaffold_63:917594-917811 -	Non-HP	N
ppt-MIR1063e	1063	High confidence	scaffold_71:1575437-1575672 -	Ppv2-0_Cluster_15283	scaffold_71:1575463-1575643 -	Non-HP	N
ppt-MIR1063d	1063	High confidence	scaffold_71:1581249-1581604 +	Ppv2-0_Cluster_15284	scaffold_71:1581278-1581301 +	Non-HP	N
ppt-MIR1063d	1063	High confidence	scaffold_71:1581249-1581604 +	Ppv2-0_Cluster_15285	scaffold_71:1581545-1581714 +	HP	20
ppt-MIR1063h	1063	High confidence	scaffold_75:1216009-1216235 +	Ppv2-0_Cluster_15841	scaffold_75:1216036-1216061 +	Non-HP	N
ppt-MIR1063a	1063	High confidence	scaffold_80:712778-712996 +	Ppv2-0_Cluster_16533	scaffold_80:712805-712829 +	Non-HP	N
ppt-MIR1063b	1063	High confidence	scaffold_125:310485-310770 +	Ppv2-0_Cluster_21788	scaffold_125:310512-310917 +	Non-HP	N
ppt-MIR1063c	1063	High confidence	scaffold_441:170781-171143 +	Ppv2-0_Cluster_37875	scaffold_441:170810-170833 +	Non-HP	N
ppt-MIR1063c	1063	High confidence	scaffold_441:170781-171143 +	Ppv2-0_Cluster_37876	scaffold_441:171097-171117 +	Non-HP	N
ppt-MIR1064	1064	-	scaffold_202:191936-192050 +	Ppv2-0_Cluster_28014	scaffold_202:191943-192044 +	HP	22
ppt-MIR1065	1065	High confidence	scaffold_75:65996-66207 -	Ppv2-0_Cluster_15756	scaffold_75:66022-66178 -	Non-HP	N
ppt-MIR1066	1066	-	scaffold_54:45985-46103 +	Ppv2-0_Cluster_12530	scaffold_54:46005-46086	Non-HP	21
ppt-MIR1067	1067	High confidence	scaffold_220:156632-156799 +	Ppv2-0_MIR1067	scaffold_220:156609-156824 +	MIRNA	21
ppt-MIR1068	1068	High confidence	scaffold_244:13340-13543 -	Ppv2-0_Cluster_30546	scaffold_244:13343-13555 -	HP	21
ppt-MIR1069	1069	-	scaffold_72:21063-21306 -	Ppv2-0_MIR1069	scaffold_72:21038-21325 -	MIRNA	21
ppt-MIR1070	1070	-	scaffold_109:1216085-1216196 +	Ppv2-0_Cluster_20163	scaffold_109:1216102-1216182 +	Non-HP	N
ppt-MIR1071	1071	-	scaffold_312:233202-233336 -	N/A	N/A	N/A	N/A
ppt-MIR1072	1072	High confidence	scaffold_15:2064032-2064340 -	Ppv2-0_Cluster_4660	scaffold_15:2064036-2064307 -	Non-HP	N
ppt-MIR1073	1073	High confidence	scaffold_313:116800-117000 -	Ppv2-0_MIR1073	scaffold_313:116809-116997 -	MIRNA	21
ppt-MIR1074	1074	-	scaffold_105:203903-204032 -	Ppv2-0_Cluster_19689	scaffold_105:203914-204020 -	HP	21
ppt-MIR1075	1075	-	scaffold_5:1647137-1647278 +	Ppv2-0_Cluster_1833	scaffold_5:1647135-1647280 +	HP	21
ppt-MIR1076	1076	-	scaffold_245:121314-121540 +	Ppv2-0_MIR1076	scaffold_245:121321-121538 +	MIRNA	22
ppt-MIR1077	1077	-	scaffold_387:337061-337212 -	Ppv2-0_Cluster_36758	scaffold_387:337018-337252 -	HP	21
ppt-MIR1078	1078	High confidence	scaffold_117:583847-583957 +	Ppv2-0_Cluster_21022	scaffold_117:583859-583946 +	HP	21
ppt-MIR1079	1079	-	scaffold_139:409407-409557 +	Ppv2-0_MIR1079	scaffold_139:409391-409574 +	MIRNA	20
ppt-MIR1211	1211	High confidence	scaffold_227:732164-732275 -	Ppv2-0_Cluster_29607	scaffold_227:732181-732260 -	HP	21
ppt-MIR1212	1212	High confidence	scaffold_107:19126-19214 -	Ppv2-0_MIR1212	scaffold_107:19112-19238 -	MIRNA	21
ppt-MIR1214	1214	High confidence	scaffold_141:592761-592853 -	Ppv2-0_MIR1214	scaffold_141:592759-592855 -	MIRNA	21
ppt-MIR1215	1215	High confidence	scaffold_173:168162-168288 -	Ppv2-0_MIR1215	scaffold_173:168098-168342 -	MIRNA	22
ppt-MIR1216	1216	High confidence	scaffold_77:341294-341451 +	Ppv2-0_MIR1216	scaffold_77:341294-341451 +	MIRNA	21
ppt-MIR1217	1217	High confidence	scaffold_100:542222-542332 +	Ppv2-0_MIR1217	scaffold_100:542192-542372 +	MIRNA	21
ppt-MIR1218	1218	-	scaffold_101:629067-629204 +	Ppv2-0_Cluster_19302	scaffold_101:629078-629194 +	Non-HP	N
ppt-MIR1219d	1219	-	scaffold_37:2002951-2003086 +	Ppv2-0_Cluster_9545	scaffold_37:2002978-2003306 +	Non-HP	21
ppt-MIR1219c	1219	-	scaffold_37:2003197-2003348 +	Ppv2-0_Cluster_9545	scaffold_37:2002978-2003306 +	Non-HP	21
ppt-MIR1219a	1219	High confidence	scaffold_625:20799-20933 +	Ppv2-0_Cluster_39000	scaffold_625:20795-21135 +	Non-HP	21
ppt-MIR1219b	1219	High confidence	scaffold_625:21055-21132 +	Ppv2-0_Cluster_39000	scaffold_625:20795-21135 +	Non-HP	21
ppt-MIR1220a	1220	-	scaffold_161:204425-204629 +	Ppv2-0_MIR1220a	scaffold_161:204425-204629 +	MIRNA	21
ppt-MIR1220b	1220	-	scaffold_161:217252-217456 -	Ppv2-0_Cluster_25018	scaffold_161:217252-217456 -	HP	21
ppt-MIR1221	1221	High confidence	scaffold_106:956224-956400 -	Ppv2-0_MIR1221	scaffold_106:956218-956410 -	MIRNA	22
ppt-MIR1222d	1222	-	scaffold_79:1036180-1036310 +	Ppv2-0_MIR1222d	scaffold_79:1036196-1036295 +	MIRNA	21
ppt-MIR1222c	1222	-	scaffold_83:1094080-1094213 -	Ppv2-0_MIR1222c	scaffold_83:1094111-1094181 -	MIRNA	21
ppt-MIR1222b	1222	-	scaffold_124:505531-505658 -	N/A	N/A	N/A	N/A
ppt-MIR1222e	1222	-	scaffold_181:736891-737019 +	N/A	N/A	N/A	N/A
ppt-MIR1222a	1222	High confidence	scaffold_219:310476-310605 +	Ppv2-0_MIR1222a	scaffold_219:310476-310605 +	MIRNA	21
ppt-MIR1223h	1223	-	scaffold_4:1659573-1659704 +	Ppv2-0_Cluster_1419	scaffold_4:1659654-1659674 +	Non-HP	N
ppt-MIR1223d	1223	-	scaffold_4:1660341-1660474 +	Ppv2-0_Cluster_1421	scaffold_4:1660424-1660445 +	Non-HP	N
ppt-MIR1223e	1223	High confidence	scaffold_57:332551-332676 +	Ppv2-0_Cluster_13032	scaffold_57:332573-332914 +	Non-HP	21
ppt-MIR1223i	1223	High confidence	scaffold_57:332815-332938 +	Ppv2-0_Cluster_13032	scaffold_57:332573-332914 +	Non-HP	21
ppt-MIR1223b	1223	-	scaffold_154:555230-555372 -	Ppv2-0_Cluster_24485	scaffold_154:555164-555339 -	Non-HP	N
ppt-MIR1223j	1223	High confidence	scaffold_154:555557-555704 -	Ppv2-0_Cluster_24486	scaffold_154:555583-555675 -	Non-HP	N
ppt-MIR1223f	1223	High confidence	scaffold_154:556184-556307 -	Ppv2-0_Cluster_24487	scaffold_154:556212-556274 -	Non-HP	N
ppt-MIR1223c	1223	-	scaffold_448:172807-172949 -	Ppv2-0_MIR1223c	scaffold_448:172817-172935 -	MIRNA	21
ppt-MIR1223a	1223	High confidence	scaffold_448:173163-173288 -	Ppv2-0_MIR1223a	scaffold_448:173163-173288 -	MIRNA	21
ppt-MIR1223g	1223	High confidence	scaffold_448:173857-173972 -	Ppv2-0_Cluster_38010	scaffold_448:173882-173944 -	Non-HP	N
ppt-MIR2077	2077	-	scaffold_158:358113-358236 +	Ppv2-0_Cluster_24800	scaffold_158:358137-358212 +	Non-HP	21
ppt-MIR2078	2078	-	scaffold_164:899534-899709 +	Ppv2-0_Cluster_25304	scaffold_164:899580-899682	Non-HP	N
ppt-MIR2079	2079	High confidence	scaffold_207:264090-264212 +	Ppv2-0_Cluster_28318	scaffold_207:264111-264195	Non-HP	21
ppt-MIR2080	2080	-	scaffold_213:519123-519251 +	Ppv2-0_Cluster_28746	scaffold_213:519148-519228	Non-HP	21
ppt-MIR2081	2081	High confidence	scaffold_3:3507247-3507368 +	Ppv2-0_Cluster_1196	scaffold_3:3507267-3507348	Non-HP	21
ppt-MIR2082	2082	High confidence	scaffold_345:14320-14450 +	Ppv2-0_MIR2082	scaffold_345:14285-14472 +	MIRNA	21
ppt-MIR2083	2083	-	scaffold_459:1465-1785 +	Ppv2-0_Cluster_38130	scaffold_459:1390-2025 +	Non-HP	21
ppt-MIR2084	2084	-	scaffold_3:3110246-3110359 +	Ppv2-0_MIR2084	scaffold_3:3110232-3110386 +	MIRNA	21
ppt-MIR2085	2085	-	scaffold_3:3110246-3110359 -	Ppv2-0_MIR2084	scaffold_3:3110232-3110386 +	MIRNA	21

2.3.4 No evidence for widespread 5-mC or secondary siRNA accumulation from *Physcomitrella* miRNA targets

It has been proposed that high ratios of miRNA-to-target abundance promote 5-mC modification of target gene DNA in *Physcomitrella* (Khraiwesh et al. 2010). However, bisulfite-seq data from Zemach et al. (Zemach et al. 2010) indicated that *Physcomitrella* genes from a wild-type specimen are largely devoid of 5-mC in all sequence contexts (Fig. 2.3C, D). This lack of gene-body 5-mC was even more strongly apparent in a set of 50 validated miRNA targets (Fig. 2.3C, D, Table 2.6), and there was no enrichment for localized peaks of 5-mC surrounding the target sites themselves (Fig. 2.3E-G). We conclude that either the earlier hypothesis stated by Khraiwesh et al. is incorrect, or alternatively that none of the natural miRNA-to-target ratios in wild-type ten-day-old protonemata are high enough to promote this effect. Additionally, there are no supporting evidence that high ratios of miRNA-to-target abundance promote 5-mC of the target gene DNA in higher plants.

It has been reported that protein-coding *Physcomitrella* miRNA targets often spawn large amounts of secondary siRNAs both upstream and downstream of miRNA target sites (Khraiwesh et al. 2010). From a list of 50 validated protein- coding miRNA targets, 30 overlapped with one or more small RNA clusters (Table 2.6). However, nearly all of the overlapped small RNA clusters (51 out of 54) had DicerCalls of "N", indicating that these small RNAs were not likely to have been derived from DCL processing. Instead, these are likely to simply be degradation products of the mRNAs, not secondary siRNAs. Even the three remaining cases can be dismissed: one is a *PpTAS3* locus and is likely not a protein-coding mRNA to begin with, while the other two are miR156 targets where processing variants of mature miR156 itself can align (Table 2.6). RNA blots against two targets of conserved, highly abundant miRNAs, miR156 and miR166 failed to detect any small RNA accumulation (Fig. 2.3H, I). We conclude there is no convincing evidence that *Physcomitrella* miRNA targets generally spawn secondary siRNAs.

Table 2.6: Degradome-validated *P. patens* miRNA target genes and overlapping ShortStack small RNA loci

miRNA (miRBase Release 20)	Target transcript ID	Target gene ID	Target gene locus (scaffold:start-end strand)	Overlapping ShortStack locus name	ShortStack locus (scaffold:start-end strand)	ShortStack annotation	ShortStack DicerCall
ppt-miR1216	Pp1s4_301V6.1	Pp1s4_301V6	scaffold_4:2000250-2000630 -	N/A	N/A	N/A	N/A
ppt-miR1216	Pp1s4_308V6.1	Pp1s4_308V6	scaffold_4:2050723-2051636 +	N/A	N/A	N/A	N/A
ppt-miR529g	Pp1s6_75V6.1	Pp1s6_75V6	scaffold_6:1029099-1031405 -	N/A	N/A	N/A	N/A
ppt-miR904a	Pp1s7_194V6.1	Pp1s7_194V6	scaffold_7:926144-933453 +	Cluster_2458	scaffold_7:932845-932859 -	Non-HP	N
ppt-miR1038-5p	Pp1s7_389V6.2	Pp1s7_389V6	scaffold_7:2336121-2340324 -	Cluster_2539	scaffold_7:2336389-2336394 -	Non-HP	N
				Cluster_2540	scaffold_7:2336657-2336981 -	Non-HP	N
				Cluster_2541	scaffold_7:2338009-2338060 -	Non-HP	N
ppt-miR1219a	Pp1s14_392V6.1	Pp1s14_392V6	scaffold_14:2418944-2425031 -	Cluster_4454	scaffold_14:2419574-2419603 -	Non-HP	N
				Cluster_4455	scaffold_14:2419895-2419920 -	Non-HP	N
ppt-miR166j	Pp1s15_11V6.1	Pp1s15_11V6	scaffold_15:91587-98465 +	Cluster_4507	scaffold_15:91916-91932 -	Non-HP	N
ppt-miR1211-3p	Pp1s20_21V6.2	Pp1s20_21V6	scaffold_20:98904-101764 +	Cluster_5760	scaffold_20:99400-99408 +	Non-HP	N
				Cluster_5761	scaffold_20:100384-100527 +	Non-HP	N
				Cluster_5762	scaffold_20:101208-101208 +	Non-HP	N
ppt-miR904a	Pp1s28_182V6.1	Pp1s28_182V6	scaffold_28:1005981-1014789 -	Cluster_7672	scaffold_28:1011559-1011560 -	Non-HP	N
				Cluster_7673	scaffold_28:1012417-1012464 -	Non-HP	N
ppt-miR534a	Pp1s28_209V6.1	Pp1s28_209V6	scaffold_28:1287855-1290057 +	Cluster_7690	scaffold_28:1289084-1289114 .	Non-HP	N
ppt-miR1039-5p	Pp1s29_44V6.2	Pp1s29_44V6	scaffold_29:309069-313749 +	N/A	N/A	N/A	N/A
ppt-miR1065	Pp1s33_92V6.1	Pp1s33_92V6	scaffold_33:537516-541093 +	N/A	N/A	N/A	N/A
ppt-miR534a	Pp1s36_325V6.1	Pp1s36_325V6	scaffold_36:2156476-2159302 +	Cluster_9356	scaffold_36:2158272-2158291 -	Non-HP	N
ppt-miR1222b	Pp1s38_312V6.1	Pp1s38_312V6	scaffold_38:1784879-1788350 -	Cluster_9713	scaffold_38:1786393-1786407 +	Non-HP	N
ppt-miR1029	Pp1s47_331V6.1	Pp1s47_331V6	scaffold_47:2112084-2114662 -	Cluster_11487	scaffold_47:2113206-2113220 -	Non-HP	N
ppt-miR477h	Pp1s48_150V6.1	Pp1s48_150V6	scaffold_48:1186304-1186936 -	N/A	N/A	N/A	N/A
ppt-miR156a	Pp1s50_125V6.1	Pp1s50_125V6	scaffold_50:1240965-1243861 -	Cluster_11946	scaffold_50:1241478-1241497 .	Non-HP	N
ppt-miR166j	Pp1s55_99V6.1	Pp1s55_99V6	scaffold_55:691604-698131 -	Cluster_12757	scaffold_55:697248-697263 +	Non-HP	N
ppt-miR1073-5p	Pp1s58_157V6.2	Pp1s58_157V6	scaffold_58:863228-867384 -	Cluster_13225	scaffold_58:863439-863479 -	Non-HP	N
				Cluster_13226	scaffold_58:866001-866014 -	Non-HP	N
				Cluster_13227	scaffold_58:866906-867101 -	Non-HP	N
ppt-miR1043-3p	Pp1s65_226V6.4	Pp1s65_226V6	scaffold_65:1265403-1269721 +	Cluster_14342	scaffold_65:1266572-1266583 +	Non-HP	N
				Cluster_14343	scaffold_65:1267242-1267252 +	Non-HP	N
				Cluster_14344	scaffold_65:1267726-1267755 +	Non-HP	N
				Cluster_14345	scaffold_65:1268011-1268239 -	Non-HP	N
ppt-miR166j	Pp1s67_107V6.1	Pp1s67_107V6	scaffold_67:789427-798290 +	Cluster_14579	scaffold_67:792643-792658 -	Non-HP	N
ppt-miR1221-5p	Pp1s71_181V6.1	Pp1s71_181V6	scaffold_71:948251-953508 +	N/A	N/A	N/A	N/A
ppt-miR408b	Pp1s77_277V6.1	Pp1s77_277V6	scaffold_77:1461739-1463836 -	Cluster_16180	scaffold_77:1462323-1462426 -	Non-HP	N
				Cluster_16181	scaffold_77:1462794-1463172 -	Non-HP	N
ppt-miR1215	Pp1s78_106V6.1	Pp1s78_106V6	scaffold_78:602133-603931 +	N/A	N/A	N/A	N/A
ppt-miR390a	Pp1s91_47V6.1	Pp1s91_47V6	scaffold_91:275335-275565 -	Cluster_17996[a]	scaffold_91:275187-275516 .	Non-HP	21
ppt-miR902b-5p	Pp1s97_176V6.1	Pp1s97_176V6	scaffold_97:937439-940190 +	Cluster_18879	scaffold_97:939064-939101 .	Non-HP	N
				Cluster_18880	scaffold_97:939895-939919 +	Non-HP	N
ppt-miR319a	Pp1s100_27V6.1	Pp1s100_27V6	scaffold_100:171378-172854 +	N/A	N/A	N/A	N/A
ppt-miR902b-5p	Pp1s105_81V6.1	Pp1s105_81V6	scaffold_105:546983-550759 +	Cluster_19713	scaffold_105:550098-550100 +	Non-HP	N
ppt-miR477h	Pp1s108_97V6.1	Pp1s108_97V6	scaffold_108:545168-547778 -	N/A	N/A	N/A	N/A
ppt-miR477h	Pp1s112_154V6.1	Pp1s112_154V6	scaffold_112:733541-736966 +	N/A	N/A	N/A	N/A
ppt-miR538a	Pp1s118_209V6.2	Pp1s118_209V6	scaffold_118:1022005-1026970 -	N/A	N/A	N/A	N/A
ppt-miR898a-5p	Pp1s123_164V6.1	Pp1s123_164V6	scaffold_123:1209167-1214007 +	N/A	N/A	N/A	N/A
ppt-miR171a	Pp1s130_63V6.1	Pp1s130_63V6	scaffold_130:292073-295688 -	N/A	N/A	N/A	N/A
ppt-miR1073-5p	Pp1s131_71V6.4	Pp1s131_71V6	scaffold_131:284420-287269 -	Cluster_22411	scaffold_131:284510-284715 -	Non-HP	N
				Cluster_22412	scaffold_131:285174-285205 -	Non-HP	N
				Cluster_22413	scaffold_131:285443-285461 -	Non-HP	N
				Cluster_22414	scaffold_131:285777-285803 -	Non-HP	N
				Cluster_22415	scaffold_131:286105-286141 -	Non-HP	N
				Cluster_22416	scaffold_131:286420-286538 -	Non-HP	N
ppt-miR1073-5p	Pp1s131_153V6.2	Pp1s131_153V6	scaffold_131:781375-784642 +	Cluster_22458	scaffold_131:781653-781667 +	Non-HP	N
				Cluster_22459	scaffold_131:782359-782605 +	Non-HP	N
				Cluster_22460	scaffold_131:782945-782984 +	Non-HP	N
				Cluster_22461	scaffold_131:783281-783292 +	Non-HP	N
				Cluster_22462	scaffold_131:783608-784131 +	Non-HP	N
				Cluster_22463	scaffold_131:784446-784494 +	Non-HP	N
ppt-miR1027a	Pp1s137_58V6.1	Pp1s137_58V6	scaffold_137:281893-282243 +	N/A	N/A	N/A	N/A
ppt-miR319a	Pp1s143_30V6.1	Pp1s143_30V6	scaffold_143:215487-218455 -	N/A	N/A	N/A	N/A
ppt-miR1028b-5p	Pp1s163_129V6.1	Pp1s163_129V6	scaffold_163:922382-923413 -	Cluster_25216	scaffold_163:923186-923203 +	Non-HP	N
ppt-miR904a	Pp1s173_134V6.1	Pp1s173_134V6	scaffold_173:751281-759185 -	Cluster_25958	scaffold_173:757033-757047 +	Non-HP	N
ppt-miR166j	Pp1s188_95V6.1	Pp1s188_95V6	scaffold_188:681910-687652 -	Cluster_27053	scaffold_188:682793-682807 +	Non-HP	N
				Cluster_27054	scaffold_188:687607-687621 +	Non-HP	N
ppt-miR156a	Pp1s194_53V6.1	Pp1s194_53V6	scaffold_194:250978-253882 -	Cluster_27414[b]	scaffold_194:251554-251573 .	Non-HP	20
ppt-miR156a	Pp1s194_57V6.1	Pp1s194_57V6	scaffold_194:256580-259970 -	Cluster_27415[b]	scaffold_194:257132-257151 +	Non-HP	20
ppt-miR171a	Pp1s205_1V6.1	Pp1s205_1V6	scaffold_205:3903-6350 -	Cluster_28197	scaffold_205:4948-4969 +	Non-HP	N
ppt-miR902b-5p	Pp1s262_39V6.1	Pp1s262_39V6	scaffold_262:281843-284517 +	N/A	N/A	N/A	N/A
ppt-miR536c	Pp1s267_8V6.1	Pp1s267_8V6	scaffold_267:70563-71817 -	N/A	N/A	N/A	N/A
ppt-miR538a	Pp1s267_56V6.1	Pp1s267_56V6	scaffold_267:347789-352289 +	N/A	N/A	N/A	N/A
ppt-miR1219a	Pp1s280_7V6.1	Pp1s280_7V6	scaffold_280:19445-26077 +	N/A	N/A	N/A	N/A
ppt-miR160a	Pp1s339_47V6.1	Pp1s339_47V6	scaffold_339:275729-278989 -	Cluster_35176	scaffold_339:277185-277317 .	Non-HP	N
ppt-miR902b-5p	Pp1s371_62V6.1	Pp1s371_62V6	scaffold_371:311162-313926 +	Cluster_36230	scaffold_371:312573-312748 +	Non-HP	N
ppt-miR319a	Pp1s391_54V6.1	Pp1s391_54V6	scaffold_391:248774-251777 -	Cluster_36859	scaffold_391:249981-249996 -	Non-HP	N

[a] *PpTAS3e locus*

[b] *small RNAs mapped to this locus may include processing variants of mature miR156 itself*

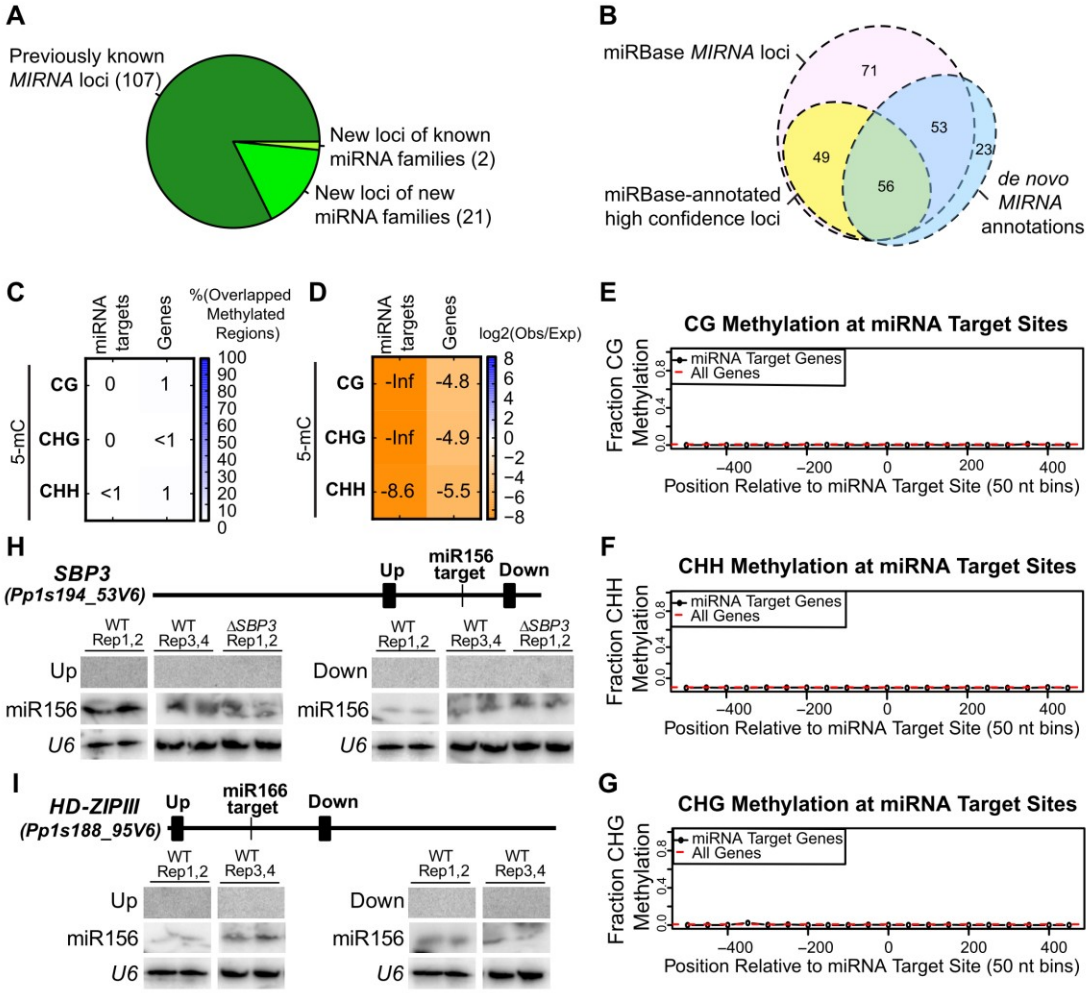

Figure 2.3: Refinement of *Physcomitrella MIRNA* annotations and functions.

(*A*) Classification of *MIRNA* loci confidently identified in this study. **(*B*)** Euler diagram comparing our *de novo MIRNA* annotations with previously annotated *Physcomitrella MIRNA* loci from miRBase 20. **(*C*)** Percent overlaps of regions of dense DNA methylation with miRNA targets and *Physcomitrella* genes. Calculated as in Fig. 2.2A. **(*D*)** Enrichment/depletion analysis of methylated regions of genome with miRNA targets and *Physcomitrella* genes. Calculated as in Fig. 2.2C. **(*E*)** Mean DNA methylation in the CG context surrounding 50 miRNA target genes (relative to their miRNA target site at position zero; black) and around arbitrarily chosen sites for all *Physcomitrella* genes (red). **(*F*)** As in *E* except for DNA methylation in the CHG context. **(*G*)** As in *E* except for DNA methylation in the CHH context. **(*H*)** Northern blot of small RNAs surrounding miR156 target site in *PpSBP3*. Filled rectangles on the gene schematic indicate probe positions. **(*I*)** As in *H* except for the miR166 target *Pp1s188_96V6*.

2.3.5 Discovery and mutagenesis of a *Physcomitrella minimal Dicer-Like (mDCL) gene*

Next, we revisited annotation of the *Physcomitrella DCL* gene family. Dicers emerged early in eukaryotes and independently diverged in plants and animals (Mukherjee et al. 2013). Plants contain four ancient clades of *DCL* genes, with members in each clade being sub-functionalized for different types of small RNAs (Margis et al. 2006). *Physcomitrella* has been reported to have no members of the *DCL2* clade, single members of both the *DCL3* and *DCL4* clades, and two members of the *DCL1* clade (*PpDCL1a* and *PpDCL1b*) (Axtell et al. 2007). Mutants in all four genes have been described (Cho et al. 2008; Khraiwesh et al. 2010; Arif et al. 2012). Upon review of *Physcomitrella DCL* genes, we made two surprising discoveries. First, we noticed that the DCL1b sequence used in 2007 (Axtell et al. 2007) to construct a phylogeny of DCL proteins was truncated and did not contain all of the domains typical of DCL proteins. The current transcript annotations at the *PpDCL1b* locus split the locus into three separate transcripts, despite the fact that RNA-seq data (Chen et al. 2012b) clearly indicate the presence of a single larger transcript (Fig. 2.4). Genome alignment of the full-length cDNA for *PpDCL1b* reported by Khraiwesh et al. (2010) revealed numerous discrepancies, including multiple nonsense changes, frameshifts, and unalignable regions (Fig. 2.4). Thus, we conclude that *PpDCL1b* is an expressed, spliced pseudogene incapable of producing a functional protein.

The second surprise was the identification of a *minimal DCL* (*mDCL*) gene encoding only the PAZ domain and two RNaseIII domains (Fig. 2.5A, B). (The *PpmDCL* locus is not located near the *PpDCL1b* pseudogene). PpmDCL is not a member of any of the canonical four clades of plant DCL proteins (Fig. 2.5A). The protozoan parasite *Giardia intestinalis* has been shown to produce a functional Dicer with a similarly minimal domain composition (MacRae et al. 2006). In addition, the ciliated protozoan *Tetrahymena thermophila* also produces a Dicer protein that lacks an N-terminal helicase domain (although it also lacks a PAZ domain); *Tetrahymena DCL1* is required for accumulation of scan RNAs that direct programmed DNA deletion events (Malone et al. 2005; Mochizuki and Gorovsky 2005). However, to the best of our knowledge, functional Dicer proteins lacking a helicase domain have not been previously described in any multicellular organisms. We hypothesized that *PpmDCL* contributes to production of endogenous *Physcomitrella* siRNAs. To test this hypothesis, we used homologous

recombination to obtain two independent *Ppmdcl* mutant lines (Fig. 2.6).

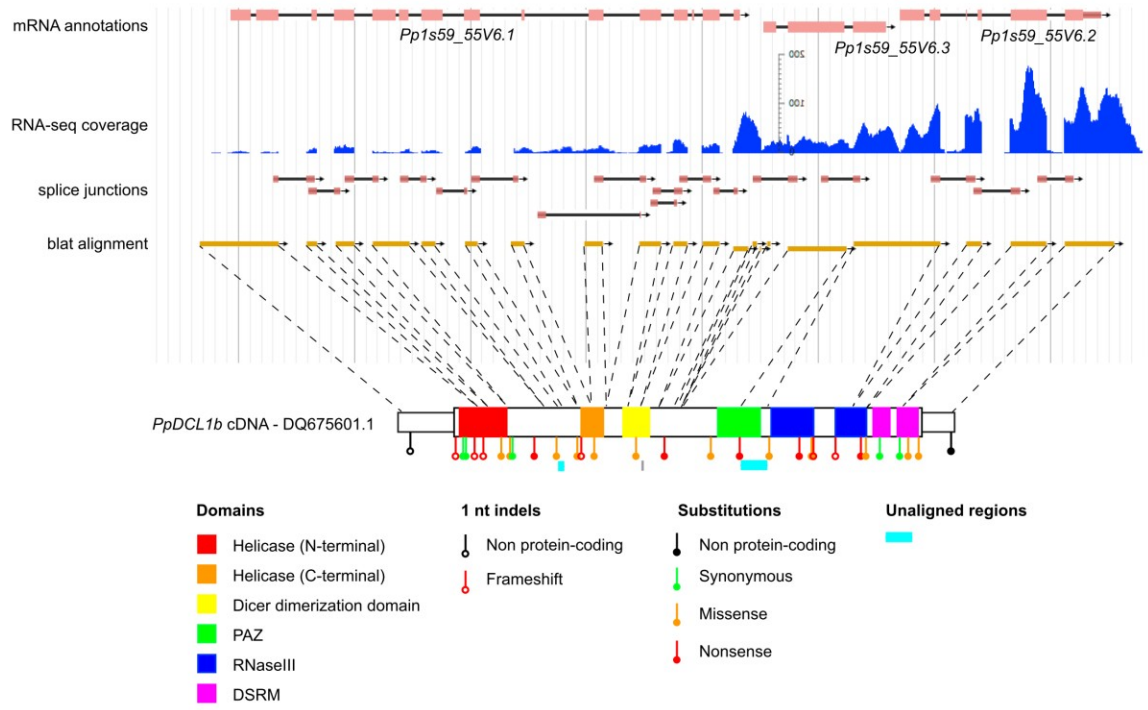

Figure 2.4: *PpDCL1b* is a pseudogene.
Genome-browser screen-shot showing transcript models (top), RNA-seq data (second), splice junctions (third), blat alignment data (fourth), and GenBank-submitted cDNA with highlighted differences (bottom). Examining the RNA-seq data with splice junctions shows that the full cDNA is expressed. Blat alignment reveals that half of the PAZ domain is missing (cyan). Also, there are various frameshifts and substitutions in the GenBank submitted *PpDCL1b* cDNA which led us to conclude that *PpDCL1b* as a pseudogene.

2.3.6 Heterochromatic siRNA mutants and *Ppmdcl* mutants have a similar accelerated growth phenotype

Phylogenetic analysis indicated that *Physcomitrella* has single genes encoding Pol I, Pol II, and Pol III largest sub-units, as well as three other largest sub-unit genes that are most closely related to the *Arabidopsis* Pol IV and Pol V largest subunits (Fig. 2.5C). The phylogeny did not resolve clear *Physcomitrella* Pol IV and Pol V largest sub-

unit homologs, but analysis of domain organizations clarified this issue (Fig. 2.5C, D). A high density of multiple GW/WG/GWG motifs in the C- terminal domain is a characteristic of the Pol V, but not the Pol IV largest sub-unit (Haag and Pikaard 2011), thus we named the sole *Physcomitrella* gene with dense GW/WG/GWG motifs *PpNRPE1a*. This annotation is consistent with a previous analysis based upon BLAST (Arif et al. 2013). *PpNRPE1b* was named based on the very high identity to the *PpNRPE1a* locus, and it is positioned around 600kb away. However, *PpNRPE1b* appears to lack a dense array of GW/WG/GWG motifs at the C-terminus (Fig. 2.5D). The *PpNRPE1b* gene has previously been suggested to be a Pol IV largest sub-unit homolog (Arif et al. 2013). We hypothesized (and later confirmed; see below), that the third gene encoded the largest sub-unit of a *Physcomitrella* Pol IV homolog, and named it *PpNRPD1*. Using homologous recombination, we obtained a single mutant line each for *Ppnrpe1a* and *Ppnrpd1* (Figs. 2.7, 2.8). Attempts to isolate a *Ppnrpe1b* mutant failed for unknown reasons.

As previously shown (Zong et al. 2009), *Physcomitrella* contains *RDR* genes in the α and γ clades (Fig. 2.5E). Only members of the α clade have been shown to affect small RNA biogenesis in plants. *Pprdr6* mutants have an accelerated juvenile to adult gametophyte transition phenotype, and lose the accumulation of *trans*-acting siRNAs (Talmor-Neiman et al. 2006a). We named the only other *Physcomitrella* member of the α clade *PpRDR2* (Fig. 2.5E). Phylogenetic analysis indicated that *PpRDR2* is closer to *Arabidopsis RDR1* than *RDR2*, and this gene has previously been suggested to be an *RDR1* homolog (Arif et al. 2013). However, our subsequent functional analysis demonstrated that the function of *PpRDR2* is homologous to that of *Arabidopsis RDR2*, justifying our naming decision. Two independent *Pprdr2* mutant lines were created using homologous recombination (Fig. 2.9).

Expression levels in protonemata, as estimated by RNA-seq data (Chen et al. 2012b), were moderate for all of the *Physcomitrella* genes we studied, with the exception of the *PpNRPE1b* for which we were unable to obtain a mutant (Fig. 2.5F). This suggests that the protonematal stage of growth is a valid time point to assay for effect of these mutations on small RNA populations.

We previously observed that *Ppdcl3* mutants display an accelerated juvenile to adult transition in gametophyte growth (Cho et al. 2008). In flowering plants, *DCL3*, *RDR2*, Pol IV, and Pol V are known to collaborate in the heterochromatic siRNA

pathway, so we hypothesized that *Pprdr2*, *Ppnrpd1*, and *Ppnrpe1a* mutants would also display the same phenotype. We found that this was indeed the case (Fig. 2.5G). We also found that *Ppmdcl* plants had an accelerated juvenile to adult transition (Fig. 2.5G), suggesting that *PpmDCL* also contributes to the heterochromatic siRNA pathway.

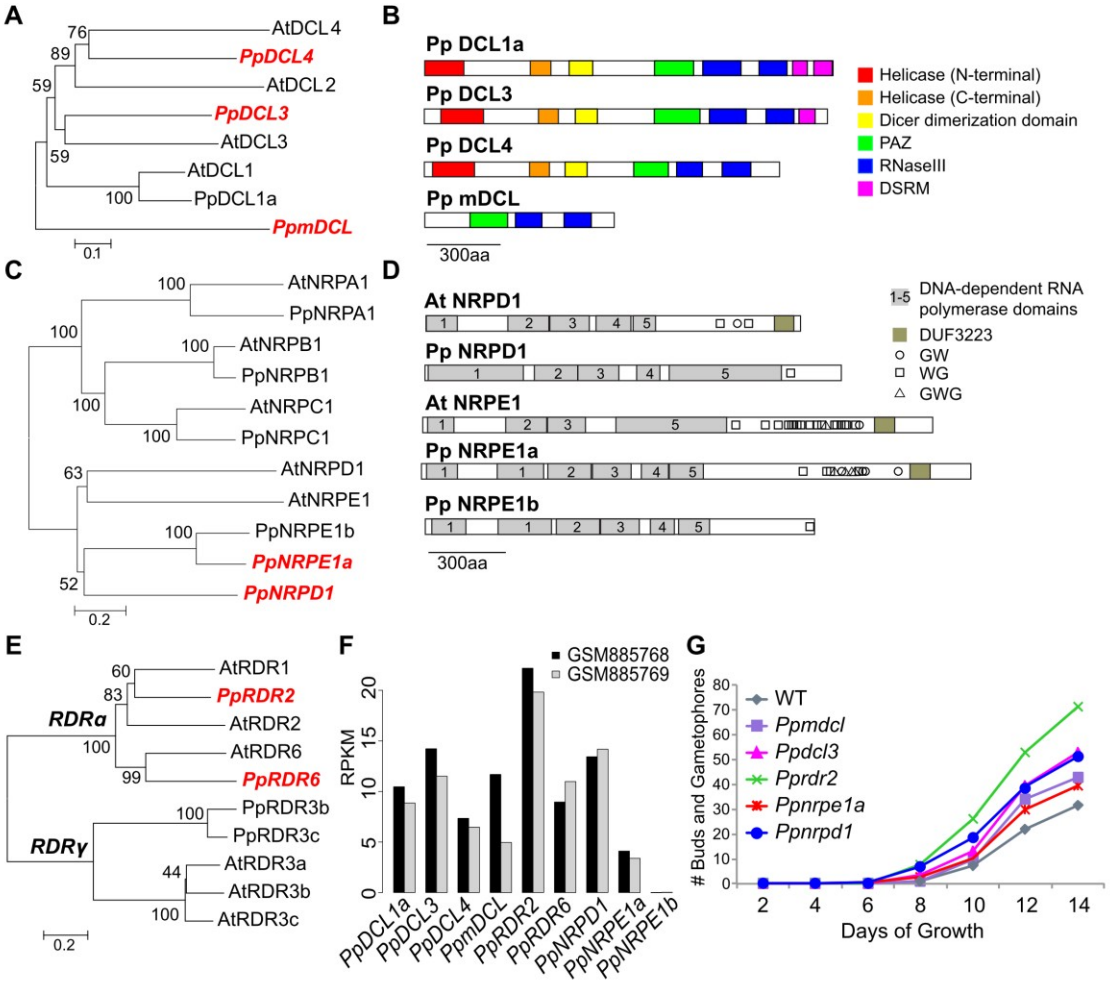

Figure 2.5: Relationships and phenotypes of *Physcomitrella DCL*, DNA-dependent RNA-polymerase, and *RDR* genes.

(*A*) Phylogenetic analysis of *Arabidopsis thaliana* (At) and *Physcomitrella* (Pp) DCL proteins. Entries in red italics indicate mutants used in this study. Numbers are bootstrap percentages from 1,000 replicates. Scale bar indicates substitutions per site. **(*B*)** Domain structures of *Physcomitrella* DCL proteins. **(*C*)** As in *A*, except for largest sub-units of DNA-dependent RNA polymerases. **(*D*)** As in *B* except for *Arabidopsis thaliana* (At) and *Physcomitrella* (Pp) largest sub- units of DNA-dependent RNA polymerases. DUF3223: Domain of unknown function, similar to a sequence found in a putative ribosomal RNA processing protein, DEFECTIVE CHLOROPLASTS AND LEAVES (DeCL) **(*E*)** As in *A* except for RNA-dependent RNA polymerases. **(*F*)** mRNA accumulation for the indicated genes in protonemata according to RNA-seq data (RPKM: Reads per kilobase per million) **(*G*)** Rates of buds and gametophore production. Seven-day old protonemal tissues were inoculated on BCD media, and total numbers of buds and gametophores were counted every two days.

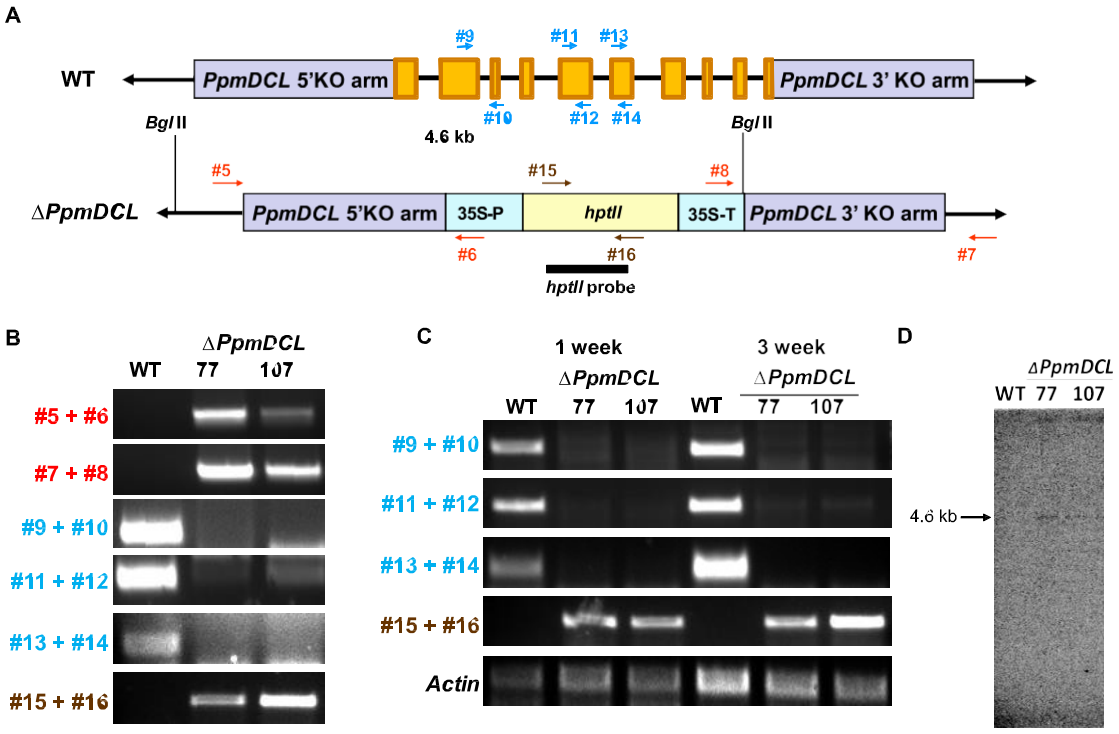

Figure 2.6: Targeted Knock Out of *PpmDCL*.
Regions upstream and downstream of the open reading frame of *PpmDCL* were cloned into the pUQ vector and transformed into *Physcomitrella*.
(A) The schematic of knock out by homologous recombination. The numbered arrows indicate approximate locations of primers (Table 2.8). 35S-P, CaMV 35S promoter; 35S-T, CaMV 35S Terminator **(B)** Genotyping of transformed plants by genomic DNA PCRs using the indicated primer sets. **(C)** Transcript analysis by RT-PCR using indicated primer sets. The position of primers used in (*B*) and (*C*) were marked in (*A*). *Actin* served as a control. **(D)** DNA blot analysis of *ΔPpmDCL*. The *BglII* digested genomic DNAs were blotted and hybridized with *hptII* probe. This result shows the vector was inserted into a single site in the genome The *BglII* recognition sites and *hptII* probe site are depicted in (*A*).

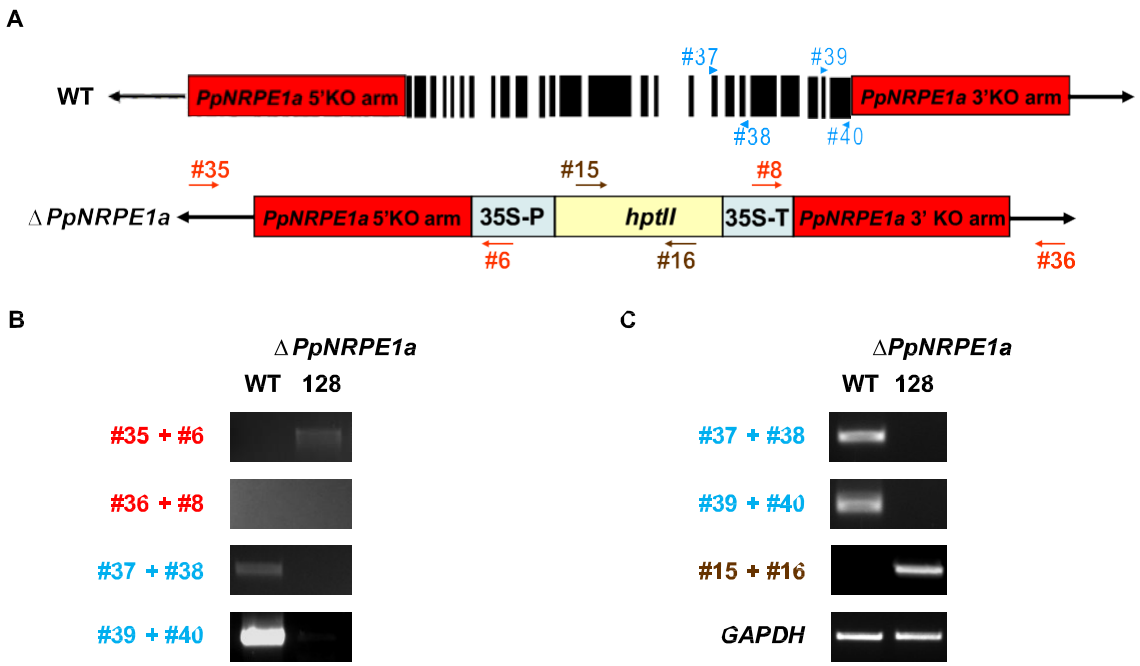

Figure 2.7: Targeted Knock Out of *PpNRPE1a*.
Regions upstream and downstream of the open reading frame of *PpNRPE1a* were cloned into the pUQ vector and transformed into *Physcomitrella*.
(A) The schematic of knock-out by homologous recombination. The numbered arrows indicate approximate locations of primers (Table 2.8). 35S-P, CaMV 35S promoter; 35S-T, CaMV 35S Terminator **(B)** Genotyping of transformed plants by genomic DNA PCRs using the indicated primer sets. **(C)** Transcript analysis by RT-PCR using indicated primer sets. The position of primers used in (*B*) and (*C*) were marked in (*A*). *GAPDH* served as a control.

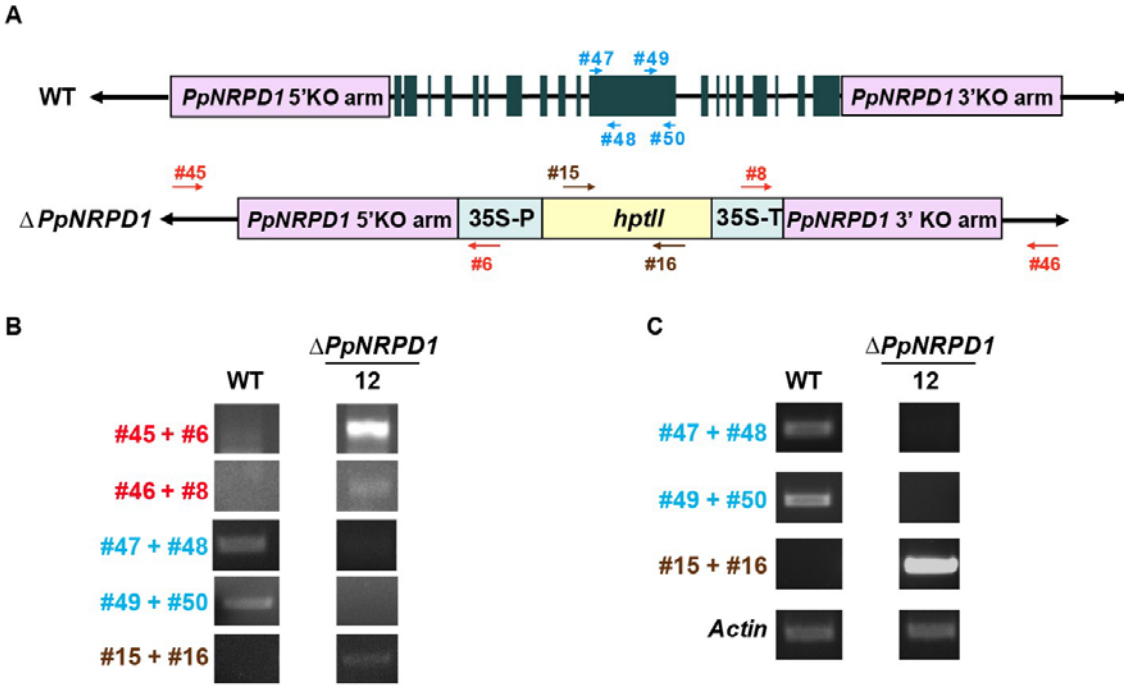

Figure 2.8: Targeted Knock Out of *PpNRPD1*.
Regions upstream and downstream of the open reading frame of *PpNRPD1* were cloned into the pUQ vector and transformed into *Physcomitrella*.
(A) The schematic of knock out by homologous recombination. The numbered arrows indicate approximate locations of primers (Table 2.8). 35S-P, CaMV 35S promoter; 35S-T, CaMV35S terminator **(B)** Genotyping of transformed plants by genomic DNA PCRs using the indicated primer sets. Both 5' and 3' recombination in the line 12 was confirmed. The two internal primers targeting internal regions which were supposed to be deleted, did not produce PCR products showing some disruption of the locus. **(C)** Transcript analysis by RT-PCR using indicated primer sets. *PpNRPD1* transcript was absent in line 12. The position of primers used in (*B*) and (*C*) were marked in (*A*). *Actin* served as a control.

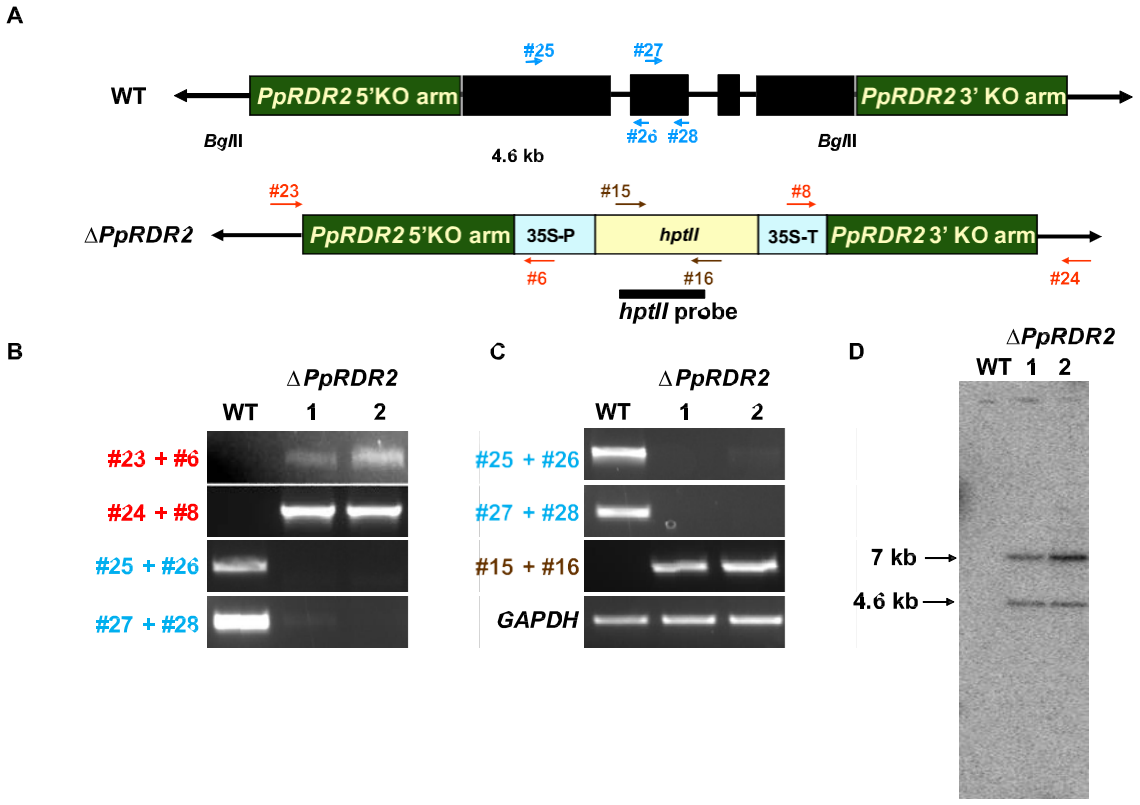

Figure 2.9: Targeted Knock Out of *PpRDR2*.
Regions upstream and downstream of the open reading frame of *PpRDR2* were cloned into the pUQ vector and transformed into *Physcomitrella*.
(A) The schematic of knock out by homologous recombination. The numbered arrows indicate approximate locations of primers (Table 2.8). 35S-P, CaMV 35S promoter; 35S-T, CaMV 35S Terminator **(B)** Genotyping of transformed plants by genomic DNA PCRs using the indicated primer sets. **(C)** Transcript analysis by RT-PCR using indicated primer sets. The position of primers used in (*B*) and (*C*) were marked in (*A*). *GAPDH* served as a control. **(D)** DNA blot analysis of *ΔPpRDR2*. The *BglII* digested genomic DNAs were blotted and hybridized with hptII probe. The 7kb band corresponds to the fragment size for the tandem repeat of the targeting cassette. This result shows the vector was inserted into a single site with a tandem repeat in the genome. The *BglII* recognition sites and *hptII* probe site are depicted in (A).

2.3.7 *PpmDCL* promotes accumulation of 23 nt RNAs from heterochromatic siRNA loci

We tested the hypothesis that the *Pprdr2*, *Ppnrpd1*, *Ppnrpe1a*, and *Ppmdcl* mutants affected 23-24 nt siRNA accumulation by constructing and sequencing multiple small RNA-seq libraries from ten-day old protonemata (Table 2.1). Also included were *Ppdcl4* and *Pprdr6* mutants (known to affect secondary siRNAs (Talmor-Neiman et al. 2006b; Arif et al. 2012), and *Ppdcl3* mutants (which our previous analysis implicated in 23-24 nt siRNA accumulation (Cho et al. 2008)). All mutants were represented by two to four biological replicates (Table 2.1).

None of the mutants examined had major effects on the abundance or size distributions of RNAs produced by *MIRNA* or 20-22 nt HP loci (Fig. 2.10A, B). Only *Pprdr6* mutants had major effects on RNA abundance from 20-22 nt siRNA loci (Fig. 2.10C). However, several of the mutants tested had major effects on small RNAs from 23-24 nt HP and 23-24 nt siRNA loci. *Pprdr2* and *Ppnrpd1* mutants had the most severe effects; siRNAs of all size were essentially absent from 23- 24 nt siRNA loci, and only some residual 24 nt RNAs remained from 23-24nt HP loci (Fig. 2.10D, E). Consistent with our earlier smaller-scale observations (Cho et al. 2008), 22-24 nt siRNAs were lost in *Ppdcl3* mutants, while the abundance of 21 nt RNAs remained similar to wild-type (Fig. 2.10D, E). *Ppnrpe1a* mutants had a slight increase in overall 23 and 24 nt siRNA abundance (Fig. 2.10D, E). *Ppmdcl* mutants had unique small RNA profile alterations at 23-24 nt HP and 23-24 nt siRNA loci; the levels of 23 nt siRNAs were decreased, but the levels of 21 nt and 24 nt siRNAs were increased (Fig. 2.10D, E). We conclude that *PpmDCL* affects the heterochromatic siRNA pathway by promoting the production of 23 nt siRNAs at the expense of 21 nt and 24 nt siRNAs.

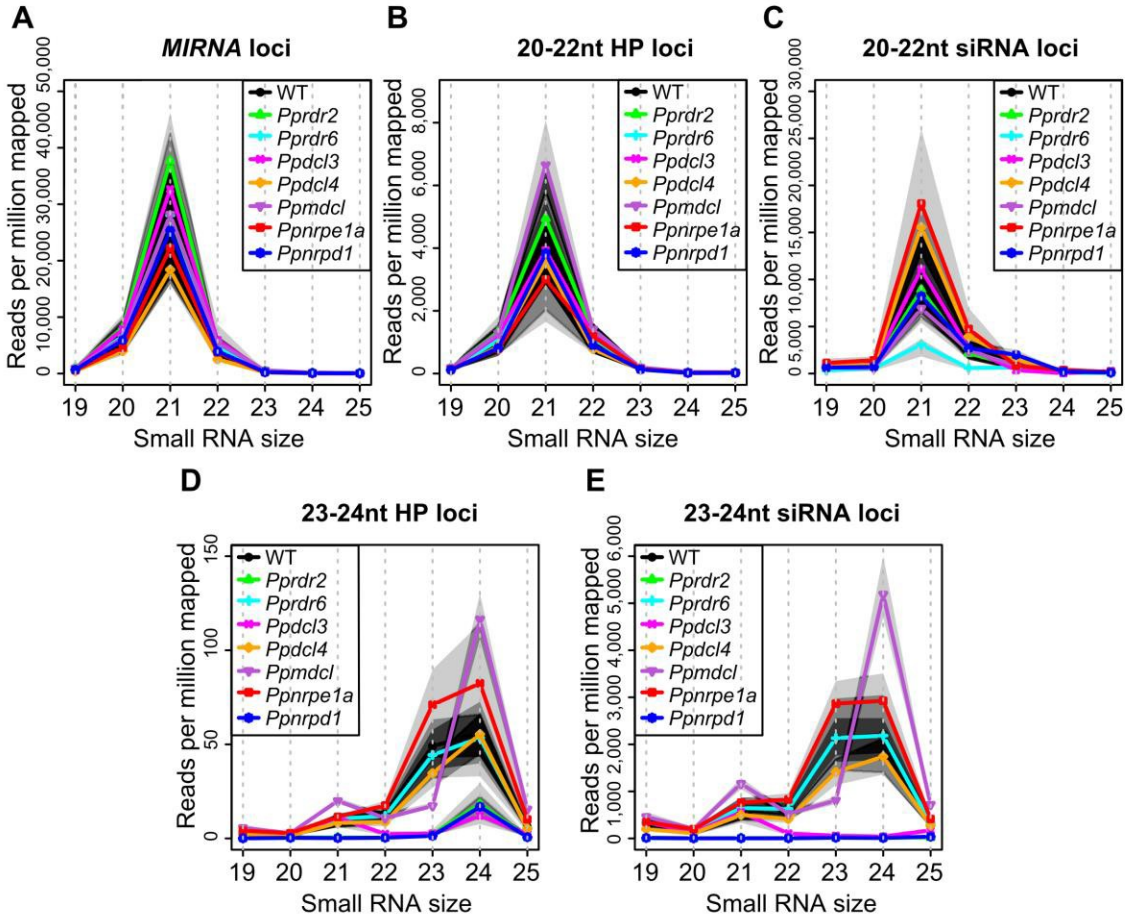

Figure 2.10: *PpmDCL* promotes 23 nt RNA accumulation and represses 24 nt RNA accumulation at heterochromatic siRNA loci.

(*A*) RNA abundance by size within *MIRNA* loci for the indicated genotypes. Solid lines indicate the median values for all biological replicate libraries. Shaded regions indicate the maximum and minimum values across all replicate libraries. **(*B*)** As in *A* except for 20-22nt HP loci. **(*C*)** As in *A* except for 20-22nt siRNA loci. **(*D*)** As in *A* except for 23-24nt HP loci. **(*E*)** As in *A* except for 23-24nt siRNA loci.

2.3.8 Differential expression analysis reveals distinct sub-groups of heterochromatic siRNA loci

A differential expression analysis was performed by tallying small RNA alignments of all sizes from each library within each of our annotated *DCL*- derived small RNA loci. A multi-dimensional scatterplot (MDS) of these data was prepared to illustrate

overall differences in small RNA accumulation between each of the samples (Fig. 2.11). Biological replicates for each genotype were generally consistent with each other, indicated by their tight groupings on the MDS (Fig. 2.11). *Pprdr6*, *Ppdcl4*, and *Ppmdcl* mutants clustered closely with wild- type, while *Pprdr2* and *Ppnrpd1* formed a second cluster of libraries distinct from wild-type and from all of the other mutants (Fig. 2.11). A third, looser cluster was formed by the *Ppdcl3* and *Ppnrpe1a* mutants.

Loci were considered differentially expressed (DE) in a particular mutant if they had at least a two-fold change compared to the wild-type, at a false discovery rate of less than 0.01 (Table 2.7, Table2.7_Differential_expression_analysis.xlsx). Very few DE loci were found in *Ppdcl4* and *Pprdr6* mutants, indicating that the secondary siRNA pathway does not make a major contribution to most of the endogenous small RNA loci under study (Fig. 2.12A). Large numbers of down-regulated 23-24 nt siRNA loci were found in *Pprdr2*, *Ppnrpd1*, and *Ppnrpe1a* mutants (Fig. 2.12A). A much smaller number of down-regulated 23-24 nt siRNA loci were apparent in *Ppdcl3* mutants (Fig. 2.12A). The modest numbers of up-regulated loci observed in *Ppdcl3*, *Ppnrpd1*, and *Pprdr2* mutants were mostly *MIRNA*s, 20-22 nt HP loci, and 20-22 nt siRNA loci (Fig. 2.12A). It is possible that the heterochromatic siRNAs dependent on *Ppdcl3*, *Ppnrpd1*, and *Pprdr2* compete with miRNA and 20-22 nt siRNA accumulation. Alternatively, because small RNA-seq quantification is proportional rather than absolute, small RNAs from these loci may appear up-regulated only because of the gross absence of 23-24 nt siRNAs in these samples. Interestingly, relatively small numbers of 23-24 nt siRNA loci were up-regulated in *Ppmdcl* and *Ppnrpe1a* mutants (Fig. 2.12A), suggesting the existence of distinct subsets of heterochromatic siRNA loci.

We next integrated DE calls for loci between the various mutants. The most common category was loci that were not DE in any of the five mutants (n = 7,735). We plotted and analyzed the next 11 most common patterns of loci that were DE in at least one of the five mutants (Fig. 2.12B). *Pprdr6* and *Ppdcl4* were not involved in any of the top 11 patterns, and so were omitted from the figure. Loci down-regulated in both *Pprdr2* and *Ppnrpd1* mutants were most numerous (Group 1, Fig. 2.12B). Group 2 loci were down-regulated in *Pprdr2*, *Ppnrpd1*, and *Ppnrpe1a* mutants, while groups 3 and 4 were comprised of loci down-regulated in *Pprdr2* only and *Ppnrpe1a* only, respectively. Except for groups 8, 10, and 11 the remainder of the most common patterns were loci that were down-regulated in one or more of the *Pprdr2*, *Ppnrpd1*, *Ppnrpde1a*, and *Ppdcl3* mutants.

Ppnrpe1a mutants had an interesting pattern of DE loci. 515 loci were down-regulated solely in this mutant while unchanged in the other four mutants (Group 4, Fig. 2.12B). 119 loci were up-regulated in *Ppnrpe1a* mutants but down-regulated in *Pprdr2* and *Ppnrpd1* mutants (Group 8, Fig. 2.12B). Another 110 loci were up- regulated in *Ppnrdpe1a* mutants and not DE in any other genotype (Group 10, Fig. 2.12B). Thus, there appear to be distinct subsets of heterochromatic siRNA loci uniquely affected by *PpNRPE1a*. *Ppmdcl* mutants were also interesting; the only group of loci in the top 11 patterns affected was group 11, which comprised 95 loci up-regulated in *Ppmdcl* mutants but unchanged in the other four mutants (Fig. 2.12B). *PpmDCL* therefore negatively regulates a unique subset of heterochromatic siRNA loci in terms of overall small RNA abundance. Co-occupancy analysis of the loci in these groups of interest relative to general genomic features did not strongly differentiate them (Fig. 2.13A, B). All of the groups were enriched for overlaps with 5-mC, repeats, and transposons, and depleted for overlaps with genes and gene-proximal regions (Fig. 2.13A, B). We did however note that group 4, comprising the 515 loci down-regulated only in *Ppnrpe1a* mutants, had slightly more association with genes than did the other groups (Fig. 2.13A, B).

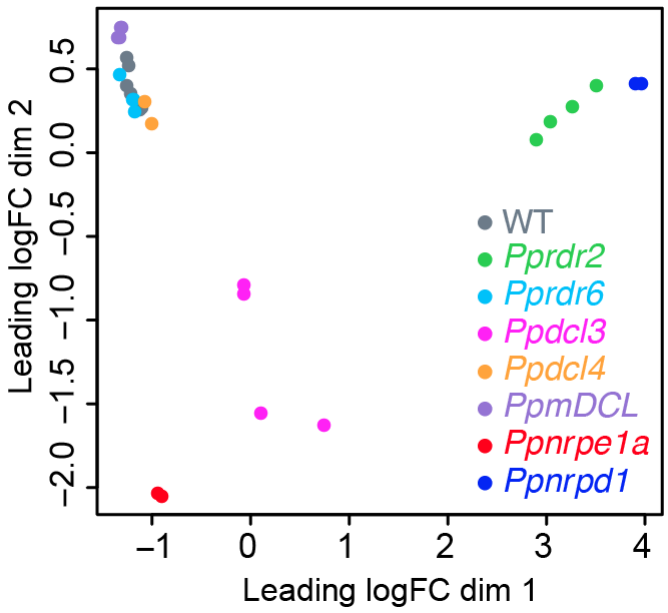

Figure 2.11: Biological replicates for each genotype are consistent with each other. Multidimensional scatter plot showing the overall relationship between each mutant and wild-type (WT) biological replicate small RNA-seq library. Leading fold-change (FC) is the

(root-mean-square) average of the largest absolute log2- fold-changes between each pair of samples.

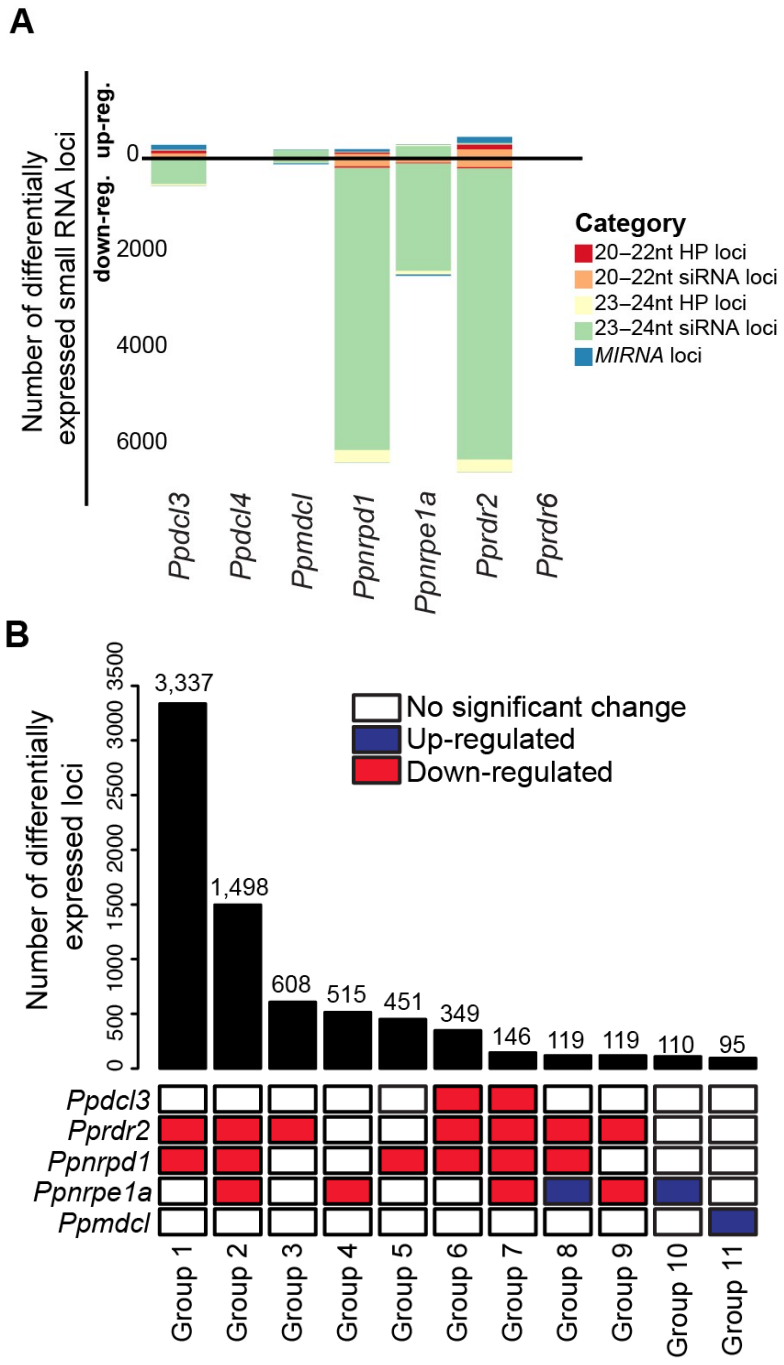

Figure 2.12: Differential expression analysis of *Physcomitrella* small RNAs in mutants.

(*A*) Numbers of differentially down-regulated or up-regulated small RNA loci in each of the indicated mutants compared to the wild-type. Differential expression for a locus is defined as with at least 2-fold-change with a FDR<0.01. **(*B*)** Numbers of differentially expressed loci (bar chart; upper panel) represented by different mutant combinations (heatmap; lower panel).

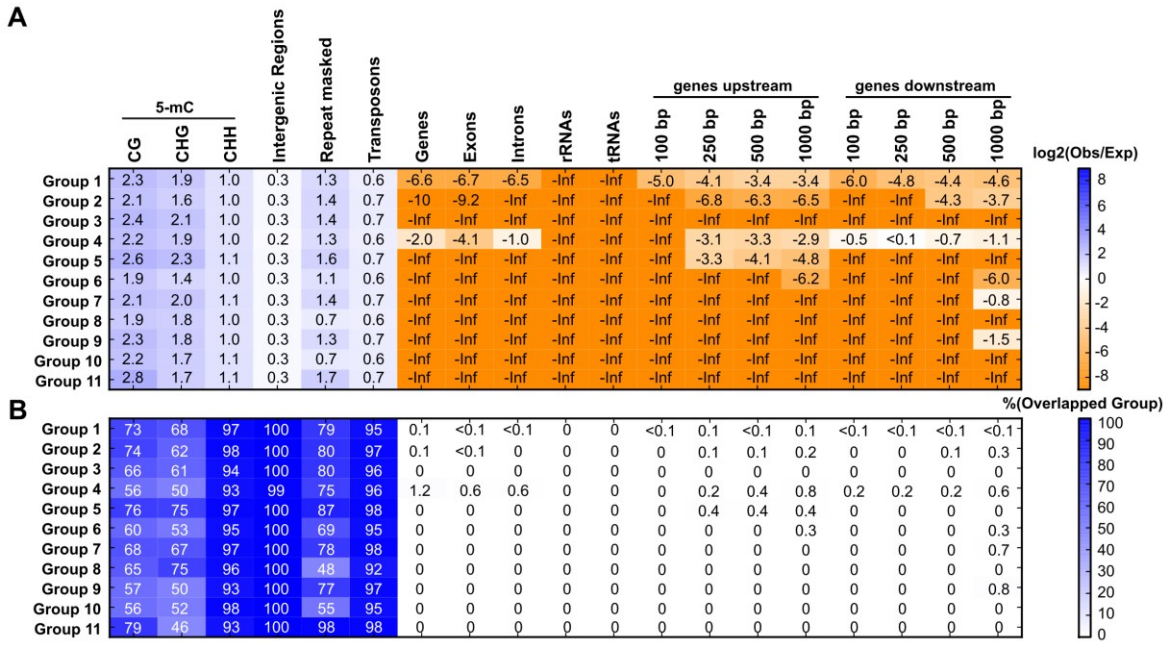

Figure 2.13: Genomic features of heterochromatic siRNA loci.

(*A*) Observed overlap/expected overlap ratios for different mutant groups defined in *C* relative to various genomic features. Calculated as in Fig. 2.2C. **(*B*)** Overlap percentage for mutant groups defined in *C* relative to various genomic features. Calculated as in Fig. 2.2A.

2.4 Discussion

We analyzed more than 10^8 mapped small RNA-seq reads from wild-type *Physcomitrella* and used these data to produce a comprehensive set of small RNA gene annotations. Setting aside degradation products that are unlikely to be part of the DCL/AGO regulatory system, most *Physcomitrella* small RNA genes produce 23-24 nt siRNAs. These loci are enriched for overlaps with repeats, regions of dense 5-mC, and nearly always avoid protein-coding genes. The *Physcomitrella* 23-24 nt siRNA loci are also strongly dependent upon *PpRDR2*, *PpNRPD1* (the presumed largest sub-unit of a Pol IV complex), and *PpDCL3* for small RNA production. Altogether, these data lead us to conclude that *Physcomitrella* utilizes a heterochromatic siRNA pathway fundamentally similar to that of flowering plants. Therefore, the potential absence of heterochromatic siRNAs in gymnosperms (Dolgosheina et al. 2008; Morin et al. 2008) and lycophytes (Banks et al. 2011) could reflect secondary loss of the pathway in those specific lineages. However, more recent data show that, for gymnosperms, endogenous 24 nt siRNAs can be found, albeit in tissue-specific patterns (Nystedt et al. 2013; Zhang et al. 2013b). Thus, we favor the hypothesis that heterochromatic siRNAs are a universal feature of land plant transcriptomes.

Physcomitrella heterochromatic siRNAs do however have some atypical features compared to those in flowering plants. Small RNA-seq samples from most wild- type tissues of most flowering plants are dominated by 24 nt heterochromatic siRNAs. In contrast, heterochromatic siRNAs are weakly expressed in *Physcomitrella* protonemata, where 21 nt miRNA expression dominates the small RNA profile in terms of abundance. Also in contrast to flowering plants, whose heterochromatic siRNAs are mostly 24 nts, *Physcomitrella* heterochromatic siRNA loci produce a mixture of 23 nt and 24 nt RNAs at nearly equal levels, with much lower levels of 21 nt and 22 nt RNAs. Our genetic analysis indicates that the *PpmDCL* gene is responsible specifically for 23 nt siRNA accumulation from these loci; in *Ppmdcl* mutants, 23 nt RNAs are strongly reduced while 24 nt RNAs are strongly increased at heterochromatic siRNA loci. At the same loci, loss of *PpDCL3* function eliminates 22 nt, 23 nt, and 24 nt RNA accumulation. We speculate

that PpmDCL is dependent upon PpDCL3 due to its lack of an N-terminal helicase domain. We also speculate that PpmDCL competes with PpDCL3 for small RNA precursors produced by Pol IV and PpRDR2. In *Ppmdcl* mutants, PpDCL3 processes the excess precursors to make mostly 24 nt RNAs. In *Ppdcl3* mutants, PpmDCL cannot function, leading to the loss of both the PpmDCL-dependent 23 nt and PpDCL3-dependent 24 nt RNAs. Further investigation is required to test this hypothesis. We find an apparent *mDCL* gene in the lycophyte *Selaginella moellendorffii*, but not in any angiosperm genomes, suggesting that the use of a minimal Dicer-Like gene for heterochromatic siRNA biogenesis may be a feature unique to basal land plants.

Our data also demonstrate that miRNA functions in *Physcomitrella* are not as unusual as previously proposed by Khraiwesh et al. (2010). Despite sequencing wild-type small RNAs to a depth of more than 10^8 mapped reads, we find no strong evidence for widespread secondary siRNA biogenesis from miRNA targets. We also find no evidence suggestive of miRNA-directed DNA methylation of miRNA target genes in wild-type plants. Khraiwesh et al. (2010) reported that the *PpDCL1b* gene promotes miRNA target-mRNA cleavage and prevents miRNA-directed DNA methylation of target genes. However, our analysis clearly shows that *PpDCL1b* is a pseudogene incapable of producing a DCL protein as the genome alignment of the full-length cDNA for *PpDCL1b* reported by Khraiwesh et al. (2010) revealed numerous discrepancies, including multiple nonsense changes, frameshifts, and unalignable regions. *PpDCL1a* and *PpDCL1b* are highly similar in nucleotide sequence. Therefore, we speculate that the phenotypes reported for *Ppdcl1b* mutants are attributable to disruption of *PpDCL1a*, either because the *PpDCL1b* pseudogene produces a *trans*-acting factor that regulates *PpDCL1a*, or because of inadvertent targeting of *PpDCL1a* during homologous recombination.

We believe our annotation of *Physcomitrella* small RNA-producing genes is comprehensive and useful, but we have identified two areas in which future improvements can clearly be made. The first area is the annotation of non- *MIRNA* hairpin (HP loci). In contrast to *MIRNA*s, for which there are a suite of community-accepted annotation criteria (Meyers et al. 2008), there are at present no commonly agreed upon criteria for annotating non-*MIRNA* hairpins that produce small RNAs. Our method, using ShortStack version 1.0.1, annotated 643 HP loci, 483 of which were dominated by 23 nt and 24 nt RNAs (Fig. 2.1J). However, small RNA abundance from these 483 loci was all but eliminated in the *Pprdr2* mutant (Fig. 2.10C). The simplest

explanation for *PpRDR2*-dependency is that the precursors of these RNAs were double-stranded RNAs, not hairpins. Thus, it is possible that many of the 23-24 nt HP loci in fact do not derive from hairpin RNAs, and merely have fortuitous overlap with inverted repeats. Future development of the ShortStack method will include a focus on improving annotations of non-*MIRNA* hairpin loci to reduce false positives. The second area of future improvement is the reference genome itself. While our work was being prepared for submission, a much-improved version of the *Physcomitrella* nuclear genome (version 3.0) was released for unrestricted use on Phytozome 10. The new assembly has closed many gaps and assembled the scaffolds into 27 pseudochromosomes. Future work will revisit annotations of *Physcomitrella* small RNA producing genes in light of the improved genome assembly. In particular, we expect that this will reveal chromosome-scale patterns in small RNA production that could not be seen using the previous draft assembly.

2.5 Methods

2.5.1 Small RNA-seq and reference annotation of wild-type *Physcomitrella* small RNA genes

Total RNA was extracted using the miRNeasy Mini kit (Qiagen) per the manufacturer's instructions from ten-day-old protonemata grown on cellophane-overlaid BCD medium (Ashton and Cove 1977) supplemented with 5 mM ammonium tartrate and cultured at 25°C, 16h day/8h night. Small RNA libraries were constructed using the TruSeq Small RNA kit (Illumina) per the manufacturer's instructions and sequenced on a HiSeq 2500 (Illumina) instrument using 50 nt single-end runs. Small RNA-seq data from the wild-type libraries (Table 2.1) were analyzed with ShortStack version 1.0.1 (Axtell 2013b). First, each library was adapter trimmed and aligned to the reference genome (version 1.6 nuclear assembly downloaded from Phytozome v9.0 (Goodstein et al. 2012) combined with the plastid and mitochondrial genomes) using settings --adapter TGGAATTC --align_only. The alignments were then merged using the samtools (Li et al. 2009) merge command and the resulting alignment file was used for *de novo* ShortStack analysis under default settings. Full results are listed in Table 2.2 (Table2.2_Pp_WT_ShortStack_smallRNA_loci_v1.6.xlsx), and the annotations are also hosted at http://plantsmallrnagenes.psu.edu/Physcomitrella_patens. Raw wild-type small RNA-seq data, processed data, and alignments were deposited to NCBI GEO (GSE44900), and are also available at http://plantsmallrnagenes.psu.edu/Physcomitrella_patens.

2.5.2 Co-occupancy Analyses

Regions of dense 5-mC occupancy in the CG, CHG, and CHH contexts were calculated in 50 nt intervals based on protonematal bisulfite-seq data from NCBI GEO accession GSM497264 (Zemach et al. 2010). A given 50 nt interval was considered densely methylated in a particular context if there were more than six reads of all Cs in that context. Distribution of coverage depths suggested that at least 6 reads were required in a bin in order to make any judgement (Fig. 2.14) Then, in qualifying bins (6 reads or more), the distribution of conversion events for all three contexts were examined and they were all found to be bimodal. 60% threshold for CG and CHG

contexts and 20% threshold for CHH context were chosen as these thresholds seemed to separate the bimodal distributions nicely (Fig. 2.15). Repeat-masked regions were obtained from the version 1.6 repeat-masked genome assembly via Phytozome. Transposon locations were derived from (Rensing et al. 2008) based on a gff3 file kindly provided by Stefan Rensing. Intergenic, genic, exonic, intronic, gene-upstream, and gene-downstream locations were calculated based on the version 1.6 transcriptome assembly gff3 file obtained from Phytozome. rRNA gene locations were based on regions of significant similarity (BLASTn e-value of <= 1E-10) to the rRNA consensus sequences. tRNA gene locations were based on genome-wide analysis with tRNAscan-SE version 1.3.1 under default parameters (Lowe and Eddy 1997). All of these annotations are browsable at plantsmallrnagenes.psu.edu/Physcomitrella_patens/jbrowse/ .

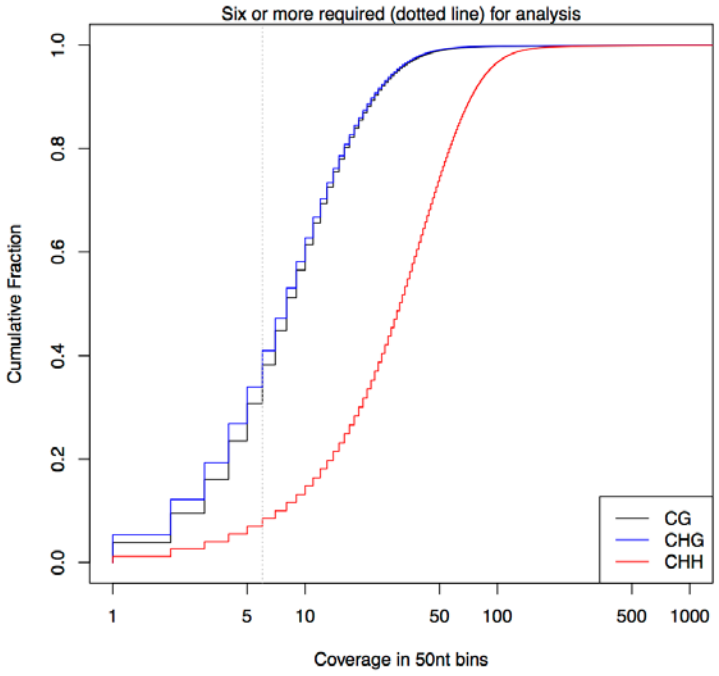

Figure 2.14: Coverage depths for CG, CHG and CHH methylation in 50 nt bins.

Distribution of coverage depths calculated using Zemach et al.'s pre-processed file of conversion calls in 50 nt bins. Black, blue and red lines indicates CG, CHG and CHH methylations, respectively. Vertical gray, dotted line indicates 6-read-coverage. (Zemach et al. 2010).

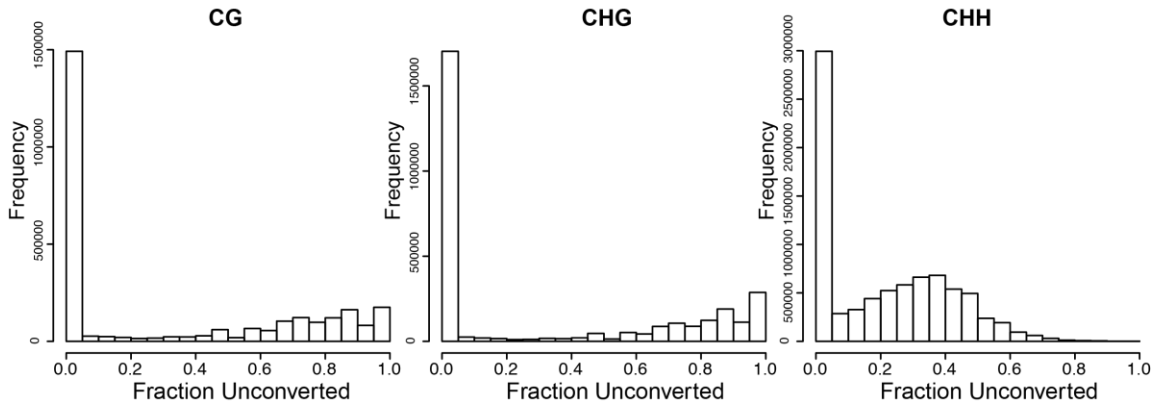

Figure 2.15: Distribution of conversion events for CG, CHG and CHH methylation. Fraction of methylated cytosines (unconverted) in qualifying 50 nt bins is shown for CG, CHG, and CHH methylation (data by Zemach et al. 2010).

The absolute numbers of overlapping loci and the total of non-redundant overlapping nucleotides for each pair-wise comparison of feature types were calculated. Enrichment/depletion was calculated based on the ratio of the observed to the expected number of overlapping nucleotides. The expected number of overlapping nucleotides for any pair-wise comparison is given by E = (x/g) * (y/g) * g, where E is the expected number of overlapping nucleotides under the null hypothesis of random location, x is the total number of non-redundant nucleotides for feature type 1, y is the total number of non-redundant nucleotides for feature type 2, and g is the total genome size.

2.5.3 miRNA and miRNA Target Analyses

MIRNA hairpin sequences and mature miRNA sequences identified by our *de novo* annotation effort were compared to the prior annotations in miRBase 20. The 23 loci that had not been previously annotated were registered with miRBase. We also used miRBase's 'confidence' community annotation system (Kozomara and Griffiths-Jones 2014) to up-vote and comment on the existing annotations. A set of 50 high-confidence miRNA targets was curated from Addo-Quaye et al. (2009) (Table 2.6) and compared to the 5-mC data (see above for processing methods) from Zemach et al. (2010).

2.5.4 Small RNA Blots

Small RNA blots were performed as described (Cho et al. 2012) with modification. Total RNAs from ten-day-old samples were extracted using Tri-Reagent (Sigma-Aldrich), and small RNAs were fractionated as described (Pall and Hamilton 2008). 20 µg of total RNAs were separated on 20% PAGE gel, trans-blotted onto the Hybond X (GE Healthcare) membrane, and cross-linked using 1-ethyl-3-(3-dimethylamonipropyl) carbodiimide (Pall and Hamilton 2008). Probes were independently labeled with T4 polynucleotide kinase (New England Biolabs) and mixed before hybridization. Hybridization, washing, and detection were performed as described (Cho et al. 2012). The probe sequences are listed in Table 2.8.

Table 2.8: Oligonucleotide sequences used in this study

#	Use	Sequences (5'→3')
1	*mDCL* KO vector construction, 5'KO arm (Forward)	CGCCTAGGATTTAAATAGATGTGTATTAATTACACCAACAC
2	*mDCL* KO vector construction, 5'KO arm (Reverse)	CGAAGCTTAATGATGATACAGGGGTGACAACGG
3	*mDCL* KO vector construction, 3'KO arm (Forward)	CGAGATCTCTTTATAGAAGGCATCTAGGAAGTC
4	*mDCL* KO vector construction, 3'KO arm (Reverse)	CGACGCGTATTTAAATTACAATAGATTAATTTTCATACAAA
5	*mDCL* KO identification of checking for 5' recombination (Forward)	ACCTCCAACGAGATGAGAACTACGC
6	KO identification checking for 5' recombination, 35S Promoter Internal (Reverse)	AGATAGCTGGGCAATGGAATCCGA
7	*mDCL* KO identification of checking for 3' recombination (Reverse)	AATATCCGCGCAGGTTAAGTTCCTAGC
8	KO identification checking for 3' recombination, 35S Terminator Internal (Forward)	GGGTTTCGCTCATGTGTTGAGCAT
9	*mDCL* KO genotyping (Internal Forward1)	GAAGCACTCGATGGTGGTGG
10	*mDCL* KO genotyping (Internal Reverse1)	ACTGCAGATGTTCGCCGTACGTAG
11	*mDCL* KO genotyping (Internal Forward2)	GGGCAAGTCATTGGACTCAAAACC
12	*mDCL* KO genotyping (Internal Reverse2)	CTTCCTCTTGGTACACCGCTC
13	*mDCL* KO genotyping (Internal Forward3)	GCATGTGAAGGGAACCACTCATAC
14	*mDCL* KO genotyping (Internal Reverse3)	CGTCTTGGTATTTAGCAGTTCAGC
15	Identification of *hptII* gene in mutants (Forward)	TGTTTATCGGCACTTTGCATCGGC
16	Identification of *hptII* gene in mutants (Reverse)	AGCTGCATCATCGAAATTGCCGTC
17	*Actin* (Forward)	ATCTGGAATGGTCAAGGCCGGTTT
18	*Actin* (Reverse)	TCATCTTCTCCCTGTTCGCCTTCG
19	*RDR2* KO vector construction, 5'KO arm (Forward)	CAAGCTTGGGACAAGGGAAGAGGTTCTCAAA
20	*RDR2* KO vector construction, 5'KO arm (Reverse)	AACTCGAGACACCCACCACATTCCTCAGTCAT
21	*RDR2* KO vector construction, 3'KO arm (Forward)	CCAGATCTACTGCTACACAGCGAGGATTTCTG
22	*RDR2* KO vector construction, 3'KO arm (Reverse)	CCACGCGTTCAAGCAATGGGATAGGAGGCCAA
23	*RDR2* KO identification of checking for 5' recombination (Forward)	GAGAGATGCAGTTTCGCAGCAGTA
24	*RDR2* KO identification of checking for 3' recombination (Reverse)	TGGCTATATGTATGGTAATAAGGGACC
25	*RDR2* KO genotyping (Internal Forward1)	ACAATGATCAGGGCATGGATGGGA
26	*RDR2* KO genotyping (Internal Reverse1)	ACCCGCTGCGAGCATATCTATCAA
27	*RDR2* KO genotyping (Internal Forward2)	TGATAGATATGCTCGCAGCGGGTT
28	*RDR2* KO genotyping (Internal Reverse2)	AAACCAAGCAGTCAACCATGTGCC
29	*GAPDH* F	CCTCTTGCAAAGGTGATCAACGAC
30	*GAPDH* R	ACCACACGGTTGCTGTAACCCCAC
31	*NRPE1a* KO vector construction, 5'KO arm (Forward)	GGAAGCTTCCGGAAGAATTTGGCTAATCCGCA
32	*NRPE1a* KO vector construction, 5'KO arm (Reverse)	GGCTCGAGCGAGCGATAAGCATTAAAGCAACG
33	*NRPE1a* KO vector construction, 3'KO arm (Forward)	GGAGATCTTGCGTGAAACCTATTTGAGATGGA
34	*NRPE1a* KO vector construction, 3'KO arm (Reverse)	GGACGCGTGCCACAAGTCCAAGACATTAGAACT
35	*NRPE1a* KO identification of checking for 5' recombination (Forward)	TCTGTTGTTGCTGATGCAGGTCAG

36	*NRPE1a* KO identification of checking for 3' recombination (Reverse)	GTGTCTTCAAGCTAGACATATTTAGAAATGG
37	*NRPE1a* KO genotyping (Internal Forward1)	GGGACAAATTTTCTTTTGTGTCAGTTA
38	*NRPE1a* KO genotyping (Internal Reverse1)	AATACCAAACCAAGTCTCTGTGAG
39	*NRPE1a* KO genotyping (Internal Forward2)	AACTTGGTGGCAGGCTTTCTGACG
40	*NRPE1a* KO genotyping (Internal Reverse2)	TCAAGATCCTCATGATCAATAGGC
41	*NRPD1* KO vector construction, 5'KO arm (Forward)	GGCCTAGGTGTCATTTAGGATAGTGCGGG
42	*NRPD1* KO vector construction, 5'KO arm (Reverse)	GGCTCGAGCCTTCAAGCACAAAAACAAAG
43	*NRPD1* KO vector construction, 3'KO arm (Forward)	GGAGATCTGATTGGTTACCTTCGCAATGCCAT
44	*NRPD1* KO vector construction, 3'KO arm (Reverse)	GGACGCGTGCAATTTGATGGCTCCTTGT
45	*NRPD1* KO identification of checking for 5' recombination (Forward)	TGTGAAGGCAGTTAATGGTGA
46	*NRPD1* KO identification of checking for 3' recombination (Reverse)	GGAGATGGATACTATGATTGATGG
47	*NRPD1* KO genotyping (Internal Forward1)	AGATACATGAAGGGGCATATTTTAGC
48	*NRPD1* KO genotyping (Internal Reverse1)	GTCGTTCAATATTAAAAGCCGTGAC
49	*NRPD1* KO genotyping (Internal Forward2)	TTGGATAAGGTTGCTGTCGATAGG
50	*NRPD1* KO genotyping (Internal Reverse2)	ACCATACCGTGATGATAAAGTGTG
51	Small RNA gel blot of ppt-miR156 (probe)	GTGCTCACTCTCTTCTGTCA
52	Small RNA gel blot U6 (probe)	TTGTGCGTGTCATCCTTGCGCA
53	Small RNA gel blot *SBP3* up target region F (probe)	GTATCCCTGCCCTTCAACTTCAGGTTGGTTTTATGTTTGTC GAAACAGCT
54	Small RNA gel blot *SBP3* up target region R (probe)	AGCTGTTTCGACAAACATAAAACCAACCTGAAGTTGAAGG GCAGGGATAC
55	Small RNA gel blot *SBP3* down target region F (probe)	TGAGTCTGTGGGGCTGAATTGTGGGCTAGCTGCGACTGG TTACGGGGCTC
56	Small RNA gel blot *SBP3* down target region R (probe)	GAGCCCCGTAACCAGTCGCAGCTAGCCCACAATTCAGCC CCACAGACTCA
57	Small RNA gel blot *HD-ZIPIII* up target site F (probe)	CAACGCAAGGAAGCAACAAGGCTGGTCAGTGTTAATGCAA AGCTGACAGC
58	Small RNA gel blot *HD-ZIPIII* up target site R (probe)	GCTGTCAGCTTTGCATTAACACTGACCAGCCTTGTTGCTT CCTTGCGTTG
59	Small RNA gel blot *HD-ZIPIII* down target site F (probe)	GATTACTGTACTTTGAGATACACTACAATTTTGGAGGATGG AAACCTGGT
60	Small RNA gel blot *HD-ZIPIII* down target site R (probe)	ACCAGGTTTCCATCCTCCAAAATTGTAGTGTATCTCAAAGT ACAGTAATC

2.5.5 Phylogenetic Analysis

Sequence alignments were generated using ClustalW with default parameters (Thompson et al. 1994) and used for phylogenetic analysis with MEGA4 software (Tamura et al. 2007) using the neighbor-joining method. Phylogenetic distances were evaluated using the Poisson correction model (Nei and Kumar 2000). Positions with alignment gaps were eliminated for pairwise alignments. Topology reliability was checked using bootstrap analysis with 1,000 replicates. Accession numbers: AtDCL1 (At1g01040), AtDCL2 (At3g03300), AtDCL3 (At3g43920), AtDCL4 (At5g20320), PpDCL1a (ABV31244.1), PpDCL1b (DQ675601), PpDCL3 (ABV31245), PpDCL4 (EF670438), PpmDCL (KF179046), AtNRPA1 (At3G57660), AtNRPB1 (At4G35800.1), AtNRPC1(At5G60040.2), AtNRPD1 (At1G63020), AtNRPE1 (At2G40030), PpNRPA1 (Pp1s338_40V6), PpNRPB1 (Pp1s460_26V6.1), PpNRPC1 (Pp1s26_192V6.1), PpNRPE1a (KF908782), PpNRPE1b (KF908783), PpNRPD1 (Pp1s193_6V6.1), AtRDR1(At1G14790), AtRDR2 (At4G11130), AtRDR3a (At2G19910), AtRDR3b

(At2G19920), AtRDR3c (At2G19930), AtRDR6 (At3G49500), PpRDR2 (Pp1s178_112V6.1), PpRDR6 (ABF82438.1), PpRDR3b (Pp1s218_13V6.1), PpRDR3c (Pp1s386_30V6.1).

2.5.6 Construction of Vectors

For the construction of knock-out vectors, two approximately one kb regions 5' and 3' from the open reading frame of desired genes were amplified using specific primer sets (Table 2.8) and inserted into the pUQ vector (Cho et al. 2008), as previously described (Cho et al. 2012).

2.5.7 DNA Blot Analysis

Genomic DNAs were extracted using a Phytopure DNA Extraction kit (GE Healthcare. For DNA blot analysis, the *Bgl*II digested genomic DNAs of *Ppmdcl* and *Pprdr2* were blotted onto a Hybond NX nylon membrane (GE Healthcare) and hybridized following a standard protocol (Sambrook and Russell 2001). For a probe, PCR amplified *hptII* fragment was radio-labelled with [α-32P] dCTP using an NEblot Kit (New England Biolabs) per the manufacturer's instructions.

2.5.8 Real-Time PCR

Total RNAs were extracted from ten-day-old protonemata using the miRNeasy Mini kit. RT-PCR reactions were performed as previously described (Cho et al. 2012). Primer sequences are listed in Table 2.8.

2.5.9 Differential Expression Analysis

Small RNA-seq samples from the various mutants (Table 2.1) were trimmed (--adapter TGGAATTC), aligned to the version 1.6 genome (including plastid and mitochondrial genomes), and analyzed using ShortStack version 1.1.0 in 'count' mode using the wild-type *de novo* small RNA gene annotations as the --count file. Counts from separate sequencing runs of the same libraries were combined (Table 2.1) and used for differential expression analysis with the R package edgeR (Robinson et al. 2010). Libraries were normalized with the "calcNormFactors" function, and analyzed with the "exactTest" function analysis for each mutant in comparison with wild-type. Differentially expressed genes at a 1% FDR were retrieved using the "decideTestsDGE" function, and further filtered to retain only those with two-fold or greater deviation from wild- type.

Table 2.7 (Table2.7_Differential_expression_analysis.xlsx) contains the full details and results of these analyses.

2.5.10 Data Access

cDNA sequences for *PpmDCL*, *PpNRPE1a*, and *PpNRPE1b* have been deposited to NCBI under accessions KF179046, KF908782, and KF908783, respectively. Small RNA-seq data has been deposited to NCBI GEO under accessions GSE44900 (wild-type) and GSE51419 (mutants). The full set of *Physcomitrella* small RNA gene annotations and associated data are also available and browsable at http://plantsmallrnagenes.psu.edu/Physcomitrella_patens.

Chapter 3

Summary and Prospects

3.1 Summary

3.1.1 Available resources for annotating small RNA genes in plants

In plants, a particularly wide variety of small regulatory RNAs is produced by DCLs and utilized as sequence-specific guides by AGO proteins. The known DCL/AGO-associated small RNAs are 20-24 nts in length. Several major types have been described, including miRNAs, secondary short interfering RNAs (secondary siRNAs), and heterochromatic siRNAs. In Chapter 1, I introduced the critical components involved in small RNA biogenesis and discussed the challenges and limitations of small RNA gene annotation based on our current knowledge of small RNAs. Firstly, there are complications in miRNA annotations as we observed inconsistency between empirical data and miRBase in terms of annotation of the mature miRNA. It also appears that *MIRNA* hairpins tend to produce more than a single product which might change the current definition of the mature, guide miRNA, especially if miRNA variants other than the most abundant miRNA (Jeong et al. 2013) are found to be functional as well (Fig. 1.5). Reliable annotation of the functional miRNAs becomes more complicated since the newly emerging data provide a compelling line of evidence which creates room for potentially critical factors, such as the presence of AGO-loaded miRNA*s, importance of 3'-end modifications of mature miRNAs after dicing, and the existence of miRNA superfamilies (Manavella et al. 2012; Zhai et al. 2013, 2011; Shivaprasad et al. 2012; Li et al. 2012).

Secondly, although a lot of effort has been made to annotate *MIRNA* loci, recent data suggests that *MIRNA* loci account for only a very small proportion of the total genome that actively produces 20-24 nt small RNAs (Fig. 1.6B, left). The fact that the majority of expressed plant small RNAs are not miRNAs highlights the 'annotation gap' between the current knowledge of small RNA expression and annotations of small RNAs. Third, there are no community-accepted standards for annotating hpRNA loci, even when they are potentially abundant and functional. Fourth, annotation of heterochromatic siRNAs, which are responsible for repressing heterochromatin, is

subject to variation as a new line of evidence suggests that small RNAs other than 24 nt in length might well serve as heterochromatic siRNAs (Stroud et al. 2013; Nobuta et al. 2008).

Huge amounts of small RNA alignment data have been produced using small RNA-seq, and progress at using these alignments to create small RNA gene annotations has been made (Table 1.2). We believe that our newly developing web server (plantsmallrnagenes.psu.edu) utilizing ShortStack is quite useful not only for providing the small RNA-seq alignments but also for creating reliable comprehensive reference annotations considering the complications indicated above.

3.1.2 Comprehensive annotation of *Physcomitrella* small RNA loci

The work in Chapter 2 provides evidence that *Physcomitrella* expresses heterochromatic siRNAs that have a largely similar biogenesis pathway as in flowering plants. The reproducibility of our genetic analysis, as evident by the consistency between biological replicates, looks promising in providing reliable conclusions about small RNA gene annotation (Fig. 2.11). Our differential expression analysis utilizing extensive small RNA-seq from wild-type and RNAi-defective mutants led us to conclude that *Physcomitrella* heterochromatic siRNA loci are dependent on *PpRDR2*, *PpNRPD1*, and *PpDCL3* for small RNA accumulation (Fig. 2.12A, B). Unlike angiosperm heterochromatic siRNA loci, which predominantly produce 24 nt siRNAs, heterochromatic siRNA loci in *Physcomitrella* produce mixtures of 23-24 nt siRNAs. However, these 23-24 nt siRNA loci are enriched for overlaps with repeats and dense 5-mC regions and avoid protein-coding genes (Fig. 2.13), similar to what has been observed for heterochromatic siRNAs in flowering plants. Therefore, the most parsimonious scenario is that, as for miRNAs, the heterochromatic siRNA pathway is an ancestral trait that was present in the last common ancestor of bryophytes and all other subsequently diverged lineages of plants. The major difference is the use of the novel *mDCL* gene to produce 23 nt heterochromatic siRNAs in *Physcomitrella* (Fig. 2.10D, E).

We identified 130 *MIRNA* loci where 23 of these loci are novel compared to annotations present in miRBase release 20 (Fig. 2.3A, Table 2.5). We also conclude that *Physcomitrella* miRNA functions are not as unusual as has previously been suggested (Khraiwesh et al. 2010); we find that *PpDCL1b* is a pseudogene, and we find no evidence that *Physcomitrella* miRNAs spawn abundant secondary siRNAs from protein-

coding target mRNAs, nor direct 5-mC deposition at target chromatin (Fig. 2.3C-I). The results of the work in Chapter 2 further contribute to our expanding knowledge of small RNA producing loci in the deep-branching moss *Physcomitrella patens*. Finally, our publically available and browsable annotations of *Physcomitrella* small RNA genes provide a useful resource for further study of all classes of small RNAs in this model organism.

3.2 Prospects

3.2.1 Availability of the reference genome

A key goal in genomics is the complete annotation of the expressed regions of the genome. In plants, substantial portions of the genome make regulatory small RNAs produced by DCL proteins and utilized by AGO proteins. Currently, these include miRNAs and various types of endogenous siRNAs. Small RNA-seq, enabled by cheap and fast DNA sequencing, has produced an enormous volume of data on plant miRNA and siRNA expression in recent years. Despite growing efforts to improve small RNA gene annotations, there are no community-accepted standards for discerning different types of plant small RNA loci. Instead, small RNA genes are mainly characterized by custom-built genome analyses. Some areas should be revisited in order to improve our abilities to better characterize small RNA genes in plants.

Probably the most critical potential limitation in identifying genes using small RNA-seq data, is the availability of the reference genome annotation. Recently, a much-improved version of the *Physcomitrella* nuclear genome (version 3.0) has been released on Phytozome 10. The new assembly provides 27 pseudochromosomes instead of scaffolds with many gaps closed. However, we performed our small RNA-seq analysis using the previous v1.6 draft assembly because of the embargo of v3.0 on publications using the whole genome-wide analysis. Once the “reserved analyses” rights are waived, we will revisit annotations of *Physcomitrella* small RNA producing genes in light of the improved genome assembly. Using the new genome assembly will allow us to elucidate the chromosome-scale patterns of small RNA production that could not be seen using the previous draft assembly.

3.2.2 Improving mapping strategies

Another area of future improvement includes “fine-tuning” of the mapping strategies. In our hands, we realized that most siRNA loci are identified by only one abundant siRNA read. This is somewhat unexpected compared to the known features of siRNAs observed in other species, where borders of the repeat-associated siRNA loci are defined by dispersed mixture of diverse siRNA reads. The fact that our identified siRNA loci appear as ‘unique’ might be an artifact due to our mappings settings. The current ShortStack version selects one random locus for multi-mapper reads during alignment to the reference genome. First, the reference genome used in this study is composed of scaffolds, so the assembly includes a lot of gaps. Second, it might be important to consider additional criteria when selecting one of the repetitive loci for the multi-mapped reads. Our lab is currently developing a simulator to test new strategies in mapping settings in order to minimize false positives. For instance, one alternative strategy to develop a more reliable assignment for a given multi-mapped read could be based on the assumption that a multi-mapped read most likely derives from the transcriptionally active regions of the genome. We expect that our improved assignments of multi-mappers will result in more reliable annotations of small RNA genes, particularly the ones that are derived from the repetitive regions.

3.2.3 Exploiting small RNA-seq data

The characterization of small RNA repertoires in plants would lead us discover a wide range of possible novel gene regulatory mechanisms. Enormous amounts of small RNA-seq data are now available for many plant species, and the barriers to obtaining even more data grow lower and lower. In this study, we used an extensive small RNA-seq dataset to create a reference annotation of wild-type *Physcomitrella* small RNA genes. Genetic analysis using RdDM-defective mutants revealed the differentially expressed small RNAs. However, any genomic region normally repressed in wild-type that becomes de-repressed (i.e. transcriptionally active) in the mutant cannot be detected in this analysis because the reference small RNA genes were defined based on the wild-type libraries only. In order to identify regions of genome that are normally silenced and are dependent on any of the RdDM components, we need to repeat differential expression analysis using the newly created reference set of small RNA gene

annotations obtained by the mutant library. Also, investigating the loci that are outside of the DCL size range might be useful in discovering new types of small RNAs.

3.2.4 Investigating spatial and temporal expression of small RNAs

Our current knowledge of regulatory small RNA pathways is mainly obtained from the model organisms. However, small RNA populations have become more available for many plant species with the advent of next-generation sequencing technologies. With the aid of improved annotations, novel classes of small RNAs are likely to be discovered, especially using non-model plant organisms. One example includes the miR482/2118 superfamily of miRNAs which are found to function to target *NB-LRR* innate immune receptor mRNAs in potato, tomato and tobacco (Zhai et al. 2011; Shivaprasad et al. 2012; Li et al. 2012). These studies demonstrated that members of the miR482/2118 superfamily initiate large amounts of secondary, phased siRNAs from *NB-LRR* genes. Bacterial/viral infections were shown to correlate with reduced miRNA accumulation, thereby decreasing secondary siRNA accumulation, and increasing *NB-LRR* mRNA accumulation. Investigating small RNAs in different organisms would allow us to reveal novel types and/or functions of regulatory small RNAs.

Another limitation for discovering novel small RNAs could be due to their tissue-specific expression. Recent lines of evidence show that categorizing small RNA functions based on a rigid small RNA size requirement is rather superficial and misleading. It has been shown that vegetative nucleus-specific transposable elements in *Arabidopsis* pollen accumulate high amounts of 21-22 nt siRNAs to presumably target silencing in gametes (Slotkin et al. 2009). Other studies also pointed out the importance of cell-type-specific small RNA expression in ovules and developing endosperms (Mosher et al. 2009; Olmedo-Monfil et al. 2010). A recent study monitored a multi-generational time-course of establishment of silencing of an active transposon. It demonstrated a switch from the production of RDR6-dependent 21-22 nt siRNAs to 24 nt siRNAs, which appear to be responsible for the ‘initiation’ and the ‘maintenance’ phases of repression, respectively. In the early phases, silencing appears to be largely post-transcriptional and controlled by RDR6-dependent 21-22 nt siRNAs, whereas it is switched to be controlled at the transcriptional level, which is associated with 24 nt siRNAs, in the later generations (Marí-Ordóñez et al. 2013). In our study,

heterochromatic siRNAs are weakly expressed in wild-type ten-day old protonemata, instead, 21 nt miRNAs dominate the small RNA population. Future studies investigating the small RNA profiles in other life stages of *Physcomitrella* will provide deeper insights into the evolutionary history of regulatory small RNAs.

3.2.5 Identifying factors involved in small RNA biogenesis pathways

One of the critical future improvements is to focus on identifying new components that are involved in small RNA biogenesis pathways. It is clear that some AGO1-loaded miRNAs undergo 3'-truncation and oligo-U tailing which might potentially be important for downstream targeting specificities (Zhai et al. 2013). Not only 3'modifications but also spatio-temporal expression of pri-miRNAs have been suggested to be critical for miRNA function (Válóczi et al. 2006; Meng et al. 2011).

The current model for the heterochromatic siRNA biogenesis pathway has improved by the recent findings providing a link between repressive histone marks and heterochromatic siRNA biogenesis. Two recent studies show evidence that H3K9 methylation mark-specific SHH1/DTF1 guide the positioning of Pol IV, which in turn transcribes the precursors of 24 nt heterochromatic siRNAs (Válóczi et al. 2006; Meng et al. 2011). Similarly, methyl-DNA binding SUVH2/SUVH9 proteins were shown to recruit Pol V to loci with pre-existing DNA methylation marks to induce subsequent transcription (Johnson et al. 2014). Altogether, recruitment of Pol V seems to act as a self-reinforcing loop between repressive histone marks and DNA methylation to maintain transcriptional silencing. However, there are a number questions yet to be answered. For instance, Pol V-occupancy does not fully overlap with Pol IV-occupancy. Are there any direct feedback mechanisms between Pol IV- and Pol V-dependent loci? What is the biological significance of Pol V-only-dependent loci since Pol V-occupancy is independent of Pol IV-dependent heterochromatic siRNA production? Investigating the interplay between 5-mC, Pol IV-occupancy, Pol V-occupancy and repressive histone marks is of great importance for future research in understanding how heterochromatin silencing is maintained.

References

Abrahante JE, Daul AL, Li M, Volk ML, Tennessen JM, Miller EA, Rougvie AE. 2003. The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev Cell* **4**: 625–637.

Addo-Quaye C, Snyder JA, Park YB, Li Y-F, Sunkar R, Axtell MJ. 2009. Sliced microRNA targets and precise loop-first processing of MIR319 hairpins revealed by analysis of the Physcomitrella patens degradome. *RNA N Y N* **15**: 2112–2121.

Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Gruissem W, Meins F, Hohn T, et al. 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Res* **34**: 462–471.

Ambrose BA, Purugganan MD. 2012. *Annual Plant Reviews, The Evolution of Plant Form*. John Wiley & Sons.

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**: 203–207.

Arazi T, Talmor-Neiman M, Stav R, Riese M, Huijser P, Baulcombe DC. 2005. Cloning and characterization of micro-RNAs from moss. *Plant J* **43**: 837–848.

Arif MA, Fattash I, Ma Z, Cho SH, Beike AK, Reski R, Axtell MJ, Frank W. 2012. DICER-LIKE3 activity in Physcomitrella patens DICER-LIKE4 mutants causes severe developmental dysfunction and sterility. *Mol Plant* **5**: 1281–1294.

Arif MA, Frank W, Khraiwesh B. 2013. Role of RNA Interference (RNAi) in the Moss Physcomitrella patens. *Int J Mol Sci* **14**: 1516–1540.

Ashton NW, Cove DJ. 1977. The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, Physcomitrella patens. *Mol Gen Genet MGG* **154**: 87–95.

Aukerman MJ, Sakai H. 2003. Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* **15**: 2730–2741.

Axtell MJ. 2013a. Classification and Comparison of Small RNAs from Plants. *Annu Rev Plant Biol* **64**: 137–159.

Axtell MJ. 2013b. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA N Y N*.

Axtell MJ, Bartel DP. 2005. Antiquity of MicroRNAs and Their Targets in Land Plants. *Plant Cell* **17**: 1658–1673.

Axtell MJ, Bowman JL. 2008. Evolution of plant microRNAs and their targets. *Trends Plant Sci* **13**: 343–349.

Axtell MJ, Jan C, Rajagopalan R, Bartel DP. 2006. A two-hit trigger for siRNA biogenesis in plants. *Cell* **127**: 565–577.

Axtell MJ, Snyder JA, Bartel DP. 2007. Common functions for diverse small RNAs of land plants. *Plant Cell* **19**: 1750–1769.

Backman TWH, Sullivan CM, Cumbie JS, Miller ZA, Chapman EJ, Fahlgren N, Givan SA, Carrington JC, Kasschau KD. 2008. Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Res* **36**: D982–985.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al. 2011. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**: 960–963.

Bartel B, Bartel DP. 2003. MicroRNAs: At the Root of Plant Development? *Plant Physiol* **132**: 709–717.

Bass BL. 2000. Double-stranded RNA as a template for gene silencing. *Cell* **101**: 235–238.

Baulcombe D. 2004. RNA silencing in plants. *Nature* **431**: 356–363.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**: 363–366.

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu J-K. 2005. Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in Arabidopsis. *Cell* **123**: 1279–1291.

Bouche N, Lauressergues D, Gasciolli V, Vaucheret H. 2006. An antagonistic function for Arabidopsis DCL2 in development and a new function for DCL4 in generating viral siRNAs. *EMBO J* **25**: 3347–3356.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**: 1089–1103.

Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. 2003. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113**: 25–36.

Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O. 2008. Widespread translational inhibition by plant miRNAs and siRNAs. *Science* **320**: 1185–1190.

Burleigh JG, Barbazuk WB, Davis JM, Morse AM, Soltis PS. 2012. Exploring Diversification and Genome Size Evolution in Extant Gymnosperms through Phylogenetic Synthesis. *J Bot* **2012**: e292857.

CAI X, HAGEDORN CH, CULLEN BR. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**: 1957–1966.

Cao X, Jacobsen SE. 2002. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proc Natl Acad Sci U S A* **99 Suppl 4**: 16491–16498.

Carbonell A, Fahlgren N, Garcia-Ruiz H, Gilbert KB, Montgomery TA, Nguyen T, Cuperus JT, Carrington JC. 2012. Functional analysis of three Arabidopsis ARGONAUTES using slicer-defective mutants. *Plant Cell* **24**: 3613–3629.

Carmell MA, Girard A, van de Kant HJG, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**: 503–514.

Carmell MA, Hannon GJ. 2004. RNase III enzymes and the initiation of gene silencing. *Nat Struct Mol Biol* **11**: 214–218.

Cerutti H, Casas-Mollano JA. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* **50**: 81–99.

Chan SW-L, Zilberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE. 2004. RNA silencing genes control de novo DNA methylation. *Science* **303**: 1336.

Chávez Montes RA, Rosas-Cárdenas de FF, De Paoli E, Accerbi M, Rymarquis LA, Mahalingam G, Marsch-Martínez N, Meyers BC, Green PJ, de Folter S. 2014. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat Commun* **5**. http://www.nature.com/ncomms/2014/140423/ncomms4722/full/ncomms4722.html (Accessed June 3, 2014).

Chen C-J, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-Berthet E, Cognat V, Colot V, Voinnet O, Heard E, et al. 2012a. ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics* **28**: 3147–3149.

Chen H-M, Li Y-H, Wu S-H. 2007. Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc Natl Acad Sci* **104**: 3318–3323.

Chen X. 2004. A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science* **303**: 2022–2025.

Chen X, Liu J, Cheng Y, Jia D. 2002. HEN1 functions pleiotropically in Arabidopsis development and acts in C function in the flower. *Dev Camb Engl* **129**: 1085–1094.

Chen Y-R, Su Y -s., Tu S-L. 2012b. Distinct phytochrome actions in nonvascular plants revealed by targeted inactivation of phytobilin biosynthesis. *Proc Natl Acad Sci* **109**: 8310–8315.

Chiou T-J, Aung K, Lin S-I, Wu C-C, Chiang S-F, Su C. 2006. Regulation of Phosphate Homeostasis by MicroRNA in Arabidopsis. *Plant Cell Online* **18**: 412–421.

Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, Axtell MJ. 2008. Physcomitrella patens DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. *PLoS Genet* **4**: e1000314.

Cho SH, Coruh C, Axtell MJ. 2012. miR156 and miR390 regulate tasiRNA accumulation and developmental timing in Physcomitrella patens. *Plant Cell* **24**: 4837–4849.

Chuang CF, Meyerowitz EM. 2000. Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **97**: 4985–4990.

Coruh C, Shahid S, Axtell MJ. 2014. Seeing the forest for the trees: annotating small RNA producing genes in plants. *Curr Opin Plant Biol* **18C**: 87–95.

Cove D. 2005. The Moss Physcomitrella Patens. *Annu Rev Genet* **39**: 339–358.

Cove D, Bezanilla M, Harries P, Quatrano R. 2006. Mosses as Model Systems for the Study of Metabolism and Development. *Annu Rev Plant Biol* **57**: 497–520.

Cuperus JT, Carbonell A, Fahlgren N, Garcia-Ruiz H, Burke RT, Takeda A, Sullivan CM, Gilbert SD, Montgomery TA, Carrington JC. 2010. Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat Struct Mol Biol* **17**: 997–1003.

Daxinger L, Kanno T, Bucher E, van der Winden J, Naumann U, Matzke AJM, Matzke M. 2009. A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. *EMBO J* **28**: 48–57.

Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O. 2006. Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. *Science* **313**: 68–71.

Dlakić M. 2006. DUF283 domain of Dicer proteins has a double-stranded RNA-binding fold. *Bioinforma Oxf Engl* **22**: 2711–2714.

Dolgosheina EV, Morin RD, Aksay G, Sahinalp SC, Magrini V, Mardis ER, Mattsson J, Unrau PJ. 2008. Conifers have a unique small RNA silencing signature. *RNA* **14**: 1508–1515.

Dunoyer P, Brosnan CA, Schott G, Wang Y, Jay F, Alioua A, Himber C, Voinnet O. 2010. An endogenous, systemic RNAi pathway in plants. *EMBO J* **29**: 1699–1712.

Dunoyer P, Lecellier C-H, Parizotto EA, Himber C, Voinnet O. 2004. Probing the MicroRNA and Small Interfering RNA Pathways with Virus-Encoded Suppressors of RNA Silencing. *Plant Cell* **16**: 1235–1250.

Elbashir SM, Lendeckel W, Tuschl T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**: 188–200.

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, et al. 2007. High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PloS One* **2**: e219.

Fattash I, Voss B, Reski R, Hess WR, Frank W. 2007. Evidence for the rapid expansion of microRNA-mediated regulation in early land plant evolution. *BMC Plant Biol* **7**: 13.

Fei Q, Xia R, Meyers BC. 2013. Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. *Plant Cell Online* **25**: 2400–2415.

Finnegan EJ, Margis R, Waterhouse PM. 2003. Posttranscriptional gene silencing is not compromised in the Arabidopsis CARPEL FACTORY (DICER-LIKE1) mutant, a homolog of Dicer-1 from Drosophila. *Curr Biol CB* **13**: 236–240.

Fujii H, Chiou T-J, Lin S-I, Aung K, Zhu J-K. 2005. A miRNA involved in phosphate-starvation response in Arabidopsis. *Curr Biol CB* **15**: 2038–2043.

Fusaro AF, Matthew L, Smith NA, Curtin SJ, Dedic-Hagan J, Ellacott GA, Watson JM, Wang M-B, Brosnan C, Carroll BJ, et al. 2006. RNA interference-inducing hairpin RNAs in plants act through the viral defence pathway. *EMBO Rep* **7**: 1168–1175.

Gandikota M, Birkenbihl RP, Höhmann S, Cardon GH, Saedler H, Huijser P. 2007. The miRNA156/157 recognition element in the 3’ UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. *Plant J Cell Mol Biol* **49**: 683–693.

Gasciolli V, Mallory AC, Bartel DP, Vaucheret H. 2005. Partially Redundant Functions of Arabidopsis DICER-like Enzymes and a Role for DCL4 in Producing trans-Acting siRNAs. *Curr Biol* **15**: 1494–1500.

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**: 199–202.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186.

Grivna ST, Beyret E, Wang Z, Lin H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**: 1709–1714.

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5’ end formation in Drosophila. *Science* **315**: 1587–1590.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–840.

Haag JR, Pikaard CS. 2011. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat Rev Mol Cell Biol* **12**: 483–492.

Hamilton A, Voinnet O, Chappell L, Baulcombe D. 2002. Two classes of short interfering RNA in RNA silencing. *EMBO J* **21**: 4671–4679.

Hamilton AJ, Baulcombe DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**: 950–952.

Hammond SM. 2005. Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett* **579**: 5822–5829.

Hammond SM, Boettcher S, Caudy AA, Kobayashi R, Hannon GJ. 2001. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**: 1146–1150.

Hardcastle TJ, Kelly KA, Baulcombe DC. 2012. Identifying small interfering RNA loci from high-throughput sequencing data. *Bioinforma Oxf Engl* **28**: 457–463.

Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, Schwach F, Doonan JH, Baulcombe DC. 2010. The Arabidopsis RNA-Directed DNA Methylation Argonautes Functionally Diverge Based on Their Expression and Interaction with Target Loci. *Plant Cell Online*. http://www.plantcell.org/content/early/2010/02/19/tpc.109.072199 (Accessed October 15, 2013).

Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE. 2006. Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat Genet* **38**: 721–725.

Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. 2005. RNA Polymerase IV Directs Silencing of Endogenous DNA. *Science* **308**: 118–120.

Horwich MD, Li C, Matranga C, Vagin V, Farley G, Wang P, Zamore PD. 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol CB* **17**: 1265–1272.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**: 69–82.

Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC. 2007. Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting. *Plant Cell Online* **19**: 926–942.

Huettel B, Kanno T, Daxinger L, Bucher E, van der Winden J, Matzke AJM, Matzke M. 2007. RNA-directed DNA methylation mediated by DRD1 and Pol IVb: a versatile pathway for transcriptional gene silencing in plants. *Biochim Biophys Acta* **1769**: 358–374.

Jeong D-H, Thatcher SR, Brown RSH, Zhai J, Park S, Rymarquis LA, Meyers BC, Green PJ. 2013. Comprehensive Investigation of MicroRNAs Enhanced by Analysis of Sequence Variants, Expression Patterns, ARGONAUTE Loading, and Target Cleavage. *PLANT Physiol* **162**: 1225–1245.

Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. 2007. CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res* **35**: D829–D833.

Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan G-L, Walbot V, Sundaresan V, Vance V, Bowman LH. 2009. Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res* **19**: 1429–1440.

Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ, et al. 2014. SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* **advance online publication**. http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12931.html (Accessed April 7, 2014).

Johnston RJ, Hobert O. 2003. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426**: 845–849.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* **14**: 787–799.

Jones-Rhoades MW, Bartel DP, Bartel B. 2006. MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol* **57**: 19–53.

Kallman T, Chen J, Gyllenstrand N, Lagercrantz U. 2013. A Significant Fraction of 21-Nucleotide Small RNA Originates from Phased Degradation of Resistance Genes in Several Perennial Species. *PLANT Physiol* **162**: 741–754.

Kanno T, Bucher E, Daxinger L, Huettel B, Kreil DP, Breinig F, Lind M, Schmitt MJ, Simon SA, Gurazada SGR, et al. 2010. RNA-directed DNA methylation and plant development require an IWR1-type transcription factor. *EMBO Rep* **11**: 65–71.

Kanno T, Huettel B, Mette MF, Aufsatz W, Jaligot E, Daxinger L, Kreil DP, Matzke M, Matzke AJM. 2005. Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat Genet* **37**: 761–765.

Katiyar-Agarwal S, Morgan R, Dahlbeck D, Borsani O, Villegas A, Zhu J-K, Staskawicz BJ, Jin H. 2006. A pathogen-inducible endogenous siRNA in plant immunity. *Proc Natl Acad Sci* **103**: 18002–18007.

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

Ketting RF, Fischer SEJ, Bernstein E, Sijen T, Hannon GJ, Plasterk RHA. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* **15**: 2654–2659.

Khraiwesh B, Arif MA, Seumel GI, Ossowski S, Weigel D, Reski R, Frank W. 2010. Transcriptional control of gene expression by microRNAs. *Cell* **140**: 111–122.

Kirino Y, Mourelatos Z. 2007a. Mouse Piwi-interacting RNAs are 2’-O-methylated at their 3’ termini. *Nat Struct Mol Biol* **14**: 347–348.

Kirino Y, Mourelatos Z. 2007b. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. *RNA N Y N* **13**: 1397–1401.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**: D68–73.

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**: D152–D157.

Kurihara Y, Watanabe Y. 2004. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A* **101**: 12753–12758.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**: 853–858.

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313**: 363–367.

Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AMS, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**: 385–389.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**: 204–220.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**: 862–864.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843–854.

Lee T, Gurazada SGR, Zhai J, Li S, Simon SA, Matzke MA, Chen X, Meyers BC. 2012. RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. *Epigenetics* **7**: 781–795.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**: 415–419.

Lee Y, Jeon K, Lee J-T, Kim S, Kim VN. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* **21**: 4663–4670.

Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23**: 4051–4060.

Li F, Pignatta D, Bendix C, Brunkard JO, Cohn MM, Tung J, Sun H, Kumar P, Baker B. 2012. MicroRNA regulation of plant innate immune receptors. *Proc Natl Acad Sci* **109**: 1790–1795.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lin S-Y, Johnson SM, Abraham M, Vella MC, Pasquinelli A, Gamberi C, Gottlieb E, Slack FJ. 2003. The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev Cell* **4**: 639–650.

Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**: 2077–2080.

Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–476.

Lippman Z, Martienssen R. 2004. The role of RNA interference in heterochromatic silencing. *Nature* **431**: 364–370.

Liu C, Axtell MJ, Fedoroff NV. 2012a. The helicase and RNaseIIIa domains of Arabidopsis Dicer-Like1 modulate catalytic parameters during microRNA biogenesis. *Plant Physiol* **159**: 748–758.

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N-H. 2012b. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **24**: 4333–4345.

Liu Q, Rand TA, Kalidas S, Du F, Kim H-E, Smith DP, Wang X. 2003. R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* **301**: 1921–1925.

Llave C, Kasschau KD, Rector MA, Carrington JC. 2002a. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**: 1605–1619.

Llave C, Xie Z, Kasschau KD, Carrington JC. 2002b. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* **297**: 2053–2056.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Lu C, Kulkarni K, Souret FF, MuthuValliappan R, Tej SS, Poethig RS, Henderson IR, Jacobsen SE, Wang W, Green PJ, et al. 2006. MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res* **16**: 1276–1288.

Ma J-B, Ye K, Patel DJ. 2004. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**: 318–322.

MacLean D, Moulton V, Studholme DJ. 2010. Finding sRNA generative locales from high-throughput sequencing data with NiBLS. *BMC Bioinformatics* **11**: 93.

MacRae IJ, Doudna JA. 2007. Ribonuclease revisited: structural insights into ribonuclease III family enzymes. *Curr Opin Struct Biol* **17**: 138–145.

MacRae IJ, Zhou K, Doudna JA. 2007. Structural determinants of RNA recognition and cleavage by Dicer. *Nat Struct Mol Biol* **14**: 934–940.

MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA. 2006. Structural Basis for Double-Stranded RNA Processing by Dicer. *Science* **311**: 195–198.

Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL. 2005. Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of Tetrahymena thermophila. *Mol Cell Biol* **25**: 9151–9164.

Manavella PA, Koenig D, Rubio-Somoza I, Burbano HA, Becker C, Weigel D. 2013. Tissue-specific silencing of Arabidopsis SU(VAR)3-9 HOMOLOG8 by miR171a. *Plant Physiol* **161**: 805–812.

Manavella PA, Koenig D, Weigel D. 2012. Plant secondary siRNA production determined by microRNA-duplex structure. *Proc Natl Acad Sci* **109**: 2461–2466.

Margis R, Fusaro AF, Smith NA, Curtin SJ, Watson JM, Finnegan EJ, Waterhouse PM. 2006. The evolution and diversification of Dicers in plants. *FEBS Lett* **580**: 2442–2450.

Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. 2013. Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet* **45**: 1029–1039.

Marques JT, Kim K, Wu P-H, Alleyne TM, Jafari N, Carthew RW. 2010. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. *Nat Struct Mol Biol* **17**: 24–30.

Martienssen RA, Colot V. 2001. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**: 1070–1074.

Matsuda S, Ichigotani Y, Okuda T, Irimura T, Nakatsugawa S, Hamaguchi M. 2000. Molecular cloning and characterization of a novel human gene (HERNA) which encodes a putative RNA-helicase. *Biochim Biophys Acta* **1490**: 163–169.

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJM. 2009. RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol* **21**: 367–376.

Matzke MA, Birchler JA. 2005. RNAi-mediated pathways in the nucleus. *Nat Rev Genet* **6**: 24–35.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet* **15**: 394–408.

Meng Y, Shao C, Wang H, Chen M. 2011. The Regulatory Activities of Plant MicroRNAs: A More Dynamic Perspective. *Plant Physiol* **157**: 1583–1595.

Mette MF, Aufsatz W, van der Winden J, Matzke MA, Matzke AJM. 2000. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J* **19**: 5194–5201.

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, et al. 2008. Criteria for annotation of plant MicroRNAs. *Plant Cell* **20**: 3186–3190.

Mochizuki K, Gorovsky MA. 2005. A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* **19**: 77–89.

Morin RD, Aksay G, Dolgosheina E, Ebhardt HA, Magrini V, Mardis ER, Sahinalp SC, Unrau PJ. 2008. Comparative analysis of the small RNA transcriptomes of Pinus contorta and Oryza sativa. *Genome Res* **18**: 571–584.

Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, et al. 2009. Evolution of Genome Size and Complexity in Pinus. *PLoS ONE* **4**: e4332.

Mosher RA, Melnyk CW, Kelly KA, Dunn RM, Studholme DJ, Baulcombe DC. 2009. Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature* **460**: 283–286.

Mosher RA, Schwach F, Studholme D, Baulcombe DC. 2008. PolIVb Influences RNA-Directed DNA Methylation Independently of Its Role in siRNA Biogenesis. *Proc Natl Acad Sci* **105**: 3145–3150.

Mukherjee K, Campos H, Kolaczkowski B. 2013. Evolution of Animal and Plant Dicers: Early Parallel Duplications and Recurrent Adaptation of Antiviral RNA Binding in Plants. *Mol Biol Evol* **30**: 627–641.

Nakano M. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res* **34**: D731–D735.

Navarro L, Dunoyer P, Jay F, Arnold B, Dharmasiri N, Estelle M, Voinnet O, Jones JDG. 2006. A Plant miRNA Contributes to Antibacterial Resistance by Repressing Auxin Signaling. *Science* **312**: 436–439.

Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press.

Nickrent DL, Parkinson CL, Palmer JD, Duff RJ. 2000. Multigene Phylogeny of Land Plants with Special Reference to Bryophytes and the Earliest Land Plants. *Mol Biol Evol* **17**: 1885–1895.

Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, Arteaga-Vazquez M, Sidorenko L, Jeong D-H, Yen Y, et al. 2008. Distinct size distribution of endogenous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. *Proc Natl Acad Sci U S A* **105**: 14958–14963.

Nogueira FTS, Madi S, Chitwood DH, Juarez MT, Timmermans MCP. 2007. Two small regulatory RNAs establish opposing fates of a developmental axis. *Genes Dev* **21**: 750–755.

Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK. 2013. The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs1[W][OA]. *Plant Physiol* **162**: 116–131.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

Ohara T, Sakaguchi Y, Suzuki T, Ueda H, Miyauchi K, Suzuki T. 2007. The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. *Nat Struct Mol Biol* **14**: 349–350.

Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada J-P. 2010. Control of female gamete formation by a small RNA pathway in Arabidopsis. *Nature* **464**: 628–632.

Onodera Y, Haag JR, Ream T, Costa Nunes P, Pontes O, Pikaard CS. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**: 613–622.

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D. 2003. Control of leaf morphogenesis by microRNAs. *Nature* **425**: 257–263.

Pall GS, Hamilton AJ. 2008. Improved northern blot method for enhanced detection of small RNA. *Nat Protoc* **3**: 1077–1084.

Pantano L, Estivill X, Marti E. 2011. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics* **27**: 3202–3203.

De Paoli E, Dorantes-Acosta A, Zhai J, Accerbi M, Jeong D-H, Park S, Meyers BC, Jorgensen RA, Green PJ. 2009. Distinct extremely abundant siRNAs associated with cosuppression in petunia. *RNA N Y N* **15**: 1965–1970.

Papp I, Mette MF, Aufsatz W, Daxinger L, Schauer SE, Ray A, van der Winden J, Matzke M, Matzke AJM. 2003. Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol* **132**: 1382–1390.

Park J-E, Heo I, Tian Y, Simanshu DK, Chang H, Jee D, Patel DJ, Kim VN. 2011. Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature* **475**: 201–205.

Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, Poethig RS. 2005. Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A* **102**: 3691–3696.

Park W, Li J, Song R, Messing J, Chen X. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol CB* **12**: 1484–1495.

Pikaard CS, Haag JR, Ream T, Wierzbicki AT. 2008. Roles of RNA polymerase IV in gene silencing. *Trends Plant Sci* **13**: 390–397.

Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi M-A, Lerbs-Mache S, Colot V, Lagrange T. 2005. Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev* **19**: 2030–2040.

Qian K, Auvinen E, Greco D, Auvinen P. 2012. miRSeqNovel: An R based workflow for analyzing miRNA sequencing data. *Mol Cell Probes* **26**: 208–211.

Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. 2006. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* **20**: 3407–3425.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y, et al. 2008. The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science* **319**: 64–69.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Röther S, Meister G. 2011. Small RNAs derived from longer non-coding RNAs. *Biochimie* **93**: 1905–1915.

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev* **20**: 2214–2222.

Saito K, Sakaguchi Y, Suzuki T, Suzuki T, Siomi H, Siomi MC. 2007. Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes Dev* **21**: 1603–1608.

Sambrook J, Russell DW. 2001. *Molecular Cloning: A Laboratory Manual*. CSHL Press.

Schaefer D, Zryd JP, Knight CD, Cove DJ. 1991. Stable transformation of the moss Physcomitrella patens. *Mol Gen Genet MGG* **226**: 418–424.

Schaefer DG, Zrÿd JP. 1997. Efficient gene targeting in the moss Physcomitrella patens. *Plant J Cell Mol Biol* **11**: 1195–1206.

Schauer SE, Jacobsen SE, Meinke DW, Ray A. 2002. DICER-LIKE1: blind men and elephants in Arabidopsis development. *Trends Plant Sci* **7**: 487–491.

Shahid S, Axtell MJ. 2014. Identification and annotation of small RNA genes using ShortStack. *Methods San Diego Calif* **67**: 20–27.

Shahid S, Axtell MJ. Identification and annotation of small RNA genes using ShortStack. *Methods* **In Press**.

Shivaprasad PV, Chen H-M, Patel K, Bond DM, Santos BACM, Baulcombe DC. 2012a. A MicroRNA Superfamily Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs. *Plant Cell Online* tpc.111.095380.

Shivaprasad PV, Chen H-M, Patel K, Bond DM, Santos BACM, Baulcombe DC. 2012b. A MicroRNA Superfamily Regulates Nucleotide Binding Site–Leucine-Rich Repeats and Other mRNAs. *Plant Cell Online* tpc.111.095380.

Slotkin RK, Freeling M, Lisch D. 2005. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet* **37**: 641–644.

Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. *Cell* **136**: 461–472.

Soppe WJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, Peeters AJ. 2000. The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol Cell* **6**: 791–802.

Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V. 2012. The UEA sRNA Workbench: A Suite of Tools for Analysing and Visualising Next Generation Sequencing microRNA and Small RNA Datasets. *Bioinforma Oxf Engl*. http://www.ncbi.nlm.nih.gov/pubmed/22628521 (Accessed June 22, 2012).

Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome. *Cell* **152**: 352–364.

Sunkar R, Girke T, Zhu J-K. 2005. Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Res* **33**: 4443–4454.

Sunkar R, Zhu J-K. 2004. Novel and Stress-Regulated MicroRNAs and Other Small RNAs from Arabidopsis. *Plant Cell Online* **16**: 2001–2019.

Tabara H, Sarkissian M, Kelly WG, Fleenor J, Grishok A, Timmons L, Fire A, Mello CC. 1999. The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell* **99**: 123–132.

Tabara H, Yigit E, Siomi H, Mello CC. 2002. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans. *Cell* **109**: 861–871.

Talmor-Neiman M, Stav R, Klipcan L, Buxdorf K, Baulcombe DC, Arazi T. 2006a. Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. *Plant J Cell Mol Biol* **48**: 511–521.

Talmor-Neiman M, Stav R, Klipcan L, Buxdorf K, Baulcombe DC, Arazi T. 2006b. Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. *Plant J Cell Mol Biol* **48**: 511–521.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

Tanurdzic M, Vaughn MW, Jiang H, Lee T-J, Slotkin RK, Sosinski B, Thompson WF, Doerge RW, Martienssen RA. 2008. Epigenomic Consequences of Immortalized Plant Cell Suspension Culture. *PLoS Biol* **6**: e302.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinforma Oxf Engl* **25**: 1105–1111.

Troitsky AV, Ignatov MS, Bobrova VK, Milyutina IA. 2007. Contribution of genosystematics to current concepts of phylogeny and classification of bryophytes. *Biochem Biokhimiia* **72**: 1368–1376.

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**: 320–324.

Válóczi A, Várallyay E, Kauppinen S, Burgyán J, Havelda Z. 2006. Spatio-temporal accumulation of microRNAs is highly coordinated in developing plant tissues. *Plant J Cell Mol Biol* **47**: 140–151.

Vaucheret H. 2009. AGO1 homeostasis involves differential production of 21-nt and 22-nt miR168 species by MIR168a and MIR168b. *PloS One* **4**: e6442.

Vaucheret H. 2008. Plant ARGONAUTES. *Trends Plant Sci* **13**: 350–358.

Voinnet O. 2009. Origin, biogenesis, and activity of plant microRNAs. *Cell* **136**: 669–687.

Voinnet O. 2001. RNA silencing as a plant immune system against viruses. *Trends Genet TIG* **17**: 449–459.

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**: 1833–1837.

Wang Y, Juranek S, Li H, Sheng G, Wardle GS, Tuschl T, Patel DJ. 2009. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* **461**: 754–761.

Wassenegger M, Heimes S, Riedel L, Sänger HL. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**: 567–576.

Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H. 2006. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev* **20**: 1732–1743.

Waterhouse PM, Wang MB, Lough T. 2001. Gene silencing as an adaptive defence against viruses. *Nature* **411**: 834–842.

Welker NC, Maity TS, Ye X, Aruscavage PJ, Krauchuk AA, Liu Q, Bass BL. 2011. Dicer's helicase domain discriminates dsRNA termini to promote an altered reaction mode. *Mol Cell* **41**: 589–599.

Wesley SV, Helliwell CA, Smith NA, Wang MB, Rouse DT, Liu Q, Gooding PS, Singh SP, Abbott D, Stoutjesdijk PA, et al. 2001. Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J Cell Mol Biol* **27**: 581–590.

Wierzbicki AT. 2012. The role of long non-coding RNA in transcriptional gene silencing. *Curr Opin Plant Biol* **15**: 517–522.

Wierzbicki AT, Cocklin R, Mayampurath A, Lister R, Rowley MJ, Gregory BD, Ecker JR, Tang H, Pikaard CS. 2012. Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes Dev* **26**: 1825–1836.

Wierzbicki AT, Haag JR, Pikaard CS. 2008a. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**: 635–648.

Wierzbicki AT, Haag JR, Pikaard CS. 2008b. Noncoding transcription by RNA Polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* **135**: 635–648.

Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. 2009. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nat Genet* **41**: 630–634.

Wightman B, Ha I, Ruvkun G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**: 855–862.

Wu L, Zhou H, Zhang Q, Zhang J, Ni F, Liu C, Qi Y. 2010. DNA methylation mediated by a microRNA pathway. *Mol Cell* **38**: 465–475.

Wu-Scharf D, Jeong B, Zhang C, Cerutti H. 2000. Transgene and transposon silencing in Chlamydomonas reinhardtii by a DEAH-box RNA helicase. *Science* **290**: 1159–1162.

Xia R, Meyers BC, Liu Z, Beers EP, Ye S, Liu Z. 2013. MicroRNA Superfamilies Descended from miR390 and Their Roles in Secondary Small Interfering RNA Biogenesis in Eudicots. *Plant Cell Online* **25**: 1555–1572.

Xie F, Xiao P, Chen D, Xu L, Zhang B. 2012. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol Biol*.

Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC. 2005. Expression of Arabidopsis MIRNA genes. *Plant Physiol* **138**: 2145–2154.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004a. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* **2**: E104.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004b. Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLoS Biol* **2**: e104.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004c. Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLoS Biol* **2**: e104.

Yan KS, Yan S, Farooq A, Han A, Zeng L, Zhou M-M. 2003. Structure and conserved RNA binding of the PAZ domain. *Nature* **426**: 468–474.

Yang L, Wu G, Poethig RS. 2012. Mutations in the GW-repeat protein SUO reveal a developmental function for microRNA-mediated translational repression in Arabidopsis. *Proc Natl Acad Sci U S A* **109**: 315–320.

Yang X, Li L. 2011. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinforma Oxf Engl* **27**: 2614–2615.

Ye R, Wang W, Iki T, Liu C, Wu Y, Ishikawa M, Zhou X, Qi Y. 2012. Cytoplasmic Assembly and Selective Nuclear Import of Arabidopsis ARGONAUTE4/siRNA Complexes. *Mol Cell* **46**: 859–870.

Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW, Steward R, Chen X. 2005. Methylation as a crucial step in plant microRNA biogenesis. *Science* **307**: 932–935.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919.

Zhai J, Jeong D-H, De Paoli E, Park S, Rosen BD, Li Y, Gonzalez AJ, Yan Z, Kitto SL, Grusak MA, et al. 2011. MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev* **25**: 2540–2553.

Zhai J, Zhao Y, Simon SA, Huang S, Petsch K, Arikit S, Pillay M, Ji L, Xie M, Cao X, et al. 2013. Plant MicroRNAs Display Differential 3’ Truncation and Tailing Modifications That Are ARGONAUTE1 Dependent and Conserved Across Species. *Plant Cell* **25**: 2417–2428.

Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. 2004. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118**: 57–68.

Zhang H, Ma Z-Y, Zeng L, Tanaka K, Zhang C-J, Ma J, Bai G, Wang P, Zhang S-W, Liu Z-W, et al. 2013a. DTF1 is a core component of RNA-directed DNA methylation and may assist in the recruitment of Pol IV. *Proc Natl Acad Sci* 201300585.

Zhang J, Zhang S, Han S, Li X, Tong Z, Qi L. 2013b. Deciphering small noncoding RNAs during the transition from dormant embryo to germinated embryo in Larches (Larix leptolepis). *PloS One* **8**: e81452.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126**: 1189–1201.

Zhang X, Zhao H, Gao S, Wang W-C, Katiyar-Agarwal S, Huang H-D, Raikhel N, Jin H. 2011. Arabidopsis Argonaute 2 regulates innate immunity via miRNA393(∗)-mediated silencing of a Golgi-localized SNARE gene, MEMB12. *Mol Cell* **42**: 356–366.

Zhong X, Hale CJ, Law JA, Johnson LM, Feng S, Tu A, Jacobsen SE. 2012. DDR complex facilitates global association of RNA Polymerase V to promoters and evolutionarily young transposons. *Nat Struct Mol Biol* **19**: 870–875.

Zilberman D, Cao X, Jacobsen SE. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299**: 716–719.

Zilberman D, Cao X, Johansen LK, Xie Z, Carrington JC, Jacobsen SE. 2004. Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr Biol CB* **14**: 1214–1220.

Zong J, Yao X, Yin J, Zhang D, Ma H. 2009. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447**: 29–39.

VITA

Ceyda Coruh ceydacoruh@gmail.com

EDUCATION

Ph.D. Candidate in Plant Biology
The Pennsylvania State University, University Park, PA (August 2014 - expected)
Intercollege Graduate Program in Plant Biology

M.S. in Biological Sciences and Bioengineering
Sabanci University, Istanbul, Turkey (July 2007)
Faculty of Engineering and Natural Sciences

B.S. in Biological Sciences and Bioengineering
Sabanci University, Istanbul, Turkey (July 2005)
Faculty of Engineering and Natural Sciences

PUBLICATIONS (Ph.D.)

Coruh C*, Cho SH*, Shahid S, Liu Q, Wierzbicki A, Axtell MJ (2014) Comprehensive annotation of *Physcomitrella patens* small RNA loci reveals 23nt heterochromatic siRNAs dependent on a minimal Dicer-Like gene. *Submitted to Genome Biology*. *co-first authors

Coruh C, Shahid S, Axtell MJ (2014) Seeing the forest for the trees: Annotating small RNA producing genes in plants. *Current Opinion in Plant Biology*, 18: 87-95.

Cho SH, **Coruh C**, Axtell MJ (2012) miR156 and miR390 Regulate tasiRNA Accumulation and Developmental Timing in *Physcomitrella patens*. *Plant Cell*, 24: 4837-4849.

Ma Z, **Coruh C**, Axtell MJ (2010) *Arabidopsis lyrata* small RNAs: Transient *MIRNA* and small interfering RNA loci within the *Arabidopsis* genus. *Plant Cell*, 22: 1090-1103.

Cho SH, Addo-Quaye C, **Coruh C**, Arif MA, Ma Z, Frank W, Axtell MJ (2008) *Physcomitrella patens* DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. *Plos Genetics*, 4: e1000314.

PUBLICATIONS (M.S.)

Durmaz E*, **Coruh C***, Dinler G, Grusak MA, Peleg Z, Saranga Y, Fahima T, Yazici A, Ozturk L, Cakmak I, Budak H (2011) Expression and Cellular Localization of *ZIP1* Transporter Under Zinc Deficiency in Wild Emmer Wheat. *Plant Mol Biol Rep*, 29: 582-596. *co-first authors

Isik Z, Parmaksiz I, **Coruh C**, Geylan-Su YS, Cebeci O, Beecher B, Budak H, (2007) Organellar genome analysis of rye (*Secale cereale*) representing diverse geographic regions. *Genome*, 50(8): 724-34.