TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Charting Small RNA Landscape – an Exciting Journey in the Postgenomic Era

Yu Wang

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzende:		Univ.-Prof. Dr. C.-C. Schön
Prüfer der Dissertation:		
	1.	Univ.-Prof. Dr. H.-W. Mewes
	2.	Univ.-Prof. Dr. H. Schoof (Rheinische Friedrich-Wilhelms-Universität Bonn)

Die Dissertation wurde am 31.01.2011 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 06.06.2011 angenommen.

Contents

摘要

这篇博士论文着重研究小调控核糖核酸，包括微核糖核酸（miRNA）和小干扰核糖核酸（siRNA）。我用计算方法研究了小核糖核酸的生成机理和它们的调控机能。完成的科研项目涵盖了微核糖核酸基因的识别，微核糖核酸调控区域的分析，病毒小干扰核糖核酸的种群分析以及微核糖核酸标靶的相互作用。这篇论文涉及的模式生物有拟南芥（*Arabidopsis thaliana*），高粱（*Sorghum bicolor*），9 种植物病毒和人（*Homo sapiens*）。这些生物的多样性反映了小核糖核酸在各种不同的生命形式中都有广泛的调控功能。

Abstract

This thesis focuses on small regulatory RNAs, which include microRNA (miRNA) and small interfering RNA (siRNA). Using computational methods, this investigation included the biogenesis of small RNAs and their regulatory roles. The scope of work covered miRNA gene identification, miRNA promoter analysis, viral small interfering RNA (vsRNA) population analysis and miRNA target interactions. The model organisms studied in this thesis are *Arabidopsis thaliana*, *Sorghum bicolor*, 9 plant viruses and *Homo sapiens*. The phylogenetic heterogeneity of these organisms suggests wide spread regulatory roles of small RNAs in various formats of life.

Kurzbeschreibung

In dieser Doktorarbeit wurden kleine regulatorische RNA-Moleküle, insbesondere sogenannte Mikro-RNAs (miRNAs) und kleine interferierende RNAs (siRNAs) bearbeitet. Unter Verwendung computer-gestützter Analysen evaluierte ich neben der Biogenese auch die regulatorischen Rollen kleiner RNAs. Die Studien umfassen die Identifizierung von miRNA Genen und deren Promoteranalysen, die Untersuchung viraler siRNA-Populationen sowie die Interaktionen von miRNAs mit Zielgenen. Die in dieser Arbeit untersuchten Arten umfassen die Ackerschmalwand (*Arabidopsis thaliana*), die Hirse (*Sorghum bicolor*), 9 Pflanzenviren und den Menschen *(Homo sapiens*). Die grosse phylogenetische Vielfalt, die die untersuchten Organismen abdecken, unterstreicht die Bedeutung und weite Verbreitung kleiner RNA-Moleküle und deren regulatorischen Wirkungen.

Acknowledgement

Writing this thesis won't be possible without generous help and support of many people. I was fortunate to be accepted by Prof. Mewes to work at MIPS and got the opportunity to write this Ph.D. thesis. My stay at MIPS was an eye opener. I enjoyed enormously the pleasure of scientific discovery.

A good scientific taste never comes naturally. Without proper guidance, one can easily get lost at technique details of scientific daily work. Working at Dr. Klaus Mayer's group, I was fascinated by the beauty of biology. I thank Klaus for his biological insight and the academic freedom he provided so that I could wander through the wonderful land of small RNAs. I would like to thank colleagues from MIPS, Rémy Bruggmann, Irmtraud Dunger, Heidrun Gundlach, Georg Haberer, Gabi Kastenmüller, Wanseon Lee, Milheale Martis, Corinna Montrone, Michael Seidel, Manuel Spannagl, Xi Wang, Philip Wong, for the fun we had together and for the enjoyable scientific discussions and other small talks. Georg translated my thesis abstract into German, I am grateful for that.

Most of my works presented in this thesis are results of collaborations with biologists and other colleagues. I would like to thank all my co-authors, including Simone Brabletz, Thomas Brabletz, Ulrike Burk, Sabine Dietmann, Livia Donaire, Tobias Hindemitt, César Llave, Dominik Lutter, Andreas Ruepp, and Fabian Theis.

Prof. Heiko Schoof agreed to be in my Ph.D. committee. Prof. Chris-Carolin Schön chaired the defense of my thesis. I thank both of them for their time and discussions of my thesis.

When I was a kid, I got many children's books of scientists from my parents. I thank them for preparing me to do scientific work at a young age. The love from my family is one of the most important supports for this Ph.D thesis. I thank my mother and my sister for their patience. My brother-in-law, David Miller, helped me to proofread the thesis. I appreciate it. Of course all the remaining mistakes in the thesis are my own.

Preface

Biology has changed. There was a time that a biologist, for example Barbara McClintock, knew every plant in her field[1]. Once she told a group of students from the Department of Biology at Harvard University to "take the time and look". The students were puzzled. How can one find time to look and to think? "They argued that the new technology of molecular biology is self-propelling. It doesn't leave time. There's always the next experiment, the next sequencing to do". And that was almost half a century ago.

In his 2002 Nobel Prize Lecture, Sydney Brenner stated, "We are drowning in the sea of information and starving for knowledge"[2]. Indeed, in the information age, acquiring tremendous amount of (biological) experimental data becomes almost routine and nearly automated. But analyzing them is not. This is how, I, with an educational background in computation and engineering, have been afforded the opportunity to work extensively with large amounts of genomic and deep sequencing data. This thesis is a summary of my recent adventures in the wonderful land of small RNAs.

Some of my work has been done with *Arabidopsis thaliana*, a model organism for plant molecular biologists. *Arabidopsis thaliana* has a relatively small genome, about 120 megabytes (MB). This is extraordinary in plants since many of them have huge genomes, which continuously present a computational challenge. In *Arabidopsis thaliana*, there are about 24028 protein coding genes (TAIR9 annotation). All together, the coding parts of these genes are about 42 MB, which consist of 35% of the genome. The remaining genome is occupied by introns (20MB, 16.67% of the genome), untranslated regions (UTRs) of protein coding genes (16MB, 13.3% of the genome), transposable elements and tandem repeats (26MB, 21.67% of the genome) and some other genomic regions (~33% of the genome).

This research focuses on the non-coding genetic elements or regions on a genome, in particular small

[1] Evelyn Fox Keller, *a feeling for the organism: The Life and Work of Barbara McClintock*, page 198 (W. H. Freeman; 10th anniversary edition, 1994)

[2] Sydney Brenner, *Nature's gift to Science*, *Les Prix Nobel, The Nobel Prizes 2002*, page 279, (Nobel Foundation, Stockholm, 2003)

RNAs. Some small RNAs, such as microRNAs are, like protein coding genes, encoded on DNA. These are the subjects of the second part of the thesis (Chapter 2), where I present my work on miRNAs and their promoter regions. In the third part (Chapter 3), I discuss viral small RNAs, which are mostly derived from plant RNA viruses. The fourth part (Chapter 4) is dedicated to microRNAs targets, which are mostly protein coding genes. MicroRNAs normally downregulate the protein output of their targets. We found evidences indicating that microRNAs may coordinately regulate human protein complexes. The first part of the thesis (Chapter 1) is a general introduction in which I will use *Arabidopsis thaliana* as a model to describe the non-coding RNA landscape. The last part (Chapter 5) is a summary and discussion.

Chapter1 Introduction

This chapter is based on a Chapter in the book of “Genetics and Genomics of the Brassicaceae” in the series “Plant Genetics and Genomics – Crops and Models” (Haberer et al., 2010).

1.1 The *Arabidopsis thaliana* non-coding RNA landscape

Non-coding RNAs (ncRNA) recently emerged as an important landmark on the genomic landscape of eukaryotes. The widespread involvement of ncRNAs in various cellular activities has not been appreciated until recently (reviewed in (Mercer et al., 2009; Prasanth and Spector, 2007)). It is the fanatic pace of technological advancement that enables us to probe the cellular universe in an unmatched resolution. Many new discoveries have been made in the last years about ncRNAs in the model organism *Arabidopsis thaliana*, which are the subjects of this introduction.

As contrasting to messenger RNA (mRNA), non-coding RNA refers to a RNA molecule in a cell which does not encode for protein. Roughly, ncRNAs can be functionally classified into two categories (Figure 1.1). The first category consists of so called housekeeping ncRNAs(Szymanski and Barciszewski, 2002) or infrastructural ncRNAs(Mattick and Makunin, 2006) which are ubiquitously expressed at stable levels in different cellular contexts, and are responsible for the viability of the cell. The representatives of these ncRNAs are rRNA, tRNA, snoRNA, vault RNA(vRNA, (Kickhoefer et al., 2003)), etc. (the upper half of Figure 1.1). The functions of infrastructural RNAs have been well studied (reviewed in (Mattick and Makunin, 2006; Storz, 2002)) and are therefore not discussed further here. Many of these ncRNA families can be found in Rfam, a large collection of multiple sequence alignments and covariance models for ncRNAs(Gardner et al., 2009).

The second category is regulatory ncRNA, including small RNAs, macro ncRNAs, antisense transcripts and many other ncRNAs that are regulators at transcriptional or post-transcriptional level (the lower half of Figure 4). Small regulatory RNAs have been identified in various organisms, where they generally function by base-pairing with complementary sequences in other RNAs or DNA. In eukaryotes, small RNAs are less than ~35 nucleotides (nt) long, which can be subdivided into several classes according to origins. Small/Short interfering RNA (siRNA) is derived from a double strand RNA, either from inverted repeats, transposons, or natural antisense transcripts, which is then named as repeat-associated siRNA (rasiRNA), Long interspersed element-1 specific siRNAs (L1-siRNA), and natural antisense transcript siRNA (nat-siRNA), respectively. Other small RNAs include microRNA (miRNA), piwi-associate RNA (piRNA), 21U RNA and tiny ncRNA (tncRNA) (Reviewed in (Ghildiyal and Zamore, 2009)).

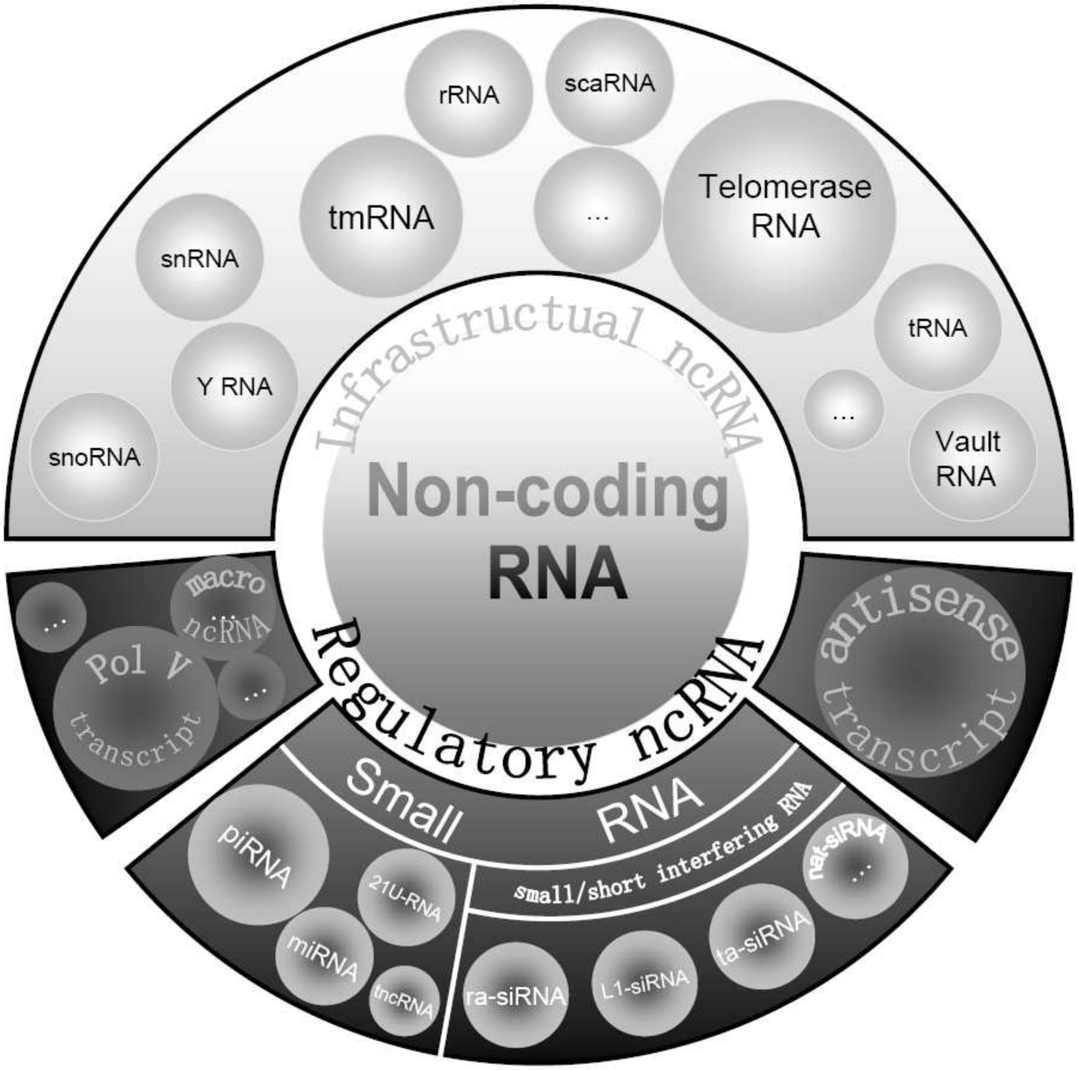


Figure 1.1

Classification of non-coding RNA families

Non-coding RNA families can be roughly classified into the following functional categories: infrastructural RNAs (represented by the upper half of the figure) and regulatory RNAs (represented by the lower half of the figure). This article focuses on recently discovered non-coding regulatory RNAs, which include small RNAs, antisense transcripts, macro ncRNA, Pol V transcripts and other (long) ncRNAs. Small RNAs consist of piwi-associated RNA, microRNA, short interfering RNA, 21U RNA and tiny non-coding RNA.

Abbreviations: L1-siRNA, Long Interspersed Element-1 specific siRNAs; macro ncRNA, macro non-coding RNA; miRNA, microRNA; nat-siRNA, nature antisense transcripts derived short interfering RNA; piRNA, piwi-associated RNA; rRNA, ribosomal RNA; ra-siRNA, repeat-associated short interfering RNA; scaRNA, small Cajal body specific RNA; siRNA, short interfering RNA; snRNAs, small nuclear RNAs; snoRNAs, small nucleolar RNAs; tRNA, transfer RNA; ta-siRNA, trans-acting short interfering RNA; tmRNA, transfer-messenger RNA; tncRNA, tiny non-coding RNA;Y RNA, Y chromosome RNA.

In the following sections, I am going to present a beautiful genomic landscape of non-coding RNAs in the model organism *Arabidopsis thaliana*, to review how much we have learned about non-coding RNAs so far and what kind of biological process they are involved in. I start with long non-coding RNAs which include antisense transcripts and Pol V transcripts in *A. thaliana*.

1.2 Long non-coding RNA

1.2.1 Natural Antisense Transcript (NAT)

In plants, whole genome tiling array experiments not only identified many unannotated protein coding genes, but also discovered that about 7600(Yamada et al., 2003) or 12,090(Stolc et al., 2005) genes in *Arabidopsis thaliana* and about 24% genes in *Oryza sativa*(Li et al., 2007) had antisense transcripts. These results suggest that antisense expression is widely spread in plant genomes. Many *Arabidopsis thaliana* genes reside on the genome in an overlapping fashion, forming cis-NATs pairs. Based on TAIR 6.0 annotation, *Jin et al*. (Jin et al., 2008) tested expression patterns of 1057 cis-NAT pairs in *A. thaliana*. They found a subset of cis-NAT pairs displayed negatively correlated expression profiles as well as inverse differential expression changes under at least one experimental condition.

1.2.2 Transcript generated by RNA Polymerase V (Pol V)

Arabidopsis thaliana harbors two RNA polymerases, Pol VI and Pol V (Wierzbicki et al., 2008), which are implicated in small RNA mediated chromatin-based gene silencing. In *A. thaliana*, transcription of intergenic noncoding regions by Pol V promotes heterochromatin formation and silencing of nearby genes (Wierzbicki et al., 2008). Pol V transcripts were detected in six intergenic noncoding regions by RT-PCR by comparing low abundance transcripts of wild type plants with those of Pol V mutants. Pol V transcripts are at least 200nt long and don’t contain poly A tails. They physically interact with

ARGONAUTE4 (AGO4) and possibly guide AGO4 to target loci through base-paring with associated siRNAs (Wierzbicki et al., 2009). Consequently, repressive epigenetic modifications of target regions are induced and in addition nearby genes and retroelements are transcriptionally silenced as well. Although Pol V is specific to flowering plants, it is tempting to speculate that the pervasive intergenic transcripts detected in eukaryotic genomes (Kapranov et al., 2007) may play a similar role in modifying chromatin structure.

In general, the functions of the majority of long ncRNAs are not elucidated very well. There might be many new discoveries remaining to be made on the functional role of long ncRNas. In contrast, in the last decade, small RNAs, especially miRNAs, has been shown to be crucial players for plant development, defense against viral and bacterial infection and stress responses. They have been exhaustively studied in *A. thaliana*. The following sections will highlight and summarize knowledge accumulated about the molecular function of this class of noncoding RNA.

1.3 Small RNA

Small RNAs (sRNAs) are 17 to 35 nt long regulatory non-coding RNAs that include microRNAs (miRNAs), small/short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). These small molecules regulate gene expression, control transposon activities, maintain genome integrity, and help to protect the plant cell from invading viruses. Although sRNAs have been already been extensively studied, recent progresses in the deep sequencing technology (review in (Mardis, 2008; Metzker, 2010; Meyers et al., 2006)) have revealed many new species of sRNAs. I will summarize the latest finding about sRNAs and particularly, present miRNA and several other small regulatory RNAs such as salt induced natural antisense transcript siRNA (nat-siRNAs) and trans-acting siRNA (tasiRNA).

1.3.1 MicroRNA

MicroRNAs (miRNAs) are about 21nt long endogenous non-coding RNAs that regulate a large number of protein coding genes at the post-transcriptional level. miRNA was first discovered in *C. elegans* in 1992 (Lee et al., 1993) while the first plant miRNA was reported about ten years later (Llave et al., 2002a; Llave et al., 2002b; Reinhart et al., 2002). Animal miRNA and plant miRNA differ in many aspects. Many miRNAs are deeply conserved in each kingdom, for instance, *let-7* is conserved in worms, insects and mammals while miR156 is conserved in land plants, from moss to

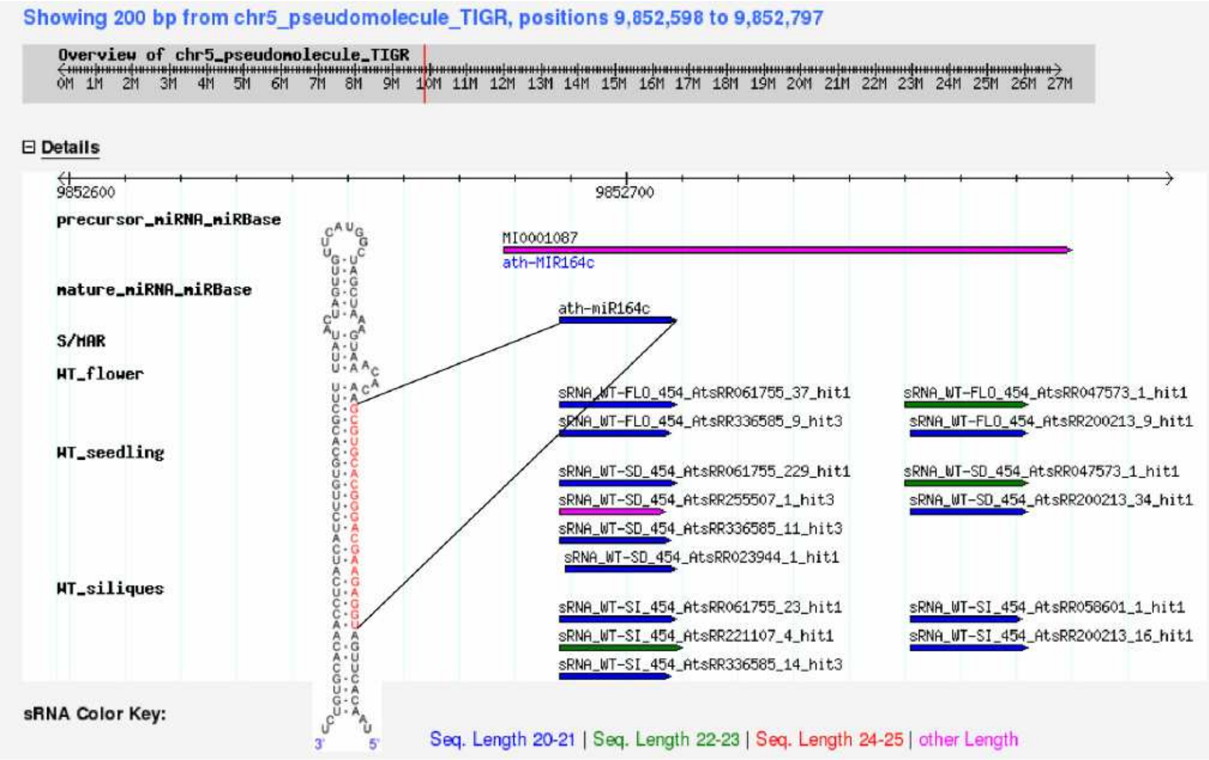

Showing 200 bp from chr5_pseudomolecule_TIGR, positions 9,852,598 to 9,852,797
Overview of chr5_pseudomolecule_TIGR
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M 25M 26M 27M
Details
9852600
9852700
precursor_miRNA_miRBase
MI0001087
ath-MIR164c
mature_miRNA_miRBase
ath-miR164c
S/MAR
WT_flower
sRNA_WT-FLO_454_AtsRR061755_37_hit1
sRNA_WT-FLO_454_AtsRR336585_9_hit3
sRNA_WT-FLO_454_AtsRR047573_1_hit1
sRNA_WT-FLO_454_AtsRR200213_9_hit1
WT_seedling
sRNA_WT-SD_454_AtsRR061755_229_hit1
sRNA_WT-SD_454_AtsRR255507_1_hit3
sRNA_WT-SD_454_AtsRR336585_11_hit3
sRNA_WT-SD_454_AtsRR023944_1_hit1
sRNA_WT-SD_454_AtsRR047573_1_hit1
sRNA_WT-SD_454_AtsRR200213_34_hit1
WT_siliques
sRNA_WT-SI_454_AtsRR061755_23_hit1
sRNA_WT-SI_454_AtsRR221107_4_hit1
sRNA_WT-SI_454_AtsRR336585_14_hit3
sRNA_WT-SI_454_AtsRR058601_1_hit1
sRNA_WT-SI_454_AtsRR200213_16_hit1
sRNA Color Key:
Seq. Length 20-21 | Seq. Length 22-23 | Seq. Length 24-25 | other Length
3'
5'

Figure 1.2

A genome browser view of ath-miR164c.

The small RNA genome browser is hosted at MIPS:http://mips.helmholtz-muenchen.de/cgi-bin/proj/plant/gbrowse/gbrowse/siRNA/. The precursor, ath-MIR164d and the mature miRNA ath-miR164c are shown in the figure, together with deep sequencing of sRNA profile in flowers, siliques and seedlings of *Arabidopsis thaliana* (accession: Columbia). Note that not only mature miRNA sequences but also miRNA* sequences were detected in different organ/tissues. The miRNA precursor has the typical hairpin structure with the mature miRNA sequence on one arm of the hairpin. Sequences are colored according to different lengths.

flowering plants. Currently the microRNA database - miRBase database (release 14) contains 10883 miRNA entries. The total number of human miRNA is at least ~800, which makes it one of the largest gene families in the human genome. Numerous miRNA prediction methods have been developed based on the secondary structure prediction, phylogenetic conservation, and thermodynamic stability of stem-loops (Figure 1.2, a representative of conserved plant miRNAs: ath-miR164c). While these computational methods are indispensable in whole genome miRNA gene scanning, the predicted miRNA candidates have to pass a “expression criteria”, which requires the detection of the predicted miRNA sequences by experimental techniques, either Northern blotting, cDNA cloning, microarray analysis or deep sequencing(Ambros et al., 2003a; Meyers et al., 2008).

miRNA biogenesis

The majority of animal miRNAs are located in the introns of protein-coding genes (Kim and Kim, 2007). These miRNAs are most likely cotranscribed along with their “host gene”, the gene to which the respective intron is associated with. The processing of intronic miRNAs does not interfere with splicing events (Kim and Kim, 2007). Many intergenic animal miRNAs form polycistronic clusters, facilitating coordinated expression. The long primary miRNA transcripts (pri-miRNAs) are transcribed by RNA polymerase II (Lee et al., 2004) or by RNA polymerase III (in human, (Borchert et al., 2006), Figure 1.4). The transcription of miRNA genes is likely controlled by *cis*-regulatory elements in the promoter regions of miRNA genes.

Animal pri-miRNAs typically consist of a stem-loop with 5’ and 3’ flanking segments. Pri-miRNAs are processed by the RNase III like enzymes Drosha and its double-strand RNA binding partner DGCR8 (in vertebrates). DGCR8 recognizes the junction between the flanking segments and the stem-loop. The Drosha-DGCR8 complex is therefore anchored precisely at the junction, placing Drosha on the stem-loop, ~11 bp away from the junction, where Drosha cleaves the stem-loop to liberate a precursor of miRNA (pre-miRNA) (Han et al., 2006). Interestingly, some of miRNA containing introns are so short that they encode only precursors of miRNAs. This type of miRNA is termed “miRtron” since their biogenesis doesn’t require Drosha cleavage (Berezikov et al., 2007; Okamura et al., 2007; Ruby et al., 2007). Animal pre-miRNAs are typically ~70nt long with a distinctive stem-loop structure, whose one stem contains miRNA and the other contains miRNA*. Pre-miRNAs are exported to cytoplasm by the nuclear transport factor Exportin-5. In the cytoplasm, Dicer, another RNase III type protein, processes pre-miRNAs to generate ~22 nt miRNA:miRNA* duplexes with 2 nt 3’ overhangs. Mature miRNAs are then selected by an unknown mechanism and incorporated into the RNA-induced silencing complex (RISC). RISC is the central element of all RNA silencing pathways. It consists of at least one Argonaute protein and a small non-coding RNA. Depending on the incorporated small RNA and the Argonaute protein, RISC can cleavage mRNA, block protein synthesis and introduce transcriptional gene silencing. miRNA*s were generally believed to be degraded quickly. However, recent reports find that in many tissues miRNA*s are constantly detected, suggesting miRNA*s may be functionally active in certain circumstances. In flies, miRNA*s are sorted into RNA interference pathway by being loaded into AGO2 (Ghildiyal et al., ; Okamura et al., 2009).

Many plant miRNA genes (reviewed in (Voinnet, 2009)) are found in intergenic regions and are believed to represent independent transcriptional units. Recently, 63 TSSs of Arabidopsis miRNAs have been identified by 5' RACE (Xie et al., 2005). A follow up computational study predicted many transcription factor binding sites, some of which are overrepresented in the plant miRNA promoter regions as compared to promoters of protein coding genes (Megraw et al., 2006). Just like normal protein coding genes, plant miRNA genes can have multiple transcription starting sites (Song et al., 2007), exon-intron structure and alternative 3' ends. It has been shown that some of ath-miR164 primary transcripts have two exons (Nikovics et al., 2006).

Similar to animal miRNA biogenesis, plant pri-miRNA is processed by DICER-Like 1 enzyme (DCL1) and other proteins including a double-stranded RNA (dsRNA) binding-domain protein, Hyponastic Leaves1 (HYL1) and a zinc-finger-domain protein Serrate (SE). In *A. thaliana,* DCL1 and HYL1 colocalize in discrete nuclear bodies, called D-bodies. Fang and Spector (Fang and Spector, 2007) showed that an introduced pri-miRNA was recruited to D-bodies, indicating that D-bodies are involved in pri-miRNA processing in the cell nucleus. HYL1 may assist DCL1 to precisely process pri-miRNAs into pre-miRNAs. Some other unidentified factors may also participate in pri-miRNA processing *in vivo*. Comparing to ~70 nt animal pre-miRNAs, plant pre-miRNAs are in general larger. The average lengths of *Arabidopsis thaliana* and *Oryza sativa* pre-miRNAs are 170nt and 145nt, respectively. It is believed that in *A. thaliana*, DCL1 is mainly responsible for processing pre-miRNAs into miRNA:miRNA* duplexes. There are two exceptions (ath-miR822 and ath-miR839) where DCL4 takes the dicing role in miRNA maturation, implying a second biogenesis pathway of miRNAs (Rajagopalan et al., 2006). After mi RNA:miRNA* duplexes are diced from pre-miRNAs, the duplexes are methylated on the ribose of the last nucleotide by the miRNA methyltransferase HEN1 and transferred to the cytoplasm by the HASTY (HST) transporter. In the cytoplasm, miRNA is mainly incorporated into the RISComplex along with ARGONAUTE1 (AGO1) or ARGONAUTE type proteins.

miRNA and target interaction

miRNAs recognize their targets by simple nucleotide base-pairing. In animals, miRNAs mostly guide RISCs to the 3'UTRs of target mRNAs and probably repress translation or destabilize mRNAs. In

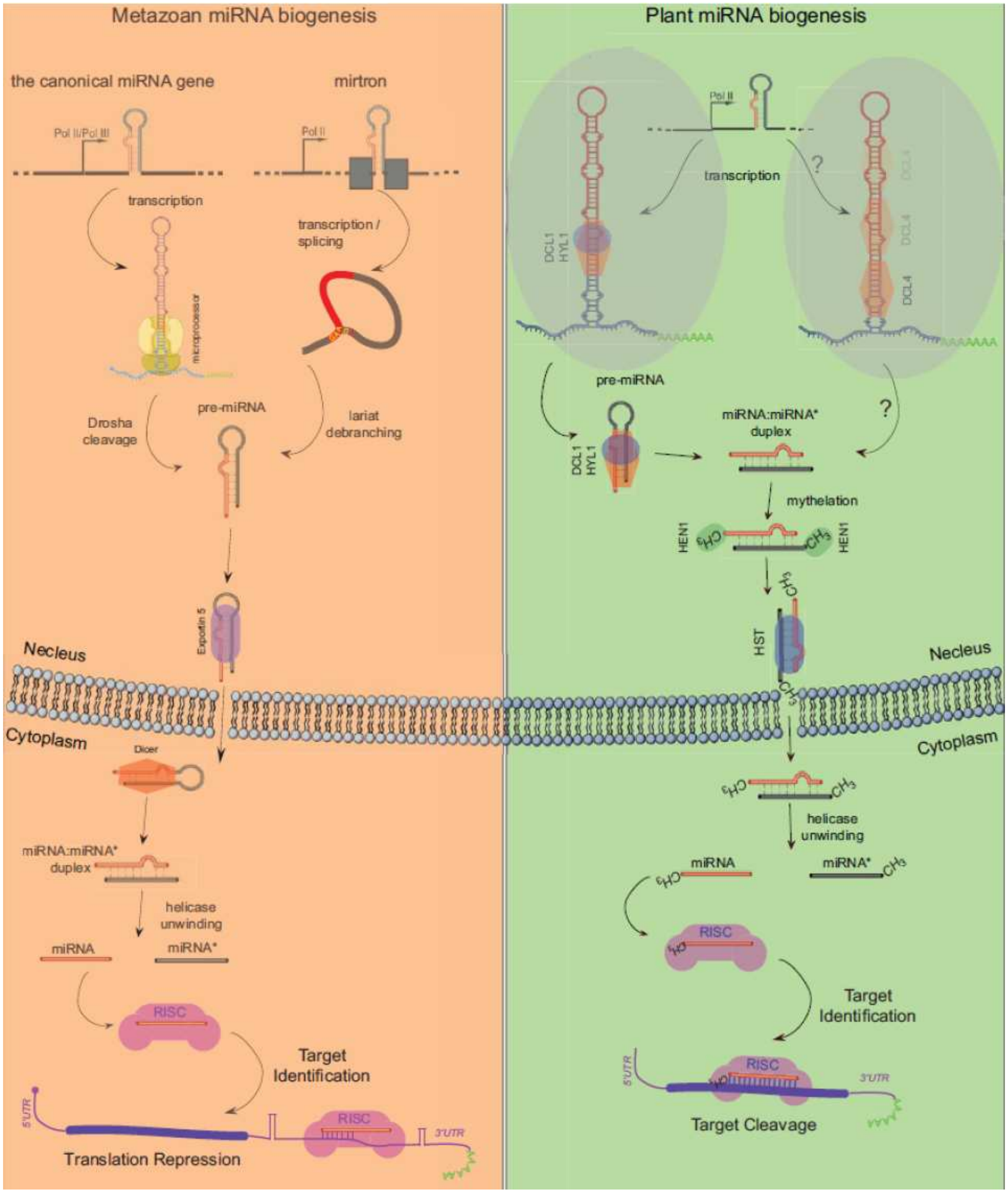


Figure 1.3

Animal and plant miRNA biogenesis.

The left panel shows two animal miRNA biogenesis pathways: the canonical miRNA pathway and the miRtron pathway. The right panel shows two plant miRNA biogenesis pathways: the DCL1 pathway and the DCL4 pathway.

plants, miRNAs usually have near perfect complementarities to their targets and guide RISCs to cleave the corresponding transcripts (reviewed by (Brodersen and Voinnet, 2009; Voinnet, 2009)). A recent study has shown that translational repression is a general mode of plant miRNA action (Brodersen et al., 2008).

Target prediction has been fairly straightforward for plant miRNAs due to the near perfect complementarities to their targets(Fahlgren and Carrington, 2010). Conversely, target prediction for animal miRNAs is intriguingly complicated because of the often limited complementarity between animal miRNAs and their targets. Based on limited experimental data, the first generation miRNA target prediction programs used evolutionary conservation as a powerful tool to assist miRNA binding sites identification. Although limited in numbers most experimentally verified binding sites show a strong complementarity bias to the 5' ends of animal miRNAs. The perfect complementarity to the region covering seven nucleotides starting from either the first or the second nucleotide at the 5' end of a miRNA has been argued as sufficient to identify a target site. However, a recent experiment seriously challenged the so-called "the perfect seed matching rule". Didiano D. and Hobert O. (Didiano and Hobert, 2006) showed that G:U base pairing is tolerated in the 'seed' region of the *lsy-6* miRNA interaction with its in vivo target *cog-1*, and that 6- to 8-base-pair perfect seed pairing is not a generally reliable predictor for an interaction of *lsy-6* with a 3' UTR. Furthermore, they demonstrated that the predicted target sites of 13 *lsy-6* target genes are not functional in their sensor system. MicroRNA *lsy-6* can only interact with its target site in specific 3' UTR contexts. Obviously, an mRNA can have its own secondary structure which may interfer miRNAs binding to their target sites. Two nonsequence specific contextual features beyond miRNA target sites were later proved to be critical determinants of miRNA-mediated 3' UTR regulation (Didiano and Hobert, 2008). In plant, the emerging picture of the limited complementarites between plant miRNAs and functional target sites (Brodersen et al., 2008; Dugas and Bartel, 2008) argues towards the need to further develop prediction tools for plant miRNA targets.

Thus far no high-throughput experimental methods for miRNA target identification are available, which has motivated the continuous development of computational tools of target prediction programs (reviewed in (Alexiou et al., 2009)). Current efforts of target verification have been focused on bringing in the cellular context of miRNAs, targets and effector proteins such as Argonautes (Chi et al., 2009; Easow et al., 2007; Hammell et al., 2008; Hong et al., 2009).

Arabidopsis thaliana miRNA families

There are 190 miRNA genes in *Arabidopsis thaliana*, consisting of 91 families. Many *A. thaliana* miRNA families are deeply conserved in land plants (Table 1, (Axtell and Bartel, 2005; Sunkar and Jagadeeswaran, 2008). Some miRNA families expanded by gene duplication (Allen et al., 2004; Maher et al., 2006) with the largest one, ath-miR169, consisting of 14 family members. One scenario for de novo genesis of microRNA genes is by inverted duplication of target gene sequences. Several recently evolved miRNA genes (only found in *A. thaliana*) and other small RNA- generating loci were shown to contain extended similarities at the arms on each side of their respective foldback precursors to some protein coding sequences, suggesting that these miRNAs and small RNA loci evolved from inverted gene duplication (Allen et al., 2004; Fahlgren et al., 2007; Rajagopalan et al., 2006).

Table 1.1

miRNA families in *Arabidopsis thaliana* and their conservation in the plant kingdom

Only miRNAs with known targets are listed in the table. For each family, the conservation is shown for Eudicots, Monocots, Magnoliids, Gymnosperms, Ferns, Lycopods, and Mosses. Note that the indicated conservation is rather conservative since the miRNA profiling might not be carried out in many plant species. One interesting observation is that conserved miRNA families tend to have more family member while non-conserved miRNAs tend to be a singleton.

miRNA Family	**No. Loci**	**Target Family**	**conservation**							**function**
			Eudicots	**Monocots**	**Magnoliids**	**Gymnosperms**	**Ferns**	**Lycopods**	**Mosses**	
miR156/miR157	12	SPL	■	■	■	■	■	■	■	retain the juvenile status of the young plants
miR158	2	PPR								
miR159	6	MYB	■	■	■	■		■		desensitize hormone signaling during germination
miR160	3	ARF	■	■	■	■	■	■	■	regulate auxin signal transduction during plant development
miR161	1	PPR								
miR162	2	DCL	■	■						maintain DCL1 at a funcationally sufficient level
miR163	1	SAMT								

miRNA Family	No. Loci	Target Family	conservation							function
			Eudicots	Monocots	Magnoliids	Gymnosperms	Ferns	Lycopods	Mosses	
miR164	3	NAC	■	■	■	■				regulate plant development and aging induced cell death
miR165/miR166	9	HD-ZIPIII	■	■	■	■	■	■	■	regulate leaf development, vascular patterning, SAM function
miR167	4	ARF	■	■	■	■			■	regulate adventitious rooting, ovule and anther development
miR168	2	AGO	■	■	■	■	■			maintain AGO1 homeostasis
miR169	14	NF-Y	■	■	■	■	■			involve in nutrient, drought response
miR170/miR171	4	SCL	■	■	■	■	■	■	■	involve in cell differentiation
miR172	5	AP2	■	■	■	■	■			regulate flowering time and floral organ identity
miR173	1	TAS1,TAS2								generate tasiRNAs which target PPR proteins
miR319		TCP	■	■	■			■	■	coordinate leaf development and physiology
miR390/miR391	3	TAS3	■	■	■		■		■	generate tasiRNAs which target ARF3,4, involve in developmental timeing and patterning
miR393	2	TIR1/AFB, bHLH	■	■		■				involve in antibacterial resistance
miR394	2	F-Box	■	■	■	■				
miR395	6	APS,AST	■	■						regulate sulphur assimilation
miR396	2	GRF	■	■	■	■	■	■		regulate leaf growth and development
miR397	2	LAC	■	■						regulate nonessential copper proteins
miR398	3	CSD, CytC oxidase	■	■	■	■				regulate copper homeostasis
miR399	6	E2-UBC	■	■						regulate phosphate homeostasis

miRNA Family	No. Loci	Target Family	conservation							function
			Eudicots	Monocots	Magnoliids	Gymnosperms	Ferns	Lycopods	Mosses	
miR400	1	PPR								
miR402	1	ROS1-Like								
miR403	1	AGO	■							
miR408	1	LAC, PLC	■	■		■		■	■	regulate nonessential copper proteins
miR447	3	2-PGK								
miR472	1	CC-NBS-LRR								
miR771	1	eIF-2								
miR773	1	MET2								
miR774	1	F-box								
miR775	1	GT								
miR776	1	PK								
miR777	1	CIP4.1-like								
miR778	1	SUVH								
miR780	1	CHX								
miR823	1	CMT3								
miR824	1	MADS-Box								
miR827	1	SPX								
miR828	1	MYB,TAS4								
miR842	1	JR/MBP								
miR844	1	PK								
miR846	1	JR/MBP								
miR856	1	CHX								
miR857	1	LAC								regulate nonessential copper proteins
miR858	1	MYB								
miR859	1	F-Box								

A large fraction of *A. thaliana* conserved miRNA targets are transcription factors involved in the regulation of plant development. Therefore, conserved miRNAs participate in a range of plant developmental processes (Table 1.1). Some miRNAs are also implicated in stress response and antiviral or antibacterial defense. In the following, key findings for individual miRNA families are summarized.

miR156/miR157

12 loci in *Arabidopsis thaliana* encode members of the miR156/miR157 family. All miR156 miRNAs are 20nt while miR157a,b,c are 21nt and miR157d is 20nt long. Nevertheless, all mature miR156/miR157 sequences are almost identical, with only one nucleotide difference. Since they most likely regulate the same set of targets, SPL (SQUAMOSA PROMOTER BINDING-LIKE) transcription factors, miR156 and miR157 are grouped into one family.

miR156 is highly expressed early in shoot development and the expression level of miRNA 156 decreases over time. High level of miR156 in young plants is associated with low SPL levels. Consequently, when miR156 abundance declines, SPL levels increase. miR156 targets ten different SPL genes, including SPL3/4/5, SPL9/15, SPL2/10/11 and SPL6/SPL13a/b. SPL genes function in distinct pathways to promote different adult phase vegetative traits and flowering. By simultaneously repressing these SPL involved pathways, miR156 acts as a master regulator to retain the juvenile status of the young plants (Wang et al., 2009; Wu et al., 2009).

miR159

miR159 is induced by abscisic acid (ABA) and drought treatment during germination in an ABI3 (abscisic acid-insensitive3) dependent manner (Reyes and Chua, 2007). All three members of the miR159 family are expressed during germination. miR159 targets several MYB transcription factors. During seedling stress response, ABA-induced miR159 cleaves MYB33 and MYB101 transcripts to desensitize hormone signaling. In miR159a and miR159b double mutants, MYB33 and MYB65 expression levels are elevated, but the expression of five other GAMYB-like family members is not altered in mir159ab plants. Phenotypic analysis demonstrated miR159a and miR159b specifically target MYB33 and MYB65 (Allen et al., 2007).

miR160

miR160 targets three ARF (auxin response transcription factor) genes, ARF10, ARF16, and ARF17. ARF genes are involved in the regulation of auxin signal transduction during plant development. Transgenic plants expressing a miRNA-resistant version of ARF10 and ARF17 and demonstrate dramatic developmental defects, such as embryo and emerging leaf symmetry anomalies, leaf shape defects, premature inflorescence development, altered phyllotaxy along the stem, contorted flowers, twisted siliques, sterility, and root growth defects (Liu et al., 2007; Mallory et al., 2005). Derepression of ARF10 increases the sensitivity of seeds and seedlings to exogenous ABA during germination and post-germinative shoot formation, indicating that repression of ARF10 by miR160 is essential for normal plant development. In root, miR160 regulate ARF10 and ARF16, together with auxin, generating a pattern consistent with root cap development(Wang et al., 2005). In adventitious root, miR160, targeting ARF17 and miR167, targeting ARF6 and ARF8, form a complex regulatory network to control adventitious root initiation(Gutierrez et al., 2009).

miR161/miR163

miR161 and miR163 are non-conserved microRNAs in *A. thaliana*. The foldback arms of MIR163 have significant similarity to a segment of three SAMT-like genes, which are miR163 targets. There is also segmental similarity between MIR161 foldback arms and PPR target genes. Together, these data indicate that MIR161 and MIR163 genes evolved relatively recently by inverted duplication events associated with active expansion of target gene families. The resulting hairpin structures were then adapted to the miRNA biogenesis pathway (Allen et al., 2004). The duplication event may have included promoter sequences of the original genes since significant similarities between the promoters of MIR161/163 genes and their corresponding target genes were also detected(Wang et al., 2006).

miR162

miR162 is expressed at a relatively low level in *A. thaliana* seedlings, leafs and inflorescences. miR162 has two loci on the genome, MIR162a and MIR162b. miR162 is believed to be processed by DCL1 from the MIR162a primary transcript. miR162b has a very low abundance in wild type plants, may compensate for the loss of function of MIR162a (Hirsch et al., 2006). Remarkably, miR162 targets DCL1 and establishes a negative feedback loop to regulate DCL1 abundance.

miR164

The miR164 family consists of three family members, miR164a, miR164b and miR164c, which target NAC (no apical meristem-NAM, Arabidopsis transcription factor like family-ATAF, and CUP-SHAPED COTYLEDON-CUC) genes. Transgenic plants with expression of miR164-resistant CUC1, CUC2 mRNA show alternation in embryonic, vegetative and floral development. Conversely, constitutive overexpression of miR164 not only phenocopies cuc1 cuc2 double mutants but also leads to new phenotypes such as leaf and stem fusions (Laufs et al., 2004; Mallory et al., 2004a; Peaucelle et al., 2007; Raman et al., 2008). miR164abc triple-mutants show a more severe disruption of shoot development than any single miR164 mutant (Sieber et al., 2007). miR164 single mutants have different phenotypes. miR164a and miR164b mutant plants express less miR164 and more NAC1. These plants have more lateral roots as compared to wild type plants (Guo et al., 2005). miR164a mutant plants also have a higher level of leaf serration, which can be abolished by CUC2 inactivation, implying that the balance between coexpressed CUC2 and miR164a determines the extent of serration (Nikovics et al., 2006). miR164c mutants have extra petals in early-arising flowers (Baker et al., 2005). These observations suggest that the redundancy among miR164 genes is not complete and each miR164 gene has also specialized functions. In *A. thaliana* leaves, miR164 expression gradually decreases with aging through negative regulation by ethylene insensitive 2 (EIN2). Consequently, a miR164 target, Oresara1 (ORE1) is up-regulated and contributes to aging related cell death (Kim et al., 2009). Together, the results demonstrate that miR164 is an important regulatory component in plant development and aging induced cell death.

miR165/miR166

In *A. thaliana* there are two miR165 genes and 7 miR166 genes. Their mature miRNA sequences differ from one another at most by a single nucleotide. The individual miR165/miR166 genes exhibit distinct temporal and spatial expression patterns in different plant organs (Table 1.2, (Jung and Park, 2007)). In leaves, histone deacetylases (HDACs), HDT1/HD2A and HDT2/HD2B, ASYMMETRIC LEAVES, AS1 and AS2 act independently to control levels and/or patterns of miR165/166 distribution for the development of adaxial-abaxial leaf polarity (Ueno et al., 2007). AS1/AS2 may negatively regulate the expression of miR165/miR166 expressions (Fu et al., 2007). In shoot apical meristem (SAM), miR165/miR166 is negatively regulated in parallel by AGO10 and tasiR-ARF pathway (Liu et al., 2008b).

miR165/miR166 target the class III homeodomain leucine-zipper (HD-ZIPIII) transcription factors, including PHB (PHABULOSA) and PHV (PHAVOLUTA). Mutations in the miR165/166 complementary sites of PHB and PHV genes cause severe developmental defects in leaves (Mallory et al., 2004b). Overexpression of miR165/miR166 causes dramatic reduction of the transcripts of five HD-ZIPIII transcription factors in *A. thaliana* (Zhou et al., 2007).

Table 1.2

Temporal and spatial expression of the miR165 and miR166 family

In *A. thaliana* there are two miR165 genes and 7 miR166 genes. Their mature miRNA sequences differ from one another by a single nucleotide. The individual miR165/miR166 genes exhibit distinct temporal and spatial expression patterns in different plant organs. Expression characteristics are indicated by different gray shading of the respective table cells.

name	temporal expression							spatial expression											
	embryos				late stages								floral organs						
	early globular	late globular	heart (abaxial)	torpedo (abaxial)	SAM	root-hypocotyl junctions	cotyledon vasculatures (6days old seedlings)	leaf veins	root tissues	root tips	rossette leaf tricomes	immature siliques	stamens	ovules	stigma	receptacles	sepals	petals
miR165a	■	■	■	■	░	■	▩	▩	▩	■	□	▩	□	■	□	□	░	░
miR165b	■	■	■	■	░	■	▩	▩	▩	■	□	□	■	□	□	□	░	░
miR166a	■	■	■	■	░	■	▩	▩	▩	■	□	□	■	□	□	□	░	░
miR166b	■	■	■	■	░	■	▩	▩	▩	■	□	■	□	■	■	□	░	░
miR166c	■	■	■	■	░	■	▩	▩	▩	■	□	□	■	□	□	□	░	░
miR166d	■	■	■	■	░	■	▩	▩	□	□	□	▩	□	■	□	□	░	░
miR166e	■	■	■	■	░	■	▩	▩	▩	■	□	□	■	□	□	□	░	░
miR166f	■	■	■	■	░	■	▩	▩	▩	■	□	□	■	□	□	□	░	░
miR166g	unknown	unknown	unknown	unknown	□	□	□	□	▩	□	■	░	■	□	■	■	░	▩

■	high level expression
▩	moderate level expression
░	low level expression
□	not detectable

miR167

miR167 controls expression patterns of both ARF6 and ARF8. MIR167a is expressed in ovules and anthers where wild type ARF6 and ARF8 transcripts are normally not detected. However, transgenic plants with modified ARF6 (mARF6) and ARF8 (mARF8) target sites accumulate both mARF6 and mARF8 transcripts in ovules and anthers. These plants have sterile flowers due to arrested ovule development and anther indehiscence, indicating that miR167 regulates both female and male floral organ development (Wu et al., 2006). miR167 is also involved in adventitious rooting. miR167 and miR160 form a complex regulatory network with their corresponding targets ARF6, ARF8 and ARF17. These ARF transcription factors not only regulate each other's expression transcriptionally but also modulate miR167 and miR160 abundance (Gutierrez et al., 2009).

miR168

miR168 has two loci on the genome, miR168a, which is highly expressed with a predominantly 21-nt miR168 species, and miR168b, which is lowly expressed with an equal amount of 21- and 22-nt miR168 species (Vaucheret, 2009). miR168 co-expresses with AGO1 and cleaves AGO1 mRNA. At post-transcriptional level, miR168 is stabilized by AGO (Vaucheret et al., 2006). This retro-regulation of miR168 and AGO1 is necessary to maintain AGO1 homeostasis, which is crucial for the proper functioning of other miRNAs and plant development (Vaucheret et al., 2004).

miR169

Using a deep sequencing approach miR169 miRNA was shown to be repressed upon phosphate limitation (Hsieh et al., 2009), while several miR169 genes were found to be repressed upon nitrogen limitation by a quantitative real-time polymerase chain reaction platform (Pant et al., 2009). miR169 targets Nuclear factor Y (NF-Y) transcription factor NFYA5 in *A. thaliana*. NFYA5 is strongly induced by drought in vascular tissues and guard cells. Conversely, miR169a and miR169c were substantially down-regulated by drought. Coexpression experiments suggest that miR169a was more efficient than other miR169 genes at repressing the NFYA5 mRNA level (Li et al., 2008).

miR170/miR171

miR171 targets three Scarecrow-like (SCL) transcription factors for cleavage (Llave et al., 2002b). The miR171 gene is transcribed in a highly cell type-specific manner and its transcription coincides with its site of endonucleolytic activity, suggesting miR171 is involved in cell differentiation (Parizotto et al., 2004).

miR172

miR172 promotes flowering and adult patterns of epidermal differentiation in leaves. miR172 down-regulates a subfamily of APETALA2 (AP2) transcription factor genes mostly through a translational repression mechanism and mRNA cleavage (Aukerman and Sakai, 2003; Chen, 2004; Schwab et al., 2005). There are five loci of miR172 genes in *A. thaliana*. Transcription of three miR172 genes, MIR172a, MIR172b and MIR172c are elevated steadily throughout the plant growth stages while the transcript levels of MIR172d and MIR172e are very low and do not exhibit any temporal regulation (Jung et al., 2007). SPL9 and SPL10 directly activate the transcription of MIR172b (Wu et al., 2009). The overall miR172 abundance is regulated by photoperiod via GIGANTEA(GI)-mediated miRNA processing (Jung et al., 2007). Together with miR156, miR172 is a part of regulatory feedback loops that integrate several distinct genetic pathways to guardian the stable transformation from the juvenile to the adult phase (Wang et al., 2009; Wu et al., 2009).

miR173

miR173, miR390 and miR828 are unique from other miRNAs in their ability to trigger trans-acting small interfering RNAs (tasiRNA) production from TAS1 and TAS2 (miR173), TAS3 (miR390), TAS4 (miR828). miR173 cleaves non-coding transcripts TAS1a, 1b, 1c and TAS2. The cleaved 3' transcripts are stabilized, most likely by SUPPRESSOR OF GENE SILENCING3 (SGS3), and converted to double strand RNAs (dsRNAs) by RNA-DEPENDENT RNA POLYMERASE6 (RDR6). Sequentially, dsRNAs are cleaved by DCL4 to produce phased 21-nt tasiRNAs, which target different mRNAs, including pentatricopeptide repeat (PPR) proteins (Allen et al., 2005; Yoshikawa et al., 2005). miR173 can route any non-TAS transcripts into the TAS pathway to yield phased siRNAs, as long as these non-TAS transcripts contain a miR173 target site. This unique feature distinguishes miR173 from other conventional miRNAs, such as miR159, miR167, miR169 and miR171, whose target sites are also introduced into the very same non-TAS transcripts, although the transcripts are cleaved on these sites but non phased siRNAs are detected (Felippes and Weigel, 2009; Montgomery et al., 2008b). It is currently unknown that how miR173 and its protein partners convey transcripts to TAS pathway.

miR319

Although miR319 and miR159 belong to one miRNA gene family by sharing 81% sequence identity, miR319 predominantly targets TCP (TEOSINTE BRANCHED/CYCLOIDEA/PCF) transcription factor genes while miR159 regulates MYB transcription factors. These specializations on targeting are due to both expression and sequence (although limiting) differences. (Palatnik et al., 2007)

A miR319a mutant, jaw-D, which overexpresses miR319a, has highly crinkled leaves, indicating that miR319 controls leaf morphogenesis (Palatnik et al., 2003). In jaw-D plants, 5 TCP genes are downregulated as compared to wild type plants. The miRNA-resistant forms of the TCP2 gene, when constitutively expressed in jaw-D plants, was able to rescue the leaf shape and curvature defects of jaw-D, suggesting that TCP mRNA degradation causes jaw-D phenotypes. miR319-regulated TCP transcription factors have also been shown to control biosynthesis of the hormone jasmonic acid and senescence (Schommer et al., 2008). Taken together, miR319 and its TCP targets regulate two sequential processes in leaf development: growth and senescence.

miR390/miR391

miR390 and miR391 are related miRNAs. The sequence difference between these miRNAs is at most 5 nt. Therefore, miR390 and miR391 are grouped into one miRNA family. However, most functional analyses of this family is focused on miR390, which binds to AGO7 (Montgomery et al., 2008a) and targets TAS3 non-coding transcripts to generate tasiRNAs to regulate several AUXIN RESPONSE FACTORs (including ARF3/ETTIN and ARF4) (Allen et al., 2005; Axtell et al., 2006; Fahlgren et al., 2006). Unlike miR173, which produces phased tasiRNAs from 3’ cleaved transcripts, miR390 has two target sites on TAS3. Phased tasiRNAs are produced between these two target sites. Interestingly, the 5’ target site has a central mismatch which prevents a 5’ cleavage of TAS3. This 5’ functional non-cleavage target site is essential for the TAS3 pathway as conversion to a cleavable miR390 target demolishes the production of tasiRNAs from TAS3a locus (Axtell et al., 2006).

miR393

miR393 is involved in antibacterial resistance. miR393 targets Transport Inhibitor Response 1 (TIR1) and and Auxin signaling F-Box proteins 2, 3 (AFB2, AFB3), which are receptors for auxin. Upon perception of pathogen-associated molecular patterns (PAMPs), for example, a flagellin-derived peptide, *A. thaliana* plants upregulates miR393, which triggers down-regulation of TIR1, AFB2 and AFB3. Repression of auxin signaling restricts bacteria growth, implying miRNA-mediated suppression of auxin signaling in plant resistance (Navarro et al., 2006). Conversely, some Pto DC3000 effectors counter plant defense by repression the induction of miR393. In virulent Pto DC3000–treated plants, induction of the PAMP-responsive primary transcripts of miR393a and miR393b is suppressed (Navarro et al., 2008).

miR395

miR395 targets genes of two different families, a low-affinity sulphate transporter (SULTR2;1) and three ATP sulphurylases (APS1, APS3 and APS4) (Allen et al., 2005; Jones-Rhoades and Bartel, 2004; Kawashima et al., 2009). Both of the gene families are parts of the sulphur assimilation pathway. Upon sulphur starvation (-S), 5 of 6 miR395 loci are induced in leaves and roots, except miR395f, whose expression level is beyond the limit of detection. All six miR395 genes are located on the chromosome 1, where miR395a,b,c are clustered together within ~5kb and miR395d,e,f are cluster together with in ~4kb. However, the close vicinity of these miR395 genes doesn't consequence in exact the same expression pattern, suggesting that these genes might be transcriptionally regulated differently among each other (Kawashima et al., 2009). Some of miR395 are strongly induced by -S in root tips and in both phloem companion cells of shoots and roots from young seedling. The induction of miR393 by -S is compromised in SULPHUR LIMITATION1 (SLIM1, a key transcription factor in the sulphur assimilation pathway) mutant plants, suggesting that SLIM1 directly or indirectly regulates miR395.

Table 1.3

Spatial expression of the miR395 family

name	whole plant			seedling				
	leaf	root		shoot		root		
		overall	tip	vascular tissues	phloem companion cells	tip	phloem companion cells	contex
miR395a	▓	▓	▓	▓	▓	▓	▓	▓
miR395b	▓	▓	▓	▓	▓	▓	▓	
miR395c	█	█	█	█	█	█	█	
miR395d		░	█	░	░	█	░	
miR395e	█	█	█	█	█	█	█	
miR395f								

█	high level expression
▓	moderate level expression
░	low level expression
□	not detectable

miR396

miR396 targets Growth-Regulating Factor (GRF) genes which are putative transcription factors with roles in leaf growth. miR396 genes are predominantly expressed in leaf and seedling. Transgenetic plants that constitutively overexpress miR396 have narrow leaves due to reduction in cell number and have lower densities of stomata (Liu et al., 2009). miR396 can either be induced by *P. syringae* pv. Tomato (DC3000hrcC), which trigger a robust basal defense response in *A. thaliana*, or by environmental stresses such as high-salinity, drought and cold (Liu et al., 2008a), indicating that induction of miR396 and consequently the down-regulation of GRF genes are inherent reactions of *A. thaliana* plants when the plants are challenged by environmental factors.

miR397

The miR397 family consists of two genes, miR397a and miR397b, both of which is localized on chromosome IV. miR397 is ubiquitously expressed in mature plants and in seedlings (Abdel-Ghany and Pilon, 2008). miR397 targets Copper-Containing Proteins such as laccases LAC2, LAC4, and LAC17 (Abdel-Ghany and Pilon, 2008). In response to copper deficiency, SQUAMOSA promoter binding protein–like7 (SPL7) activates several miRNAs including miR397 (Yamasaki et al., 2009). It has been proposed that the down-regulation of Copper-Containing Proteins by miRNAs allows plants to save copper for the most essential functions to deal with copper deficiency.

miR398

miR398 targets two Cu/Zn superoxide dismutases (cytosolic CSD1 and chloroplastic CSD2) (Sunkar et al., 2006; Yamasaki et al., 2007) and a subunit of the mitochondrial cytochrome c oxidase (COX5b-1)(Yamasaki et al., 2007). The miR398 level is negatively correlated with the CSD1 and CSD2 mRNA level in many adult plant tissues as well as in young, 2 weeks old seedlings, indicating that miR398 regulates the spatial and temporal expression pattern of CSD1 and CSD2 mRNAs (Sunkar et al., 2006). On one hand, miR398 expression is downregulated transcriptionally by oxidative stresses(Sunkar et al., 2006), ozone fumigation and biotic stress (P. syringae) (Jagadeeswaran et al., 2009). On the other hand, miR398 can be induced by sucrose, resulting in decreased CSD1 and CSD2 mRNA and protein accumulation (Brodersen et al., 2008; Dugas and Bartel, 2008). Intriguingly, plants expressing CSD1 and CSD2 mRNAs with engineered miR398 sites of more mismatches display increased mRNA accumulation, whereas CSD1 and CSD2 protein accumulation remain sensitive to miR398 levels, suggesting that miR398 can act as a translational repressor when target site complementarity is reduced (Dugas and Bartel, 2008).

miR399

miR399 targets the mRNA of a putative ubiquitinconjugating enzyme (UBC). Remarkably, the UBC mRNA has five miR399 target sites in the 5' UTR. All six miR399 genes are induced in the vascular tissues of roots and leaves by low-phosphate stress to repress the UBC expression level (Aung et al., 2006; Bari et al., 2006; Chiou et al., 2006; Fujii et al., 2005). The primary transcripts of miR399 are also strongly induced, ranging from 1000 to 10000 fold changes, by low inorganic phosphate (Pi) and rapidly repressed after addition of Pi (Bari et al., 2006). In grafted Arabidopsis plants, where shoots constitutively over-expressing miR399 are grafted with wild type roots, miR399 accumulates to a substantial level in roots while miR399 primary transcripts or precursors are not detectable (Lin et al., 2008; Pant et al., 2008). Taken together, these suggest that miR399 can be translocated from shoots to roots.

1.3.2 Small/Short interfering RNA

Small/Short Interfering RNAs (siRNAs) are generated by Dicer-mediated processive cleavage of double-stranded RNA (dsRNA). Like miRNAs, siRNAs are incorporated into RISC, along with Argonaute proteins. siRNAs are implicated in a variety of processes, including defense against viruses, establishment of heterochromatin, silencing of transposons and transgenes, and post-transcriptional regulation of mRNAs.

It was first reported in 1998 that a few molecules of a dsRNA can repress the expression of a gene homologous to the dsRNA in *C. elegans* (Fire et al., 1998). This phenomenon was termed RNA interference (RNAi). It became clear later when a study in plants suggested that the mechanism of dsRNA silencing might occur through the action of small ~25 nt double-stranded RNAs (Hamilton and Baulcombe, 1999). The dsRNAs were processed into ~21nt long siRNAs to induce the cleavage of a homologous transcript. Thereafter, siRNAs were widely used tools for gene silencing in mammalian cells. These siRNAs are normally referred as exogenous siRNAs, either supplied by experimental modifications or more naturally, contributed by invading viruses.

In *Arabidopsis thaliana*, recent works discovered several different types of endogenous siRNAs, including repeat associated siRNAs (rasiRNAs, including PolIV-dependent (p4)-siRNAs), natural antisense siRNAs (nat-siRNAs) and trans-acting siRNAs (tas-siRNAs).

Repeat Associated siRNAs

Endogenous siRNAs have been found by large scale cDNA clone efforts in *C. elegans* (Ambros et al., 2003b), *Drosophila* (Aravin et al., 2003) and *Arabidopsis thaliana* (Llave et al., 2002a). Different types of endogenous siRNAs were identified. Repeat associated siRNAs (rasiRNA), also occasionally referred to as Heterochromatin siRNAs (hc-siRNA) (Vaucheret, 2006), are ~24nt long, derived from repeat sequences related to transposons and are commonly found in worms, flies and plants, also in the unicellular eukaryotes such as *Trypanosoma brucei* and *Schizosaccharomyces pombe* (reviewed by (Aravin and Tuschl, 2005)). rasiRNAs are involved in establishing and maintaining heterochromatin structure and in controlling transposons.

With the advent of deep sequencing technology, scientists were able to identify millions of small RNAs in *Arabidopsis thaliana*, most of which are rasiRNAs (Kasschau et al., 2007; Lu et al., 2005; Mosher et al., 2009; Rajagopalan et al., 2006). The biogenesis of rasiRNAs most likely requires a plant specific polymerase, Pol IV, although a limited number of rasiRNAs are still presented in Pol IV mutants, which may be created from other polymerases(Mosher et al., 2009). An SNF2 domain-containing protein (CLASSY1), together with Pol IV, may be responsible for generating single strand RNA transcripts, which in turn are subject to RNA-DEPENDENT RNA POLYMERASE2 (RDR2) dependent creation of double strand RNAs (dsRNAs). The dsRNAs are then cleaved by DCL3 into 24nt rasiRNAs (Xie et al., 2004). Given their repetitive nature, rasiRNAs can usually be mapped to multiple loci on the genome, which in turn makes the identification of their true origins difficult (Figure 1.3). Thus on a whole genome scale it is currently unknown which specific repeat, transposon and other intergenic loci give rise to the large amount of small RNAs.

The surprising fact that millions of small RNAs are presented in a cell argues for the importance of small RNA function. However, in contrast to the well studied miRNAs which account for only 5% of

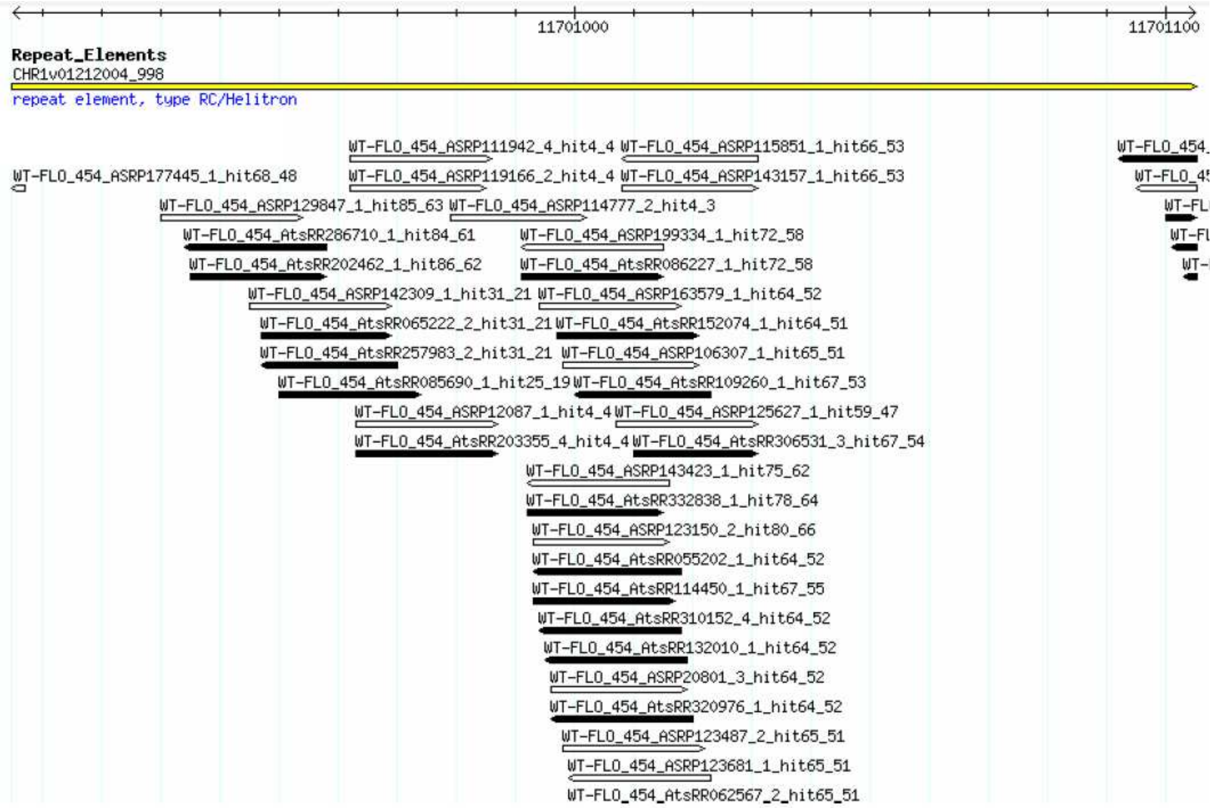

11701000
11701100
Repeat_Elements
CHR1v01212004_998
repeat element, type RC/Helitron
WT-FLO_454_ASRP111942_4_hit4_4
WT-FLO_454_ASRP115851_1_hit66_53
WT-FLO_454_
WT-FLO_454_ASRP177445_1_hit68_48
WT-FLO_454_ASRP119166_2_hit4_4
WT-FLO_454_ASRP143157_1_hit66_53
WT-FLO_45
WT-FLO_454_ASRP129847_1_hit85_63
WT-FLO_454_ASRP114777_2_hit4_3
WT-FL
WT-FLO_454_AtsRR286710_1_hit84_61
WT-FLO_454_ASRP199334_1_hit72_58
WT-Fl
WT-FLO_454_AtsRR202462_1_hit86_62
WT-FLO_454_AtsRR086227_1_hit72_58
WT-
WT-FLO_454_ASRP142309_1_hit31_21
WT-FLO_454_ASRP163579_1_hit64_52
WT-FLO_454_AtsRR065222_2_hit31_21
WT-FLO_454_AtsRR152074_1_hit64_51
WT-FLO_454_AtsRR257983_2_hit31_21
WT-FLO_454_ASRP106307_1_hit65_51
WT-FLO_454_AtsRR085690_1_hit25_19
WT-FLO_454_AtsRR109260_1_hit67_53
WT-FLO_454_ASRP12087_1_hit4_4
WT-FLO_454_ASRP125627_1_hit59_47
WT-FLO_454_AtsRR203355_4_hit4_4
WT-FLO_454_AtsRR306531_3_hit67_54
WT-FLO_454_ASRP143423_1_hit75_62
WT-FLO_454_AtsRR332838_1_hit78_64
WT-FLO_454_ASRP123150_2_hit80_66
WT-FLO_454_AtsRR055202_1_hit64_52
WT-FLO_454_AtsRR114450_1_hit67_55
WT-FLO_454_AtsRR310152_4_hit64_52
WT-FLO_454_AtsRR132010_1_hit64_52
WT-FLO_454_ASRP20801_3_hit64_52
WT-FLO_454_AtsRR320976_1_hit64_52
WT-FLO_454_ASRP123487_2_hit65_51
WT-FLO_454_ASRP123681_1_hit65_51
WT-FLO_454_AtsRR062567_2_hit65_51

Figure 1.3

A genome browser view of rasiRNAs.

A cluster of rasiRNAs map to a repeat element of Helitron on Chromosome1. Here two set of rasiRNAs are shown in two different colors, white and black. Note that these rasiRNAs can be mapped to the genome multiple times (from 4 times to 72 times). It is therefore impossible to assign their genomic origins unambiguously.

sRNA mass and less than 0.1% of the sequence complexity (Mosher et al., 2009), the majority of rasiRNAs remains to be assigned for clear biological roles. Below, I summarize several representative cases of rasiRNAs that are involved in maintaining genome integrity and regulating gene expressions.

24nt rasiRNAs with 5' terminal adenosine are preferentially incorporated into RISC complexes along with AGO4 proteins (Mi et al., 2008). These rasiRNAs are responsible for silencing transposons by guiding DNA methyltransferases together with AGO4. Additionally, AGO4 has the Asp-Asp-His catalytic motif which enables AGO4, guided by rasiRNAs, to cleave target RNA transcripts. From some loci secondary siRNAs are created, which forms a positive feedback loop to reinforce the silencing effects (Qi et al., 2006).

Most 24nt rasiRNAs are lost in the vegetative nucleus (VN) of *Arabidopsis thaliana* pollen while 21nt rasiRNAs from Athila retrotransposons are generated in VN and accumulate both in VN and sperm. Many transposable elements are reactivated in the VN which may due to the downregulation of the heterochromatin remodeler decrease in DNA methylation 1 (DDM1) and components of the rasiRNAs biogenesis machinery, such as Pol IV, RDR2 and DCL3 (Slotkin et al., 2009). It has been suggested that epigenetic reprogramming in VN may reveal intact TEs in the genome and 21nt Athila siRNAs are created in VN in order to regulate Athila activity in sperms(Slotkin et al., 2009). Intriguingly, during the seed development, rasiRNAs are created specifically from maternal chromosomes and massively accumulate in endosperm (Mosher et al., 2009). It is possible that paternal chromosomes from sperms might be restricted in their contribution to rasiRNA production due to an unfavorable chromatin structure.

Natural antisense siRNAs

The first Natural Antisense Transcript siRNA (nat-siRNA) is discovered in salt-stressed plants of *A. thaliana* (Borsani et al., 2005). Salt induced stress triggers transcription of SRO5 (SIMILAR TO RCD ONE 5), whose 3'UTR overlaps 760nt with a stress related gene P5CDH (Delta1-pyrroline-5-carboxylate dehydrogenase) on the opposite strand. Subsequentially a 24nt nat-siRNA is generated that introduces cleavage of the P5CDH transcript. The biogenesis of the 24nt nat-siRNA involves DCL2, RDR6, SGS3 and Pol IV. As a result of initial cleavage, 21nt nat-siRNAs generated by DCL1 were formed to further cleave the constitutive transcript P5CDH.

By searching public small RNA databases, Jin et al. (Jin et al., 2008) found 828 small RNAs matching 165 NAT overlapping regions, suggesting that nat-siRNA regulation of antisense genes might be an frequent gene regulatory mechanism in *A. thaliana*.

Trans-acting siRNAs

Trans-acting siRNAs (tasiRNAs) are derived from TAS (Tans-Acting siRNA) genes in *A. thaliana*. TAS genes are long ncRNAs that are cleaved by miR173 (TAS1 and TAS2), miR390 (TAS3) and miR828 (TAS4) (Adenot et al., 2006; Fahlgren et al., 2006; Garcia et al., 2006). The ends of the cleaved RNA transcripts are stabilized by SUPPRESSOR OF GENE SILENCING3 (SGS3). The stabilized cleavage products are then converted to double strand RNAs by RNA-DEPENDENT RNA POLYMERASE6 (RDR6). The dsRNAs is processed by DCL4 into in-phased 21-nt tasiRNAs.

In general, the 5’ half of tasiRNAs shows a high level of complementarity with their target mRNAs. Like plant miRNAs, tasiRNAs regulate the expression of their target mRNAs by guiding mRNA cleavage. TAS3 derived tasiRNAs target AUXIN RESPONSE FACTORs (including ARF3 and ARF4, (Adenot et al., 2006; Fahlgren et al., 2006; Garcia et al., 2006)) to regulate developmental timing and patterning.

1.4 Charting small RNA Landscape

In higher eukaryotes, protein-coding genes occupy only a mere fraction of a genome, in the case of *Arabidopsis thaliana*, exons take up only 35% of the genome. The quantitative majority output of a genome is RNA transcripts that have no coding potential. These RNA transcripts include long non-coding RNAs and small RNAs, as discussed in here. Most importantly, what we have functionally characterized so far might just be the tip of an iceberg. It is commonly accepted now that a large amount of Transcripts of Unknown Functions (TUFs) exist in eukaryotic cells (see the RNA review issue of *Cell*, 20 February, 2009 Volume 136, Issue 4). TUFs are usually located in unannotated regions of genomes or antisense to known protein coding genes. They can be polyadenylated, nonpolyadenylated or bimorphic, many of which have little coding potential. This new picture of eukaryotic transcriptome justifies that more research efforts are need to elucidate the function of these non-coding RNA transcripts. In the following parts of the thesis, I will present my work on small RNAs, one of the major classes functional non-coding RNAs.

Reference:

Abdel-Ghany, S.E., and Pilon, M. (2008). MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis. J Biol Chem *283*, 15932-15945.

Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouche, N., Gasciolli, V., and Vaucheret, H. (2006). DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. Curr Biol *16*, 927-932.

Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M., and Hatzigeorgiou, A.G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics *25*, 3049-3055.

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell *121*, 207-221.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet *36*, 1282-1290.

Allen, R.S., Li, J., Stahle, M.I., Dubroue, A., Gubler, F., and Millar, A.A. (2007). Genetic analysis reveals functional redundancy and the major target genes of the Arabidopsis miR159 family. Proc Natl Acad Sci U S A *104*, 16371-16376.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M*., et al.* (2003a). A uniform system for microRNA annotation. RNA *9*, 277-279.

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003b). MicroRNAs and other tiny endogenous RNAs in C. elegans. Curr Biol *13*, 807-818.

Aravin, A., and Tuschl, T. (2005). Identification and characterization of small RNAs involved in RNA silencing. FEBS Lett *579*, 5830-5840.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev Cell *5*, 337-350.

Aukerman, M.J., and 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.

Aung, K., Lin, S.I., Wu, C.C., Huang, Y.T., Su, C.L., and Chiou, T.J. (2006). pho2, a phosphate overaccumulator, is caused by a nonsense mutation in a microRNA399 target gene. Plant Physiol *141*, 1000-1011.

Axtell, M.J., and Bartel, D.P. (2005). Antiquity of microRNAs and their targets in land plants. Plant Cell *17*, 1658-1673.

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell *127*, 565-577.

Baker, C.C., Sieber, P., Wellmer, F., and Meyerowitz, E.M. (2005). The early extra petals1 mutant uncovers a role for microRNA miR164c in regulating petal number in Arabidopsis. Curr Biol *15*, 303-315.

Bari, R., Datt Pant, B., Stitt, M., and Scheible, W.R. (2006). PHO2, microRNA399, and PHR1 define a phosphate-signaling pathway in plants. Plant Physiol *141*, 988-999.

Berezikov, E., Chung, W.J., Willis, J., Cuppen, E., and Lai, E.C. (2007). Mammalian mirtron genes. Mol Cell *28*, 328-336.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R., and Zhu, J.K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. Cell *123*, 1279-1291.

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

Brodersen, P., and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. Nat Rev Mol Cell Biol *10*, 141-148.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science *303*, 2022-2025.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479-486.

Chiou, T.J., Aung, K., Lin, S.I., Wu, C.C., Chiang, S.F., and Su, C.L. (2006). Regulation of phosphate homeostasis by MicroRNA in Arabidopsis. Plant Cell *18*, 412-421.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol *13*, 849-851.

Didiano, D., and Hobert, O. (2008). Molecular architecture of a miRNA-regulated 3' UTR. RNA *14*, 1297-1317.

Dugas, D.V., and Bartel, B. (2008). Sucrose induction of Arabidopsis miR398 represses two Cu/Zn superoxide dismutases. Plant Mol Biol *67*, 403-417.

Easow, G., Teleman, A.A., and Cohen, S.M. (2007). Isolation of microRNA targets by miRNP immunopurification. RNA *13*, 1198-1204.

Fahlgren, N., and Carrington, J.C. (2010). miRNA Target Prediction in Plants. Methods Mol Biol *592*, 51-57.

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

Fahlgren, N., Montgomery, T.A., Howell, M.D., Allen, E., Dvorak, S.K., Alexander, A.L., and Carrington, J.C. (2006). Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. Curr Biol *16*, 939-944.

Fang, Y., and Spector, D.L. (2007). Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living Arabidopsis plants. Curr Biol *17*, 818-823.

Felippes, F.F., and Weigel, D. (2009). Triggering the formation of tasiRNAs in Arabidopsis thaliana: the role of microRNA miR173. EMBO Rep *10*, 264-270.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Fu, Y., Xu, L., Xu, B., Yang, L., Ling, Q., Wang, H., and Huang, H. (2007). Genetic interactions between leaf polarity-controlling genes and ASYMMETRIC LEAVES1 and 2 in Arabidopsis leaf patterning. Plant Cell Physiol *48*, 724-735.

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

Garcia, D., Collier, S.A., Byrne, M.E., and Martienssen, R.A. (2006). Specification of leaf polarity in Arabidopsis via the trans-acting siRNA pathway. Curr Biol *16*, 933-938.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R*., et al.* (2009). Rfam: updates to the RNA families database. Nucleic Acids Res *37*, D136-140.

Ghildiyal, M., Xu, J., Seitz, H., Weng, Z., and Zamore, P.D. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA *16*, 43-56.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.

Guo, H.S., Xie, Q., Fei, J.F., and Chua, N.H. (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. Plant Cell *17*, 1376-1386.

Gutierrez, L., Bussell, J.D., Pacurar, D.I., Schwambach, J., Pacurar, M., and Bellini, C. (2009). Phenotypic plasticity of adventitious rooting in Arabidopsis is controlled by complex regulation of AUXIN RESPONSE FACTOR transcripts and microRNA abundance. Plant Cell *21*, 3119-3132.

Haberer G.*, Wang Y.*, Mayer K. The noncoding landscape of the genome of Arabidopsis thaliana (2010), a book chapter of "Genetics and Genomics of the Brassicaceae" in the series "Plant Genetics and Genomics – Crops and Models", Springer, * as the joint first authors

Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950-952.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nat Methods *5*, 813-819.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887-901.

Hirsch, J., Lefort, V., Vankersschaver, M., Boualem, A., Lucas, A., Thermes, C., d'Aubenton-Carafa, Y., and Crespi, M. (2006). Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts. Plant Physiol *140*, 1192-1204.

Hong, X., Hammell, M., Ambros, V., and Cohen, S.M. (2009). Immunopurification of Ago1 miRNPs selects for a distinct class of microRNA targets. Proc Natl Acad Sci U S A *106*, 15085-15090.

Hsieh, L.C., Lin, S.I., Shih, A.C., Chen, J.W., Lin, W.Y., Tseng, C.Y., Li, W.H., and Chiou, T.J. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol *151*, 2120-2132.

Jagadeeswaran, G., Saini, A., and Sunkar, R. (2009). Biotic and abiotic stress down-regulate miR398 expression in Arabidopsis. Planta *229*, 1009-1014.

Jin, H., Vacic, V., Girke, T., Lonardi, S., and Zhu, J.K. (2008). Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. BMC Mol Biol *9*, 6.

Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell *14*, 787-799.

Jung, J.H., and Park, C.M. (2007). MIR166/165 genes exhibit dynamic expression patterns in regulating shoot apical meristem and floral development in Arabidopsis. Planta *225*, 1327-1338.

Jung, J.H., Seo, Y.H., Seo, P.J., Reyes, J.L., Yun, J., Chua, N.H., and Park, C.M. (2007). The GIGANTEA-regulated microRNA172 mediates photoperiodic flowering independent of CONSTANS in Arabidopsis. Plant Cell *19*, 2736-2748.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L*., et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol *5*, e57.

Kawashima, C.G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y.N., Saito, K., Takahashi, H., and Dalmay, T. (2009). Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. Plant J *57*, 313-321.

Kickhoefer, V.A., Emre, N., Stephen, A.G., Poderycki, M.J., and Rome, L.H. (2003). Identification of conserved vault RNA expression elements and a non-expressed mouse vault RNA gene. Gene *309*, 65-70.

Kim, J.H., Woo, H.R., Kim, J., Lim, P.O., Lee, I.C., Choi, S.H., Hwang, D., and Nam, H.G. (2009). Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in Arabidopsis. Science *323*, 1053-1057.

Kim, Y.K., and Kim, V.N. (2007). Processing of intronic microRNAs. EMBO J *26*, 775-783.

Laufs, P., Peaucelle, A., Morin, H., and Traas, J. (2004). MicroRNA regulation of the CUC genes is required for boundary size control in Arabidopsis meristems. Development *131*, 4311-4322.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051-4060.

Li, L., Wang, X., Sasidharan, R., Stolc, V., Deng, W., He, H., Korbel, J., Chen, X., Tongprasit, W., Ronald, P*., et al.* (2007). Global identification and characterization of transcriptionally active regions in the rice genome. PLoS One *2*, e294.

Li, W.X., Oono, Y., Zhu, J., He, X.J., Wu, J.M., Iida, K., Lu, X.Y., Cui, X., Jin, H., and Zhu, J.K. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. Plant Cell *20*, 2238-2251.

Lin, S.I., Chiang, S.F., Lin, W.Y., Chen, J.W., Tseng, C.Y., Wu, P.C., and Chiou, T.J. (2008). Regulatory network of microRNA399 and PHO2 by systemic signaling. Plant Physiol *147*, 732-746.

Liu, D., Song, Y., Chen, Z., and Yu, D. (2009). Ectopic expression of miR396 suppresses GRF target gene expression and alters leaf growth in Arabidopsis. Physiol Plant *136*, 223-236.

Liu, H.H., Tian, X., Li, Y.J., Wu, C.A., and Zheng, C.C. (2008a). Microarray-based analysis of stress-regulated microRNAs in Arabidopsis thaliana. RNA *14*, 836-843.

Liu, P.P., Montgomery, T.A., Fahlgren, N., Kasschau, K.D., Nonogaki, H., and Carrington, J.C. (2007). Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J *52*, 133-146.

Liu, Q., Yao, X., Pi, L., Wang, H., Cui, X., and Huang, H. (2008b). The ARGONAUTE10 gene modulates shoot apical meristem maintenance and leaf polarity establishment by repressing miR165/166 in Arabidopsis. Plant J.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. (2002a). Endogenous and silencing-associated small RNAs in plants. Plant Cell *14*, 1605-1619.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002b). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053-2056.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005). Elucidation of the small RNA component of the transcriptome. Science *309*, 1567-1569.

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. Genome Res *16*, 510-519.

Mallory, A.C., Bartel, D.P., and Bartel, B. (2005). MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. Plant Cell *17*, 1360-1375.

Mallory, A.C., Dugas, D.V., Bartel, D.P., and Bartel, B. (2004a). MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. Curr Biol *14*, 1035-1046.

Mallory, A.C., Reinhart, B.J., Jones-Rhoades, M.W., Tang, G., Zamore, P.D., Barton, M.K., and Bartel, D.P. (2004b). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. EMBO J *23*, 3356-3364.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet *9*, 387-402.

Mattick, J.S., and Makunin, I.V. (2006). Non-coding RNA. Hum Mol Genet *15 Suppl 1*, R17-29.

Megraw, M., Baev, V., Rusinov, V., Jensen, S.T., Kalantidis, K., and Hatzigeorgiou, A.G. (2006). MicroRNA promoter element discovery in Arabidopsis. RNA *12*, 1612-1619.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat Rev Genet *10*, 155-159.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J.*, et al.* (2008). Criteria for annotation of plant MicroRNAs. Plant Cell *20*, 3186-3190.

Meyers, B.C., Souret, F.F., Lu, C., and Green, P.J. (2006). Sweating the small stuff: microRNA discovery in plants. Curr Opin Biotechnol *17*, 139-146.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C.*, et al.* (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116-127.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008a). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell *133*, 128-141.

Montgomery, T.A., Yoo, S.J., Fahlgren, N., Gilbert, S.D., Howell, M.D., Sullivan, C.M., Alexander, A., Nguyen, G., Allen, E., Ahn, J.H.*, et al.* (2008b). AGO1-miR173 complex initiates phased siRNA formation in plants. Proc Natl Acad Sci U S A *105*, 20055-20062.

Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. Nature *460*, 283-286.

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., Voinnet, O., and Jones, J.D. (2006). A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. Science *312*, 436-439.

Navarro, L., Jay, F., Nomura, K., He, S.Y., and Voinnet, O. (2008). Suppression of the microRNA pathway by bacterial effector proteins. Science *321*, 964-967.

Nikovics, K., Blein, T., Peaucelle, A., Ishida, T., Morin, H., Aida, M., and Laufs, P. (2006). The balance between the MIR164A and CUC2 genes controls leaf margin serration in Arabidopsis. Plant Cell *18*, 2929-2945.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89-100.

Okamura, K., Liu, N., and Lai, E.C. (2009). Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. Mol Cell *36*, 431-444.

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. Nature *425*, 257-263.

Palatnik, J.F., Wollmann, H., Schommer, C., Schwab, R., Boisbouvier, J., Rodriguez, R., Warthmann, N., Allen, E., Dezulian, T., Huson, D.*, et al.* (2007). Sequence and expression differences underlie functional specialization of Arabidopsis microRNAs miR159 and miR319. Dev Cell *13*, 115-125.

Pant, B.D., Buhtz, A., Kehr, J., and Scheible, W.R. (2008). MicroRNA399 is a long-distance signal for the regulation of plant phosphate homeostasis. Plant J *53*, 731-738.

Pant, B.D., Musialak-Lange, M., Nuc, P., May, P., Buhtz, A., Kehr, J., Walther, D., and Scheible, W.R. (2009). Identification of nutrient-responsive Arabidopsis and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing. Plant Physiol *150*, 1541-1555.

Parizotto, E.A., Dunoyer, P., Rahm, N., Himber, C., and Voinnet, O. (2004). In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. Genes Dev *18*, 2237-2242.

Peaucelle, A., Morin, H., Traas, J., and Laufs, P. (2007). Plants expressing a miR164-resistant CUC2 gene reveal the importance of post-meristematic maintenance of phyllotaxy in Arabidopsis. Development *134*, 1045-1050.

Prasanth, K.V., and Spector, D.L. (2007). Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev *21*, 11-42.

Qi, Y., He, X., Wang, X.J., Kohany, O., Jurka, J., and Hannon, G.J. (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. Nature *443*, 1008-1012.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev *20*, 3407-3425.

Raman, S., Greb, T., Peaucelle, A., Blein, T., Laufs, P., and Theres, K. (2008). Interplay of miR164, CUP-SHAPED COTYLEDON genes and LATERAL SUPPRESSOR controls axillary meristem formation in Arabidopsis thaliana. Plant J *55*, 65-76.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. Genes Dev *16*, 1616-1626.

Reyes, J.L., and Chua, N.H. (2007). ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. Plant J *49*, 592-606.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Schommer, C., Palatnik, J.F., Aggarwal, P., Chetelat, A., Cubas, P., Farmer, E.E., Nath, U., and Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. PLoS Biol *6*, e230.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of microRNAs on the plant transcriptome. Dev Cell *8*, 517-527.

Sieber, P., Wellmer, F., Gheyselinck, J., Riechmann, J.L., and Meyerowitz, E.M. (2007). Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. Development *134*, 1051-1060.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzic, M., Becker, J.D., Feijo, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell *136*, 461-472.

Song, L., Han, M.H., Lesicka, J., and Fedoroff, N. (2007). Arabidopsis primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. Proc Natl Acad Sci U S A *104*, 5437-5442.

Stolc, V., Samanta, M.P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D.C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S*., et al.* (2005). Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. Proc Natl Acad Sci U S A *102*, 4453-4458.

Storz, G. (2002). An expanding universe of noncoding RNAs. Science *296*, 1260-1263.

Sunkar, R., and Jagadeeswaran, G. (2008). In silico identification of conserved microRNAs in large number of diverse plant species. BMC Plant Biol *8*, 37.

Sunkar, R., Kapoor, A., and Zhu, J.K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. Plant Cell *18*, 2051-2065.

Szymanski, M., and Barciszewski, J. (2002). Beyond the proteome: non-coding regulatory RNAs. Genome Biol *3*, reviews0005.

Ueno, Y., Ishikawa, T., Watanabe, K., Terakura, S., Iwakawa, H., Okada, K., Machida, C., and Machida, Y. (2007). Histone deacetylases and ASYMMETRIC LEAVES2 are involved in the establishment of polarity in leaves of Arabidopsis. Plant Cell *19*, 445-457.

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes Dev *20*, 759-771.

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., Mallory, A.C., and Bartel, D.P. (2006). AGO1 homeostasis entails coexpression of MIR168 and AGO1 and preferential stabilization of miR168 by AGO1. Mol Cell *22*, 129-136.

Vaucheret, H., Vazquez, F., Crete, P., and Bartel, D.P. (2004). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. Genes Dev *18*, 1187-1197.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell *136*, 669-687.

Wang, J.-W., Wang, L.-J., Mao, Y.-B., Cai, W.-J., Xue, H.-W., and Chen, X.-Y. (2005). Control of root cap formation by MicroRNA-targeted auxin response factors in Arabidopsis. Plant Cell *17*, 2204--2216.

Wang, J.W., Czech, B., and Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana. Cell *138*, 738-749.

Wang, Y., Hindemitt, T., and Mayer, K.F. (2006). Significant sequence similarities in promoters and precursors of Arabidopsis thaliana non-conserved microRNAs. Bioinformatics *22*, 2585-2589.

Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell *135*, 635-648.

Wierzbicki, A.T., Ream, T.S., Haag, J.R., and Pikaard, C.S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet *41*, 630-634.

Wu, G., Park, M.Y., Conway, S.R., Wang, J.W., Weigel, D., and Poethig, R.S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. Cell *138*, 750-759.

Wu, M.F., Tian, Q., and Reed, J.W. (2006). Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development *133*, 4211-4218.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. Plant Physiol *138*, 2145-2154.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol *2*, E104.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M.*, et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science *302*, 842-846.

Yamasaki, H., Abdel-Ghany, S.E., Cohu, C.M., Kobayashi, Y., Shikanai, T., and Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in Arabidopsis. J Biol Chem *282*, 16369-16378.

Yamasaki, H., Hayashi, M., Fukazawa, M., Kobayashi, Y., and Shikanai, T. (2009). SQUAMOSA Promoter Binding Protein-Like7 Is a Central Regulator for Copper Homeostasis in Arabidopsis. Plant Cell *21*, 347-361.

Yoshikawa, M., Peragine, A., Park, M.Y., and Poethig, R.S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev *19*, 2164-2175.

Zhou, G.K., Kubo, M., Zhong, R., Demura, T., and Ye, Z.H. (2007). Overexpression of miR165 affects apical meristem formation, organ polarity establishment and vascular development in Arabidopsis. Plant Cell Physiol *48*, 391-404.

Chapter 2 MicroRNA Genes

Three important miRNA research topics are: first, how to identify a miRNA gene in an organism. Second, how are miRNA genes regulated? And third, how do miRNA genes interact with other genes? Hereafter, I will address the first topic from a genomic aspect. I will use *Sorghum bicolor* as a model to demonstrate how to annotate miRNA genes in a newly sequenced genome (Paterson et al. 2009). Then, I will present my work on miRNA promoters, which includes newly evolved miRNA genes in *Arabidopsis thaliana* and their targets(Wang et al. 2006). The promoter analysis of miRNA genes is related to the second topic of the miRNA research. The third topic will be discussed in the chapter 4 of the thesis.

2.1 Introduction

Like protein coding genes, miRNA genes are located on DNA and can be transcribed by either RNA polymerase II (Pol II) or III (Pol III), depending on their genomic locations. Many miRNAs are transcribed by Pol II, including most plant miRNA genes, animal intronic/exonic miRNA genes and some of viral miRNA genes. Some miRNA genes are proximal to Alu repeats or tRNAs, which are transcribed together by Pol III. These two different transcriptional regulatory models of miRNA genes are discussed separately in the following sections.

2.1.1 miRNA genes transcribed by RNA polymerase II

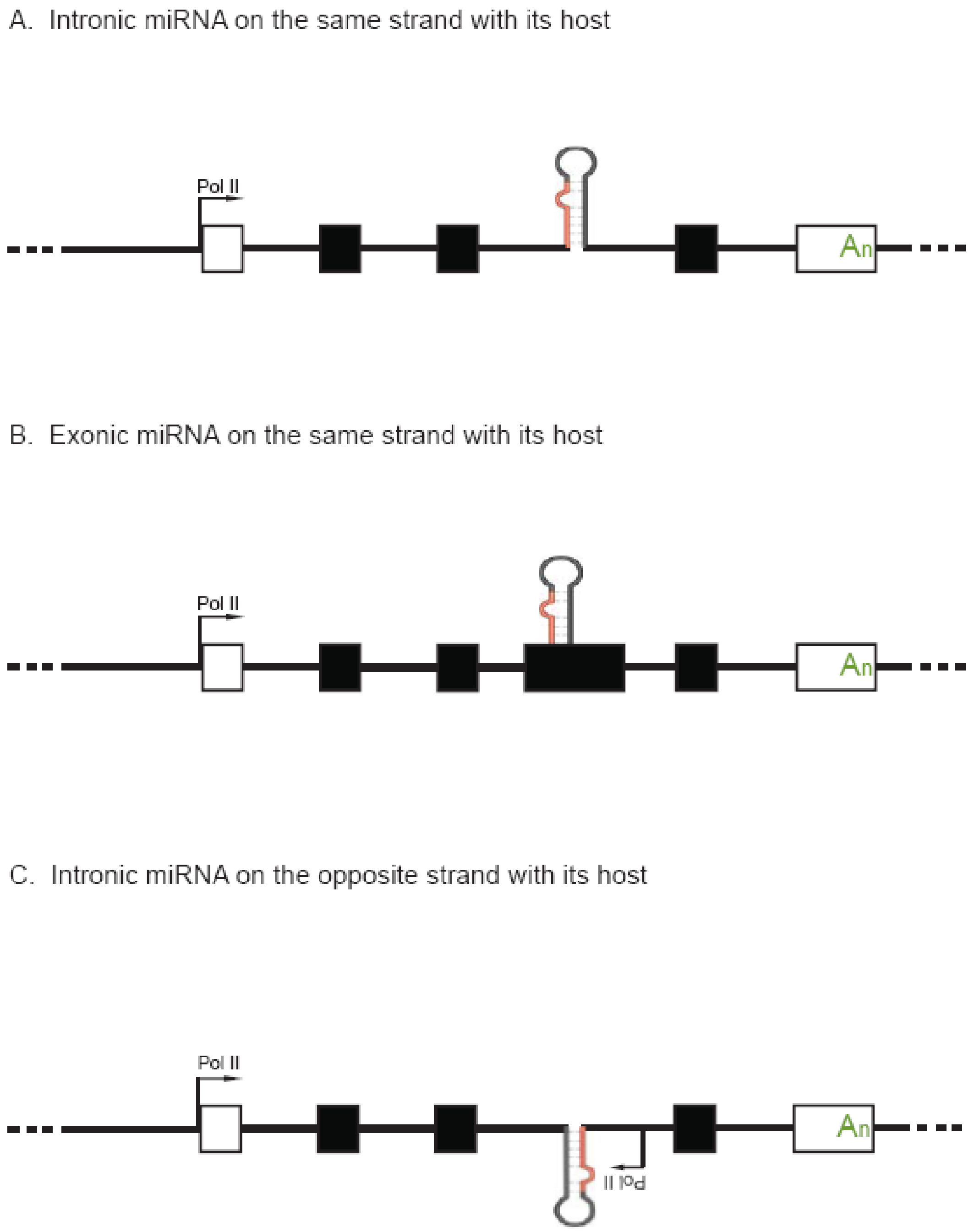
In plant, most of miRNA genes are located in the intergenic regions while many animal miRNA genes are located in introns, exons or untranslated regions (UTRs) of protein-coding gene or non-coding RNA genes (Table 2.1). More than half of human miRNA genes are found in the introns of protein coding genes. These miRNA genes are normally transcribed together with their hosts by Pol II (Figure 2.1A and B, (Lee et al., 2004)) except intronic miRNA genes located on the opposite strand (Figure 2.1C). In the former scenario, miRNA genes are most likely co-regulated with their hosts. They share the same transcriptional machinery. Intronic miRNA genes are likely to be processed by Drosha without interfering the alternative splicing event of their host. In the later scenario, the miRNA gene is transcribed by Pol II with its own upstream promoter and other regulatory elements (Figure 2.1C). For example, mouse miRNAs mmu-miR-1-2 and mmu-miR-133-a1 are located on the opposite strand between the 12th and 13th exons of Mind bomb1. The primary transcript of this miRNA cluster is 2.5 kilobase and is regulated by its own SRF-dependent enhancer in heart and MyoD-dependent enhancer in skeletal muscle(Zhao et al., 2007b).

Table 2.1

Genomic locations of miRNA genes in different organisms

Species	intergenic	exon		intron		3'UTR	
		+	-	+	-	+	-
Caenorhabditis elegans	94	0	0	17	18	3	0
Drosophila melanogaster	55	0	0	20	3	0	0
Homo sapiens	201	9	1	200	32	27	5
Mus musculus	214	5	0	116	17	16	9
Danio rerio	287	0	0	41	8	0	1
Arabidopsis thaliana	182	0	0	0	0	2	0
Oryza sativa	242	0	0	0	0	0	0

Intergenic miRNA genes or miRNA gene clusters normally have their own promoters and regulatory elements (Figure 2.2). Many plant miRNA genes are independent transcriptional units, which may have multiple transcription starting sites, exon-intron structure (as shown in the next section by a sorghum miRNA gene, sbi-MIR444) and alternative 3’ ends. Many animal miRNA genes form clusters on a genome (Figure 2.2B), either on intergenic regions or on introns. The largest human miRNA cluster consists of 40 miRNAs. Clustered miRNA genes are generally believed to be transcribed together. However, there are a few cases where cluster miRNA genes have different expression patterns, for example, in *Arabidopsis thaliana*, ath-miR395abc cluster and ath-miR395def cluster (Table 1.3), suggesting that within a miRNA gene cluster, individual miRNA gene might be regulated differently.



A. Intronic miRNA on the same strand with its host

B. Exonic miRNA on the same strand with its host

C. Intronic miRNA on the opposite strand with its host

Figure 2.1

Schematic representation of transcriptional regulatory models of intronic and exonic miRNAs

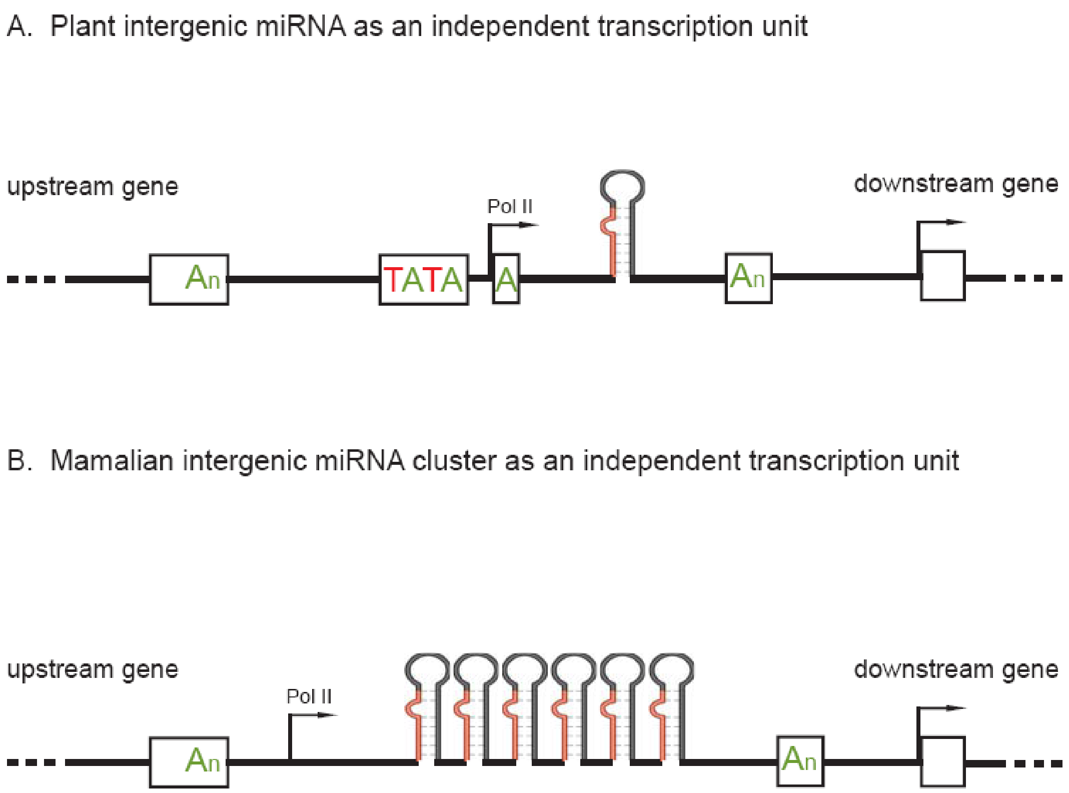

A. Plant intergenic miRNA as an independent transcription unit
upstream gene
An
TATA
A
Pol II
An
downstream gene
B. Mamalian intergenic miRNA cluster as an independent transcription unit
upstream gene
An
Pol II
An
downstream gene

Figure 2.2

Schematic representation of transcriptional regulatory models of intergenic miRNAs.

2.1.2 miRNA genes transcribed by RNA polymerase III

Some newly identified human miRNA genes (~50) are interspersed among Alu repeats (Figure 2.3A and (Borchert et al., 2006)). These miRNA genes were shown to be transcribed together with Alu repeats by Pol III, a transcriptional scenario which is very similar to the transcription of γ herpesvirus 68 (MHV68) miRNA genes (Figure 2.3B and (Pfeffer et al., 2005)). Some MHV68 miRNAs are located immediately downstream of tRNA sequences (Figure 2.3B), suggesting that the pre-miRNAs are transcribed by pol III. The Pol III tRNA promoter of a conventional tRNA gene consists of an A box (positions +8 to +19) and a B box (positions +52 to +62), followed by a run of oligo(T) as a terminator signal (Figure 2.3B). The discovery that miRNA genes are transcribed by Pol III by using promoters of nearby repetitive elements indicates an important role for repetitive elements in miRNA gene expression(Borchert et al., 2006; Pfeffer et al., 2005).

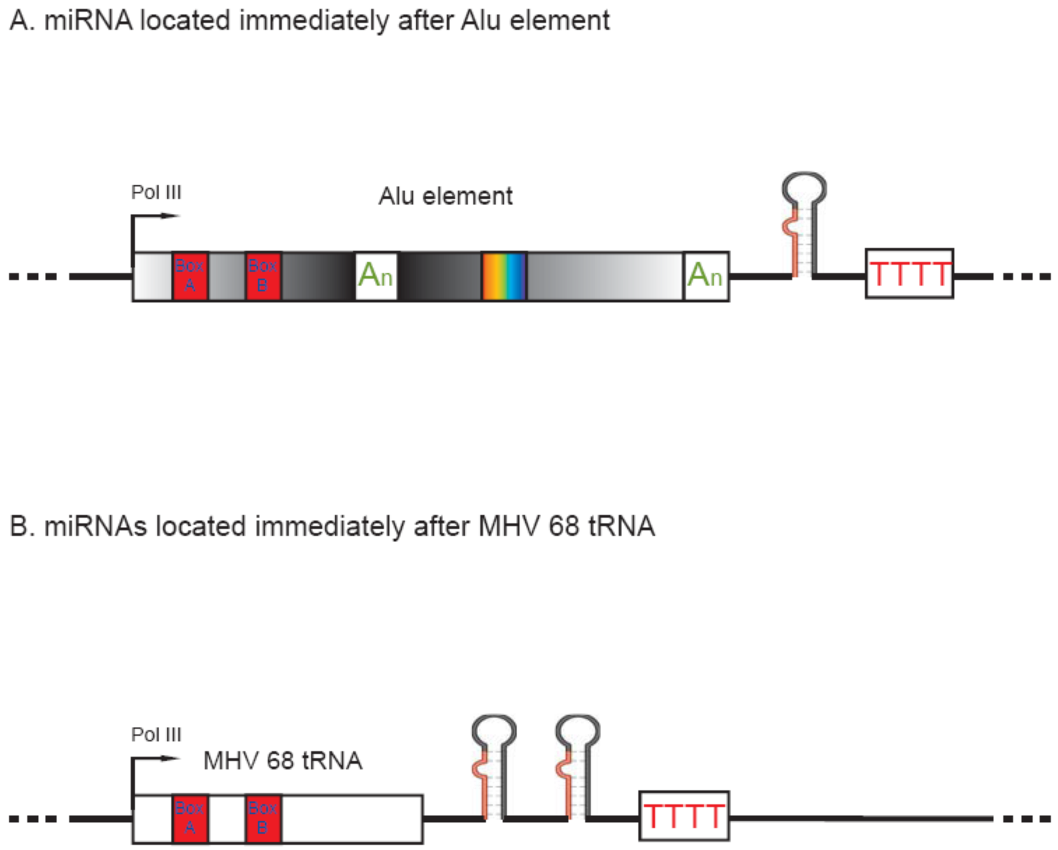

A. miRNA located immediately after Alu element
Pol III
Alu element
Box A
Box B
An
An
TTTT
B. miRNAs located immediately after MHV 68 tRNA
Pol III
MHV 68 tRNA
Box A
Box B
TTTT

Figure 2.3

Schematic representation of transcriptional regulatory models of miRNAs, which are transcribed by Pol III.

The knowledge of genomic locations of known miRNA genes from different organisms can be useful for finding new miRNA genes and studying their functions in other organisms. In the next section, I will present a case study on how to annotate new microRNA genes in a newly sequenced genome. I used a comparative genomics approach to annotate 149 miRNA genes on the newly sequenced *Sorghum bicolor* genome.

2.2 *Sorghum bicolor* miRNAs

Sorghum bicolor was sequenced as a model organism for C4 photosynthesis and tropical grasses. It is the drought and heat tolerance of sorghum that motivated plant scientists to decipher the *Sorghum bicolor* genome(Paterson et al., 2009). Compare to the first sequenced crop, rice, sorghum is able to grow in hot and semidry places. Intriguingly, when rice plants are under drought stress, one rice miRNA gene, osa-MIR169g is induced. I found that in *Sorghum bicolor*, the orthlogous miRNA gene of osa-MIR169g has been duplicated(Paterson et al., 2009). Furthermore, miRNA 169 targets members of nuclear factor Y

(NF-Y) family, which were related to drought tolerance in *Arabidopsis thaliana* and maize. In this part of the thesis, I will use *Sorghum bicolor* as a model to demonstrate how to annotate new miRNA genes on a newly sequenced genome and how the understanding of miRNA functions can be integrated to the physiology of the newly sequenced organism.

2.2.1 Annotating miRNA genes on the *Sorghum bicolor* genome

I set out to find sorghum miRNAs and their targets in the *Sorghum bicolor* genome, the second fully sequenced grass genome. I took the advantage of a comparative genomic approach to annotate sorghum miRNAs by using Rice miRNAs. This is a well accepted approach to annotate miRNAs in a newly sequenced genome. This approach has been widely tested in animal miRNA annotation for evolutionarily close related species, for example, two worm genomes, 12 fly genomes and mammalian genomes. Rice and Sorghum were derived from a common ancestor 70 million years ago(Paterson et al., 2009). I reasoned that it was appropriate to use Rice miRNA for the annotation task. Many Rice miRNA genes have been validated by deep sequencing experiments and some other experimental approaches. It was decided not to use the whole plant miRNA sets to avoid unnecessary inclusion of some mis-annotated miRNA genes from other plant species.

Mapping known sorghum miRNA genes

I mapped a known set of 72 Sorghum miRNA precursors (miRBase release 11.0) to the *Sorghum bicolor* genome. These 72 Sorghum miRNA precursors were annotated on incomplete Sorghum genomic sequences by using *Arabidopsis thaliana* and *Oryza sativa* miRNAs. 67 Sorghum precursors can be anchored on the *Sorghum bicolor* genome. sbi-MIR169e,h, 172d, 394b and 395c are discarded as authentic sorghum miRNA genes due to reasons stated below. sbi-MIR169e doesn't have a significant match to any genomic regions of the sorghum genome. sbi-MIR169h has almost identical sequence as sbi-MIR169g. But sbi-MIR169g can be perfectly anchored on the genome while 169h can't. sbi-MIR172d can't be fully mapped to the genome. sbi-MIR394b is a wrong annotation of miR394. It has almost exact same sequence as sbi-MIR394a, apart from three insertions of n in the sequence. sbi-MIR395c is 208 nt long. It completely overlaps with a shorter 395 miRNA precursor, sbi-MIR395d, which is 104 nt long. It looks like that 395c and d are the same miRNA gene but with different precursor length. Therefore, I decide to take the shorter precursor as the correct annotation for the miRNA gene 395 at this locus.

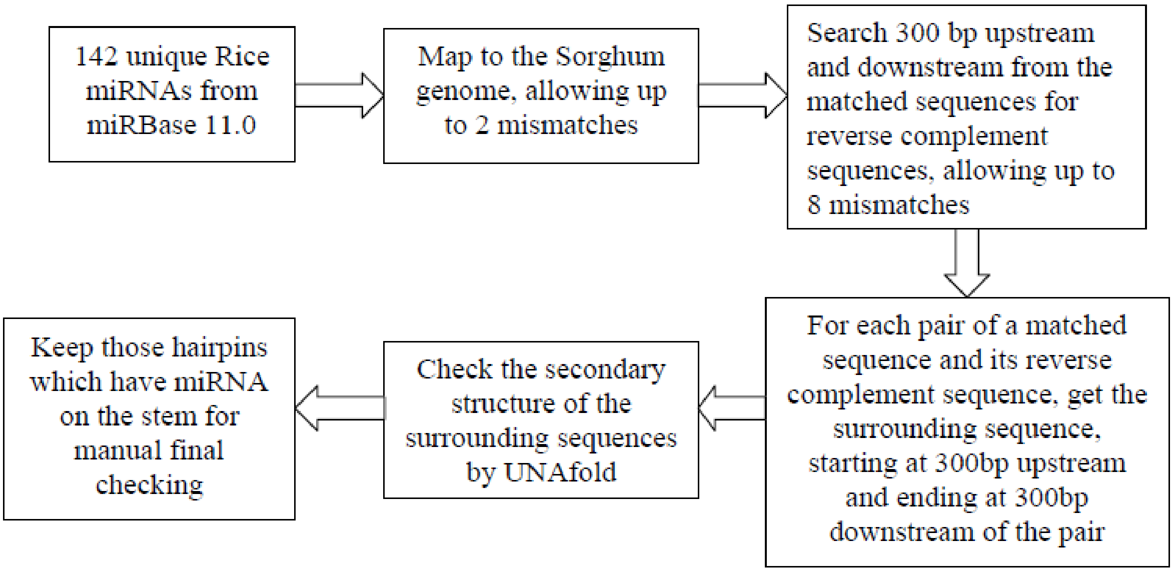

142 unique Rice miRNAs from miRBase 11.0
Map to the Sorghum genome, allowing up to 2 mismatches
Search 300 bp upstream and downstream from the matched sequences for reverse complement sequences, allowing up to 8 mismatches
For each pair of a matched sequence and its reverse complement sequence, get the surrounding sequence, starting at 300bp upstream and ending at 300bp downstream of the pair
Check the secondary structure of the surrounding sequences by UNAfold
Keep those hairpins which have miRNA on the stem for manual final checking

Figure 2.4

miRNA annotation pipeline for finding miRNA genes on the *Sorghum bicolor* genome.

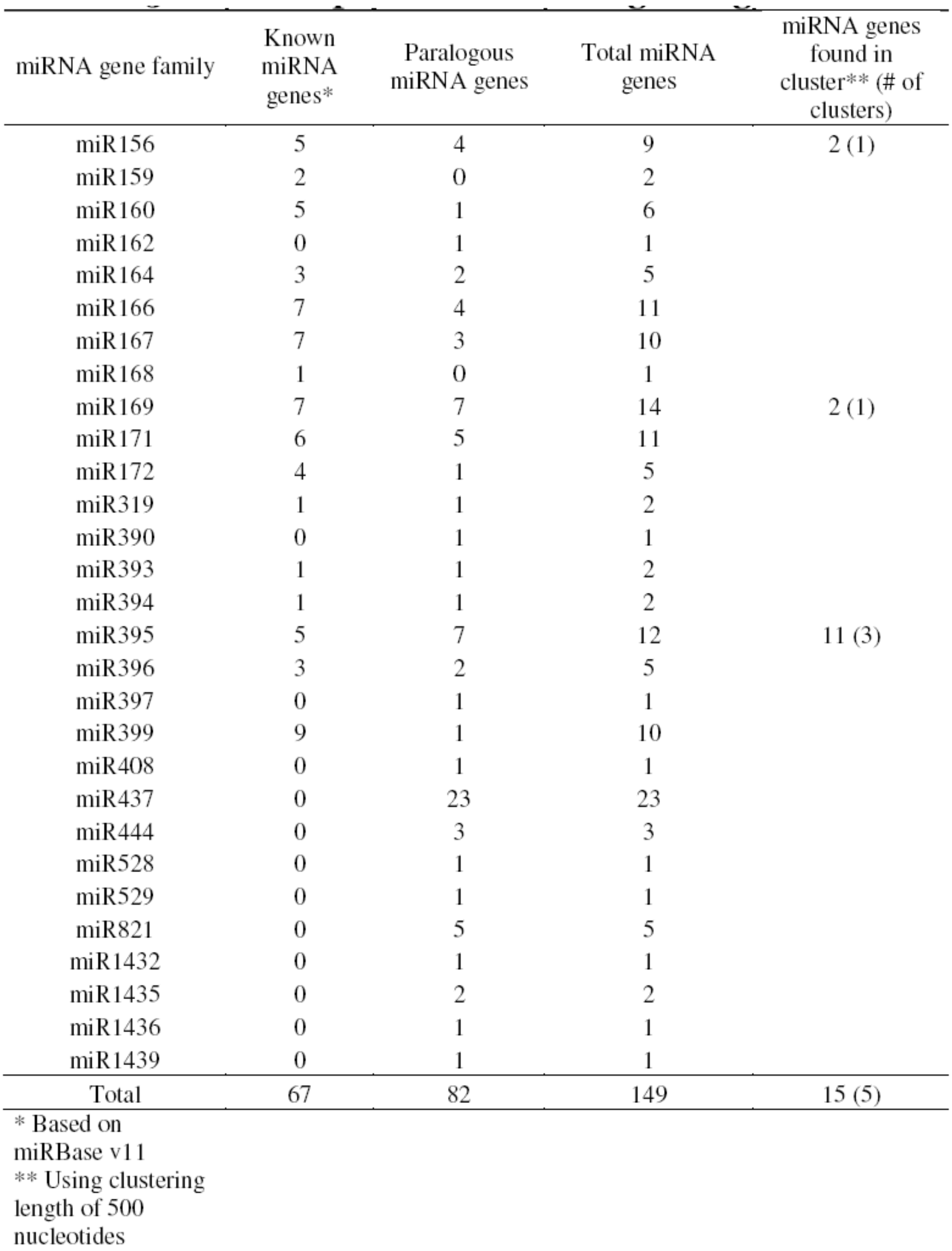
Annotating new sorghum miRNA genes

I used *Oryza sativa* miRNAs (miRBase release 11.0) to identify sorghum miRNA orthologs. This is a well accepted strategy for annotating miRNA genes on a newly sequenced genome (Meyers et al., 2008). In miRBase release 11.0, there are 353 Oryza sativa miRNA genes (miRNA precursors) and 373 mature miRNAs. The number of unique rice miRNA mature sequences is 142. I developed a computational pipeline for miRNA gene annotation (Figure 2.4). The key structural constrain for a miRNA precursor is the hairpin-like structure of the sequence, where a miRNA sequence should sit on a stem of a hairpin structure. I mapped 142 unique rice miRNA sequences to the sorghum genome and checked the surrounding sequences for the proper secondary structure. All the candidate structures were finally subjected to a manual verification. The annotated Sorghum bicolor miRNAs are listed in Table 2.2.

Exon skipping in sbi-miR444 gene

Lu Chen *et al*. (Lu et al., 2008) have previously identified three sorghum natural antisense miRNAs (nat-miRNAs). These miRNAs are located at the antisense strand of their target genes (MADS-box transcription factors) and contain long introns in their precursor sequences. Three sbi-miR444 precursors

Table 2.2 miRNA genes present in the sorghum genome



miRNA gene family	Known miRNA genes*	Paralogous miRNA genes	Total miRNA genes	miRNA genes found in cluster** (# of clusters)
miR156	5	4	9	2 (1)
miR159	2	0	2	
miR160	5	1	6	
miR162	0	1	1	
miR164	3	2	5	
miR166	7	4	11	
miR167	7	3	10	
miR168	1	0	1	
miR169	7	7	14	2 (1)
miR171	6	5	11	
miR172	4	1	5	
miR319	1	1	2	
miR390	0	1	1	
miR393	1	1	2	
miR394	1	1	2	
miR395	5	7	12	11 (3)
miR396	3	2	5	
miR397	0	1	1	
miR399	9	1	10	
miR408	0	1	1	
miR437	0	23	23	
miR444	0	3	3	
miR528	0	1	1	
miR529	0	1	1	
miR821	0	5	5	
miR1432	0	1	1	
miR1435	0	2	2	
miR1436	0	1	1	
miR1439	0	1	1	
Total	67	82	149	15 (5)

* Based on miRBase v11
** Using clustering length of 500 nucleotides

were mapped to the Sorghum genome. Interestingly, one miR444 locus (sbi-miR444.p1 and p2) produces two precursors due to exon skipping (Figure 2.5). sbi-miR444.p1 and p2 are located on chromosome 4. sbi-miR444.p1 has two exons which are shared by sbi-miR444.p2. Additionally, sbi-miR444.p2 has a short exon (33bp), located between the other two exons (Figure 2.5B). Both p1 and p2 form canonical microRNA hairpin structures (Figure 2.5A), but the additional exon makes sbi-miR444.p2 to have lower free energy (-92.40 kcal/mol) as compared to that of p1 (-78.60 kcal/mol), suggesting a more stable secondary structure.

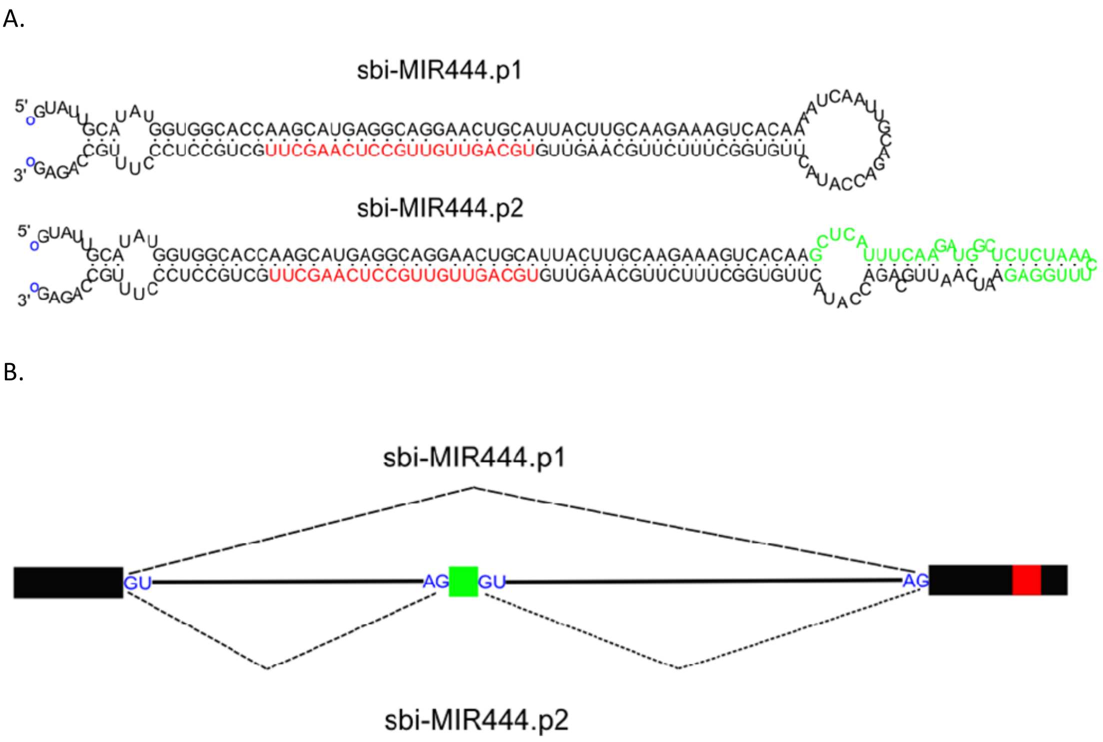

A.
sbi-MIR444.p1
5'
3'
GGUGGCACCAAGCAUGAGGCAGGAACUGCAUUACUUGCAAGAAAGUCACAA
CCUCCGUCGUUCGAACUCCGUUGUUGACGUGUUGAACGUUCUUUCGGUGUU
sbi-MIR444.p2
5'
3'
GGUGGCACCAAGCAUGAGGCAGGAACUGCAUUACUUGCAAGAAAGUCACAAG
CCUCCGUCGUUCGAACUCCGUUGUUGACGUGUUGAACGUUCUUUCGGUGUUC
B.
sbi-MIR444.p1
GU
AG
GU
AG
sbi-MIR444.p2

Figure 2.5

sbi-MIR444 gene

A. Secondary structure of sbi-MIR444.p1 and sbi-MIR444.p2 gene, the mature sbi-miR444 sequences are marked with red color and the exon is marked with green color.

B. Gene structures of sbi-MIR444.p1 and sbi-MIR444.p2 gene, alternative splicing sites are indicated.

Conservation of miRNA genes in Sorghum and Rice

miRNA precursor forms a stable hairpin secondary structure where miRNA is normally found on the stems of the hairpin. The functional region of a miRNA precursor is the miRNA sequence and sometimes also the miRNA* sequence. It has been observed that the conserved part of miRNA precursor largely overlaps with miRNA:miRNA* region while the loop region of the hairpin is not conserved when *Arabidopsis thaliana* miRNAs were compared with those of rice(Jones-Rhoades and Bartel, 2004). The evolutionary distance between Arabidopsis and rice is between 120 and 200 million years(Wolfe et al., 1989) while the distance between sorghum and rice is 70 million years(Paterson et al., 2009). Furthermore, unlike the comparison with Arabidopsis, which is a dicot, the comparison between sorghum and rice is within monocots. Therefore, it is interesting to check the miRNA precursor sequence conservation between orthologous miRNA genes of rice and sorghum.

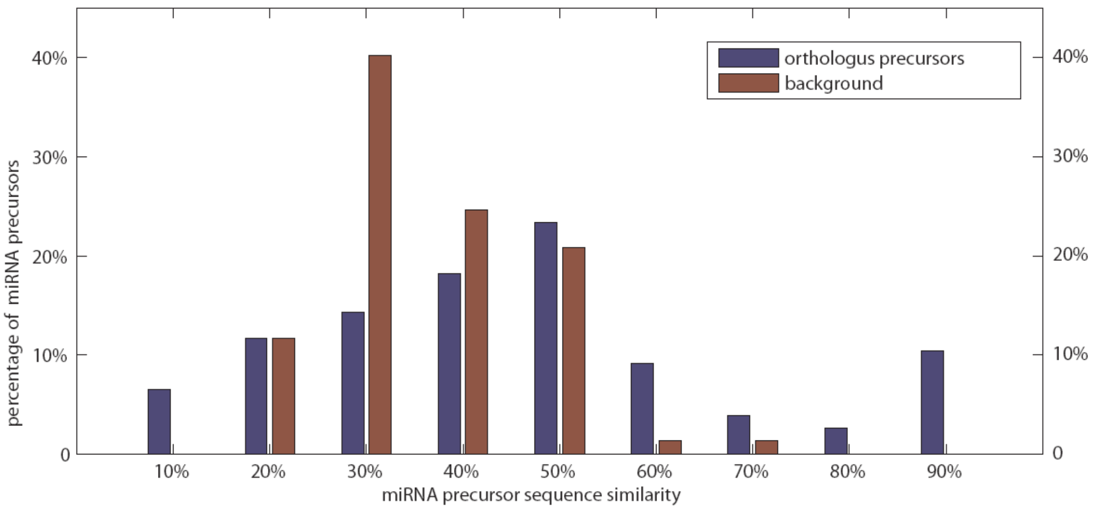


Figure2.6

Comparison of precursor similarity between orthologous miRNA precursors and a background model. Precursor similarity is measured by alignable nucleotides between two precursors.

I used BLAST to compare the sequence similarities between the orthologous pairs (Figure 2.6). The precursor sequence similarity of orthologous pairs is obviously extended beyond the miRNA:miRNA* regions (Figure 2.6). The background model is calculated by comparing similarities between non orthologous miRNA genes within the same family. Interestingly, there are five orthologous pairs which have no more than 10 percent precursor sequence similarity. Comparing them with the background model, I find these five orthologous pairs have some mismatches in the miRNA:miRNA* region. On the other hand, there are eight orthologous pairs which have more than 90 percent precursor sequence similarity. As an example, I plot the sequence alignment of sbi-MIR171.p1 and osa-MIR171a in Figure 2.7. It is striking that two miRNA precursors are so much conserved at sequence level (Figure 2.7B). There is only one nucleotide difference on the stem regions. When I align the 500 nt orthologous sequences, centered at the miRNA precursors, I find near perfect similarity at the center (Figure 2.7A)

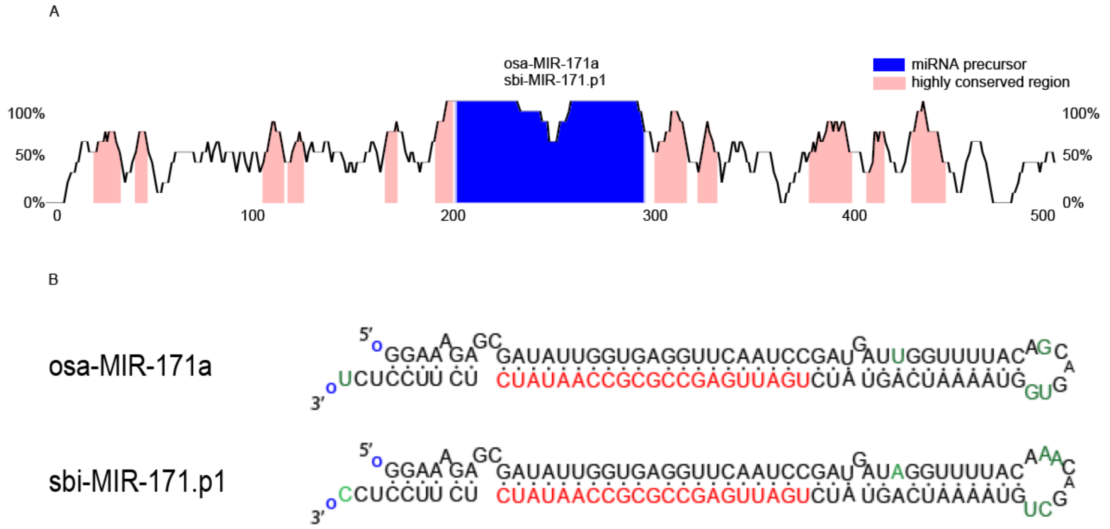


Figure 2.7

Near perfect conservation of osa-MIR171a and sbi-MIR171.p1

A. Vista alignment of 500bp sequences, centered on osa-MIR-171a and sbi-MIR171.p1. The precursor sequences are marked by blue and some other highly conserved regions are mark by pink.

B. Secondary structures of osa-MIR171a and sbi-MIR171.p1. The mature miRNA sequences are indicated in red and non-conserved nucleotides are indicated in green.

where the precursors are located, the similarity drops obviously outside the precursor regions. There are also highly conserved regions outside the precursors, which might host functional *cis* elements related to miRNA transcriptional control.

2.2.2 miR169 targets drought related NF-Y gene family

In a survey of drought related miRNA genes, osa-MIR169g was the only miRNA gene of miR169 family which was induced by drought and the induction of osa-MIR169g was more prominent in roots than in shoots(Zhao et al., 2007a). Intriguingly, when I checked the orthologous relationship between *Oryza sativa* and *Sorghum bicolor* miRNA genes, I found the orthologous Sorghum miRNA gene of osa-MIR169g was duplicated into two miRNA genes, sbi-MIR169c and sbi-MIR169d (Figure2.8). The location of miRNA genes are on the syntenic blocks of *Sorghum bicolor* chromosome 6 and *Oryza sativa* chromosome 4. The protein coding genes surrounding osa-MIR169g are in a block collinear with Sorghum ones which surrounded sbi-MIR169c and sbi-MIR169d (Figure 2.8, (Paterson et al., 2009)).

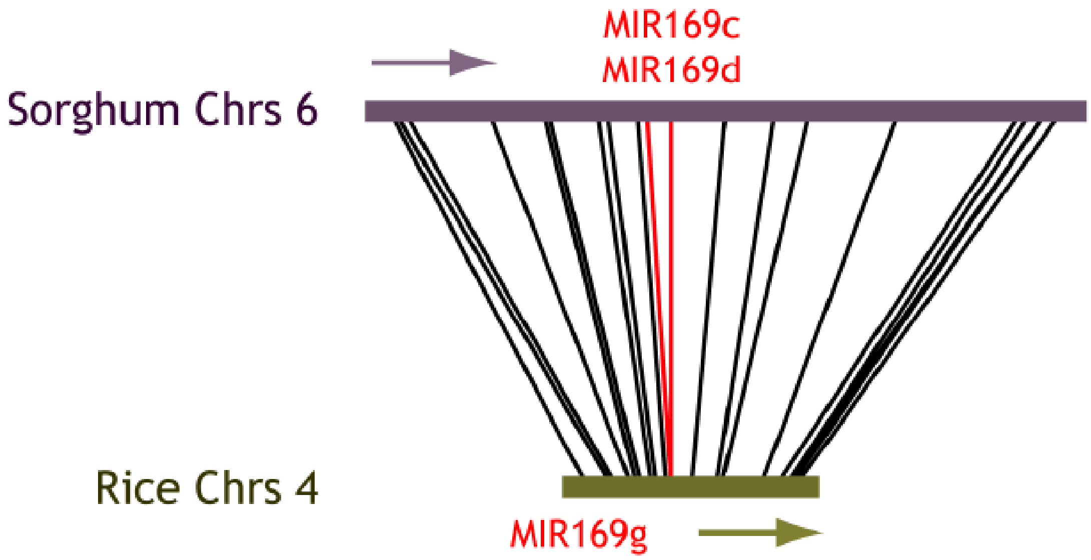


Figure 2.8

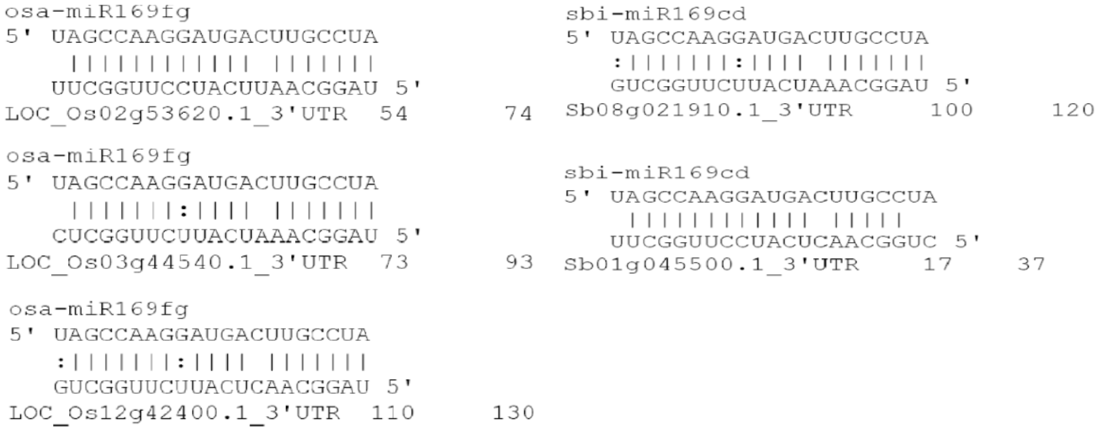
Orthologous relationships between osa-MIR169g and sbi-MIR169c,d

miRNA genes are marked with red color on the chromosomes. The solid black lines indicate orthologous relationships between protein coding genes. The solid red lines indicate orthologous relationships between miRNA genes.

In silico analysis of miR169 targets indicates that nuclear transcription factor Y genes (NF-Y) have near perfect complementarity to miR169 (Figure 2.9). In Rice, LOC_Os02g53620 is NF-Y subunit A-3, LOC_Os03g44540 is NF-Y subunit A-10, LOC_Os12g42400 is NF-Y subunit A-2. In Sorghum, Sb08021910 is similar to NF-Y subunit B family protein and Sb01g045500 is similar to RAPB (Rice HAP B subunit protein) (Yao et al., 1999), which also belongs to NF-Y family(Nelson et al., 2007).

NF-Y B subunits were implicated in drought tolerance in *Arabidopsis thaliana* and *Zea mays* (Nelson et al., 2007). Arabidopsis NF-Y B1 constitutively expressed plants exhibited less severe silting than wild type plants under drought stress. Better drought tolerance was also observed in transgenic maize plants of Maize NF-Y B2, which is most closely related to Arabidopsis NF-Y B1(Nelson et al., 2007).

Arabidopsis NF-Y A5 subunit is regulated transcriptionally and posttranscriptionally to promote drought resistance(Li et al., 2008). NF-Y A5 was highly induced by drought and a part of induction is through transcriptional regulation. Furthermore, NF-Y A5 contains a target site of miR169. Controversially, unlike in Rice, Arabidopsis miR169 was downregulated by drought stress(Li et al., 2008) which contributed in accumulation of NF-Y A5 in drought stressed Arabidopsis plants.



```
osa-miR169fg
5' UAGCCAAGGAUGACUUGCCUA
    |||||||||||| |||||||
   UUCGGUUCCUACUUAACGGAU 5'
LOC_Os02g53620.1_3'UTR  54      74

osa-miR169fg
5' UAGCCAAGGAUGACUUGCCUA
    |||||||:|||| |||||||
   CUCGGUUCUUACUAAACGGAU 5'
LOC_Os03g44540.1_3'UTR  73      93

osa-miR169fg
5' UAGCCAAGGAUGACUUGCCUA
   :|||||||:|||| |||||||
   GUCGGUUCUUACUCAACGGAU 5'
LOC_Os12g42400.1_3'UTR  110     130

sbi-miR169cd
5' UAGCCAAGGAUGACUUGCCUA
   :|||||||:|||| |||||||
   GUCGGUUCUUACUAAACGGAU 5'
Sb08g021910.1_3'UTR     100     120

sbi-miR169cd
5' UAGCCAAGGAUGACUUGCCUA
     ||||||||||||| |||||
   UUCGGUUCCUACUCAACGGUC 5'
Sb01g045500.1_3'UTR     17     37
```

Figure 2.9

miR169 targets in *Sorghum bicolor* and *Oryza sativa*.

One interesting observation is that all miR169 target sites on NF-Y genes are located in 3' UTRs (Figure 2.9 and (Li et al., 2008)), which are not common for plant miRNA target sites. Most of target sites of plant microRNAs are in the coding regions of genes[1].

Taken together, these data suggest that miR169 and NF-Y family members are part of complex regulatory network which that contribute to drought resistance in both monocots and dicots.

In this part of thesis, the topic of miRNA conservation was discussed. Many miRNA genes are very well conserved in plant kingdom (Table 1.1). Additionally, it has been shown that miRNA precursors can also be very well conserved between Rice and Sorghum (Figure 2.6 and 2.7). Apart from conserved miRNA genes, the evolution of miRNA genes is a dynamic process. New miRNA genes are constantly evolving. What are the properties of newly evolved miRNA genes? This is the topic of the next section.

[1] There is a good reason for this, which might be related to the evolution of miRNA genes in plants. More details will be presented in the next section.

2.3 Significant sequence similarities in promoters and precursors of *Arabidopsis thaliana* non-conserved microRNAs[2]

Some plant microRNAs have been shown to be *de novo* generated by inverted duplication from their target genes(Allen et al., 2004). Subsequent duplication events potentially generate multigene microRNA families(Maher et al., 2006). Here I provide supportive evidence for the inverted duplication model of plant microRNA evolution. First, the precursors of four *Arabidopsis thaliana* non-conserved microRNA families, miR157, miR158, miR165, miR405 and miR447 share nearly identical nucleotide sequences throughout the whole miRNA precursor between the family members. The extent and degree of sequence conservation is suggestive of recent evolutionary duplication events. Furthermore, sequence similarities are not restricted to the transcribed part but extend into the promoter regions. Thus the duplication event most likely included the promoter regions as well. Conserved elements in upstream regions of miR163 and its targets were also detected. This implies that the inverted duplication of target genes has included the promoters of the target genes. Sequence conservation within promoters of miRNA families as well as between miRNA and its potential progenitor gene can be exploited for understanding the regulation of microRNA genes.

2.3.1 Introduction

Currently 117 *Arabidopsis thaliana* miRNAs have been identified (miRBase, release 8.0, (Griffiths-Jones, 2004)). These can be classified into 46 miRNA families. 21 are represented by single genes and 25 are defined as multigene families. For 13 (61.9%) single miRNA genes (MIR161, MIR163, MIR170, MIR173, MIR400-404, and MIR406-408) no homologous counterpart in *Oryza sativa* has been described. In contrast for only five (20%) multigene families (MIR157, MIR158, MIR165, MIR405 and MIR447) no homologous counterparts in *Oryza sativa* have been reported. I will refer to the latter five miRNA families as nonconserved miRNA families throughout this section.

Allen *et al.* (2004) investigated the sequence similarity between 91 *Arabidopsis* miRNAs and their targets.

[2] This work has been published in 2006. I intentionally keep the originally text from the paper unchanged so that I can reevaluate this paper at the last part of the thesis.

They found that MIR163 and MIR161 have exhaustive similarities in the two foldback arms to some of their target genes (Allen et al., 2004). Additionally, these two miRNA genes are located proximal to some of their target genes. Based on these observations an evolutionary scenario involving an inverted duplication event and active expansion of target gene families has been suggested. After the *de novo* creation of the miRNA, it may have evolved into a multigene miRNA family by subsequent duplication events. The initial duplication event may have also included the primordial gene promoter (Allen et al., 2004). This scenario has been described as the Inverted Duplication Model for miRNA gene evolution in plants (Allen et al., 2004).

Majority of plant miRNAs have extensive similarities to their targets and most of these target sites are in the coding regions. It is unlikely that a plant miRNA target site can evolve by random mutation on the coding region of a gene since there is high selective pressure to maintain the correct protein sequences. The Inverted Duplication model can neatly explain how new miRNAs and corresponding target sites can evolve simultaneously. Much less plant miRNA target sites are found in untranslated regions (UTRs). Two explanations are offered for consideration. First, on average, UTRs are much shorter than coding regions in plants. Assuming that miRNA target sites can be uniformly distributed on a gene, the chance to find a target site on UTRs is much lower than on coding regions. Second, the evolutionary constraint on coding regions is higher than on UTRs. Therefore, once a target site appears on a coding region, it is conserved better than a target site on an UTR.

Gene duplication has played a vital role in the evolution of genes(Hurles, 2004). It has been shown that duplication events most likely do not only involve the transcribed part of the respective loci but also include at least part of the promoter regions(Haberer et al., 2004; Hurles, 2004). Regulatory diversification through acquisition or derivation of distinct control elements may trigger divergence of regulatory specificity between closely related family members(Haberer et al., 2004; Xie et al., 2005). Appropriate candidates for the study of the effects of duplication and retention of conserved elements within promoter regions are evolutionary young miRNA families. Having no orthologous counterparts in monocotyledonous plants, as this is the case for nonconserved *Arabidopsis* miRNA families, is a prime indicator for recently arisen miRNA families.

The next section reports findings on the degree of sequence conservation within nonconserved *Arabidopsis* multigene miRNA families, namely miR157, miR158, miR165, miR405 and miR447. In addition to the analysis of the transcribed region I analyze the promoter regions of the miRNA families for sequence similarity and potentially conserved sequence motifs. I find extensive sequence similarities and short highly conserved sequence motifs between the promoters of miRNA family members. Furthermore, I showed that the 5' upstream sequence of *Arabidopsis* MIR163 and its target promoters are partially conserved. These findings are supportive for the inverted duplication model for miRNA gene evolution. These results and analysis pave the way to a comprehensive *in silico* assisted study of miRNA promoters and their relationship to their respective target genes.

2.3.2 miRNA and target promoter analysis

For promoter comparison and analysis I take 1000 base pair (bp) upstream from the miRNA precursors. The extent of the miRNA's primary transcript is not known for all miRNA families under investigation. To ensure the comparison of equivalent parts of the upstream sequences the primary transcript is excluded from the analysis. Known primary transcript sequences have been marked in upper case letters and are displayed in dark blue in the graphical output (see Figure 2.12) in contrast to the remaining part of the upstream sequences (lower case letters and light blue). This facilitates clarity in display of the studied sequences.

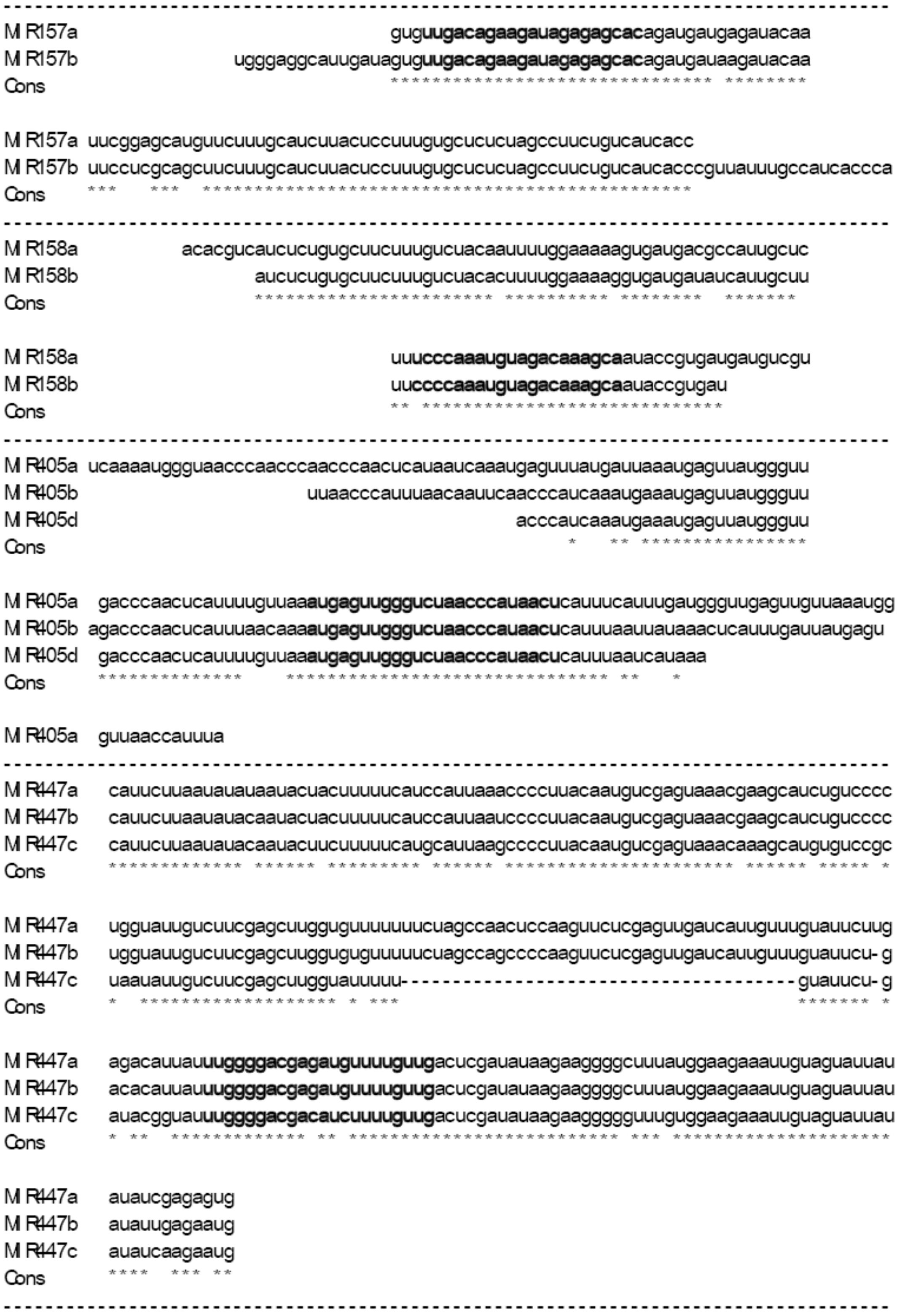
I also take 1000 bp upstream from the assumed transcription start site as a default promoter sequence of target genes. As an evidence for the position of the transcription start site I used full length cDNAs annotated in the MIPS *Arabidopsis thaliana* database (MAtDB; Schoof *et al.*, 2002).

MiRNA gene coordinates were obtained from miRBase, release 8.0 (Griffiths-Jones, 2004) (http://microrna.sanger.ac.uk/sequences/index.shtml). The upstream sequences of miRNAs and corresponding target genes were retrieved from MAtDB (Schoof et al., 2002) (http://mips.gsf.de/proj/thal/db/index.html).

Similarity comparisons for the sequences under investigation were carried out using DIALIGN (Morgenstern, 2004), a local alignment tool. Alignment algorithms are not able to detect conserved regions after rearrangement events (e.g. inversions or extensive deletions and insertions). Furthermore these algorithms are developed to align long and less conserved sequence regions rather than to detect short and highly conserved motifs. To complement this limitation I also used MotifSampler (Thijs et al., 2001), a motif discovery algorithm based on Gibbs Sampling, which does not rely on colinearity of conserved sequences and is able to identify very short but highly conserved sequence motifs. MotifSampler is a stochastic algorithm and therefore results for different runs of the program may vary but tend to cluster at conserved sequence motifs. For that reason I carried out 50 repeated runs of MotifSampler for each analysis and visualize the results as the percentage of runs a nucleotide has been detected as being part of a conserved motif (Figure 2.12). Both analyses with DIALIGN, and MotifSampler respectively were performed using a modified version of CREDO (Hindemitt and Mayer, 2005)(http:// mips .gsf.de/proj/regulomips/) a web-based tool that integrates the analysis and results of these two programs amongst others. Detailed results and parameters are available online on http:// mips.gsf.de/proj/regulomips/microRNA.

2.3.3 *Arabidopsis thaliana* non-conserved miRNA families

Evolutionary conservation of miRNAs has been exhaustively exploited for the *in silico* prediction and detection of plant miRNAs (Bonnet et al., 2004; Jones-Rhoades and Bartel, 2004; Wang et al., 2004). Although the majority of *Arabidopsis* miRNAs are conserved in *Oryza sativa*, there are individual *Arabidopsis* miRNAs (e.g. MIR163) and even miRNA families (MIR157, MIR158, MIR165, MIR405, and MIR447) for which orthologs in *Oryza sativa* have not been reported. These nonconserved multigene miRNA families may have evolved and expanded after the separation of the monocotelydoneous and dicotelydoneous plant lineages. Alternatively these individual miRNAs or parts of them have not been retained within *Oryza sativa* or the monocotelydoneous lineage in general.



```
-----------------------------------------------------------------------------
MIR157a                  guguugacagaagauagagagcacagaugaugagauacaa
MIR157b  ugggaggcauugauaguguugacagaagauagagagcacagaugauaagauacaa
Cons                    ******************************* ********

MIR157a  uucggagcauguucuuugcaucuuacuccuuugugcucucuagccuucugucaucacc
MIR157b  uuccucgcagcuucuuugcaucuuacuccuuugugcucucuagccuucugucaucacccguuauuugccaucaccca
Cons     ***   ***  ***********************************************
-----------------------------------------------------------------------------
MIR158a  acacgucaucucugugcuucuuugucuacaauuuuggaaaaagugaugacgccauugcuc
MIR158b          aucucugugcuucuuugucuacacuuuuggaaaggugaugauaucauugcuu
Cons             ********************** ********** ********   *******

MIR158a                      uuucccaaauguagacaaagcaauaccgugaugaugucgu
MIR158b                      uuccccaaauguagacaaagcaauaccgugau
Cons                         ** *****************************
-----------------------------------------------------------------------------
MIR405a  ucaaaauggguaacccaacccaacccaacucauaaucaaaugaguuuaugauuaaaugaguuauggguu
MIR405b                       uuaacccauuuaacaauucaacccaucaaaugaaaugaguuauggguu
MIR405d                                           acccaucaaaugaaaugaguuauggguu
Cons                                                   *   ** ****************

MIR405a   gacccaacucauuuuguuaaaugaguugggucuaacccauaacucauuucauuugauggguugaguuguuaaaugg
MIR405b  agacccaacucauuuaacaaaaugaguugggucuaacccauaacucauuuaauuauaaacucauuugauuaugagu
MIR405d   gacccaacucauuuuguuaaaugaguugggucuaacccauaacucauuuaaucauaaa
Cons      **************    ******************************* **   *

MIR405a   guuaaccauuua
-----------------------------------------------------------------------------
MIR447a  cauucuuaauauauaauacuacuuuuucauccauuaaaccccuuacaaugucgaguaaacgaagcaucugucccc
MIR447b  cauucuuaauauacaauacuacuuuuucauccauuaaucccc uuacaaugucgaguaaacgaagcaucugucccc
MIR447c  cauucuuaauauacaauacuucuuuuucaugcauuaagccccuuacaaugucgaguaaacaaagcauguguccgc
Cons     ************* ****** ********* ****** ********************** ****** ***** *

MIR447a  ugguauugucuucgagcuugguguuuuuuucuagccaacuccaaguucucgaguugaucauuguuuguauucuug
MIR447b  ugguauugucuucgagcuuggugugu uuuucuagccagccccaaguucucgaguugaucauuguuuguauucu-g
MIR447c  uaauauugucuucgagcuugguauuuuu-------------------------------------guauucu-g
Cons     *  ******************* * ***                                      ******* *

MIR447a  agacauuauuuggggacgagauguuuuguugacucgauauaagaaggggcuuuauggaagaaauuguaguauuau
MIR447b  acacauuauuuggggacgagauguuuuguugacucgauauaagaaggggcuuuauggaagaaauuguaguauuau
MIR447c  auacgguauuuggggacgacaucuuuuguugacucgauauaagaagggggguuuguggaagaaauuguaguauuau
Cons     * **  ************* ** ************************* *** **********************

MIR447a  auaucgagagug
MIR447b  auauugagaaug
MIR447c  auaucaagaaug
Cons     ****  *** **
-----------------------------------------------------------------------------
```

Figure 2.10

Sequence alignments of unconserved miRNA precursors.

Each row represents one sequence. miRNAs are indicated in bold font and the conserved section among miRNA families is given.

Since non-functional portions of miRNA precursors are obviously under different evolutionary pressure as compared to the miRNA:miRNA* part, It is hypothesized that for miRNA gene families the degree of conservation within the whole miRNA precursor is an indicator for the age of the constituting duplication event. To test this, I analyzed the degree of conservation between precursors of members of the nonconserved multigene families by multiple sequence alignments. The alignments of MIR157a, b, MIR158a, b, MIR405a, b, d and MIR447a, b, c are shown in Figure 2.10. In contrast to what has been found for conserved miRNA precursors (Jones-Rhoades and Bartel, 2004), sequence identities among the individual miRNA gene family members are not restricted to the miRNA:miRNA* parts but extend strikingly to the remaining stem-loop regions. Conserved miRNA families usually contain diverged stem loop sequences beside the conserved miRNA:miRNA* fraction (Jones-Rhoades and Bartel, 2004). The exhaustive similarity between individual miRNA family members and extension of this similarity beyond the conserved miRNA:miRNA* fraction support the hypothesis that nonconserved miRNA multigene families arose and evolved by recent duplication events.

For protein coding genes there is increasing evidence that duplication events are not restricted to transcribed regions only but also involve flanking promoter regions (Haberer et al., 2004; Hurles, 2004). Thus I investigated whether sequence similarities extend to the promoters of the nonconserved miRNA multigene families. Upstream promoter sequences from MIR157, MIR158, MIR165, MIR405 and MIR447 precursors were subjected to a similarity and motif analysis by applying DIALIGN and MotifSampler. The results of MIR157 are shown in Figure 2.12.

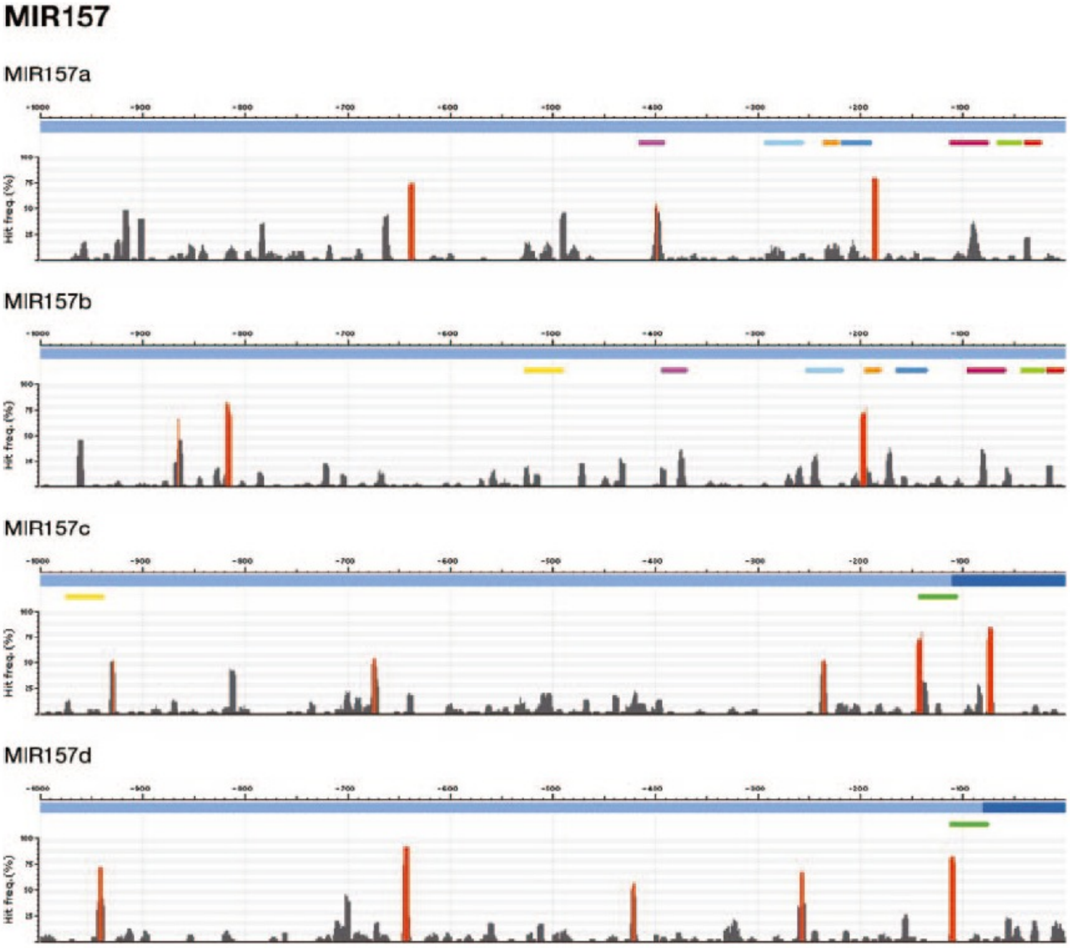


Figure 2.12

Sequence similarities within the promoter regions of MIR157.

The light blue horizontal bars represent upstream sequences. Colored bars represent conserved regions detected by DiAlign. The peaks represent the number of hits per nucleotide normalized by the number of total hits per sequence, which were calculated 50 times by MotifSampler. A peak is plotted in red when the height of the peak is >50%.

Within the miR157 family, the upstream regions of 157a and 157b share extensive sequence similarity (Figure 2.12). MIR157c and MIR157d show less similarity (15.6% alignable) than is present between MIR157a and MIR157b (44.8% alignable sequence). This observation is in agreement with the observed conservation within the precursors of the MIR157 family. MIR157a and MIR157b share nearly identical stem loop sequences (Figure 2.11) while MIR157c and MIR157d have only sequence similarity in the area surrounding the miRNA:miRNA*   region (data not shown). MIR157a and b are closely located within 7.8

kb on Chromosome 1. An expressed gene At1g66790 lies between MIR157a and b. Two flanking LTR/copia repeat elements indicate that At1g66790 was probably inserted in this genomic region by a transposon. Therefore, it is likely that the insertion had caused the recent duplication which created MIR157a and MIR157b. On the other hand, MIR157c, d are located in segmentally duplicated regions on Chromosome 1 and Chromosome 3, respectively. This segmental duplication event may be older than the duplication event which created MIR157a and b.

Conservation in promoter regions is also observed for the MIR158 family and the MIR165 family. While MIR405b and MIR405d share nearly 50% alignable sequences within their promoter regions, the upstream sequence of MIR405a contains only two alignable regions as identified by DIALIGN, which are not well conserved. The MIR447 family shows the most striking sequence similarity on their promoter sequence. More than 60% promoter sequences are alignable.

Sequence conservation between the MIR gene family members and the extensive sequence similarities within promoter sequences are supportive of comparably recent duplication events which led to the formation of the non-conserved *Arabidopsis* miRNAs families. With the exception of the regions flanking the miRNA:miRNA*, the precursor sequences of miR165 family are diverged[3]. Nevertheless the upstream sequences within the miR165 family contain significant motifs which are supportive for the common origin of MIR165a and MIR165b.

2.3.4 Newly evolved *Arabidopsis thaliana* miRNA genes and their targets

Recent findings on MIR163 led to the hypothesis that plant miRNAs may be *de novo* created by inverted duplication of its targets and subsequent evolutionary drift to lead to functional miRNAs in their present form (Allen et al., 2004). My results given above imply that duplication of miRNAs also involves at least part of the promoters. Thus far it has not been shown whether the initial formation of miRNAs by a mechanism involving inverted duplication also involves duplication of parts of the promoters and to what extent promoter elements and regions are retained between the target gene promoter and the

[3] See the last part of the thesis for a discussion of miR165

respective

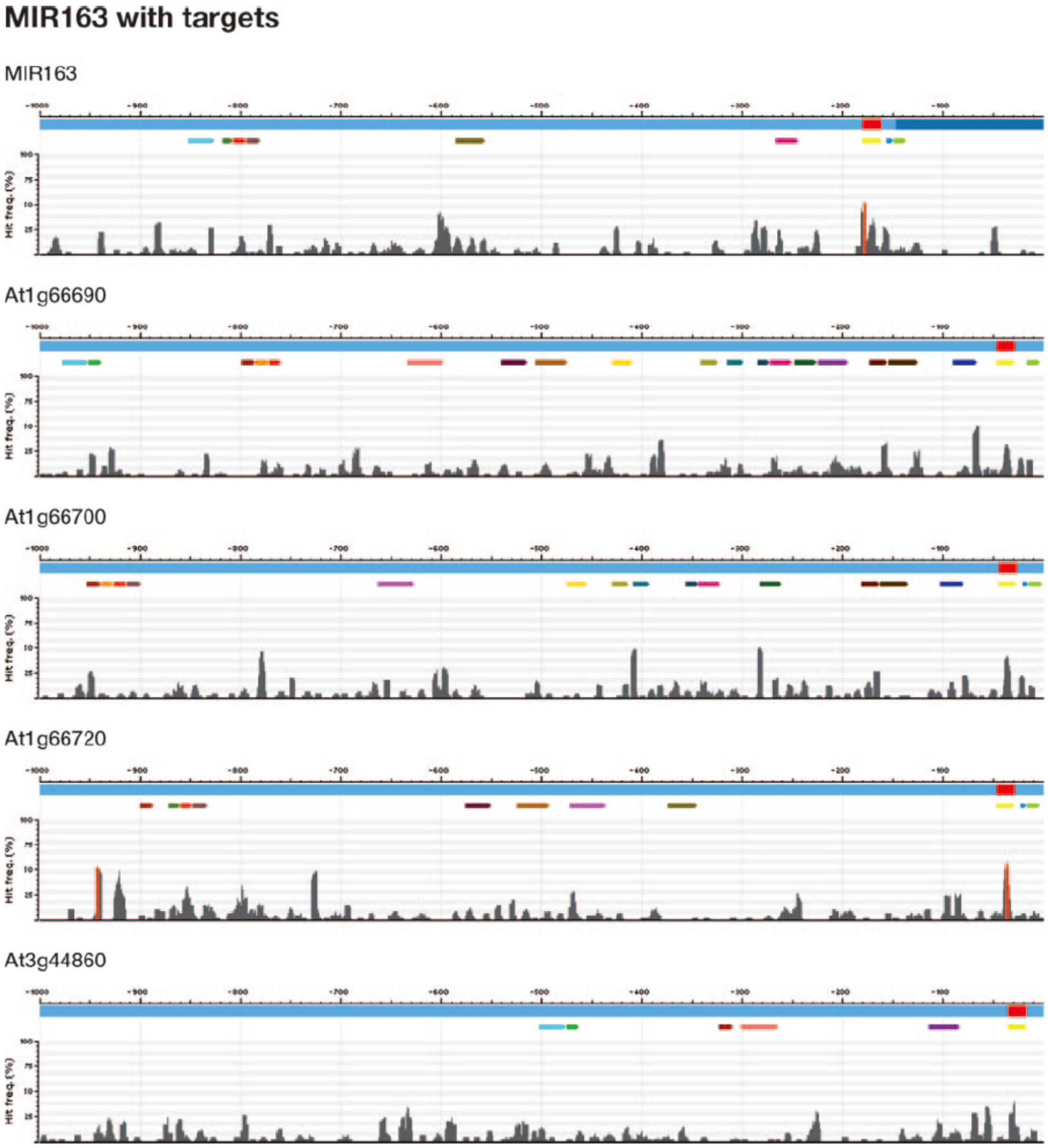


Figure 2.13

Sequence similarities within the promoter regions of MIR163 and its target genes.

The light blue horizontal bars represent upstream sequences. The dark blue bars at the right end of light blue bars indicates the 50-UTR of target genes or 50 ends of pri-miR163 which is not part of the microRNA precursor. The red horizontal bars represent the identified TATA box. Colored bars represent conserved regions detected. The peaks represent the number of hits per nucleotide normalized by the number of total hits per sequence, which were calculated 50 times by MotifSampler. A peak is plotted in red when the height of the peak is >50%.

miRNA promoter (Allen et al., 2004). In order to answer these questions I aligned the pri-miRNA163 plus 1000bp of promoter sequence to the respective UTRs and promoter sequences from the miR163 target genes, At1g66690, At1g66700, At1g66720 and At3g44860 (Figure 2.13). Numerous conserved regions as well as conserved short sequence motifs are present between the 5' upstream sequences of MIR163 and its target genes. Beside the conserved transcribed region additional conserved regions located within the promoter have been detected. The region depicted in red (Figure 2.13) harbours the TATA box consensus sequence which can be found in all target genes as well as for MIR163. Our findings on the localisation of the TATA box are consistent with recent results published by (Xie et al., 2005). Many additional promoter regions which are shared between the target genes and MIR163 have been detected (Figure 2.13). These regions have supposedly been retained since the initial inverted duplication event during the formation of MIR163. Thus they most likely represent functionally important regions which might be important for correct transcriptional regulation of MIR163 as well as the target genes. However, further experiments are needed to elucidate the functionality of these motifs.

In contrast to MIR163, there weren't any significant motifs for MIR161 and its target genes, although the stem loop of MIR161 has also been shown to have a similarity to its target genes (Allen et al., 2004). The same promoter analysis I was performed for non-conserved miRNA multigene families MIR157, MIR158, MIR165 and their corresponding targets as listed in ASRP Database (Gustafson et al., 2005). The results are included in our website. No significant sequence similarities were detected by DiAlign for MIR157, MIR158 and MIR165 and their targets, respectively. I then compared the similarity of the promoter regions of non-conserved miRNA families with those of all miRNA families, non-conserved miRNA families and their targets. I used the similarity of the promoter regions of all miRNA families and random promoter regions as the control group. The similarity of the promoter regions of non-conserved families is considerably higher than other groups (Figure 2.14), which is supportive of our hypothesis that the duplication of non-conserved miRNAs had included the promoter regions of miRNAs.

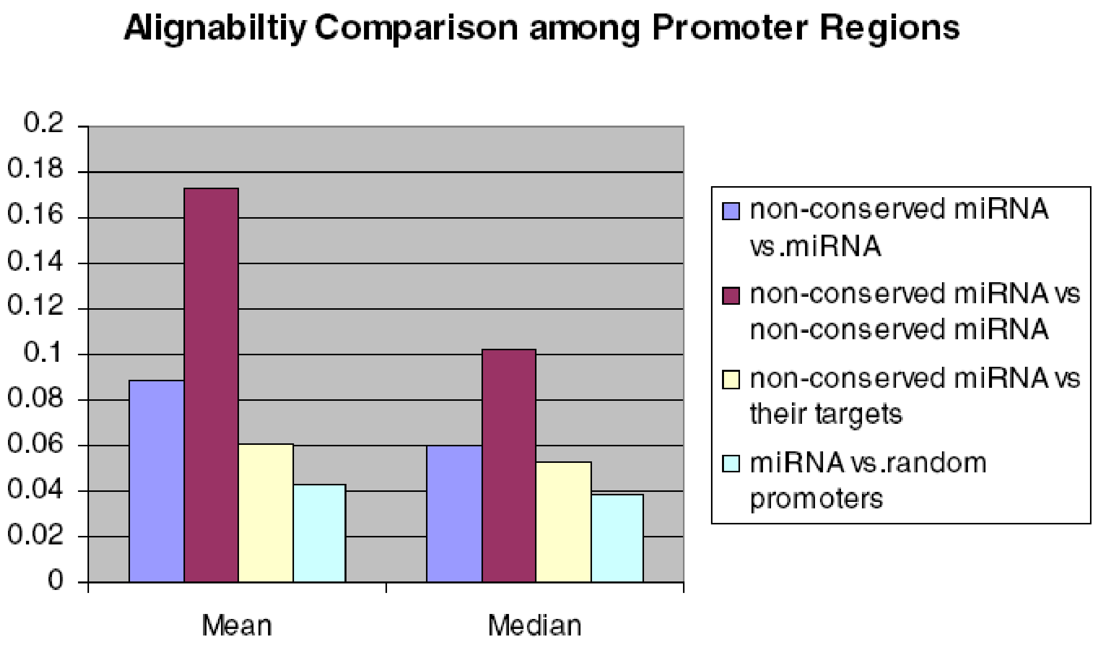


Figure 2.14

Compare the alignability of the promoter regions of non-conserved miRNA families with those of all miRNA families, non-conserved miRNA families and their targets.

The role of miRNAs as a part of gene regulatory machinery has been demonstrated by recent intensive research (Bartel, 2004; Baulcombe, 2005). However, the transcriptional regulation of miRNA genes themselves has not been the subject of extensive research. My analysis, which detected similarities within the precursor as well as the conserved regions within the promoter sequences, provides additional evidence to support the inverted duplication model for plant microRNA evolution. In addition, the detected conserved miRNA upstream motifs represent prime candidates to study the regulation of *Arabidopsis* MIR157, 158, 163, 165, 405 and 447. These finding suggests that the initial inverted duplication event, which had created miRNA from its targets, has included the promoter region of target genes. For expanded miRNA families regulatory diversification through modulation of *cis* regulatory elements might be a mechanism leading to fixation of sequence redundant miRNA genes. In analogy to duplicated protein-coding genes (Hurles, 2004) promoter elements seem to be shared within paralogous MIR genes. This permits to sketch a powerful *in silico* approach which might be considered for the experimental and *in silico* analysis of microRNA promoters.

This chapter mainly focus on miRNA genes. The topics include finding miRNA genes on a genome and promoter analysis of miRNA genes and their targets. In the next chapter, I will present another kind of small RNAs, viral derived small RNAs (vsRNAs).

Reference

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet *36*, 1282-1290.

Axtell, M.J., Snyder, J.A., and Bartel, D.P. (2007). Common functions for diverse small RNAs of land plants. Plant Cell *19*, 1750-1769.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Baulcombe, D. (2005). RNA silencing. Trends Biochem Sci *30*, 290-293.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004). Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc Natl Acad Sci U S A *101*, 11511-11516.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Chaudhury, A. (2005). Plant genetics: hothead healer and extragenomic information. Nature *437*, E1; discussion E2.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32*, D109-111.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. (2005). ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res *33*, D637-640.

Haberer, G., Hindemitt, T., Meyers, B.C., and Mayer, K.F. (2004). Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol *136*, 3009-3022.

Hindemitt, T., and Mayer, K.F. (2005). CREDO: a web-based tool for computational detection of conserved sequence motifs in noncoding sequences. Bioinformatics *21*, 4304-4306.

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. PLoS Biol *2*, E206.

Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell *14*, 787-799.

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

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051-4060.

Li, W.X., Oono, Y., Zhu, J., He, X.J., Wu, J.M., Iida, K., Lu, X.Y., Cui, X., Jin, H., and Zhu, J.K. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. Plant Cell *20*, 2238-2251.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053-2056.

Lolle, S.J., Victor, J.L., Young, J.M., and Pruitt, R.E. (2005). Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis. Nature *434*, 505-509.

Lu, C., Jeong, D.H., Kulkarni, K., Pillay, M., Nobuta, K., German, R., Thatcher, S.R., Maher, C., Zhang, L., Ware, D., *et al.* (2008). Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc Natl Acad Sci U S A *105*, 4951-4956.

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. Genome Res *16*, 510-519.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J., *et al.* (2008). Criteria for annotation of plant MicroRNAs. Plant Cell *20*, 3186-3190.

Morgenstern, B. (2004). DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res *32*, W33-36.

Nelson, D.E., Repetti, P.P., Adams, T.R., Creelman, R.A., Wu, J., Warner, D.C., Anstrom, D.C., Bensen, R.J., Castiglioni, P.P., Donnarummo, M.G., *et al.* (2007). Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. Proc Natl Acad Sci U S A *104*, 16450-16455.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., *et al.* (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457*, 551-556.

Peng, P., Chan, S.W., Shah, G.A., and Jacobsen, S.E. (2006). Plant genetics: increased outcrossing in hothead mutants. Nature *443*, E8; discussion E8-9.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., *et al.* (2005). Identification of microRNAs of the herpesvirus family. Nat Methods *2*, 269-276.

Ray, A. (2005). Plant genetics: RNA cache or genome trash? Nature *437*, E1-2; discussion E2.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., and Mayer, K.F. (2002). MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res *30*, 91-93.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113-1122.

Wang, X.J., Reyes, J.L., Chua, N.H., and Gaasterland, T. (2004). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biol *5*, R65.

Wang, Y., Hindemitt, T., and Mayer, K.F. (2006). Significant sequence similarities in promoters and precursors of Arabidopsis thaliana non-conserved microRNAs. Bioinformatics *22*, 2585-2589.

Wolfe, K.H., Gouy, M., Yang, Y.W., Sharp, P.M., and Li, W.H. (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci U S A *86*, 6201-6205.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and Carrington, J.C. (2005). Expression of Arabidopsis MIRNA genes. Plant Physiol *138*, 2145-2154.

Yao, Q., Xing, Y., Wang, Z., Zhang, J., Peng, R., and Hong, M. (1999). Cloning and sequencing of a rice (Oryza sativa L.) RAPB cDNA using yeast one-hybrid system. Sci China C Life Sci *42*, 354-361.

Zhao, B., Liang, R., Ge, L., Li, W., Xiao, H., Lin, H., Ruan, K., and Jin, Y. (2007a). Identification of drought-induced microRNAs in rice. Biochem Biophys Res Commun *354*, 585-590.

Zhao, Y., Ransom, J.F., Li, A., Vedantham, V., von Drehle, M., Muth, A.N., Tsuchihashi, T., McManus, M.T., Schwartz, R.J., and Srivastava, D. (2007b). Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. Cell *129*, 303-317.

Chapter 3 Viral Small RNA

This part of thesis focuses on viral derived small interfering RNAs (vsRNAs/viRNAs). It is based on (Donaire , 2009).

3.1 Introduction

Viruses are a very special form of life. A virus can replicate itself but only inside the cells of its host, which can be any organism. Viral infection normally triggers host defense systems to eliminate the infecting virus. Among various host reactions, RNA silencing is a mechanism used by both plant and invertebrates to counterattack viral infection (reviewed in (Ding and Voinnet, 2007)). RNA silencing depends on small RNAs, which, in the viral infection scenario, are derived from viruses.

In plants, the RNAi machinery is activated upon viral infection. Dicer like proteins target viral genomes to produce vsRNAs of various lengths (normally between 20-25nt). Traditional sequencing has been employed to probe the vsRNA populations of plant RNA and DNA viruses. However, the coverage of traditional sequencing was limited. We used multiplexed, high-throughput pyrosequencing (Roche 454) to profile populations of vsRNAs from plants infected with viruses from different genera. Populations of vsRNAs of 20 to 24 nucleotides (nts) were abundant and diverse, and reflected an effective processing of viral genomes by several Dicer-like (DCL) enzymes. Sense and antisense vsRNAs spread throughout the entire virus in a characteristic overlapping configuration in a way that virtually all nucleotide positions within the viral genome were represented in the dataset. Our data suggest that every genomic position is a putative cleavage site for vsRNA formation, although each viral genome contains regions that serve as preferential sources of vsRNA production. Hotspots of 21-, 22-, and 24-nt vsRNAs usually originated from the same genomic regions, indicating similar target affinities between DCL enzymes. Our results favor the idea that the bulk of vsRNAs originate from multiple DCL-catalyzed cleavage events on perfectly base paired double-stranded RNA, as opposed to cleavage of highly base paired structures from single-stranded viral RNA. vsRNAs generally display an inclination to begin with uridine, adenosine or cytosine, which anticipate potential interactions with various AGO complexes. Our data of vsRNA extend the current view of the distribution and composition of vsRNA in virus-infected plants, and helps us to define a more comprehensive scenario of vsRNA biogenesis and their regulatory function in plants. The result is present here in the next section.

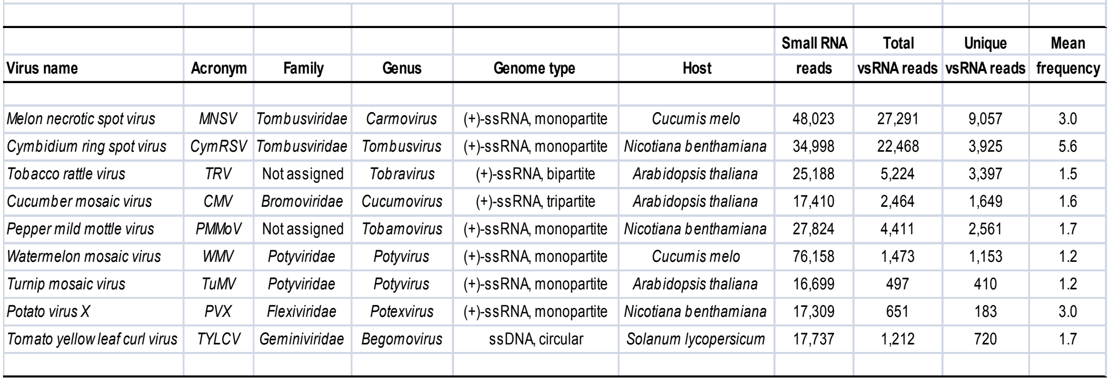
3.2 Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes

The final outcome of virus infections in plants depends on fine-tuned compatible and defensive interactions between hosts and viruses(Maule et al., 2002). Among the cell mechanisms underlying the molecular basis of these interactions, RNA silencing plays a critical role by providing a complex matrix of gene regulation that target both host and viral genomes(Ding and Voinnet, 2007; Dunoyer and Voinnet, 2005). In plants, RNA silencing is triggered by RNA with double-stranded (ds) structures which serve as a substrate for Dicer-like ribonucleases (DCL) to produce two major classes of small RNAs (sRNA): 21-nucleotide (nt) microRNAs (miRNAs) and small interfering RNAs (siRNAs) of ~21 to 25 nts(Brodersen and Voinnet, 2006). Host-encoded RNA-dependent RNA polymerases (RDR) are required for the production of each siRNA class by converting single-stranded (ss) RNA into dsRNA(Kasschau et al., 2007). sRNAs associate with distinct Argonaute (AGO)-containing effector complexes to guide them to their RNA or DNA target molecules(Hutvagner and Simard, 2008; Sontheimer and Carthew, 2005; Vaucheret, 2008). In plants, loading of sRNAs into a particular AGO complex is preferentially, but not exclusively, dictated by their 5’ terminal nucleotides. In the model plant *Arabidopsis thaliana*, AGO1 and AGO2 exhibit predominant binding affinity for 21-nt sRNAs having a 5’ terminal uridine (U) and adenosine (A), respectively, and AGO1 is involved in posttranscriptional gene silencing by guiding target mRNA degradation and translational inhibition. sRNAs recruited into AGO4 complexes are mostly of 24 nts, have a 5’ terminal A and control DNA methylation and transcriptional gene silencing of transposons and DNA repeats. Finally, AGO5 is associated with sRNAs of all size classes initiating with a 5’ cytosine (C)(Mi et al., 2008; Montgomery et al., 2008).

Plant viruses activate the RNA silencing machinery in infected cells through the formation of viral dsRNA by any of several mechanisms that include the activity of virus-encoded RNA polymerases, intermolecular basepairing between plus and minus viral RNAs, and imperfect folding of self-complementary sequences within viral ssRNA(Ding and Voinnet, 2007). In addition, at least three functional RDRs enzymes have been recognized as antiviral effectors implicated in biosynthetic pathways of viral sRNAs (vsRNA), suggesting that RDRs use viral RNA as a template to synthesize negative complementary strands. Upon DCL-mediated processing, vsRNA of 20 to 25 nts in length are generated as a natural consequence of virus infection to guide autosilencing of viral RNAs as part of an antiviral self-defense response in plants(Ding and Voinnet, 2007; Pantaleo et al., 2007). Multiple AGO genes might be involved in anti-virus defense and, at least, two AGO proteins (AGO2 and AGO5) have been shown to bind virus-derived sRNAs(Takeda et al., 2008). Moreover, the regulatory potential of these molecules presumably involves functional interactions with host transcripts through perfect or near-perfect basepairing (Moissiard and Voinnet, 2006). Compelling evidence indicates that the biogenesis of vsRNAs of different size classes involves the same DCL-dependent pathways responsible for the formation of endogenous siRNAs(Bouché et al., 2006; Gasciolli et al., 2005; Xie et al., 2004; Yoshikawa et al., 2005). In Arabidopsis, DCL4, DCL2 and DCL3 target viral genomes in a hierarchical fashion to yield vsRNAs of 21-, 22- and 24-nts, respectively. Antiviral immunity is conferred by DCL4-dependent, 21-nt vsRNAs with DCL2 acting as a DCL4 surrogate(Blevins et al., 2006; Bouché et al., 2006). 24-nt vsRNAs produced by DCL3 might be related to the perception of non-cell autonomous silencing signals.

Several studies using RNA and DNA viruses revealed that vsRNAs originate from multiple genomic regions(Chellappan et al., 2004; Ho et al., 2006; Moissiard and Voinnet, 2006; Molnár et al., 2005; Szittya et al., 2002). Although this observation may be diagnostic of a highly varied pool of vsRNAs in the infected tissue, the overall composition of the populations of vsRNAs generated by most plant viruses remains largely unknown. We used high-throughput DNA pyrosequencing to profile vsRNA across different plant virus genomes (Table 3.1).

Table 3.1. Viruses and host plants used for construction of vsRNA libraries from virus-infected plants



						Small RNA	Total	Unique	Mean
Virus name	Acronym	Family	Genus	Genome type	Host	reads	vsRNA reads	vsRNA reads	frequency
Melon necrotic spot virus	*MNSV*	*Tombusviridae*	*Carmovirus*	(+)-ssRNA, monopartite	*Cucumis melo*	48,023	27,291	9,057	3.0
Cymbidium ring spot virus	*CymRSV*	*Tombusviridae*	*Tombusvirus*	(+)-ssRNA, monopartite	*Nicotiana benthamiana*	34,998	22,468	3,925	5.6
Tobacco rattle virus	*TRV*	Not assigned	*Tobravirus*	(+)-ssRNA, bipartite	*Arabidopsis thaliana*	25,188	5,224	3,397	1.5
Cucumber mosaic virus	*CMV*	*Bromoviridae*	*Cucumovirus*	(+)-ssRNA, tripartite	*Arabidopsis thaliana*	17,410	2,464	1,649	1.6
Pepper mild mottle virus	*PMMoV*	Not assigned	*Tobamovirus*	(+)-ssRNA, monopartite	*Nicotiana benthamiana*	27,824	4,411	2,561	1.7
Watermelon mosaic virus	*WMV*	*Potyviridae*	*Potyvirus*	(+)-ssRNA, monopartite	*Cucumis melo*	76,158	1,473	1,153	1.2
Turnip mosaic virus	*TuMV*	*Potyviridae*	*Potyvirus*	(+)-ssRNA, monopartite	*Arabidopsis thaliana*	16,699	497	410	1.2
Potato virus X	*PVX*	*Flexiviridae*	*Potexvirus*	(+)-ssRNA, monopartite	*Nicotiana benthamiana*	17,309	651	183	3.0
Tomato yellow leaf curl virus	*TYLCV*	*Geminiviridae*	*Begomovirus*	ssDNA, circular	*Solanum lycopersicum*	17,737	1,212	720	1.7

3.2.1 Composition of the vsRNA populations in infected plants

vsRNA populations from *Arabidopsis thaliana* plants infected with TRV, TuMV or CMV, from *Nicotiana benthamiana* plants infected with CymRSV, PVX or PMMoV, from *Cucumis melo* plants infected with MNSV or WMV and from *Solanum lycopersicum* plants infected with TYLCV were profiled using high-throughput pyrosequencing technology (Table 3.1). sRNA was prepared from systemically infected tissues, ligated to 5’ and 3’ adapters, amplified by RT-PCR and subjected to multiplexed sequencing. Adapters were designed to specifically ligate vsRNAs containing 5’ monophosphate and 3’ hydroxyl ends, consistent with DCL-catalyzed cleavage products(Kasschau et al., 2007). A total of 281,346 reads with recognizable flanking adapter sequences and with lengths ranging between 17 and 28 nts were searched against the corresponding viral genomes. Only sequences that matched perfectly were further analyzed. A low proportion of reads contained single-position mismatches and was discarded from our analysis. Each vsRNA read could be unambiguously assigned to one unique genome position. In total, 65,691 reads were considered as vsRNAs representing 23,055 unique, although frequently overlapping, sequences (Table 3.1).

vsRNAs between 17 to 28 nts were recovered from our libraries although reads in the range of 20 to 24 nts constituted 92 % (59,984) of the total (Figure 1). vsRNA of 21 nts was clearly the predominant class followed by 22-nt vsRNA, together accounting for 77.5 % of total reads. The only exception was CymRSV that accumulated 22-nt vsRNAs to higher levels relative to 21-nt species (Figure 3.1). This agrees with the fact that sRNAs of 21 nts are preferentially sequestered by the p19 silencing suppressor encoded by tombusviruses. In general, the strong bias in size distribution was consistent with the hierarchical action of DCL4, DCL2 and DCL3 in the biogenesis of vsRNAs(Bouché et al., 2006; Deleris et al., 2006). 20-nt vsRNAs from MNSV-infected tissue accounted for 18 % of the total and accumulated nearly fourfold more than 22-nt vsRNAs (Figure 3.1). We presume that 20-nt vsRNA could be either a size-specific DCL cleavage product or a variant of the most abundant 21-nt vsRNA resulting from DCL slippage or end degradation. vsRNA from plants infected with CymRSV, TRV, CMV or PMMoV showed asymmetrical distribution in strand polarity with dominance of sense vsRNAs compared to antisense species, while MNSV-derived

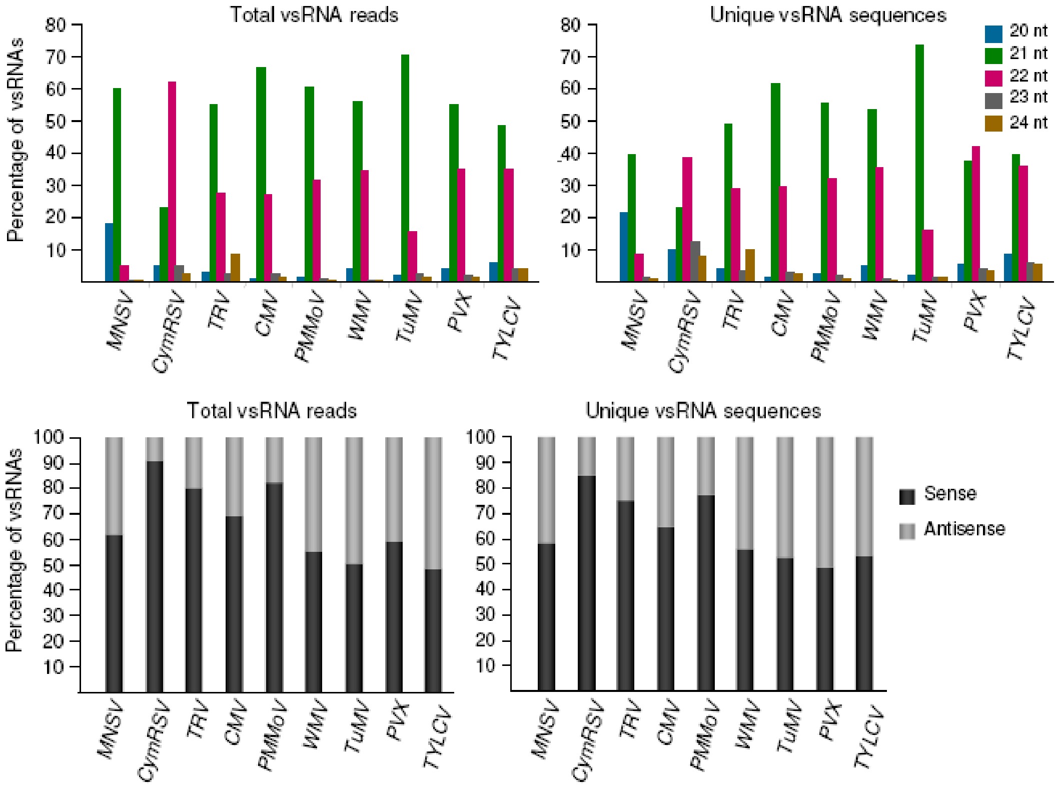


Figure 3.1

Size (top) and polarity (bottom) distribution of sequenced vsRNAs from virus-infected plants. Histograms represent the percentage of total or unique vsRNA reads within each category.

sRNAs showed only a modest enrichment of sense species (Figure 3.1). In contrast, vsRNAs from WMV, TuMV, PVX and TYLCV almost equally derived from positive and negative viral strands. In most virus tested, vsRNAs displayed significant, albeit modest, differences in base composition at their 5’ termini ($P < 0.0454$) with a preference to begin with an U or an A and a clear tendency to avoid a guanidine (G). Similar results were obtained from analyses with data sets containing unique vsRNA sequences (Figure 3.1).

The vast majority of the unique vsRNAs of 20 to 24 nts were sequenced in the range of 1 to 30 times. On average, 69% were only sequenced once, and 84% were represented in the sequenced pool by one or two reads (Figure 3.2A). Highly repetitive reads that likely accounted for abundant vsRNA species were also recorded. The very low sequence frequencies suggested that libraries were not saturating so deeper sequencing efforts are required in order to achieve saturation of unique vsRNAs species.

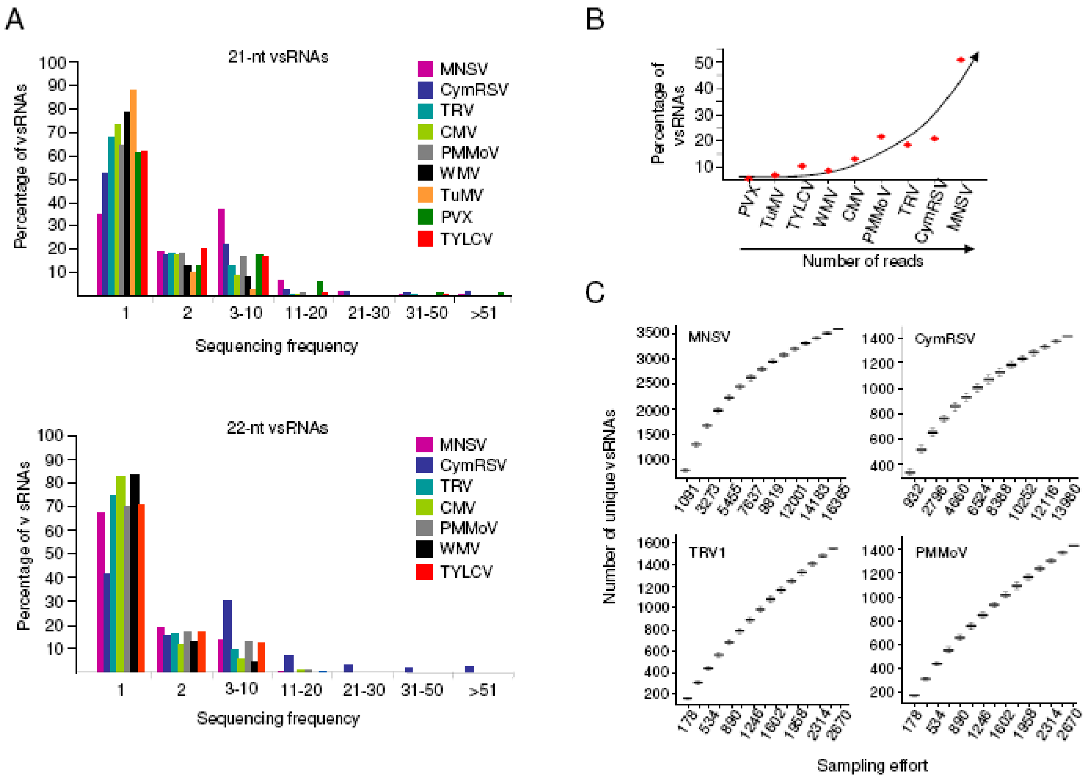


Figure 3.2

Estimation of saturation of vsRNAs and species richness in the sequenced pools.

(A) Sequencing frequencies of vsRNAs. Frequencies (x-axis) refer to as the number of times each unique vsRNA of 20 to 24 nts was sequenced within each library. (B) Percentage of vsRNAs within each library with respect to the maximum number of possible vsRNAs. Libraries are arranged on the x-axis from left to right according to their number of sequenced reads. (C) Quantification of species richness and sequencing effort. Shown are individual-based rarefaction curves of vsRNAs where the unique vsRNAs (y-axis) are plotted (Box-and-Whisker Plot) as a function of the sequencing effort (x-axis). Curves represent the means of repeated re-sampling of all pooled reads within each library. Note that the asymptote is reached only when all reads have been recovered. Data in (B) and (C) are relative to 21-nt vsRNAs, except for CymRSV that are to 22 nts.

3.2.2 Profiling viral small RNA diversity

In order to study spatial distribution and sequence diversity within the populations of vsRNAs, unique vsRNA sequences were mapped across the corresponding viral genomes and the genomic coverage for each virus was calculated. Using all size classes, 100 % of the MNSV genome was represented in the sequenced set of sense and antisense vsRNAs as all nucleotide positions in the genome were occupied by at least one unique sequenced vsRNA (Figure 3.3). vsRNAs from plants infected with CymRSV or PMMoV encompassed about 98 % of the corresponding viral genomes, while TRV, CMV or TYLCV genomes were covered by 80 to 96 % (Figure 3.3). Genomic coverage was also calculated using individual sizes, under the assumption that the vast majority of the most abundant 21- and 22-nt vsRNAs were authentic mature DCL4- and DCL2-cleavage products, respectively. 24-nt vsRNAs were investigated in CymRSV and TRV which generated sizeable populations of this class. Sense and antisense 21-nt vsRNAs produced by MNSV extended over 99.7 % of the viral genome, while 95 % of the CymRSV genome was reflected in the subset of sense and antisense 22-nt vsRNAs (Figure 3.3). 21-nt vsRNAs extended over 90 and 80 % of the TRV and CMV genome, respectively. vsRNAs of 24 nts comprised about 53 and 62 % of the TRV and CymRSV genomes, respectively. These data suggested that DCL targeting occurred widespread at multiple as opposed to isolated positions along the virus genome to form a tapestry of overlapping vsRNAs. Also, as shown in Figure 3.3, the profiles of 21-nt and 22-nt vsRNAs were largely similar, indicating a resemblance between DCL4- and DCL2-mediated targeting. For instance, regions with higher density of unique 21-nt vsRNA species overlapped with regions containing a larger number of unique 22-nt vsRNA sequences (Figure 3.3). Libraries from plants infected with PVX, TuMV contained a relatively low number of unique reads, although unique sequences of both polarities showed a tendency to spread uniformly throughout the TuMV and PVX genomes. In contrast, WMV produced little vsRNA diversity at the 5’ third of the viral RNA. This unusual pattern was not associated with the genome expression strategy of the potyviruses since the related TuMV produced equivalent number of unique vsRNA of all sizes across the entire genome. In conclusion, our results expanded beyond previous reports based on RNA hybridizations and small scale cloning of vsRNAs to highlight the genome-wide, massive formation of 20- to 24-nt vsRNAs in a characteristic overlapping configuration.

High-resolution mapping revealed that, in most cases, any given nucleotide position in the viral genome was occupied by numerous unique vsRNA species of both orientations (Figure 3.3). In regions with higher density of unique species, vsRNAs overlapped with neighboring sequences to form a ladder of vsRNAs with their 5’ termini spaced at 1-nt increments (Figure 3.4). This spacing is a hallmark for those viruses widely represented in our vsRNA collection and was suggestive of DCL-mediated cleavage of viral dsRNA occurring at consecutive nucleotide position along the viral genomes. If this is true and dsRNA comprises the full-length viral RNA, one might expect that the maximum number of unique vsRNAs of each size class generated by a particular virus would be solely constrained by the size of the genome. This theoretical number can be calculated simply as: n = x-(y-1), in which x = genomic size (Kb), and y = sRNA size. Our data revealed that sequenced MNSV-derived, 21-nt vsRNAs accounted for about 46 % of the maximum

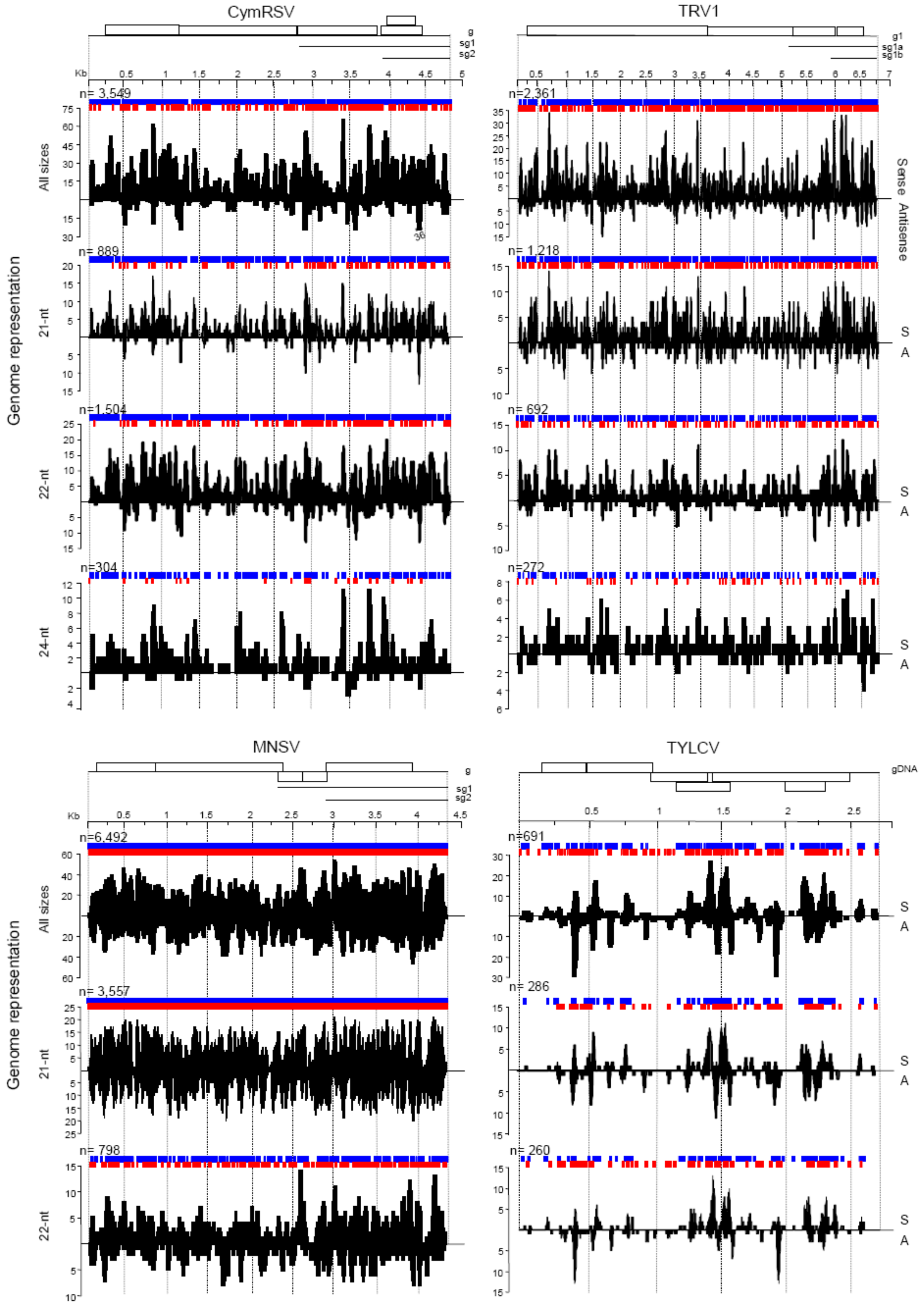


Figure 3.3

Distribution of unique vsRNAs along viral genomes.

A blue/red bar on the top of each graph stands for the genome coverage and measures the extent to which the viral genome was represented in the pool of unique vsRNA. Nucleotide genomic positions occupied by at least one unique read are indicated in blue (sense species) and red (antisense species). Graphs plot the number of unique vsRNA that hit every genomic position within each library. The number of unique vsRNA sequences (n) is indicated in each case. Bars above the axis represent sense (S) reads; those below represent antisense (AS) reads. Schematic representation of viral genomes including regions producing subgenomic RNAs are shown.

theoretical, while vsRNAs of 22-nts from the CymRSV library represented nearly 27 % (Figure 3.2B). This percentage dropped for those viruses with few vsRNA counts in the sequenced pool but raised proportionally as the number of reads increased, suggesting that the theoretical number of all possible unique vsRNAs is potentially achievable under exhaustive sequencing of the vsRNA populations (Figure 3.2B). In order to assess the effect of sampling effort on vsRNA species richness, we quantified unique vsRNA species in the sequenced pool using individual-based, species accumulation curves (Figure 3.2C)(Gotelli and Colwell, 2001). We assessed graphically whether the sequencing effort was enough to retrieve all unique sequences of a given size class by means of a resampling scheme based on common methods for the measurement of species richness in biodiversity studies(Gotelli and Colwell, 2001). From each set of sequences (21-nt vsRNAs for MNSV, TRV and PMMoV; 22-nt vsRNAs for CymRSV) we randomly sampled one-fifteenth of the total reads and counted the number of unique sequences in this group. Then we repeated the sampling, this time with groups comprising two fifteenth of the observations and counted again the number of unique sequences. We repeated the procedure for groups up to the total number of sequences and built a curve of accumulated unique sequences, which should suggest an asymptote for larger enough sampling efforts. Figure 3.2C clearly demonstrated that the number of unique sequences in our sequenced pool correlated with sampling efforts so as more individuals were sampled, more unique species were recorded. Most importantly, species richness hardly reached an asymptote, and only when practically the entire subset of reads had been sampled.

3.2.3 Distribution of viral small RNA abundance

We used direct high-throughput massive-parallel sequencing as a quantitative indicator of sRNA abundance(Kasschau et al., 2007; Lu et al., 2005; Rajagopalan et al., 2006). This approach is very convenient to achieve discrete measurement of specific sRNA species within large mixtures of overlapping sRNAs. The abundance of vsRNAs within each sequenced pool was plotted in Figure 3.5 using reads from all size classes as well as each of the most representative sizes independently. These analyses resulted in vsRNA abundance patterns that reflected a heterogeneous distribution across the entire viral genome and

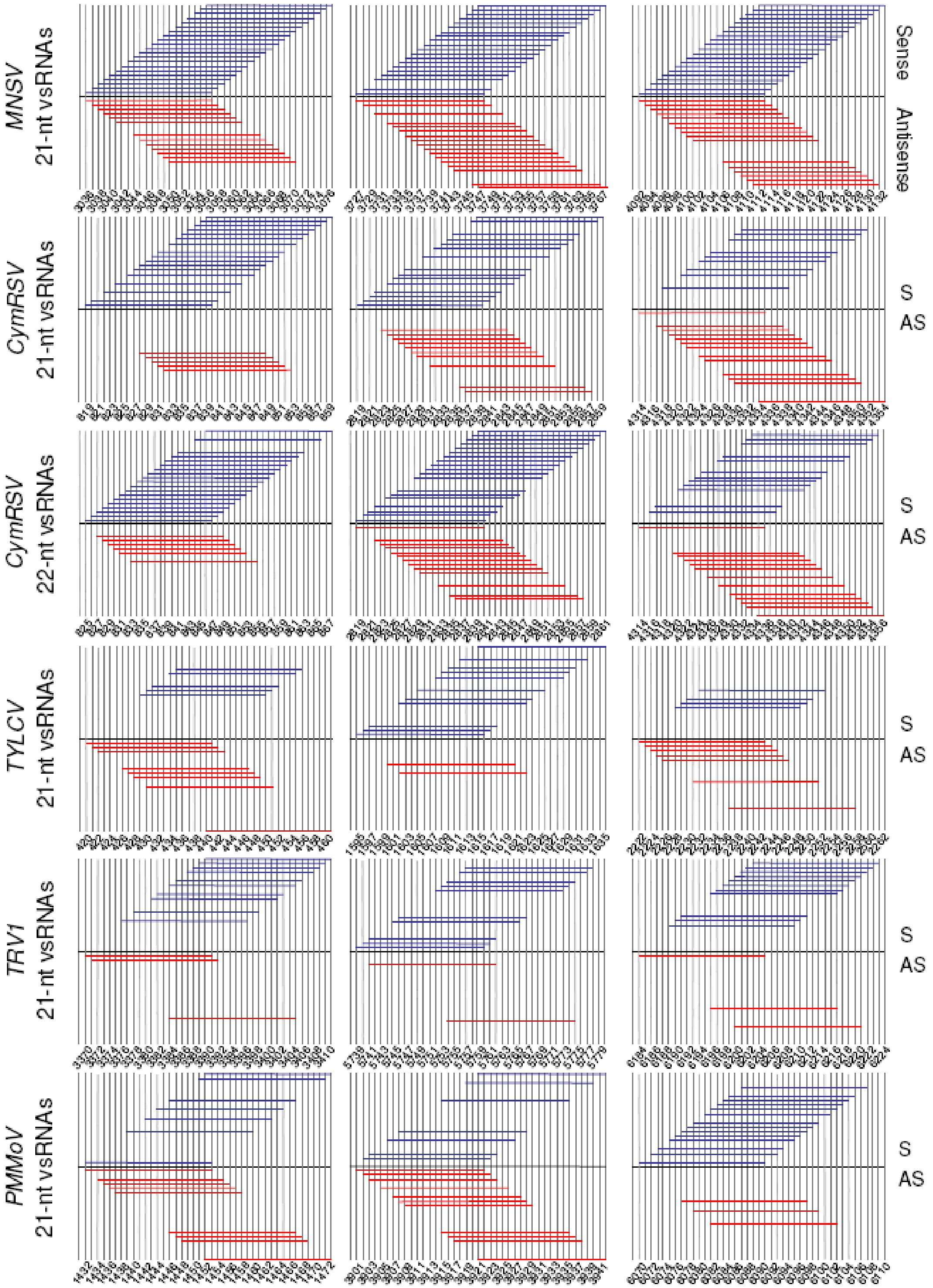


Figure 3.4

Overlapping configuration of the vsRNA populations.

Several representative regions showing overlapped vsRNAs are illustrated for each virus tested. For presentation purposes, sense (S) and antisense (AS) vsRNA with their 5' and 3' termini, respectively, located within a 21- or 22-nt window sequence are only represented. Sense vsRNA are 5'-3'; antisense vsRNAs are 3'-5'. Gridlines correspond to 1-nt increments. Nucleotide coordinates are indicated below each scheme. Note that many more vsRNA sequences (not shown) were found surrounding the regions depicted in the figure.

showed genomic regions with different densities of vsRNA reads (Figure 3.5). A significant twofold enrichment of vsRNA reads within the subgenomic RNA-forming 3’ end region was observed for CymRSV, MNSV and TRV ($t < -3.6591$, $P < 0.0336$) (Figure 3.5). In addition, hotspots of vsRNA accumulation were represented by sharp as well as broad peaks of vsRNA abundance scattered throughout the entire viral genomes. These peaks clustered multiple reads representing several overlapping unique vsRNA sequences. Strikingly, sharp peaks denoted the presence of highly abundant vsRNA reads within the cluster (Figure 3.5). The abundance distribution pattern of TYLCV- and PVX-derived vsRNAs was peculiar in that a large number of reads were gathered into a few hot spots of sense and antisense polarities (Figure 3.5). Therefore, quantitative profiling of our dataset supported our previous view using unique vsRNA sequences that each viral genome contained regions that served as preferential sources of DCL-mediated vsRNA production. In fact, for most viruses tested (those with high number of reads), the most prominent peaks of sequence diversity and abundance corresponding to vsRNAs of 21 nts usually localized within the same genomic regions as peaks corresponding to 22- or 24-nt vsRNAs (Figures 3.3 to 3.5). This observation seems to indicate that all DCL activities generating vsRNAs showed similar targeting affinities toward the same regions in a particular genome.

We next used vsRNA abundance and positions occupied by each vsRNA to explore the possibility that a fraction of vsRNAs in each library were processed in particular phase registers. Identification of predominant phase registers along the virus genome or within specific genomic regions might be suggestive of a biosynthetic pathway involving consecutive processing of viral dsRNA from preferential defined termini, reminiscent of that for endogenous trans-acting (ta)-siRNAs(Allen et al., 2005; Axtell et al., 2006; Howell et al., 2007). We randomly set five sequence windows, each corresponding to 10 cycles of DCL processing, along the virus genome, and searched for representative in-phase positions for cycles of 21 or 22 nts. All vsRNA reads with 5’ start coordinated into each of the possible 21 or 22 phasing registers within each window were determined(Howell et al., 2007). For all viruses tested, we could not detect a substantial enrichment of vsRNA reads in phase at any particular register, suggesting a random distribution of vsRNAs falling into each of the possible registers. This observation is in agreement with the ubiquitous spreading of vsRNAs.

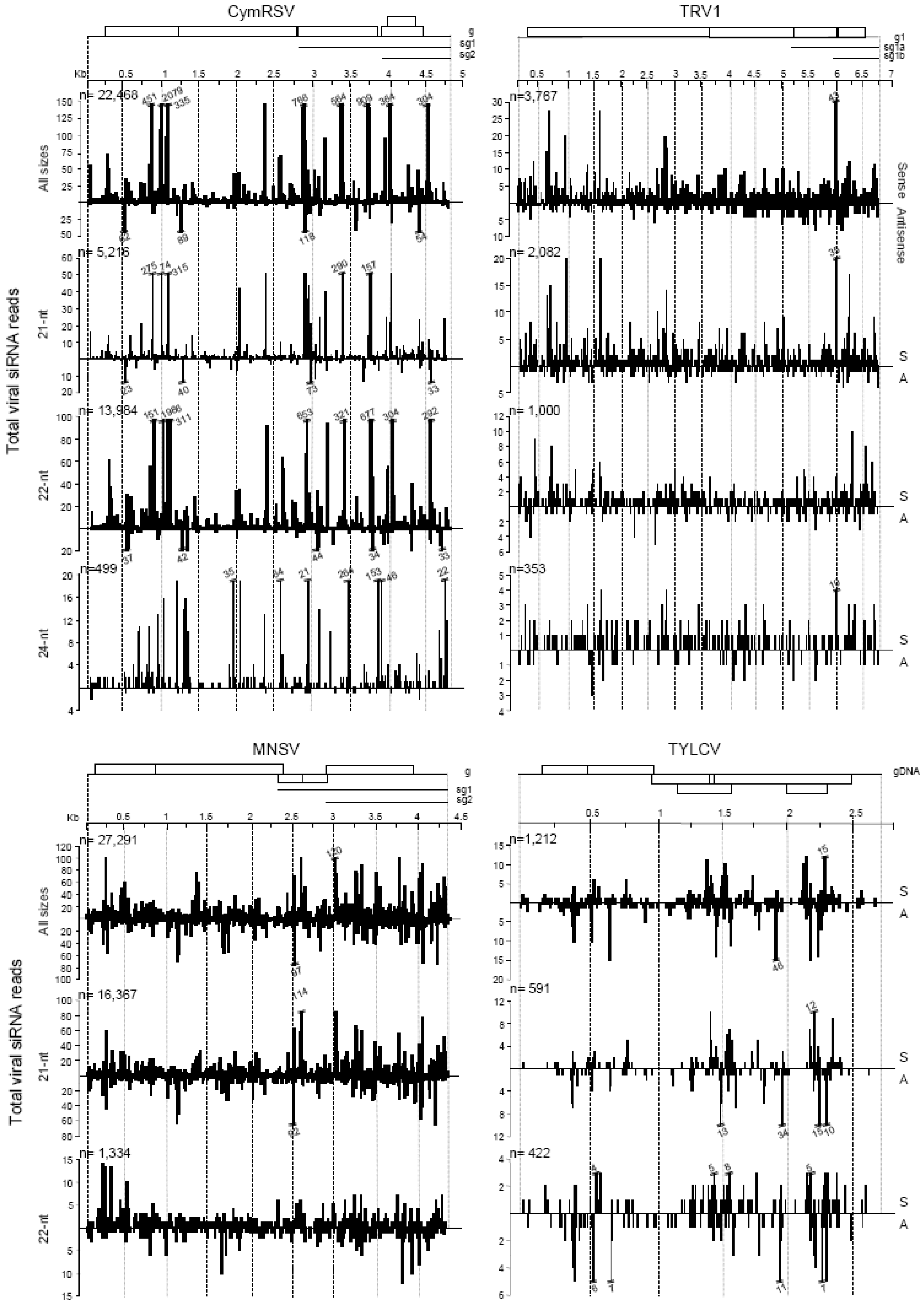


Figure 3.5

Distribution of vsRNA abundance along viral genomes.

The number of total reads with a 5′ terminus at each genomic position is plotted. For presentation purposes, maximum values plotted on the y-axis were adjusted to show average abundance values. Number on the top of each peak stands for their corresponding highest value. Figure content is as described in Figure 3.4.

3.2.4 Analysis of secondary structure within viral ssRNA

It has been proposed that hotspots of vsRNAs might reflect preferential DCL activities on foldback structures within ssRNA(Molnár et al., 2005; Szittya et al., 2002). If this is true, one might expect that sequences surrounding hotspots would adopt a stable hairpin structure that were robustly predicted by mfold program, as previously described for miRNA-like precursors(Jones-Rhoades and Bartel, 2004; Reinhart et al., 2002). To test this idea, the vsRNA sequences were compared with predicted RNA secondary structure maps generated using mfold 3.2(Zuker, 2003). The sequence and structure properties previously defined for recognition of miRNA-like precursors were used as constraints to search for putative foldback precursor of vsRNAs from viral ssRNA (Jones-Rhoades and Bartel, 2004; Meyers et al., 2008; Wang et al., 2004). As an illustrative example, we show data from CymRSV-derived secondary structures because CymRSV contained a large number vsRNA reads and nine hotspots of vsRNA accumulation were conspicuously identified along the virus genome. Each hotspot consisted of a cluster of unique, usually overlapping, vsRNAs where at least one vsRNA was sequenced more than 100 times. In general, sequences surrounding the vsRNA hotspots exhibited various degrees of intramolecular base pairing (Fig. 6). We determined the precise position of each of the vsRNA hits within the most favorable predicted stem-loops (Fig. 6). Some, but not all, of the most abundant vsRNA reads derived from hotspots 6 and 7 mapped within the longest base paired segment of the hairpin at a position consistent with cleavage at the center of the stem region (Fig. 6). Hotspots 3 and 5 also corresponded to regions predicted to form stable secondary structures, but they occupied positions within the foldbacks with limited base pairing, such that there were a large number of mismatched nucleotides and symmetric and asymmetric bulges (Fig. 6). Highly abundant reads at hotspots 1, 2 and 9 mapped in genomic regions predicted to fold into secondary structures containing an elevated number of G:U pairs, bulged or unpaired nucleotides (Fig. 6). The hotspot 8 corresponded to a region predicted to lack extensive base pairing (data not shown). Folding analysis of genomic sequences from PMMoV, TYLCV and TRV yielded similar results (data not shown) (Donaire et al., 2008). Taken together, some hotspots of vsRNAs occupied positions within the overall secondary structures that satisfied the size and structural criteria of stem–loops generating miRNAs in plants, while other hotspots occupied positions that were presumably suboptimal for DCL processing (Jones-Rhoades and Bartel, 2004; Meyers et al., 2008; Wang et al., 2004). In the latter case, regions with extensive base pairing within the predicted overall secondary structures were, in general, poorly represented in the vsRNA pool. These findings suggest that cleavage of foldbacks within viral ssRNA is unlikely to be the sole determinant to explain the formation of hotspots.

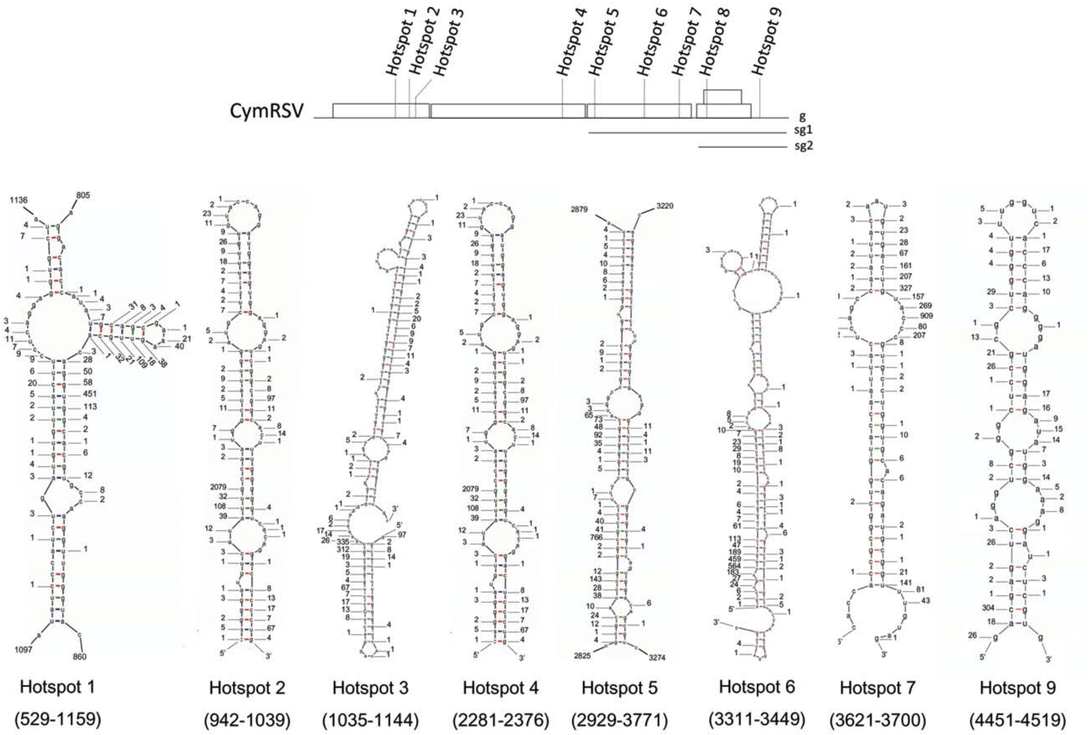


Figure 3.6

Prediction of secondary structures within CymRSV genomic ssRNA.

Predicted foldback structures correspond to sequences surrounding hotspots 3, 4, 6 and 7 as shown in Figure 3.5. The precise position of the 5' end of each sequenced vsRNA within the predicted structure in indicated by a number, which refers to as the number of reads representing each vsRNA at each position. A schematic of the CymRSV genome indicating the location of each hotspot is shown.

3.2.5 Discussion

In this study, we used a variety of plant viruses with distinct genome organization, genome expression and replication strategies as well as different host plant species to present a general, comprehensive scenario that extends the current view of the composition and distribution of vsRNAs in infected plants. Populations of vsRNAs were abundant, diverse and reflected an effective and widespread targeting of viral genomes by several interconnected RNA silencing pathways, regardless of the host plant and virus tested.

Sequence analysis of the data set revealed that the total number of vsRNA reads varied between the nine source libraries. This is likely due to intrinsic differences in the replication and virus accumulation rates between all viruses tested and in the efficiency of the RNA silencing machinery to recognize and target each viral genome in the corresponding host. We presume that these two factors may drastically influence to what extent dsRNA is formed from viral RNA templates. In addition, the levels of vsRNA accumulation in the infected tissue largely reflect the mode of action of the silencing suppressor encoded by each virus (Li and Ding, 2006). This latter situation is well-illustrated, for instance, by the potyviral HC-Pro and the TRV-16K silencing suppressors which cause a reduction in the accumulation of vsRNAs that is more accentuated in the presence of HC-Pro compared to TRV-16K. Likewise, the CMV-2b silencing suppressor specifically interferes with RDR1 to inhibit the production of 21-, 22- and 24-nt classes of secondary, CMVderived vsRNAs in the infected tissue (Diaz-Pendon et al., 2007). Nevertheless, we cannot exclude that our sequencing strategy using a multiplexed format created a bias among the libraries probably due to differences in amplicon concentrations, as the total number of host sRNA also differed between libraries.

Production of sRNAs in eukaryotes evokes two well-differentiated mechanisms that involve i) DCL processing of structured RNA or dsRNA, and ii) RDR-mediated RNA synthesis directly from ssRNA (Voinnet, 2008). The presence of a 5' monophosphate in vsRNAs, which is a characteristic feature of DCL-catalyzed cleavage products, can be inferred from ligation experiments with T4 RNA ligase, providing evidence that vsRNAs in our sequenced pool were DCL products. Moreover, the fact that the vsRNAs showed a clear bias in size distribution discounts the possibility that the vsRNA species sequenced in this study resulted mainly from non-specific RNA degradation. Our sequenced vsRNAs could be classified into the three major size classes predicted by the coordinated hierarchical actions of DCL4, DCL2 and DCL3 in vsRNA biogenesis (Bouché et al., 2006; Deleris et al., 2006). But, is dicing the sole mode of vsRNA production in plants? The accumulation of vsRNAs from several RNA and DNA plant viruses is severely compromised in loss-of-function dcl2/dcl3/dcl4 Arabidopsis mutants, indicating that DCL enzymes are major components of vsRNA biogenesis and antiviral defense (Bouché et al., 2006; Deleris et al., 2006; Diaz-Pendon et al., 2007; Donaire et al., 2008; Moissiard and Voinnet, 2006; Qu et al., 2008). In our study, deep-sequencing-based results regarding length and strand polarity distributions or spatial profiles are in tune with previous results using RNA blot analyses(Akbergenov et al., 2006; Chellappan et al., 2004; Diaz-Pendon et al., 2007; Donaire et al., 2008; Fusaro et al., 2006; Ho et al., 2006; Szittya et al., 2002; Xie et al., 2004), suggesting that our sequenced set of DCL-dependent vsRNAs fairly represents the entire population of vsRNA produced by each virus in the infected tissue. In *Caenorhabditis elegans*, secondary siRNAs are abundant and result from unprimed RNA synthesis by an RDR activity (Pak and Fire, 2007; Sijen et al., 2007). Therefore, RDR products carry 5' di- or triphosphates and are unlikely to be recovered using 5'-ligation dependent cloning methods. Whether a fraction of vsRNAs in plants results from direct RDR-dependent secondary vsRNA biosynthesis will require further experiments using cloning procedures that takes their triphosphorylated status into account.

Unique vsRNAs showed a genome-wide distribution. For viruses with the largest number of reads in the sequenced set, virtually all nucleotide positions in the genome were occupied by at least one unique vsRNA. Moreover, our results suggested that every single nucleotide position within a viral genome can be a putative cleavage site for vsRNA formation (Figure 3.4). DCL-catalyzed cleavage occurring at any of all the nucleotide positions along the viral genome would set a maximum number of unique vsRNA species of each size class to be produced by each particular virus. This theoretical number is only constrained by the size of the genome and reflects the enormous sequence diversity of the vsRNA population in the infected tissue. Extrapolating from sequencing frequencies and library saturation (evaluated using species richness within the sequenced pool), an asymptote representing the theoretical maximum number of unique vsRNA species generated by each virus might eventually be reached under exhaustive sampling of the entire vsRNA population.

Our sequence analysis indicated that vsRNAs were often biased to the genomic sense strand, as recently reported for TuMV- or TMVderived vsRNA (Ho et al., 2006). Derivation of comparable quantities of sense and antisense vsRNAs (as observed for WMV-, TuMV-, PVX- and TYLCV-derived vsRNAs) from perfectly complementary dsRNA derived from intermolecular base pairing of positive and negative viral strands is relatively easy to envision. Conversely, preferential accumulation of vsRNAs of sense polarity indirectly supports a model by which folded RNA within viral ssRNA serves as a substrate for DCL cleavage (Molnár et al., 2005). It is a common inference that these two pathways of dsRNA formation are operational during initiation and maintenance of virus-induced gene silencing in plants (Ding and Voinnet, 2007; Voinnet, 2008). DCL enzymes might be recruited into limited foldback structures to generate primary vsRNAs, while secondary vsRNA production may engage DCL processing coupled to RDR-mediated synthesis of longer complementary viral RNA. In our study, dominance of vsRNA species originating from viral positive strands was well-illustrated for CymRSV, TRV and PMMoV, which accumulated over 80% vsRNAs of sense polarity. A key question, however, is whether the asymmetrical distribution of strand polarity accurately reflects a major contribution of secondary structures from viral genomic RNA to vsRNA formation. In other words, is the cleavage of structured RNA sufficient to explain the dominance of sense vsRNA species? Our analyses of secondary structure and vsRNA distribution revealed that putative foldback structures could be predicted along viral ssRNA and that vsRNAs mapping at the center of highly base paired regions within these structures could be detected. This observation indirectly supported the idea that secondary structures could be potentially targeted by DCL activities to generate a discrete number of sense vsRNA. However, in general, there was little correlation, if any, between regions of predicted local base pairing RNA and regions where positive vsRNAs originated. Indeed, such a correlation has never been demonstrated for any viral genome (Donaire et al., 2008; Ho et al., 2006). Consequently, we suspect that it is unlikely that processing of imperfectly base paired hairpins solely accounts for the excess of sense vsRNA. Nonetheless, further research is needed to elucidate if these proposed base paired structures exist in vivo and whether they are recognized and cleaved by DCL enzymes. At the current stage, we do not have experimental data to propose an alternative explanation to the bias in the polarity ratio but strand asymmetry differences between viruses could indicate differences in the relative contribution of distinct mechanisms of dsRNA formation between viruses and the participation of specific unidentified host and/or virus-encoded factors. Recently, it has been hypothesized that the nascent viral RNA strands resulting from RDR activities might be chemically modified to prevent the negative vsRNA strand from the vsRNA duplex from entering into an AGO complex. Alternatively, intrinsic structural and/or biochemical signatures particularly associated to the viral positive strand of the vsRNA duplex might favor the selective recruitment of sense molecules into AGO proteins.

Regardless of whether primary vsRNAs are formed from viral folded ssRNA, the widespread distribution of sense and antisense vsRNAs in an overlapping configuration and their biogenesis at consecutive positions along the virus genome conciliate with DCL mediated processing of perfectly base paired, relatively long dsRNA as a principal contributor to vsRNA formation. This is consistent with the major role of several RDR-dependent pathways in the biogenesis of vsRNA and the maintenance of virus-induced RNA silencing in plants, a process linked to secondary siRNAs(Axtell et al., 2006; Diaz-Pendon et al., 2007; Donaire et al., 2008; Mourrain et al., 2000; Schwach et al., 2005; Yu et al., 2003). It is worth noting that there is no conflict between DCL processing of RDR-dependent, viral dsRNA substrates and enrichment of vsRNAs derived from positive RNA strands. This is exemplified by the fact that the largest bulk of vsRNAs of TRV or TMV, which is dominated by sense species, originates from the dicing of RDR products (Donaire et al., 2008). It would be interesting to determine whether the biogenesis of CymRSV-derived vsRNA is also RDR-dependent. Similarly, ta-siRNAs derived from Arabidopsis TAS transcripts seem to accumulate more sense than antisense species, even though they originate through DCL4-mediated processing of RDR6-dependend dsRNA (Allen et al., 2005; Axtell et al., 2006; Howell et al., 2007)

.

Despite their ubiquitous nature, regions with differential vsRNA density and diversity along each viral genome were identified. Why do some genomic regions serve as preferential sources of vsRNA formation? These areas likely hold structural features that ultimately influence accessibility, affinity or enzymatic activity of one or more components of the RNA silencing machinery required for dsRNA formation and subsequent processing into vsRNAs. For instance, vsRNAs were conspicuous at genomic regions producing subgenomic RNAs that may give rise to a substantial increase in RNA templates for synthesis of dsRNA. dsRNA substrates might also become available in the immediacy of sites for RDR template recognition and initiation of complementary-strand synthesis. RDR activities may be directed to the 3’ ends of viral RNA templates that lack molecular signatures associated to normally processed mRNA, such as those generated through random cleavage events, replication errors or vsRNA-guided cleavage (Allen et al., 2005; Axtell et al., 2006; Herr et al., 2006).

It is also tempting to speculate that vsRNA hotspots mirror a preferential DCL-mediated processing of selected regions containing hairpin secondary structure within viral ssRNA (Molnár et al., 2005; Szittya et al., 2002). In our study, genomic regions surrounding hotspots of vsRNA exhibited a number of possibilities to fold into relatively stable secondary structures. In general, highly repetitive reads were found in regions within the corresponding predicted secondary structures that lack extensive base pairing. In this situation, DCL4, DCL2 and DCL3 should act on imperfect duplexes containing a relatively higher degree of unpaired nucleotides compared to the canonical miRNA hairpin precursors. Although the affinity of these plant DCLs to folded RNA remains to be investigated, we find this scenario less probable as all these DCL enzymes are known to target relatively long, perfect dsRNA substrates, and they show only a residual processing activity on miRNA precursors (Bouché et al., 2006; Kasschau et al., 2007; Kurihara and Watanabe, 2004). In addition, vsRNAs display a clear tendency to begin with U, A and, to a lesser extent, C. This observation is in accord with the previous reporting of AGO proteins with preferred binding affinities for small RNAs having 5′ terminal U (AGO1), A (AGO2 and AGO4), and C (AGO5) (Mi et al., 2008; Montgomery et al., 2008; Takeda et al., 2008). Since the association of sRNAs with a particular AGO protein in plants is not apparently specified by the biogenesis pathways that produce the sRNAs or the structures of their precursors (Mi et al., 2008), it is reasonable to predict that vsRNAs bearing the appropriate 5′ sequence identities may be selectively loaded into multiple AGO complexes, as recently shown for CMV-derived vsRNA associated with AGO2 and AGO5 (Takeda et al., 2008). The low proportion of vsRNA beginning with a G in our data sets is consistent with the absence of AGO proteins known to prefer sRNAs having a 5′ terminal G. This is further supported by the over-representation of miRNA* sequences (arising from the opposite arm in the miRNA precursor) with a G at their 5′ ends, which are not thought to be associated with AGO complexes in plants.

The manifest potential of plant viruses to produce a diverse and abundant pool of vsRNAs and their likely association with multiple AGO effector silencing complexes have profound implications in the cross-talk interaction between plant and viruses. First, prolific generation of vsRNAs resulting from different DCL activities could reinforce silencing of virus genomes by means of RNA turnover, translational repression, or silencing signaling (Brodersen et al., 2008; Ding and Voinnet, 2007; Mlotshwa et al., 2008; Pantaleo et al., 2007). Second, owing to base complementarity to host genes, vsRNAs could hold an intrinsic regulatory potential, contributing to infection efficacy and symptom expression (Ding and Voinnet, 2007; Dunoyer and Voinnet, 2005; Mlotshwa et al., 2008; Moissiard and Voinnet, 2006). Sorting of vsRNAs into distinct AGO complexes has important functional consequences given that different AGO proteins mediate diverse effects on RNA and chromatin (Hutvagner and Simard, 2008). vsRNA recruited by AGO1- and AGO10-containing complexes may direct cleavage of target mRNAs or inhibition of mRNA translation. vsRNAs with a 5′ terminal A might associate to AGO4 and direct DNA methylation and transcriptional gene silencing at specific genomic loci that share sequence complementarity with the vsRNA.

Bioinformatic analyses for target prediction that used a mismatch/gap penalty scoring similar to that used for miRNA target prediction have been applied to identify hundreds of host genes as potential targets of vsRNAs (Moissiard and Voinnet, 2006). At the present time, the experimental evidence supporting a functional interaction between host mRNAs and vsRNAs is weak and only limited to a couple of genes among the bulk of predicted targets. Nevertheless, this finding suggests that, given the sequence diversity of the vsRNA population, an elevated number of host genes and their regulatory sequences might be targeted by vsRNA-mediated downregulation during virus infection. The challenge ahead is to determine the extent of these functional interactions between vsRNAs and their targets in a biological perspective.

Reference

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 Research *34*, 462-471.

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell *121*, 207-221.

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell *127*, 565-577.

Blevins, T., Rajeswaran, R., Shivaprasad, P.V., Beknazariants, D., Si-Ammour, A., Park, H.-S., Vazquez, F., Robertson, D., Meins, F., Hohn, T*., et al.* (2006). Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Research *34*, 6233-6246.

Bouché, N., Lauressergues, D., Gasciolli, V., and Vaucheret, H. (2006). An antagonistic function for Arabidopsis DCL2 in development and a new function for DCL4 in generating viral siRNAs. The EMBO Journal *25*, 3347-3356.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science (New York, NY) *320*, 1185-1190.

Brodersen, P., and Voinnet, O. (2006). The diversity of RNA silencing pathways in plants. Trends in Genetics: TIG *22*, 268-280.

Chellappan, P., Vanitharani, R., Pita, J., and Fauquet, C.M. (2004). Short interfering RNA accumulation correlates with host recovery in DNA virus-infected hosts, and gene silencing targets specific viral sequences. Journal of Virology *78*, 7465-7477.

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K.D., Carrington, J.C., and Voinnet, O. (2006). Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science (New York, NY) *313*, 68-71.

Diaz-Pendon, J.A., Li, F., Li, W.-X., and Ding, S.-W. (2007). Suppression of antiviral silencing by cucumber mosaic virus 2b protein in Arabidopsis is associated with drastically reduced accumulation of three classes of viral small interfering RNAs. The Plant Cell *19*, 2053-2063.

Ding, S.W., and Voinnet, O. (2007). Antiviral immunity directed by small RNAs. Cell *130*, 413-426.

Donaire L., Wang Y., Gonzalez-Ibeas D., Mayer K., Aranda M., Llave1 C., Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. (2009) Virology, Volume 392, Issue 2, 30

203-214

Donaire, L., Barajas, D., Martínez-García, B., Martínez-Priego, L., Pagán, I., and Llave, C. (2008). Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. Journal of Virology *82*, 5167-5177.

Dunoyer, P., and Voinnet, O. (2005). The complex interplay between plant viruses and host RNA-silencing pathways. Current Opinion in Plant Biology *8*, 415-423.

Fusaro, A.F., Matthew, L., Smith, N.A., Curtin, S.J., Dedic-Hagan, J., Ellacott, G.A., Watson, J.M., Wang, M.-B., Brosnan, C., Carroll, B.J.*, et al.* (2006). RNA interference-inducing hairpin RNAs in plants act through the viral defence pathway. EMBO Reports *7*, 1168-1175.

Gasciolli, V., Mallory, A.C., Bartel, D.P., and Vaucheret, H. (2005). Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. Current Biology: CB *15*, 1494-1500.

Gotelli, N.J., and Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters *4*, 379-391.

Herr, A.J., Molnàr, A., Jones, A., and Baulcombe, D.C. (2006). Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America *103*, 14994-15001.

Ho, T., Pallett, D., Rusholme, R., Dalmay, T., and Wang, H. (2006). A simplified method for cloning of short interfering RNAs from Brassica juncea infected with Turnip mosaic potyvirus and Turnip crinkle carmovirus. Journal of Virological Methods *136*, 217-223.

Howell, M.D., Fahlgren, N., Chapman, E.J., Cumbie, J.S., Sullivan, C.M., Givan, S.A., Kasschau, K.D., and Carrington, J.C. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. The Plant Cell *19*, 926-942.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nature Reviews Molecular Cell Biology *9*, 22-32.

Jones-Rhoades, M.W., and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Molecular Cell *14*, 787-799.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biology *5*, e57.

Kurihara, Y., and Watanabe, Y. (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proceedings of the National Academy of Sciences of the United States of America *101*, 12753-12758.

Li, F., and Ding, S.W. (2006). Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu Rev Microbiol *60*, 503-531.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005). Elucidation of the small RNA component of the transcriptome. Science (New York, NY) *309*, 1567-1569.

Maule, A., Leh, V., and Lederer, C. (2002). The dialogue between viruses and hosts in compatible interactions. Current Opinion in Plant Biology *5*, 279-284.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J.*, et al.* (2008). Criteria for annotation of plant MicroRNAs. The Plant Cell *20*, 3186-3190.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C.*, et al.* (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116-127.

Mlotshwa, S., Pruss, G.J., and Vance, V. (2008). Small RNAs in viral infection and host defense. Trends in Plant Science *13*, 375-382.

Moissiard, G., and Voinnet, O. (2006). RNA silencing of host transcripts by cauliflower mosaic virus requires coordinated action of the four Arabidopsis Dicer-like proteins. Proceedings of the National Academy of Sciences of the United States of America *103*, 19593-19598.

Molnár, A., Csorba, T., Lakatos, L., Várallyay, E., Lacomme, C., and Burgyán, J. (2005). Plant virus-derived small interfering RNAs originate predominantly from highly structured single-stranded viral RNAs. Journal of Virology *79*, 7812-7818.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell *133*, 128-141.

Mourrain, P., Béclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.B., Jouette, D., Lacombe, A.M., Nikic, S., Picault, N*., et al.* (2000). Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. Cell *101*, 533-542.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science *315*, 241-244.

Pantaleo, V., Szittya, G., and Burgyán, J. (2007). Molecular bases of viral RNA targeting by viral small interfering RNA-programmed RISC. Journal of Virology *81*, 3797-3806.

Qu, F., Ye, X., and Morris, T.J. (2008). Arabidopsis DRB4, AGO1, AGO7, and RDR6 participate in a DCL4-initiated antiviral RNA silencing pathway negatively regulated by DCL1. Proceedings of the National Academy of Sciences of the United States of America *105*, 14732-14737.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes & Development *20*, 3407-3425.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. Genes & Development *16*, 1616-1626.

Schwach, F., Vaistij, F.E., Jones, L., and Baulcombe, D.C. (2005). An RNA-dependent RNA polymerase prevents meristem invasion by potato virus X and is required for the activity but not the production of a systemic silencing signal. Plant Physiology *138*, 1842-1852.

Sijen, T., Steiner, F.A., Thijssen, K.L., and Plasterk, R.H.A. (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science (New York, NY) *315*, 244-247.

Sontheimer, E.J., and Carthew, R.W. (2005). Silence from within: endogenous siRNAs and miRNAs. Cell *122*, 9-12.

Szittya, G., Molnár, A., Silhavy, D., Hornyik, C., and Burgyán, J. (2002). Short defective interfering RNAs of tombusviruses are not targeted but trigger post-transcriptional gene silencing against their helper virus. The Plant Cell *14*, 359-372.

Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M., and Watanabe, Y. (2008). The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins. Plant & Cell Physiology *49*, 493-500.

Vaucheret, H. (2008). Plant ARGONAUTES. Trends in Plant Science *13*, 350-358.

Voinnet, O. (2008). Use, tolerance and avoidance of amplified RNA silencing by plants. Trends in Plant Science *13*, 317-328.

Wang, X.-J., Reyes, J.L., Chua, N.-H., and Gaasterland, T. (2004). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biology *5*, R65.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biology *2*, E104.

Yoshikawa, M., Peragine, A., Park, M.Y., and Poethig, R.S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes & Development *19*, 2164-2175.

Yu, D., Fan, B., MacFarlane, S.A., and Chen, Z. (2003). Analysis of the involvement of an inducible Arabidopsis

RNA-dependent RNA polymerase in antiviral defense. Molecular Plant-Microbe Interactions: MPMI *16*, 206-216.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research *31*, 3406-3415.

Chapter 4 MicroRNA Target Interaction

This section of thesis focuses on human microRNA targets, which most likely are protein coding genes. The work presented here is a team effort with several colleagues of my institute and an experimental group of Prof. Thomas Brabletz of University of Freiburg. In this project, I conceived the hypothesis and coordinated computational and experimental work.

4.1 MicroRNA targets

MicroRNAs regulate protein coding genes mainly in two ways. First, mRNA cleavage is accomplished by miRNAs guiding RISC complexes to near perfect complimentary target sites. Many plant miRNAs recognize and interact with their targets in this way (Chapter 1, 1.3.1). Second, translational inhibition, a major interaction mode of animal miRNAs and their targets, can be achieved by direct translational repression at various stages of mRNA translation or by mRNA destabilization (reviewed in (Filipowicz et al., 2008a)).

In animals, miRNAs are frequently found to be co-expressed in different tissues and cell types, while some form polycistronic clusters on genomes. Interactions between targets of co-expressed miRNAs (including miRNA clusters) have not yet been systematically investigated. Here we integrated information from predicted and experimentally verified miRNA targets, to characterize protein complex networks regulated by human miRNAs. We found compelling evidence that individual miRNAs or co-expressed miRNAs frequently target several components of protein complexes - indicating a coordinate posttranscriptional regulation of protein complexes by miRNAs. We experimentally verified that the miR-141-200c cluster targets different components of the CtBP/ZEB complex, suggesting a potential orchestrated regulation in epithelial to mesenchymal transition.

4.2 MicroRNAs coordinately regulate human protein complexes

4.2.1 Introduction

Hundreds of microRNA (miRNA) genes have been identified in mammalian genomes(Griffiths-Jones et al., 2008). Each miRNA may repress the translation of, and/or destabilize numerous messenger RNAs (mRNAs). Moreover, miRNA genes are frequently organized into genomic clusters(Bonci et al., 2008; Mendell, 2008; Nakada et al., 2008), which are transcribed from a common promoter as polycistronic primary transcripts, and whose coordinate functional roles remain to be investigated(Ambros, 2008). Recent large-scale, quantitative proteomics studies have demonstrated that some miRNAs probably participate in fine-tuning the production of their targets, both at the messenger RNA and the protein level (Baek et al., 2008; Selbach et al., 2008); however, the overall effect of miRNAs on many of their target proteins is often intriguingly modest. It remains unclear how these marginal effects can convey the necessary regulatory information for proper cellular activities(Flynt and Lai, 2008).

We applied a network-based strategy to systematically map coordinate regulatory interactions of single and co-expressed (including clustered) miRNAs. Previous work(Hsu et al., 2008; Liang and Li, 2007) has demonstrated that the targets of single miRNAs are more connected in the protein-protein interaction network than expected by chance. However, the use of protein-protein interaction (PPI) data provides only a rough overall picture of miRNA target interactions; it is not easy to evaluate the regulatory effects of miRNAs on such large-scaled PPI networks. Instead, as the basic functional units of the cellular machinery, experimentally verified protein complexes are natural subsets of PPI networks for investigating miRNA target interactions. Several components of protein complexes may be regulated simultaneously by a single miRNA or by several co-expressed miRNAs. Thus, although the regulation of protein synthesis is marginal for some of the miRNA targets, a cumulative effect for substantial phenotypic consequence may be achieved for those targets, which are members of the same protein complexes.

To test this hypothesis, we developed a computational framework to systematically infer protein complexes, of which several distinct components are simultaneously regulated by either single miRNAs or co-expressed miRNAs. We applied the framework to characterize the protein complex networks, which consist of 416 experimentally verified protein complexes and protein-protein interactions. These protein complex networks are regulated by 164 miRNAs and 36 known miRNA clusters in humans. We find that our framework has several advantages over previous analyses of miRNA targets and their interactions. First, high-confidence miRNA target predictions allow researchers to characterize the overall functional spectrum of miRNA-regulated protein complexes. Second, we expected that such a framework could predict miRNA targets more reliably - a hypothesis, which we supported by comparison with recent large-scale, quantitative proteomics data. Third, we demonstrated that miRNAs, which target the same protein complexes, are frequently co-expressed. Finally, we experimentally verified that the *miR141-200c* cluster simultaneously targets several protein components of the CtBP/ZEB complex, implying an efficient regulation of a protein complex by a cluster of miRNAs.

4.2.2 Results

To identify protein complexes, for which several distinct components are coordinately regulated by miRNAs, we assembled a miRNA-mRNA target network for 164 human miRNAs using 457 experimentally confirmed targets(Papadopoulos et al., 2009) and 4,236 predicted targets generated by two state-of-the-art miRNA target prediction tools, PicTar and PITA (Kertesz et al., 2007; Krek et al., 2005). The targets were mapped to a non-redundant set of experimentally verified protein complexes from the CORUM database(Ruepp et al., 2008). We compiled the protein complexes, which are more significantly associated with the target sets of single miRNAs or clustered miRNAs from 36 transcription units (Landgraf et al., 2007) than expected for random target lists based on two separate tests for statistical significance (Figure 4.1). The analysis resulted in 416 miRNA-regulated protein complexes (*P* value < 0.05; permutation test 1,000 samples; hypergeometric test with Bonferroni correction for multiple testing), which contained at least two targets of an individual miRNA or of several clustered miRNAs.

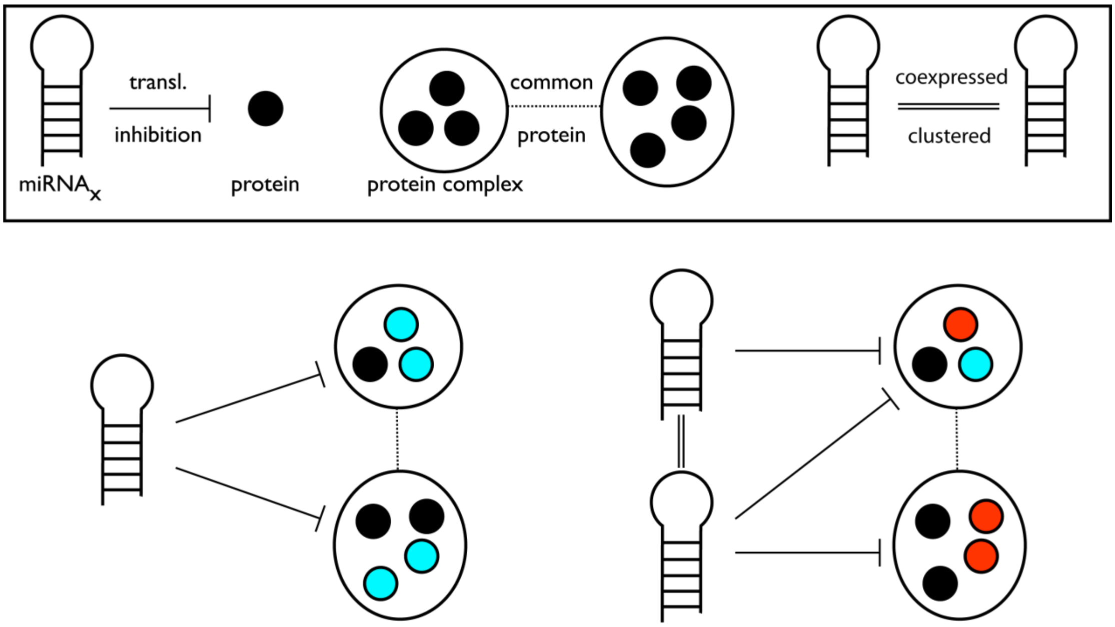


Figure 4.1

Network-based analysis of miRNA target sets.

Our strategy aims at assessing the interactions of miRNA targets and their coordinate regulation. First, targets of miRNAs and clusters of miRNAs are mapped to experimentally verified protein complexes. Second, two independent statistical tests are used to score protein complexes, for which several components are targeted by a single miRNA or by several co-expressed (clustered) miRNAs.

Functional spectrum of miRNA-regulated protein complexes

We firstly analyzed the spectrum of functions covered by our set of miRNA-regulated protein complexes. We identified the biological processes (Gene ontology categories(Ashburner et al., 2000)), which are enriched within the total set of 505 miRNA-targeted components of the protein complexes. In all, as shown in Figure 4.2a, the miRNA-regulated protein complexes are mainly involved in genomic regulatory activities, such as transcription regulation, signal transduction, chromatin regulation, development and the cell cycle. 79% of the miRNA targets in protein complexes belong to at least one of these categories. A minor but significant fraction of miRNA-regulated complexes is involved in vesicular transport (21 targets). Conversely, house-keeping functions, such as biosynthetic processes or translation are underrepresented. The results confirm earlier investigations(Cui et al., 2006) showing that miRNAs less frequently target genes involved in essential cellular processes. Experimental evidence has already been reported for the regulation of signal transduction in several metazoan species(Flynt et al., 2007; Forstemann et al., 2005; Friggi-Grelin et al., 2008; Martello et al., 2007; Silver et al., 2007) and the cell cycle(Carleton et al., 2007; Neumüller et al., 2008) by miRNAs. The regulation of the cell cycle by miRNAs is further supported by strong correlations of miRNA over-expression with different types of cancer (Volinia et al., 2006).

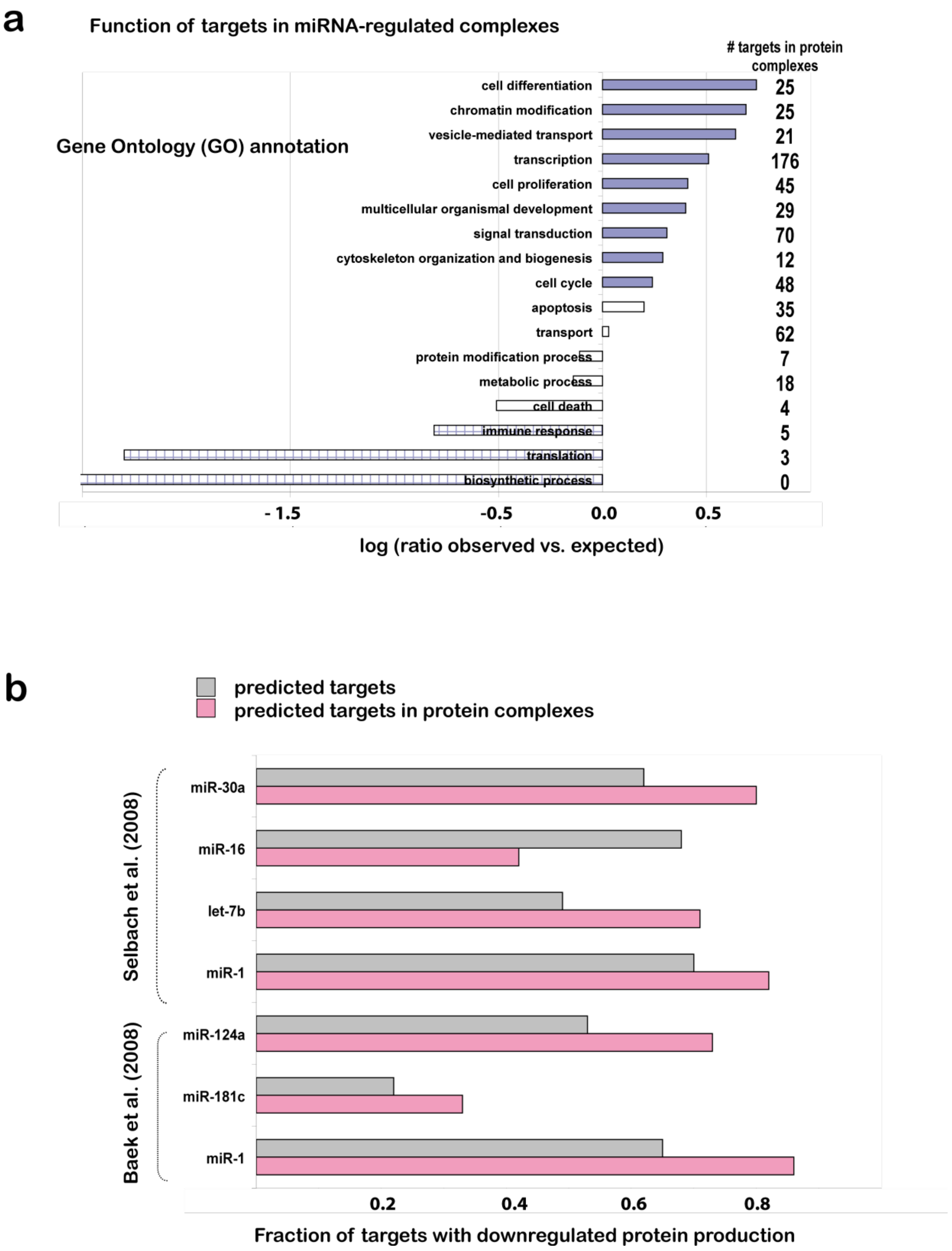


Figure 4.2

Functional analysis and validation of miRNA-regulated protein complexes.

a, Functional analysis: Enrichment of Gene Ontology (GO) categories in the target subunits of protein complexes. The log of the fraction observed (N_t/N_c) vs. expected ($N_{tall}/\mathrm{N}_{\mathrm{all}}$) is shown, where N_t is the number of target proteins with annotations t,

N_c is the number of all miRNA-regulated components, N_{tall} the number of proteins with annotations t, and N_{all} the number of all proteins. For N_t=0 the log is not defined and the bar is truncated. Categories with enrichment P value < 0.05 are indicated by color; categories with depletion *P* value < 0.05 are indicated by a pattern. **b**, Validation: Fraction of (1) predicted targets and (2) predicted targets in protein complexes with reduced protein production (log2-fold change < -0.1) for each miRNA from two experimental studies(Baek et al., 2008; Selbach et al., 2008). Less than three protein complexes were predicted for *miR-155*, *miR-124*, *miR-223*. Therefore, these three miRNAs were not included in this figure.

Validating predicted miRNA targets in protein complexes

Two recent proteomics studies measured the changes in synthesis of proteins in response to miRNA over-expression or knockdown on a genome-wide scale for selected miRNAs(Baek et al., 2008; Selbach et al., 2008). To validate our predictions, we calculated the fraction of predicted target proteins in our set of 416 miRNA-regulated protein complexes, for which the protein production was down-regulated. On average 74% of the predicted target proteins in the complexes were mildly down-regulated ($\log_2$-fold change < -0.1) in comparison to 58% of all targets predicted by PicTar and PITA (Figure 4.2b). Thus, we found that, by adding an additional layer of information from protein complexes, the reliability of target predictions increases significantly for those targets, which are components of protein complexes. For example, our data showed that the LARC (LCR-associated remodelling) complex(Mahajan et al., 2005) has three (out of 19) components, which are computationally predicted targets of *let-7*. These three components, namely *DPF2* (Zinc finger protein ubi-d4), *CHD4* (Chromodomain–helicase–DNA -binding protein 4), and *SMARCC1* (SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily C member 1) were modestly down-regulated (fold changes of -0.38, -0.21, and -0.2, respectively), when *let-7b* was over-expressed in HeLa cells(Selbach et al., 2008). LARC binds to the DNase hypersensitive 2 site in the human β-globin locus control region (LCR) and transactivates β-like globin genes(Mahajan et al., 2005). By simultaneously down-regulating three components of the LARC complex, *let-7b* might contribute to the overall transcriptional repression of the human β-globin locus.

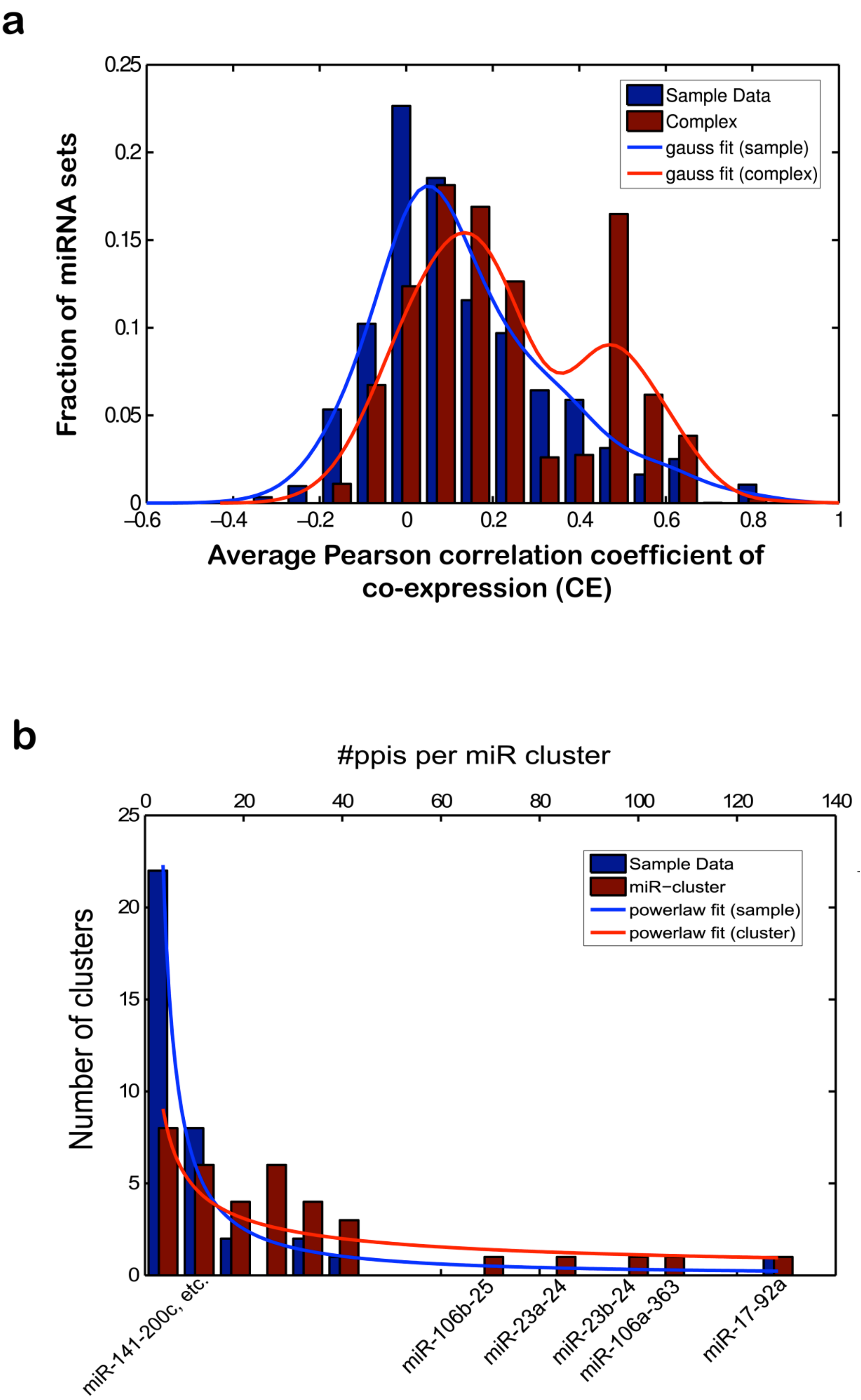


Figure 4.3

Statistical evidence of coordinate regulation by miRNAs.

a, Histogram for the average Pearson correlation of co-expression of miRNAs, which target the same protein complex (association *P* value < 0.05). b, Histogram for direct interactions of proteins targeted by *N* miRNAs within a cluster as compared to a null model of *N* randomly sampled miRNAs, respectively.

Protein complexes and miRNA expression

We next tested whether miRNAs, which target different components of the same protein complex, are more likely to be co-expressed. The average expression correlation (Co-Expression as calculated by Pearson correlation coefficients, hereafter termed CE values) of miRNAs were examined based on miRNA expression profiles obtained for 26 different organ systems and cell types(Landgraf et al., 2007). To test for statistical significance, we compared observed average CE values of the miRNA sets with corresponding values of 5,000 randomly sampled miRNA-mRNA target networks. For each randomly sampled miRNA-mRNA target network, we determined associated protein complexes and calculated the average CE value of the random miRNA sets, which have at least two targets in the protein complex. The statistical analysis revealed that the average CE value of 44% (51 out of 115) miRNA sets targeting the same protein complex was significantly higher than expected by chance (*P* value < 0.05, permutation test, 5,000 samples). 22 out of 51 significantly co-expressed miRNA sets were not in one transcription unit. 18 miRNA sets were members of the same miRNA gene family. In all, as shown in Figure 4.3a, miRNAs, which target the same complex, are significantly more co-expressed than expected by chance (*P* Value = 5.3 10^{-6}; Wilcoxon signed rank test).

Protein complex networks co-ordinately regulated by clusters of miRNAs

We systematically characterized the protein complex networks, which are simultaneously regulated by clustered miRNAs in 36 experimentally verified transcription units(Landgraf et al., 2007) (http://mips.helmholtz-muenchen.de/proj/mirnets). The interconnectivity of the target sets of the miRNA gene clusters was first assessed as follows: the number of protein-protein interactions between the target sets of each pair of miRNAs in the cluster was counted, and these values were compared to 1,000 randomly sampled sets of miRNAs. Comparing the observed number of interactions (Figure 4.3b) with the corresponding distributions of randomly sampled sets of miRNAs provides a strong indication that a significant fraction of miRNAs in clusters might coordinately regulate targets (*P* Value = 7.5 10^{-4}; Wilcoxon signed rank test,. The statistical analysis revealed 14 clusters, whose targets are significantly interconnected in the protein-protein interaction network (P value < 0.05, permutation test, 1,000 samples, Table 4.1).

Table 4.1 miRNA clusters with interconnected target sets.

miRNA cluster[15,*] (number of miRNAs) [miRNAs in the cluster with consensus target predictions]	Genomic location [chr | strand | loc(miR1)| loc(miRN)]	# targ	Ppis [#| P value]		Ppis [miR-miR] [# |P value]	
miR-17-92a (6) [*miR-17, miR-18a, miR-19a, miR-20a, miR-19b, miR-92a*]	chr13|+|90800863|90801647	661	132	0	187	0
miR-23a-24 (3) [*miR-23a, miR-27a, miR-24*]	chr19|-|13808093|13808473	576	88	0	25	0.2
miR-23b-24 (3) [*miR-23b, miR-27b, miR-24*]	chr9|+|94927055|94927932	593	100	0	29	0.132
miR-106a-363 (6) [*miR-106a, miR-18b miR-20b, miR-19b, miR-92a, miR-363*]	chrX|-|133028922|133029826	565	108	0	42	0.032
miR-29b-29a (2) [*miR-29b, miR-29a*]	chr7|-|130018752|130019555	250	26	0.002	18	0.024
miR-30e-30c (2) [*miR-30e, miR-30c*]	chr1|+|40889126|40892135	384	42	0.002	12	0.144
miR-29b-29c (2) [*miR-29b, miR-29c*]	chr1|-|204363595|204364265	246	22	0.002	16	0.072
miR-106b-25 (3) [*miR-106b, miR-93, miR-25*]	chr7|-|99335835|99336347	559	72	0.006	48	0.014
miR-183-182 (3) [*miR-183, miR-96, miR-182*]	chr7|-|129004187|129008789	443	44	0.01	37	0.074
let-7a-7d (3) [*let-7a, let-7f, let-7d*]	chr9|+|94017789|94020763	337	30	0.014	39	0.058
miR-206-133b (2) [*miR-206, miR-133b*]	chr6|+|52117110|52121776	387	36	0.02	8	0.236
miR-302b-367 (5) [*miR-302b, miR-302c, miR-302a, miR-302d, miR-367*]	chr4|-|113926627|113927317	379	32	0.022	94	0.084
miR-141-200c (2)	chr12|+|6943117|	197	14	0.026	3	0.548

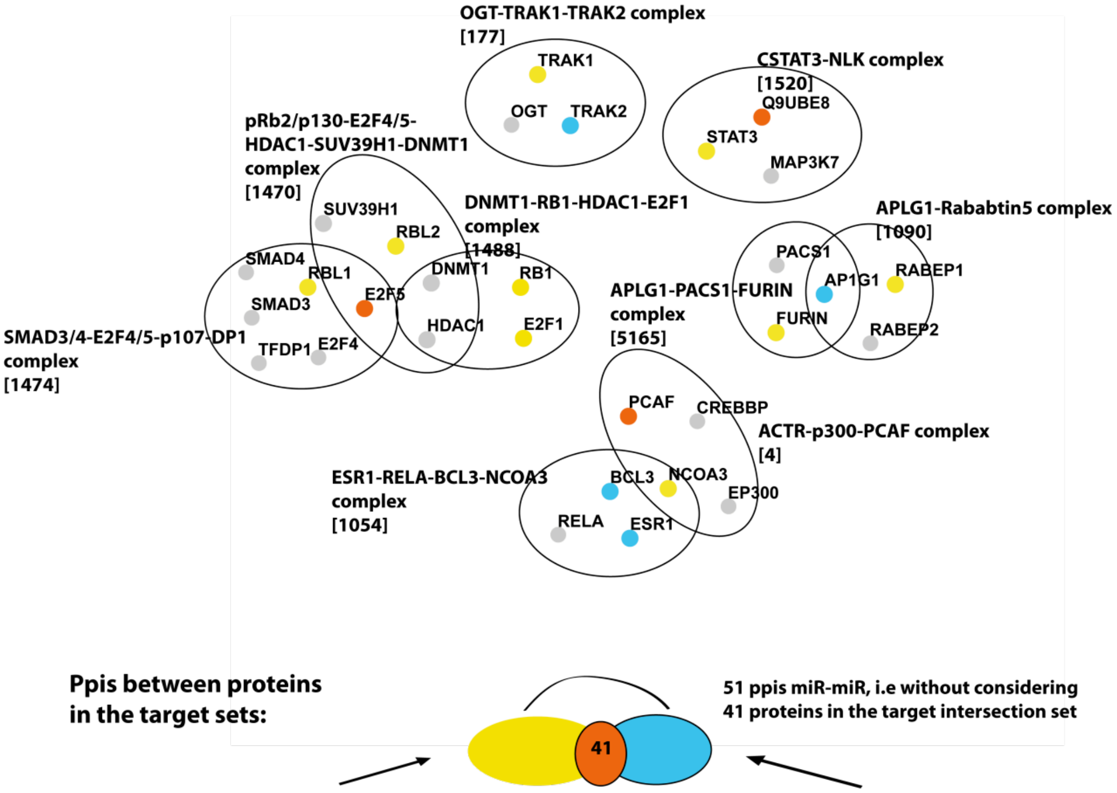
[*miR-141, miR-200c*]	6943610					
miR-199a-214 (2) [*miR-199a, miR-214*]	chr1|-|168839603| 168845421	319	26	0.03	11	0.144

*miRNA cluster is termed in the following way: *miR-first_miRNA-last_miRNA*. For instance, *miR-17-92b* cluster is consisted of six miRNAs, *miR-17* is the first miRNA in the cluster and *miR-92b* is the last miRNA in the cluster. Interconnectivity of the target sets was evaluated as: (1) number of interactions in the union target set and (2) number of interactions between target sets of all distinct miRNA pairs in a cluster (ppis [miR-miR]). The *P* values were estimated by comparing the observed value with 1,000 randomly sampled target sets of equal size

For instance, the *miR-17-92a* cluster, recently shown to be an important regulator in development and disease(Mendell, 2008), coordinately regulates single genes, for example, the proapoptotic gene *BCL2L11/BIM*. The *miR-17-92* cluster consists of six miRNAs, namely *miR-17*, *miR-18a*, *miR-19a*, *miR-20a*, *miR-19b*, *miR-92a*. *miR-17* and *miR-20a* have the same seed sequence so that they share most of predicted targets. Our network analysis reveals that miR-*17* and *miR-19b* target different components in nine protein complexes, suggesting synergistic regulation of these protein complexes (Figure 4.4). The *E2F* family of transcription factors is known to be tightly regulated by the *miR-17-92a* cluster(Mendell, 2008). Strikingly, transcriptional co-repressors of the retinoblastoma family, *Rb1/pRb*, *RBL1/p107*, and *RBL2/p130*, which are indispensable partners of *E2F* proteins, are all targeted by the *miR-17-92a* cluster (Figure 4.4, complex IDs 1470, 1474 and 1488).

CtBP/ZEB complex regulated by the *miR-141-200c* cluster

The network perspective provides fascinating insights of gene regulation by miRNA gene clusters and co-expressed miRNAs, whose target sets have not yet been analyzed at a systems-level. To explore this in detail we examined the protein complexes predicted to be coordinately regulated by the *miR-141-200c* cluster (Figure 4.5a). The *miR-141* and *miR-200c* genes are located on chromosome 12p 13.31, separated by a 338bp spacer sequence; *miR-141* and *miR-200c* belong to the *miR-200* family. The seed region of *miR-141* differs to that of *miR-200c* by one nucleotide at position 4 of the miRNA; therefore, *miR-141* and *miR-200c* have, based on the “seed” rule, different computationally predicted targets. Nevertheless, we found that the targets of *miR-141* and *miR-200c* are significantly interconnected (*P* value = 0.026,


OGT-TRAK1-TRAK2 complex [177]
TRAK1
OGT
TRAK2
CSTAT3-NLK complex [1520]
Q9UBE8
STAT3
MAP3K7
pRb2/p130-E2F4/5-HDAC1-SUV39H1-DNMT1 complex [1470]
SUV39H1
RBL2
DNMT1-RB1-HDAC1-E2F1 complex [1488]
DNMT1
RB1
SMAD4
RBL1
E2F5
SMAD3
HDAC1
E2F1
SMAD3/4-E2F4/5-p107-DP1 complex [1474]
TFDP1
E2F4
APLG1-Rababtin5 complex [1090]
PACS1
AP1G1
RABEP1
APLG1-PACS1-FURIN complex [5165]
FURIN
RABEP2
PCAF
CREBBP
ACTR-p300-PCAF complex [4]
ESR1-RELA-BCL3-NCOA3 complex [1054]
BCL3
NCOA3
EP300
RELA
ESR1
Ppis between proteins in the target sets:
51 ppis miR-miR, i.e without considering 41 proteins in the target intersection set
41

	# targets	miR-17			miR-18a			miR-19b		
		Target overlap	# ppi miR-miR	P value	Target overlap	# ppi miR- miR	P value	Target overlap	# ppi miR-miR	P value
miR-17	360									
miR-18a	71	11	18	0.074						
miR-19b	276	41	51	**0.038**	52	70	0.059			
miR-92a	206	16	55	0.057	6	14	0.051	16	36	0.195

Figure 4.4

Protein complexes regulated by the *miR-17-92a* cluster.

Protein complexes, which are significantly associated with the *miR-17-92a* cluster and are targets of *miR-17* and *miR-19b*, are shown in circles. CORUM complex IDs are given in brackets. Targets of *miR-17* are shown in yellow, targets of *miR-19b* in blue and common targets in orange. In the table, for each pair of miRNAs in the cluster (1) the number of common targets (target overlap), (2) the number of interactions between the target sets (miR-miR, i.e. without considering proteins in the target overlap), and (3) interconnection *P* values for each pair of miRNAs are listed. *miR-19a* and *miR-20a* are not listed, since their seeds and predicted target sets are nearly identical with *miR-19b* and *miR-17*, respectively.

permutation test, 1,000 samples, Table 4.1); ten experimentally verified protein complexes were found enriched in the union target set.

Very recent reports have shown that the *miR-200* family regulates epithelial to mesenchymal transition (EMT) by targeting the transcriptional repressor zinc-finger E-box binding homebox 1 (*ZEB1*) and *ZEB2*(Burk et al., 2008; Gregory et al., 2008; Korpal et al., 2008; Nakada et al., 2008; Park et al., 2008). During EMT, the *miR-141-200c* cluster and the tumor invasion suppressor gene *E-cadherin* are downregulated by *ZEB1/2*(Burk et al., 2008). *ZEB1* and *ZEB2* repress transcription through interaction with corepressor *CtBP* (C-terminal binding protein)(Postigo and Dean, 1999). Interestingly, several essential components of the *CtBP/ZEB* complex, namely *ZEB1/2*, *CtBP2*, *RCOR3* (REST corepressor 3) and *CDYL* (Chromodomain Y-like protein), are predicted targets of the *miR-141-200c* cluster. *CtBP2* has one *miR-141* target site and one *miR-200c* target site, while *ZEB1* has two *miR-200c* target sites. The *CtBP/ZEB* complex mediates the transcriptional repression of its target genes by binding to their promotors and altering the histone modification(Shi et al., 2003).

We could show that overexpression of *miR-141* and *miR-200c* led to reduced expression of *CtBP2* and *ZEB1* in human pancreatic carcinoma (PANC-1) cells (Figure 4.5b). These results are also confirmed on protein level (Figure 4.5c). As the functional consequence of miRNA overexpression, the expression of *E-cadherin* mRNA is greatly upregulated (Figure 4.5b), indicating that the repression activity of *CtBP/ZEB* complex is compromised. The interaction between the *miR-141-200c* cluster and multiple components of the *CtBP/ZEB* complex suggests a coordinated regulation of the repression activity for the *CtBP/ZEB* complex. Intriguingly, the *miR-141-200c* cluster also targets *β-catenin*, which is a shared component of cell adhesion and *Wnt* signalling(Bienz, 2005). *β-catenin* is found in the plasma membrane, where it promotes cell adhesion by binding to *E-cadherin*, in the cytoplasm, where it is easily phosphorylated and degraded in the absence of a *Wnt* signal, and in the nucleus, where it binds to *TCF* transcription factors and induces the transcription of *Wnt* target genes. Most protein-interacting motifs of *β-catenin* overlap in such a way that its interactions with each of its protein partners are mutually exclusive(Bienz, 2005). Since the *miR-141-200c* cluster and *E-cadherin* are both downregulated during EMT, it is tempting to speculate that more *β-catenin* would be made available for participating in transactivating downstream genes, which may contribute to the progress of cancer(Nakada et al., 2008).

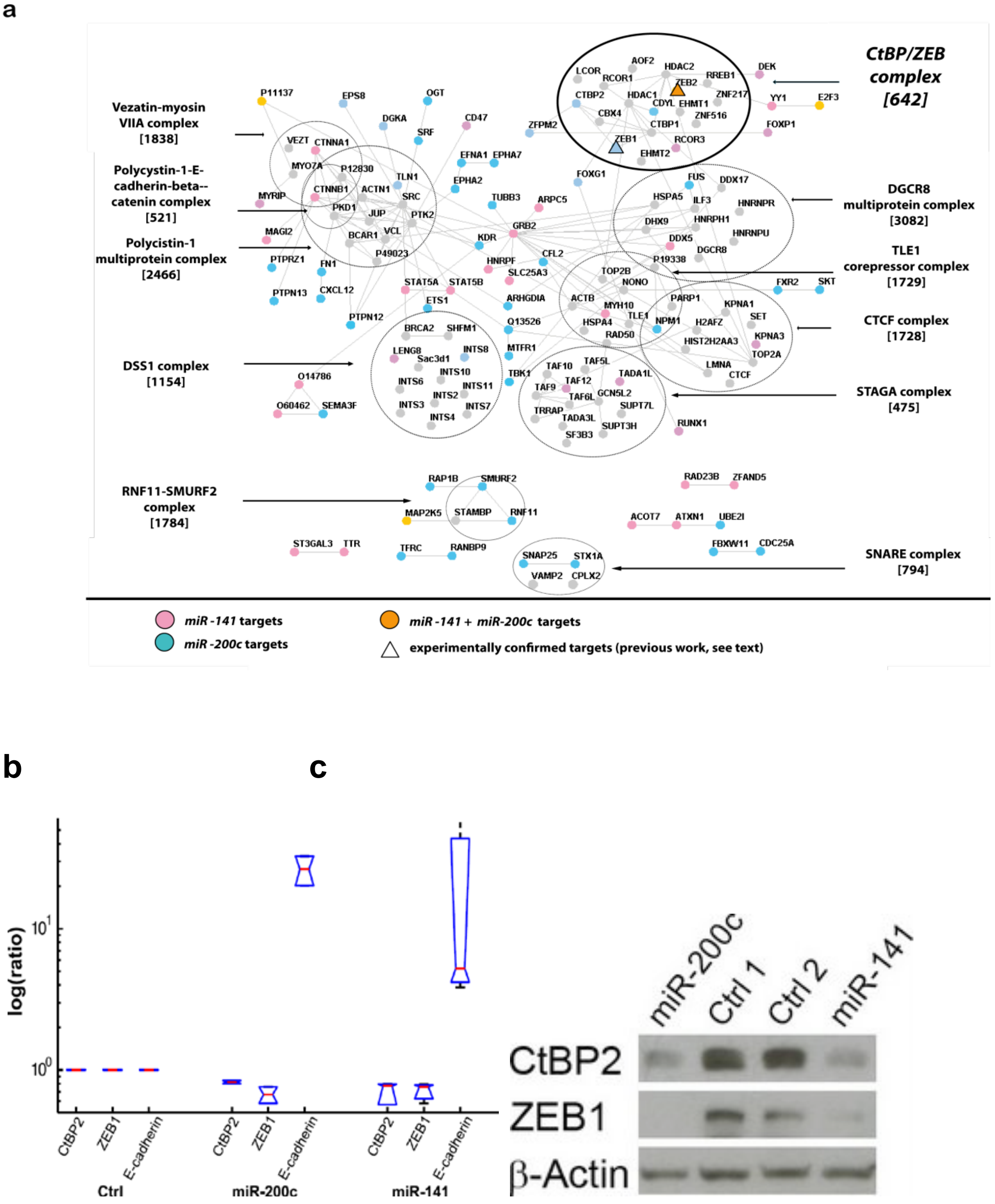


Figure 4.5

Protein complexes regulated by the *miR-141-200c* cluster. a, The *miR-141-200c* cluster is predicted to regulate 197 targets. Targets of individual miRNAs in the cluster, and common targets are indicated by different colors, additional proteins in the protein complexes are grey. Experimentally verified protein-protein interactions from the BIND, MINT, IntAct and HPRD databases are indicated as edges. CORUM complex IDs are given in brackets. b, Real-time reverse transcription–PCR of *CtBP2* and *ZEB1* after transfection of the indicated miRNAs in undifferentiated cancer cells (PANC-1). The expression levels of *E-cadherin* (of which the transcription is repressed by

CtBP/ZEB complex) are included as positive controls. c, Confirmation of the results on protein levels by immunoblots.

4.2.3 Discussion

MicroRNAs and their functions have been a fascinating research topic in recent years(Bushati and Cohen, 2007; Filipowicz et al., 2008b; Flynt and Lai, 2008). In animals, miRNA-guided regulations of gene expression are likely to involve hundreds of miRNAs and their targets. Genetic studies have successfully elucidated some miRNA activities, termed genetic switches, which have intrinsic phenotypic consequences(Bushati and Cohen, 2007; Flynt and Lai, 2008). miRNA activities can be classified based on whether their major effect is conveyed through one, a few or many targets (from tens to hundreds). All genetic switches discovered so far belong to the former class (a few targets). It is unclear how the latter class, termed target battery(Flynt and Lai, 2008), which might be subtly regulated on the protein level (Baek et al., 2008; Selbach et al., 2008), contributes to proper phenotypes.

In this part of the thesis, I have presented first comprehensive analysis of human protein complexes, which are coordinately regulated by miRNAs. Our statistical analysis suggests that, by simultaneously targeting several components of protein complexes, a single miRNA or co-expressed miRNAs may have cumulative effects. To demonstrate this, we experimentally verified that the *miR141-200c* cluster interacts with two different components of the *CtBP/ZEB* complex. The functional analysis of the miRNA-regulated protein complexes revealed a clear bias towards transcriptional regulation, signal transduction, cell cycle and chromatin regulation, for which confirmation has been reported only by individual experimental studies of selected miRNAs. Our approach provides improved candidate miRNA target lists to experimentalists, as demonstrated by a benchmark against large-scale, quantitative proteomics data.

Some ancient miRNA genes are deeply conserved in the kingdom Animalia (Christodoulou et al., ; Wheeler et al., 2009) or in the kingdom Plantae (Axtell et al., 2007) while during the evolution, young miRNA genes were constantly created, fixed or lost (Fahlgren et al., 2007; Lu et al., 2008a; Lu et al., 2008b; Rajagopalan et al., 2006). Interestingly, the genomic organization of some miRNA clusters were well preserved for millions years, implying a functional incentive to keep such configurations (Ambros, 2008;

Glazov et al., 2008). The evolution of homogeneous miRNA clusters can be explained by classical gene duplication theory (Altuvia et al., 2005). The regulatory effect of such clusters might merely be an increase of dosage. The evolution of hetergeneous miRNA clusters is more complicated. Two different miRNAs can be located near each other by various genomic events, such as recombination, transposon insertion, etc. Or large number hairpin repeats might evolve into miRNAs of different families, for example, the largest human miRNA cluster miR-379-656 (Glazov et al., 2008). This cluster consists of different miRNA families which are evolved by tandem duplication of an ancient hairpin sequence. Once a newly formed miRNA cluster proves to provide functional advantage, which might be coordinated regulation of protein complexes, the genomic organization of such a cluster can be fixed by evolution (Lu et al., 2008a).

As a general mechanism to aid the formation of macromolecular protein complexes(Keene, 2007), RNA operons, mostly sequence-specific RNA binding proteins, are recently suggested to coordinately regulate some functionally related mRNAs. In such a scenario, mRNAs of different components of a protein complex are brought together by associating with specific RNA operons. The localization of these mRNAs might also facilitate the simultaneous interaction of miRNAs and their corresponding target mRNAs. Interestingly, RNA operons bind to motifs, which are sometimes located in the 3'UTRs of mRNAs. Thus, the competition or cooperation between miRNA binding and RNA operon binding might be a research topic worth pursuing.

The protein complex networks for 164 human miRNAs and 36 genomic clusters of miRNAs can be accessed on a data-mining website (http://mips.helmholtz-muenchen.de/proj/mirnets). The concepts presented here can be used as a starting point for experimentalists to systematically evaluate miRNAs and targets interactions at a systems level.

Reference:

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T., and Margalit, H. (2005). Clustering and conservation patterns of human microRNAs. Nucleic Acids Res *33*, 2697-2706.

Ambros, V. (2008). The evolution of our thinking about microRNAs. Nature Medicine *14*, 1036-1040

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics *25*, 25-29

Axtell, M.J., Snyder, J.A., and Bartel, D.P. (2007). Common functions for diverse small RNAs of land plants. Plant Cell *19*, 1750-1769.

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71

Bienz, M. (2005). beta-Catenin: a pivot between cell adhesion and Wnt signalling. Current Biology: CB *15*, R64-67

Bonci, D., Coppola, V., Musumeci, M., Addario, A., Giuffrida, R., Memeo, L., D'Urso, L., Pagliuca, A., Biffoni, M., Labbaye, C.*, et al.* (2008). The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. Nature Medicine *14*, 1271-1277

Burk, U., Schubert, J., Wellner, U., Schmalhofer, O., Vincan, E., Spaderna, S., and Brabletz, T. (2008). A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. EMBO Reports *9*, 582-589

Bushati, N., and Cohen, S.M. (2007). microRNA functions. Annual Review of Cell and Developmental Biology *23*, 175-205

Carleton, M., Cleary, M.A., and Linsley, P.S. (2007). MicroRNAs and cell cycle regulation. Cell Cycle (Georgetown, Tex) *6*, 2127-2132

Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S., Snyman, H., Hannon, G.J., Bork, P., and Arendt, D. Ancient animal microRNAs and the evolution of tissue identity. Nature *463*, 1084-1088.

Cui, Q., Yu, Z., Purisima, E.O. and Wang, E. (2006) Principles of microRNA regulation of a human cellular signaling network. Molecular Systems Biology, 2, 46.

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

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008a). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet *9*, 102-114.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008b). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews Genetics *9*, 102-114

Flynt, A.S., and Lai, E.C. (2008). Biological principles of microRNA-mediated regulation: shared themes amid diversity. Nature Reviews Genetics *9*, 831-842

Flynt, A.S., Li, N., Thatcher, E.J., Solnica-Krezel, L., and Patton, J.G. (2007). Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate. Nature Genetics *39*, 259-263.

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., and Zamore, P.D. (2005). Normal microRNA Maturation and Germ-Line Stem Cell Maintenance Requires Loquacious, a Double-Stranded RNA-Binding Domain Protein. PLoS Biol *3*, e236.

Friggi-Grelin, F., Lavenant-Staccini, L., and Therond, P. (2008). Control of antagonistic components of the hedgehog signaling pathway by microRNAs in Drosophila. Genetics *179*, 429-439

Glazov, E.A., McWilliam, S., Barris, W.C., and Dalrymple, B.P. (2008). Origin, evolution, and biological role of miRNA cluster in DLK-DIO3 genomic region in placental mammals. Mol Biol Evol *25*, 939-948.

Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., and Goodall, G.J. (2008). The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature Cell Biology *10*, 593-601

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Research *36*, D154-158

Hsu, C.-W., Juan, H.-F., and Huang, H.-C. (2008). Characterization of microRNA-regulated protein-protein interaction network. Proteomics *8*, 1975-1979

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nature Reviews Genetics *8*, 533-543

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nature Genetics *39*, 1278-1284

Korpal, M., Lee, E.S., Hu, G., and Kang, Y. (2008). The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. The Journal of Biological Chemistry *283*, 14910-14914

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M*., et al.* (2005). Combinatorial microRNA target predictions. Nature Genetics *37*, 495-500

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M*., et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414

Liang, H., and Li, W.-H. (2007). MicroRNA regulation of human protein protein interaction network. RNA (New York, NY) *13*, 1402-1408

Lu, J., Fu, Y., Kumar, S., Shen, Y., Zeng, K., Xu, A., Carthew, R., and Wu, C.I. (2008a). Adaptive evolution of newly emerged micro-RNA genes in Drosophila. Mol Biol Evol *25*, 929-938.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R.W., Wang, S.M., and Wu, C.I. (2008b). The birth and death of microRNA genes in Drosophila. Nat Genet *40*, 351-355.

Mahajan, M.C., Narlikar, G.J., Boyapaty, G., Kingston, R.E., and Weissman, S.M. (2005). Heterogeneous nuclear ribonucleoprotein C1/C2, MeCP1, and SWI/SNF form a chromatin remodeling complex at the beta-globin locus control region. Proceedings of the National Academy of Sciences of the United States of America *102*, 15012-15017

Martello, G., Zacchigna, L., Inui, M., Montagner, M., Adorno, M., Mamidi, A., Morsut, L., Soligo, S., Tran, U., Dupont, S*., et al.* (2007). MicroRNA control of Nodal signalling. Nature *449*, 183-188

Mendell, J.T. (2008). miRiad roles for the miR-17-92 cluster in development and disease. Cell *133*, 217-222

Nakada, C., Matsuura, K., Tsukamoto, Y., Tanigawa, M., Yoshimoto, T., Narimatsu, T., Nguyen, L.T., Hijiya, N., Uchida, T., Sato, F*., et al.* (2008). Genome-wide microRNA expression profiling in renal cell carcinoma: significant down-regulation of miR-141 and miR-200c. The Journal of Pathology *216*, 418-427

Neumüller, R.A., Betschinger, J., Fischer, A., Bushati, N., Poernbacher, I., Mechtler, K., Cohen, S.M., and Knoblich, J.A. (2008). Mei-P26 regulates microRNAs and cell growth in the Drosophila ovarian stem cell lineage. Nature *454*, 241-245

Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., and Hatzigeorgiou, A.G. (2009). The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Research *37*, D155-158

Park, S.-M., Gaur, A.B., Lengyel, E., and Peter, M.E. (2008). The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes & Development *22*, 894-907

Postigo, A.A., and Dean, D.C. (1999). ZEB represses transcription through interaction with the corepressor CtBP. Proceedings of the National Academy of Sciences of the United States of America *96*, 6683-6688

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev *20*, 3407-3425.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V.*, et al.* (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Research *36*, D646-650

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63

Shi, Y., Sawada, J.-i., Sui, G., Affar, E.B., Whetstine, J.R., Lan, F., Ogawa, H., Luke, M.P.-S., Nakatani, Y., and Shi, Y. (2003). Coordinated histone modifications mediated by a CtBP co-repressor complex. Nature *422*, 735-738

Silver, S.J., Hagen, J.W., Okamura, K., Perrimon, N., and Lai, E.C. (2007). Functional screening identifies miR-315 as a potent activator of Wingless signaling. Proceedings of the National Academy of Sciences of the United States of America *104*, 18151-18156

Volinia, S., Calin, G.A., Liu, C.-G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M.*, et al.* (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. Proceedings of the National Academy of Sciences of the United States of America *103*, 2257-2261

Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., and Peterson, K.J. (2009). The deep evolution of metazoan microRNAs. Evol Dev *11*, 50-68.

Chapter 5 Summary and Discussion

The works presented in this thesis cover different topics in the field of small RNA research, including identifying miRNA genes(Paterson et al., 2009), analyzing miRNA gene promoters(Wang et al., 2006), profiling vsRNA populations(Donaire et al., 2009), and investigating miRNA target interactions(Dietmann et al., submitted to *Nucleic Acids Research*). It is now a great opportunity to discuss the results in the light of current knowledge of small RNAs.

5.1 Reevaluation

In Chapter 2, I presented the work related to miRNA genes. The first part is about how to identify miRNA genes in a newly sequenced genome. I used the *Sorghum bicolor* genome as an example. Ideally should we have used deep sequencing technology to sample different *Sorghum bicolor* tissues/organs in various development stages or under different stresses, especially drought treatment. Only by doing so, we would be able to get a more comprehensive profile of *Sorghum bicolor* small RNA/miRNA. In the *Sorghum bicolor* genome project, we used rice miRNAs to annotate sorghum miRNAs. As a consequence, we missed all the *Sorghum bicolor* specific miRNAs which may contribute significantly to sorghum's characteristic drought resistant ability. However, it is still intriguing that a drought activated miRNA gene in Rice has been duplicated in sorghum. But whether these two duplicated miRNA genes are both functional in *Sorghum bicolor* remains an open question.

The second part of Chapter 2 focuses on the promoter analysis of miRNA genes. I used the newly evolved *Arabidopsis thanliana* miRNA genes and compared their promoter regions. As previously mentioned in Chapter 1: miR157 and miR156 indeed belong to one family, and so do miR165 and miR166[1] . Therefore, miR157 and miR165 families are not non-conserved *Arabidopsis thaliana* miRNA gene families! It is just the naming of the miR157/miR165 family implies that miR157/miR165 forms one miRNA family of itself. In rice, there are no miR157 and miR165 family. The promoter analysis of miR157 and miR165 is correct. Since miR165 belongs to a conserved miRNA family, it is then not surprising that the precursors of miR165

[1] page 13 and 15

are conserved only in the miRNA:miRNA* region[2]. Ath-miR157a,b genes in *Arabidopsis thaliana* were duplicated quite recently. The promoter regions of ath-miR157a,b genes share significant similarities. The overall conclusion of the miRNA promoter analysis is still supported by unconserved miRNA gene families, miR158, miR405 and miR447.

In Chapter 4, I presented our work on miRNA target interactions. We found strong statistical evidences to support that co-expressed human miRNAs and clustered miRNAs might coordinately regulate human protein complexes. Furthermore, we experimentally showed that miR141-200c cluster indeed regulates two components of the *CtBP/ZEB* complex. miR-141 and miR-200c belong to mir-8 family. The sequence difference between mature miR-141 and miR-200c is only four nucleotides but there is one nucleotide difference in the seed region. Consequently, the targets of miR-141 and miR-200c are different according to the current understanding of animal miRNA and target interactions. Experimentally we found miR-141 and miR-200c regulated some genes with different efficiency, which can not solely be explained by the seed-rule. It would be ideal if we could experimentally verify that a cluster of miRNAs or co-expressed miRNAs, which belong to different miRNA families, target different components of a protein complex. Moreover, further experiments would be needed to show that it is because of this simultaneous targeting of miRNAs that the protein complex as a whole is regulated effectively. Nevertheless, although we studied human miRNAs and targets, the concept that coexpressed small RNAs may synergistically target protein complexes for a more efficient regulation is of course not limited to animal miRNAs.

5.2 Small RNA Deep sequencing

In Chapter 3, I presented our deep sequencing results of viral small RNAs. I showed that sense and antisense vsRNAs spread throughout the entire virus in a characteristic overlapping configuration in a way that virtually all nucleotide positions within the viral genome were sequenced. Our data suggest that every genomic position is a putative cleavage site for vsRNA formation, although each viral genome contains regions that serve as preferential sources of vsRNA production. Hotspots of 21-, 22-, and 24-nt vsRNAs usually originated from the same genomic regions, indicating similar target affinities between DCL enzymes.

[2] page 54

Plant endogenous small interfering RNAs were the very first deep sequenced small RNA populations, firstly by MPSS (Massively Parallel Sinature Sequencing) technology(Lu et al., 2005), later by 454(Kasschau et al., 2007; Rajagopalan et al., 2006) and Illumina(Fahlgren et al., 2009; Lister et al., 2008). The first wave of small RNA data avalanche immediately challenged the community not only conceptually but also technically. How to make sense out of these tremendous amount of data (Table 5.1) and how to take advantage of them to guide wet lab experiments?

Table 5.1

Deep sequencing small RNAs in *Arabidopsis thaliana*

Reference	method	Accession	Genetic Background	Organ/Tissue	Small RNA No.
Lu, C. et al. Science, 2005. 309(5740): p. 1567-9. [4].	MPSS	Columbia	wild-type	inflorescence, seedling	81,774
Henderson, I.R., et al. Nat Genet, 2006.	454	Columbia	wild-type, dcl2/3/4	inflorescence	13,743
Lu C. et al. Genome Research, doi:10.1101 /gr.5530106	MPSS, 454	Columbia	wild-type, rdr2, rdr6, dcl1-7	inflorescence	30,100
Qi, Y.J., et al. Nature, 2006. 443(7114): p. 1008-1012.	454	Landsberg erecta	wild-type	whole plant, AGO1,AGO4	45,502
Rajagopalan R., et al. Genes Dev. 2006 Dec 15;20(24):3407-25.	454	Columbia	wild-type	flower, rosette leaves, seedling, siliques	340,114
Kasschau, et al. PLoS Biol 5 (2007), e57.	454	Columbia	wild-type, rdr1-1, rdr2-1, rdr6-15, dcl1-7, dcl2-1, dcl3-1,dcl4-2	inflorescence, seedling, leaves	206,077
Zhang, X. et al. PNAS, vol. 104, no. 11, 4536-4541, 2007	454	Columbia	wild-type, nrpd1a1b, nrpd2a2b, F1 (nrpd1a1b X nrpd2a2b)	inflorescence	92,717

One interesting observation from these deep sequencing projects was that the number of shared common small RNAs from the same tissues/organs of *Arabidopsis thaliana*, sequenced by different labs or different technology, is surprisingly low. For example, out of 77,434 unique 17-nt MPSS signatures, only 13,596 matched with later sequenced 454 small RNAs(Rajagopalan et al., 2006). The modest overlap was contributed to the fact that the deep sequencing techniques were still far from saturating the small RNA pools of *Arabidopsis thaliana*. But when I compared small RNA libraries of two 454 deep sequencing projects, I found that the overlaps were so low that it is hard to believe that they were sequenced from the very same organs of *Arabidopsis thaliana*.

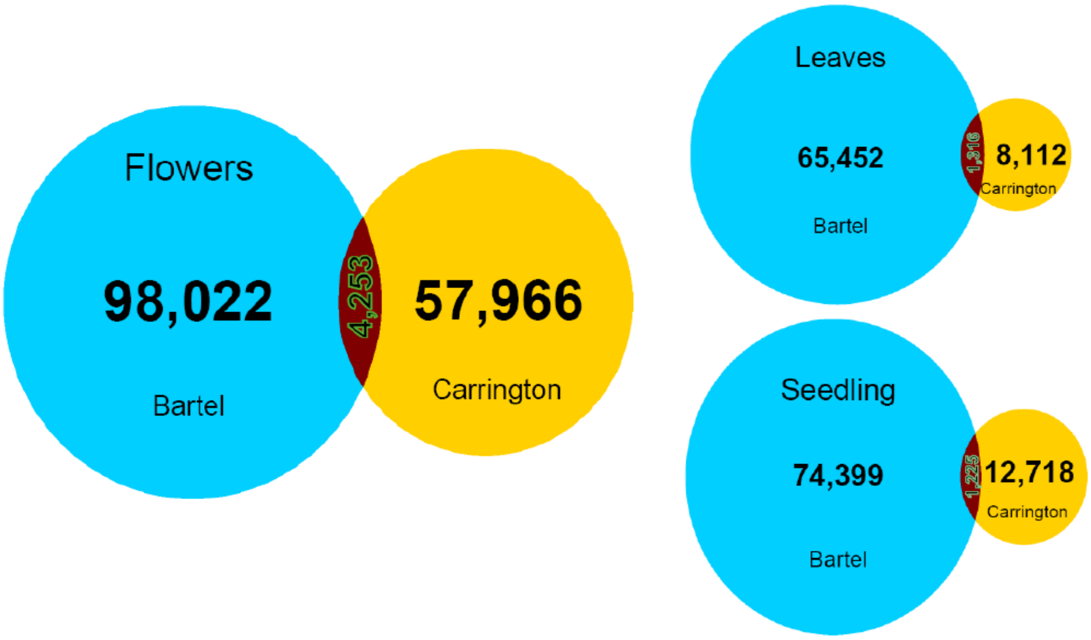


Figure 5.1

Venn diagrams of 454 sequencing results. Numbers are the unique small RNA sequences numbers in different libraries(Kasschau et al., 2007; Rajagopalan et al., 2006).

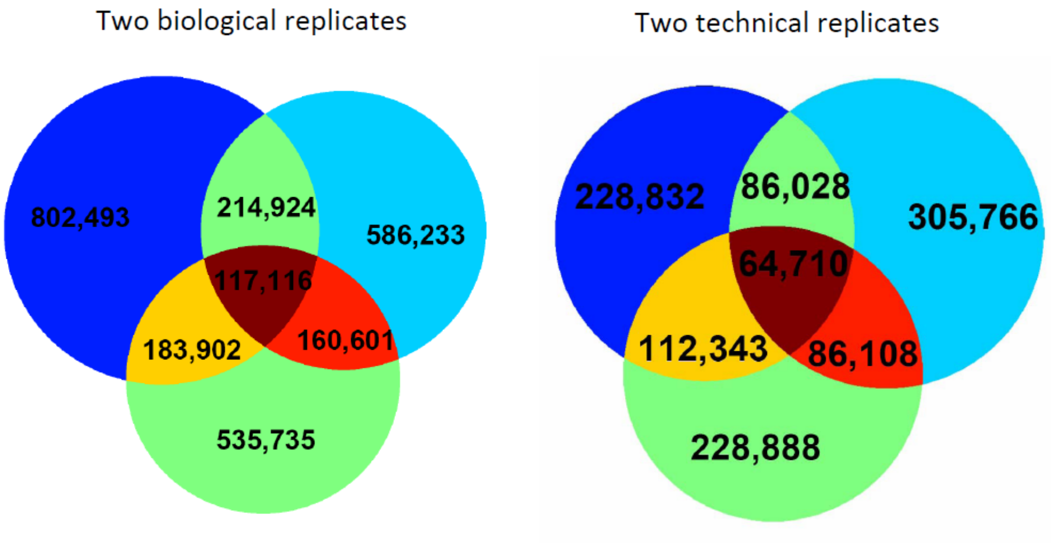


Figure 5.2

Venn diagrams of deep sequenced sRNA population from biological and technical replicates. Numbers are the unique small RNA sequences numbers in different libraries. Dark bule represents the original sample, light blue and green represent two replicates(Fahlgren et al., 2009).

It also raises a question about how deep we should sequence in order to get a complete picture of small RNA populations in *Arabidopsis thaliana*. Illumina sequencing technology pushed the sequencing depth into a new level(Fahlgren et al., 2009; Lister et al., 2008). Nevertheless, the overlaps of commonly shared small RNAs between biological replicates and technical replicates are still poor(Fahlgren et al., 2009) (Figure 5.2).

When I compared two libraries of *Arabidopsis thaliana* flower, sequenced using Illumina technology by two labs(Fahlgren et al., 2009; Lister et al., 2008), the overlap of small RNAs (Figure 5.3) is not much better than the overlap of 454 libraries (Figure 5.1). In 454 libraries, overlapped sRNAs count 2.7% of all the unique sRNA sequences. In Illumina libraries, overlapped sRNAs count 3,9% of all the unique sRNA sequences.

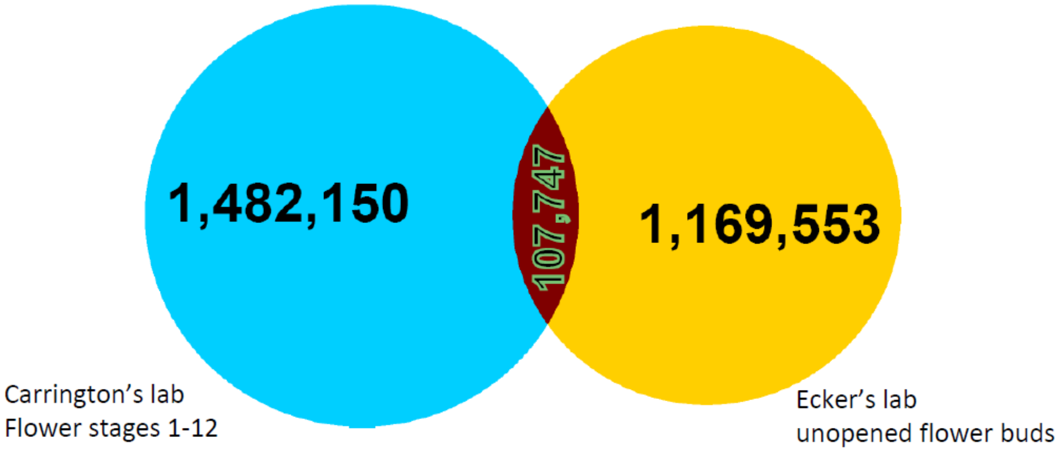


Figure 5.3
Venn diagram of deep sequenced sRNA population from flowers by two different labs(Fahlgren et al., 2009; Lister et al., 2008).

One explanation of the limited overlap is caused by the biogenesis of the majority of plant small interfering RNAs. Plant siRNAs are generated from repeat enriched genomic loci. Typically, an RNA transcript generated from one of such loci is subjected to RDR2 dependant creation of dsRNAs. The perfectly based paired dsRNA is cleaved then by DCL3. As contrasted to miRNA precursors, which have structure constraints to limit the possibilities of DCL1 cleavage, the cleavage of a dsRNA can be on every possible nucleotide of the transcript (Figure 1.3). This biogenesis scenario is very much like the biogenesis of vsRNAs as discussed in detail in the previous part of this thesis. Consequently, when a small RNA library is prepared, only a snap shot of a sRNA population is taken. At that particular moment, siRNAs generated from some dsRNAs are more or less a random collection from a pool of siRNAs which can all be mapped to those repeated enriched loci. Although the overlapping between two snap shots of siRNA population might be poor, as shown here between libraries sequenced by different labs(Fahlgren et al., 2009; Kasschau et al., 2007; Lister et al., 2008; Rajagopalan et al., 2006) (Figure 5.1 and Figure 5.3) and even among biological replicates and technical replicates (Figure 5.2)(Fahlgren et al., 2009), many of non-overlapped siRNAs of different libraries can still be derived from the same repeat enriched genomic loci. To test this hypothesis, I mapped small RNAs of flower libraries(Kasschau et al., 2007; Rajagopalan et al., 2006) sequenced by 454 to the *Arabidopsis thaliana* genome. After mapping, I checked the overlap of the genomic loci of two different flower libraries (Figure 5.4). The overlap is significantly improved, which now counts 32.4% of all sRNA mapped genomic loci. This result suggests that when comparing two siRNA populations of *Arabidopsis thaliana*, it is necessary to cluster siRNA sequences or map sRNA sequences back to the genome and then compare the genomic loci.

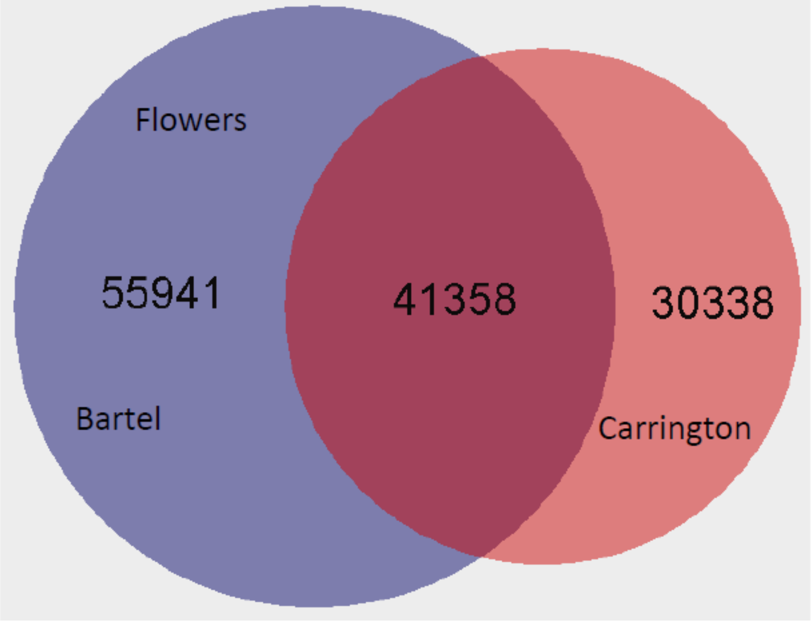


Figure 5.4
Venn diagram of sRNA mapped genomic loci of flowers using 454 sequencing from two different labs(Kasschau et al., 2007; Rajagopalan et al., 2006).

5.3 Uncharted territories

It was a rewarding journey in the last couple years to work with these fascinating small RNA molecules despite the fact that most of time they were just short strings of A, U, G, C in my computer. I am extremely grateful to have opportunities to work with experimentalists. It is only through these collaborations(Donaire et al., 2009; Paterson et al., 2009) (Dietmann et al., submitted *Nucleic Acids Research*) that these fascinating molecules become real. Where to go from here? What's next thing to do? One may ask. Here is just one of many possibilities which might again result in an exciting story.

In plant, it was commonly believed that most of plant miRNAs interact with their targets at a near perfect complimentary target site and cleave their targeted mRNAs. There are a few exceptions. First, miR390, guided by AGO7, binds to both 5' and 3' of the TAS3 transcript to generate *in phased* tasiRNAs. Interestingly, the 5' target site of miR390 has central mismatches (Figure 5.5A), preventing miR390 from cleaving TAS3 at 5'. And the configuration of the 5' target site is essential for the TAS3 pathway (Montgomery et al., 2008). This example suggests that the non-cleavage target site is functional. Second, miR399 has a target site at IPS1(Induced by Phosphate Starvation1) RNA, which has a central gap to prevent cleavage (Figure 5.5B). Instead of being cleaved, IPS1 then sequesters miR399 to prevent it from regulating PHO2 mRNA(Franco-Zorrilla et al., 2007).

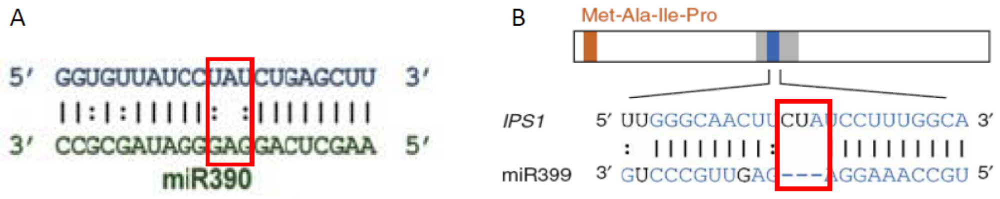
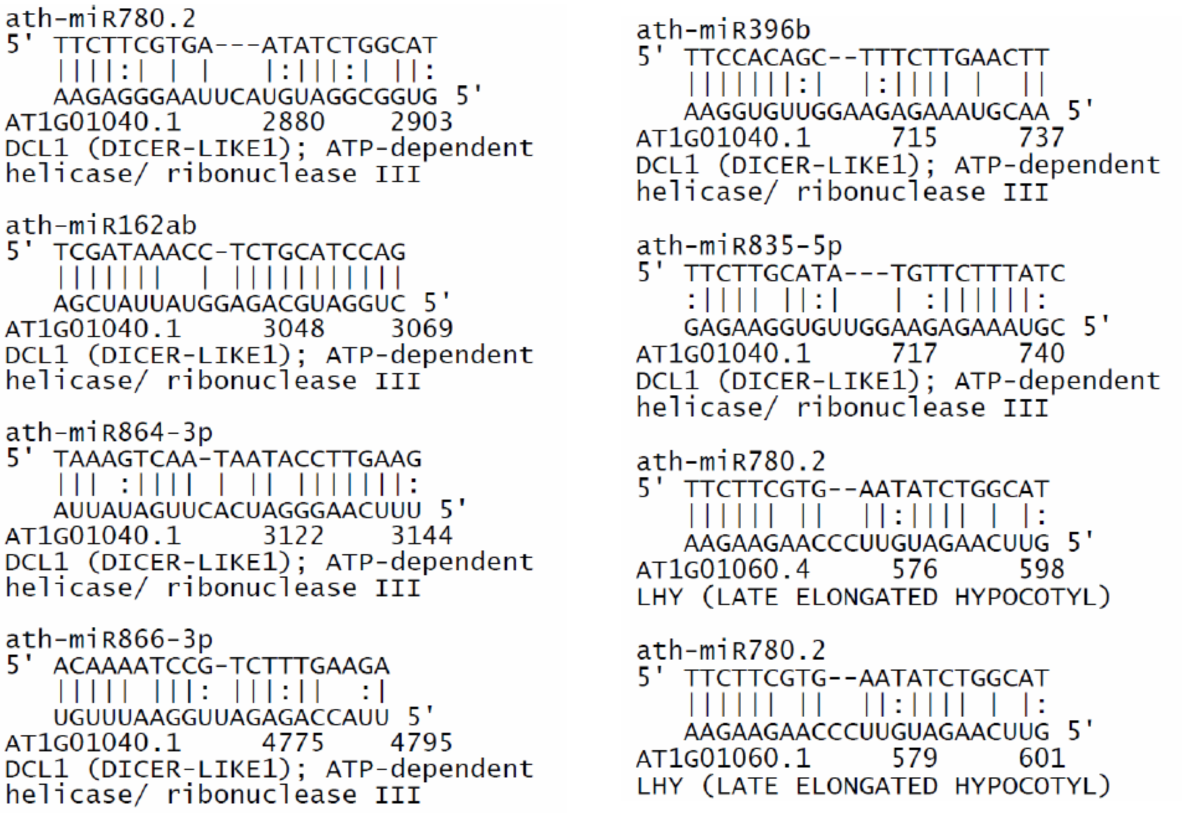

A
5' GGUGUUAUCCUAUCUGAGCUU 3'
||:|:|||||: :|||||||||
3' CCGCGAUAGGGAGGACUCGAA 5'
miR390
B
Met-Ala-Ile-Pro
IPS1 5' UUGGGCAACUUCUAUCCUUUGGCA 3'
: ||||||||: |||||||||
miR399 3' GUCCCGUUGAG---AGGAAACCGU 5'

Figure 5.5
Known plant miRNA target sites with central mismatches

Since translational repression has been shown to be a general mode of plant miRNA action(Brodersen et al., 2008) and target sites with limited complementarities can also induce translational repression(Dugas and Bartel, 2008), it is tempting to speculate that many plant miRNA target sites of central gaps and limited complementarities might have escaped detection so far. I checked the case of potential target sites with central gaps. Astonishingly, numerous target sites with central gaps do exist in *Arabidopsis* *thaliana* (Figure 5.6). In a pilot study, there are hints some of these target sites can indeed repress the translation of mRNAs (Peter Brodersen, personal communication). This is just one of many virgin territories in the wonderful land of small RNAs, awaiting exploration and admiration from scientific adventurers.



```
ath-miR780.2
5' TTCTTCGTGA---ATATCTGGCAT
   ||||:| | |   |:|||:| ||:
   AAGAGGGAAUUCAUGUAGGCGGUG 5'
AT1G01040.1      2880    2903
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III

ath-miR162ab
5' TCGATAAACC-TCTGCATCCAG
   |||||||  | |||||||||||
   AGCUAUUAUGGAGACGUAGGUC 5'
AT1G01040.1      3048    3069
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III

ath-miR864-3p
5' TAAAGTCAA-TAATACCTTGAAG
   ||| :|||| | || |||||||:
   AUUAUAGUUCACUAGGGAACUUU 5'
AT1G01040.1      3122    3144
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III

ath-miR866-3p
5' ACAAAATCCG-TCTTTGAAGA
   ||||| |||: |||:||  :|
   UGUUUAAGGUUAGAGACCAUU 5'
AT1G01040.1      4775    4795
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III
```

```
ath-miR396b
5' TTCCACAGC--TTTCTTGAACTT
   |||||||:|  |:|||| |  ||
   AAGGUGUUGGAAGAGAAAUGCAA 5'
AT1G01040.1      715     737
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III

ath-miR835-5p
5' TTCTTGCATA---TGTTCTTTATC
   :|||| ||:|   | :||||||:
   GAGAAGGUGUUGGAAGAGAAAUGC 5'
AT1G01040.1      717     740
DCL1 (DICER-LIKE1); ATP-dependent
helicase/ ribonuclease III

ath-miR780.2
5' TTCTTCGTG--AATATCTGGCAT
   |||||| ||  ||:|||| | |:
   AAGAAGAACCCUUGUAGAACUUG 5'
AT1G01060.4      576     598
LHY (LATE ELONGATED HYPOCOTYL)

ath-miR780.2
5' TTCTTCGTG--AATATCTGGCAT
   |||||| ||  ||:|||| | |:
   AAGAAGAACCCUUGUAGAACUUG 5'
AT1G01060.1      579     601
LHY (LATE ELONGATED HYPOCOTYL)
```

Figure 5.6

Potential plant miRNA target sites with central mismatches

Reference:

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

Dietmann S., Burk U., Brabletz S., Lutter D., Mayer K., Brabletz T., Theis F., Ruepp A., Wang Y., MicroRNAs coordinately regulate protein complexes (2010) , Nucleic Acids Research, under revision.

Donaire, L., Wang, Y., Gonzalez-Ibeas, D., Mayer, K.F., Aranda, M.A., and Llave, C. (2009). Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology *392*, 203-214.

Dugas, D.V., and Bartel, B. (2008). Sucrose induction of Arabidopsis miR398 represses two Cu/Zn superoxide dismutases. Plant Mol Biol *67*, 403--417.

Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S., Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W., Givan, S.A*., et al.* (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA *15*, 992-1002.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet *39*, 1033-1037.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol *5*, e57.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005). Elucidation of the small RNA component of the transcriptome. Science *309*, 1567-1569.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell *133*, 128-141.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A*., et al.* (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457*, 551-556.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev *20*, 3407-3425.

Wang, Y., Hindemitt, T., and Mayer, K.F. (2006). Significant sequence similarities in promoters and precursors of Arabidopsis thaliana non-conserved microRNAs. Bioinformatics *22*, 2585-2589.

Afterword

This Thesis has emphasized the importance of small RNAs in understanding many biological phenomena of humans and plants. I have found that its importance has been over emphasised in some cases. For example, I was amused by “Small regulatory RNAs pitch in”, *Nature*, Vol 455, 1184, in which the number of miRNA genes in an animal is correlated to the number of neurons. This would imply that, us, the most studied animal on this planet, is the most complex organism.

Each organism is a masterpiece of Natural Selection in its particular ecological niche. Michael Phelps can hardly compete with an Indo-Pacific sailfish, *Istiophorus platypterus*, which can cruise at a speed of 110 km/h. An African cheetah can easily outrun Usain Bolt. Chris Sharma, one of the best rock climber in the world, still admires the elegance of a spider’s navigation on a vertical surface. Admittedly, most of us don’t smell better than an ordinary Jasmine flower. So, why were we so arrogant to guess, at the beginning of the genomic era, that we should have the highest number of protein coding genes, and now, the highest number of miRNAs?

The fact that we can debate about organismal complexity doesn’t automatically qualify us the holy status which had been denied by Charles Darwin almost 150 years ago. Plants have evolved a very complicated (if not more) small RNA regulatory system since the last hundred millions of years. But they, as taciturn as they naturally are, don’t bother to argue with us who are more sophisticated.

▪ Publications

- Haberer G.*, **Wang Y.*** , Mayer K. The noncoding landscape of the genome of Arabidopsis thaliana (2010), a book chapter of "*Genetics and Genomics of the Brassicaceae*" in the series "*Plant Genetics and Genomics – Crops and Models*", Springer, * as the joint first authors

- Dunoyer P., Brosnan C., Schott G., **Wang Y.**, Jay F., Alioua A., Himber C. and Voinnet O. An endogenous, systemic RNAi pathway in plants (2010), *EMBO Journal, 2010 May 19;29(10):1699-712*

- Donaire L., **Wang Y.**, Gonzalez-Ibeas D., Mayer K., Aranda M., Llave1 C., Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. (2009) *Virology,* Volume 392, Issue 2, 30 September 2009, Pages 203-214

- Port M., **Wang Y.**, Schmelz H.U., Pottek T., Abend, M., A gene signature of primary tumor identifies metastasized seminoma. (2009) *Urologic Oncology: Seminars and Original Investigationsy,* doi:10.1016/j.urolonc.2009.08.008

- Paterson A., Bowers J. *et al.*, The Sorghum bicolor genome and the diversification of grasses. (2009) *Nature* **457**, 551-556 | doi:10.1038/nature07723

- Antonov A., Schmidt T., **Wang Y.,** and Mewes H. W. ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data, *Nucleic Acids Research*, 2008 Jul 1;36 (Web Server issue):W347-51

- Port M., Boltze C., **Wang Y.,** Röper B., Meineke V., and Abend M., A radiation-induced gene signature distinguishes post-Chernobyl from sporadic papillary thyroid cancers. (2007) *Radiation Research* 168, 639-649

- **Wang Y.** Mayer KF. GABI-PLASMAR: Kleine RNA-Moleküle und ihre Rolle bei der pflanzlichen Reaktion auf Umweltfaktoren. *Genomxpress Sonderausgabe – März 2007, Die Hightlights aus der zweiten GABI Förderperiode 2004-2007,* Jens Freitag edits, ISBN 978-3-00-0199953-0

- **Wang Y.**, Hindemitt T, Mayer KF. Significant sequence similarities in promoters and precursors of *Arabidopsis thaliana* non-conserved microRNAs. *Bioinformatics* (2006) Aug; doi:10.1093/bioinformatics/ btl437

- **Wang Y.***, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW. Gene selection from microarray data for cancer classification--a machine learning approach. *Comput Biol Chem*. (2005) Feb;29(1):37-46. * as the corresponding author

- Van Huffel S., **Wang Y.**, Vanhamme L., Van Hecke P., Automatic Frequency Alignment and Quantitation of single resonances in multiple magnetic resonance spectra via complex principal component analysis, *Journal of Magnetic Resonance*, vol. 158, 2002, pp. 1-14.

- **Wang Y.**, Van Huffel S., Heyvaert E., Vanhamme L. Automatic frequency correction for quantitation of magnetic resonance spectroscopic images, *Mathematics in Signal Processing V*, J. McWhirter and I. Proudler edit, Oxford University Press, Oxford, U.K. 2001

- **Wang Y.**, Van Huffel S., Heyvaert E., Vanhamme L., Mastronardi N., Van Hecke P. Magnetic Resonance Spectroscopic Qunatitation via Complex Principal Component Analysis, in *Proc. of the 5th International Conference on Signal Processing (ICSP) of 16th IFIP World Computer Congress (WCC-ICSP2000)*, Beijing, P.R. China, Aug. 2000, pp. 2074-2077.

- **Wang Y.**, Van Huffel S., Vanhamme L., Mastronardi N., Van Hecke P. Advanced Signal Processing Methods for Quantitation of Resonances in Magnetic Resonance Spectra, in *Proc. of the Thirteenth IEEE Symposium on Computer-Based Medical Systems (CBMS2000)*, Houston, USA, Jun. 2000, pp. 63-68.

- **Wang Y.**, Van Huffel S., Mastronardi N. Quantification of Resonances in Magnetic Resonance Spectra via Principal Component Analysis and Hankel Total Least Squares , in *Proc. of the Program for Research on Integrated Systems and Circuits (ProRISC99),* Mierlo, The Netherlands, Nov. 1999, pp.585-591.