UNIVERSITY OF CALIFORNIA
RIVERSIDE

Evolutionary Study of Genome Features in Cereals: a Focus on Endogenous Small RNA Generation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Sahar Nohzadeh-Malakshah

December 2012

Dissertation Committee:
Professor Renyi Liu, Chairperson
Professor Thomas Girke
Professor Timothy Close

The Dissertation of Sahar Nohzadeh-Malakshah is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

This research project would not have been possible without the support of many people. The author wishes to express her gratitude to her supervisor, Prof. Renyi Liu who was very helpful and offered invaluable assistance, support and guidance. Deepest gratitude is also due to the members of the supervisory committee members, Prof. Thomas Girke and Prof. Timothy Close without whose knowledge and assistance this study would not have been successful. Special thanks also to all her graduate friends and current and former lab members. The author would also like to convey thanks to the Department of botany and Plant Sciences at UC Riverside. The author wishes to express her love and gratitude to her beloved families for their understanding & support, through the duration of her studies.

ABSTRACT OF THE DISSERTATION

Evolutionary Study of Genome Features in Cereals: a Focus on Endogenous Small RNA Generation

by

Sahar Nohzadeh-Malakshah

Doctor of Philosophy, Graduate Program in Plant Biology

University of California, Riverside, December 2012

Professor Renyi Liu, Chairperson

The prevalence of large-scale genomic studies and technological advances in recent years is promising in the investigation of genome features in more details. The focus of this dissertation is to recruit publicly available genome data and to adopt common and novel computational methods to study the generation and evolution of two classes of small RNAs in cereals. Such a study is important since first small RNAs play a central role in several fundamental processes in cells like genome integrity, gene expression & response to stresses; and second there is a widening gap between the amount of raw and the processed and genomic data which needs to be addressed by computational biology strategies. By the aid of comparative genomics approach the sources of two different types of small RNAs were investigated. Our results provide a strong support that overlapping genes (OGs) could be a source of natural antisense

transcripts-small interfering RNAs (nat-siRNA) in cereals while most nat-siRNA generation is not well conserved in them. In addition, our data demonstrate that OGs are common and mostly species specific in maize, rice and *Brachypodium* and there is no obvious correlation between their number and the total number of genes or genome size; however, genome size and architecture does affect the frequencies and types of overlapping genes. By comparative analysis of the orientation of OGs several birth and death mechanisms were proposed among which translocation and gene creation are the major ones. Moreover, we improved the annotation of microRNA (miRNAs) genes in cereals that produce another type of small RNAs that have crucial regulatory roles in development and stress responses. A novel methodology was developed to use a large number of the recently available RNAseq data to refine gene boundaries of miRNAs and a comparative analysis were performed on them. By defining these upstream regions and using the alignment of the orthologous ones, several miRNA specific regulatory elements were identified which are conserved in cereals and are good candidates for experimental verification. Overall, this dissertation demonstrates that use of publicly available data and computational approaches would increase our understanding of small RNAs and their evolution. This can provide a foundation for the community to study their expression and function more precisely.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

ABSTRACT v

TABLE OF CONTENTS vii

CHAPTER 1. Introduction 1

1.1 Background & Context 1
1.2 Overall Research Aim & Individual Research Objectives 3
1.3 Value of Research 5
1.4 References 6

CHAPTER 2. Evolution of Overlapping Genes and Nat-siRNAs in Cereal Genomes 8

2.1 Abstract 8
2.2 Introduction 9
2.3 Results 13
2.4 Discussion 23
2.5 Materials and Methods 27
2.6 Figures & Tables 30
2.7 References 50

CHAPTER 3. Refining MicroRNA Gene Boundaries in Cereals Using RNAseq Data 51

3.1 Abstract 54
3.2 Introduction 56
3.3 Results 61
3.4 Discussion 64
3.5 Materials and Methods 65
3.6 Figures & Tables 67
3.7 References 79

Page

CHAPTER 4. Comparative Analysis of Regulatory Elements of microRNAs in Cereal Genomes 78

4.1 Abstract 83
4.2 Introduction 85
4.3 Results 87
4.4 Discussion 90
4.5 Materials and Methods 93
4.6 Figures & Tables 96
4.7 References 105

CHAPTER 5. Conclusion 108

5.1 Summary of our Findings & how they Relate to Other Works 108
5.2 Evaluation 112
5.3 Future Works 114
5.4 References 115

Chapter 1

Introduction

1.1 Background and Context

In the era of pervasive sequencing we are encountering new or improved genome data every day. The total amount of genomic data is growing approximately tenfold each year (Loh et al., 2012). Consecutive generations of sequencing technologies have increased the amount of genomic data exponentially. Over the past 10 years since the publication of the first draft of the human genome, technologies have been developed that can be used to sequence a human genome in 1 week for less than $10,000 (Loh et al., 2012). Plant genome sequencing methodology paralleled the sequencing of the human genome although with a relatively slower pace (Jackson et al., 2011). A reference genome is now available for 21 crops and crop comparative genomics is being transformed by these data and new generation of experimental and computational approaches have evolved (Morrell et al., 2011).

These advances have changed the way we conduct genetic and genomic researches. For instance in the field of plants genomic, economically important trait selection paired with the increase in the resolution of markers and the decrease in cost, will lead to improved breeding strategies. Genome sequence coupled with transcriptomics may tell us a lot about where within the genome to focus our attention in breeding programs. The application of Next Generation Sequencing (NGS) technologies for resequencing,

assuming a reference genome exists, is one of the most powerful applications for crop improvement.

A major challenge that is encountered by the scientific community is pulling out the new insights from the data sets currently available which will require not only faster computers, but also more efficient and smarter algorithms (Loh et al., 2012). The pace of innovation in genomic data creation is much higher than the pace of innovation within genomic informatics. This widening gap must be addressed before the overall field of genomics can take the leap forward that the community has foreseen and is needed for many applications, spanning from evolution to medicine. We should be able to analyze all these data effectively and without addressing this problem, these downstream informatics challenges will restrict the advancements of the entire field (Kahn, 2011). We have tried to develop one of these smart algorithms in chapter 3.

One aspect that has been transformed by the increasing numbers of reference genomes and by the estimation of sequence diversity from high-throughput resequencing, and also by the emergence of a new generation of experimental and computational approaches is comparative genomics which has been utilized throughout this thesis. Comparative genomics is traditionally thought of as the investigation of synteny (gene order) and sequence comparisons among related species (Morrell et al., 2011). It is useful in studying related genomes, the non-model organism and it can be used to bring useful traits and genome segments from wild relatives (Edwards and Batley, 2010). The future of crop improvement will be around the comparisons of individual plant genomes in different fields like genetic mapping and evolutionary analysis. If we wish to continue

increasing crop production in parallel with the growing human populations and changing climates, we should maximize the use of this genomic data and it is fundamentally important (Morrell et al., 2011).

In this thesis by the aid of comparative genomics approaches and innovative computational methodologies I have tried to analyze several crops raw sequencing data and increase our understanding of their corresponding genomes.

1.2 Overall Research Aim and Individual Research Objectives

In the context of joint pervasive sequencing data in the genomic era, plant and specifically crop genomics has two interesting key properties that motivate most of the studies that I did in this thesis. Availability of the whole genome sequences of four closely related species from the grasses in recent years- rice (Goff et al., 2002), sorghum (Paterson et al., 2009), maize (Schnable et al., 2009), and Brachypodium (2010) - on the one hand and the recent explosion of NGS data on the other hand gave us an opportunity to compare the type, frequency and characteristics of several genome features in these plants. These features that have been studied at genome-wide scale using comparative genomics approach include a) Evolution of overlapping genes. b) Identification and/or characterization of the sources of generation of two classes of plant small RNAs (natural antisense small interfering RNAss and miRNAs). c) Regulatory elements that affect miRNA gene expression. This thesis is made up of three main chapters.

Chapter 2. **Evolution of Overlapping Genes and Nat-siRNAs in Cereal Genomes**

This chapter contains two major sections. The first part states the distribution and characterization of the overlapping genes in three cereal genomes at the genome scale. Several possible mechanisms have been proposed to explain the emergence and deletion of overlapping genes over time in different species. The second part describes how overlapping genes could be a source for the generation of natural antisense transcript small RNAs (nat-siRNA) and how conserved these patterns are.

Chapter 3. **Refining microRNA Gene Boundaries in Cereals Using RNAseq Data**

This chapter proposes a novel methodology for improving microRNA gene annotation using RNAseq data. The parameters of this method have been optimized with two sets of microRNA genes with known gene boundaries. Then it was applied on two sets of microRNA with unknown gene boundaries. A comparative study of the miRNA gene boundaries was performed on two genomes as well.

Chapter 4. **Comparative Analysis of Regulatory Elements of microRNAs in Cereal Genomes**

This chapter provides some information about the frequency and conservation of microRNA-specific regulatory elements in cereal genomes.

Finally, Chapter 5 presents conclusion, evaluation and future works.

1.3 Value of this Research

The genome sequences as well as the NGS data of four cereal genomes have become available in the last couple of years. Having access to these data and by the aid of computational methods we were able to first find some support for the existence of a specific type of small RNA (nat-siRNA) in cereals, second improve the annotation of another type of small RNA (microRNA) and third study the evolution and conservation of these two types of small RNAs among several lineages. Also by bioinformatics approaches we found some candidates small RNA and the corresponding regulatory elements, which could be a good start for experimental analysis.

Moreover the huge amount of sequencing data in different organisms has caused a gap between the amount of raw genomic data and the processed and analyzed genomic data. While in animal field we see several studies have tried to combine and compare the data from different species and sources and extract as much meaningful biological data from them as possible, there are not as much studies in plants. The purpose of this work is to fill in the gap between raw and processed data in several plant genomes up to a certain extent, particularly by the aid of publicly available genome and NGS data, comparative genomics approach and common and also novel computational methodologies, we were able to make some advances in our knowledge on the conservation and evolution of small RNAs

1.4 References

(2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. In Nature , pp. 763-768.

Edwards, D., and Batley, J. (2010). Plant genome sequencing: applications for crop improvement. In Plant Biotechnol J , pp. 2-9.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). In Science , pp. 92-100.

Jackson, S.A., Iwata, A., Lee, S.H., Schmutz, J., and Shoemaker, R. (2011). Sequencing crop genomes: approaches and applications. New Phytol **191,** 915-925.

Kahn, S.D. (2011). On the future of genomic data. In Science , pp. 728-729.

Loh, P.R., Baym, M., and Berger, B. (2012). Compressive genomics. In Nat Biotechnol , pp. 627-630.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. In Nat Rev Genet , pp. 85-96.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F., Messing, J., and Rokhsar, D.S. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457,** 551-556.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E.,

Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. In Science , pp. 1112-1115.

Chapter 2

Evolution of overlapping genes and nat-siRNAs in cereal genomes

2.1 Abstract

Overlapping genes occur frequently in all kingdoms of organisms. Their widespread occurrence in large genomes is unexpected considering the large intergenic space in these genomes and the negative evolutionary pressure that may be imposed on overlapping genes. A few studies have been done on the evolution and origination of overlaps in Eukaryotes but no genome wide study on different types of overlapping genes has been conducted in plants. We have analyzed the overlapping genes in three cereal genomes: rice, maize and *Brachypodium* in which the number of identified overlapping genes were 747, 1564, 347, respectively. Even though there is no obvious correlation between the number of overlapping genes and the total number of genes or genome size, genome size does have an effect on types of overlapping genes. The larger maize genome possesses significantly more nested overlapping genes than the smaller rice and *Brachypodium* genomes. The majority of overlapping genes are species-specific, indicating frequent creation and loss of gene overlaps. Our results also show that translocation and gene creation/deletion are the major mechanisms for the origination and loss of overlapping genes. Mapping of small RNA reads to overlapping genes suggests that overlapping genes are a major source of generating nat-siRNAs in three genomes, however most nat-siRNA generation is not well conserved.

2.2 Introduction

Overlapping genes are neighbor genes that share a stretch of DNA segment, either on the same strand or on the opposite strands. Overlapping genes are found in the genomes of all kingdoms of life. Since the first example was discovered in a single-stranded DNA phage (Barrell et al., 1976), overlapping genes were found in bacteriophages (Normark et al., 1983), viruses (Samuel, 1989), and bacteria (Johnson and Chisholm, 2004; Palleja et al., 2008). Recent studies show that they occur frequently in animal and plant genomes as well (Osato et al., 2003; Veeramachaneni et al., 2004; Makalowska et al., 2005; Steigele and Nieselt, 2005; Galante et al., 2007; Henz et al., 2007; Makalowska et al., 2007; Solda et al., 2008; Zhou et al., 2009)

The frequency, orientation and conservation of overlapping genes differ dramatically among different organisms. Whereas on average only 1% of the coding region is occupied by overlapping genes in viruses (Belshaw et al., 2007), one third of genes are involved in overlapping in bacterial genomes (Johnson and Chisholm, 2004). The frequency of overlapping genes in eukaryotes is moderate. For example, overlapping genes make up 2% to 14% of all genes in fungi (Steigele and Nieselt, 2005) and vertebrates (Makalowska et al., 2007). Whereas the majority of overlapping genes occur on the same strand in viruses (Belshaw et al., 2007) and bacteria (Johnson and Chisholm, 2004), most overlapping genes in eukaryotes are on opposite strands (Solda et al., 2008). Although overlapping genes are well conserved in bacterial genomes (Johnson and Chisholm, 2004), they show much less conservation in eukaryotes and are predominantly lineage-specific (Veeramachaneni et al., 2004; Makalowska et al., 2007; Sanna et al.,

2008). For instance, only 27% of overlapping genes in human have overlapping orthologs in mouse (Makalowska et al., 2005; Sanna et al., 2008).

A few mechanisms have been proposed for the origination of overlapping genes. Genes may become overlap when one gene lost its polyadenylation signal at the 3' end and used the signal that happened to be present on the opposite strand of a neighboring gene (Shintani et al., 1999). Keese and Gibbs (1992) proposed that overlapping genes may have emerged in viruses from a process called overprinting - part of the sequence from one gene is translated *de novo* in a different reading frame or from non-coding regions (Keese and Gibbs, 1992). In vertebrate genomes, various mechanisms including translocation, development of a new splice variant, and acquisition of a terminal, non-coding exon contributed significantly to the origination of overlapping genes (Makalowska et al., 2007; Solda et al., 2008).

Being overlapping might have significant impact on the expression and function of the genes involved. It is not exactly known whether these regulatory roles are the result of the leakage of the transcriptional machinery or intentional (Dahary et al., 2005; Lapidot and Pilpel, 2006). These roles have been observed for both same strand and different strand overlaps. Same strand overlaps may provide means to coregulating gene expression tightly (Normark et al., 1983). Both negative and positive correlations have been observed for different strand overlaps. This suggests that their mechanisms of action might be diverse. Four of these mechanisms are well documented: (a) transcriptional interference in which two bulky RNA polymerase II interfere with each other and could result in an anti-correlated expression or the shutdown of both transcripts; (b) RNA

masking in which sense-antisense duplex might mask the cis-elements residing in either of the transcripts and hinder necessary processing on them; (c) double-stranded RNA (dsRNA)-dependent mechanisms such as RNA editing and RNA interference; (d) chromatin remodeling in which antisense could be involved in DNA methylation and monoallelic expression that would result in gene silencing. In addition, protein regions encoded by overlapping genes might have a propensity toward structural disorder that might alleviate evolutionary constraints imposed on their sequence by the overlap (Rancurel et al., 2009).

A major mechanism of action of overlapping genes is through the generation of small RNAs from *cis*-antisense overlapping genes. When transcripts from a pair of overlapping genes that occupy different strands (termed natural antisense transcripts, or NAT) are transcribed in the same cell, they have the potential to form double- stranded RNA molecules, which can be further cleaved by Dicer like proteins to produce small RNAs (termed NAT-siRNAs) (Borsani et al., 2005). NAT-siRNAs may be incorporated into Argonaute proteins and cause degradation of one of the overlapping gene transcripts and thus play a role in regulation of the gene expression (Borsani et al., 2005; Jin et al., 2008; Zhou et al., 2009; Lapidot and Pilpel, 2006). 3% to 8% of plant genes have the structure these cis-NATs (Jin et al., 2008; Zhou et al., 2009; Zhang et al., 2012). It has been shown that NAT-siRNAs can be generated from 30% to 64% of cis NATs in plants (Jin et al., 2008; Zhou et al., 2009; Zhang et al., 2012). Zhang et al. (2012) just recently analyzed Arabidopsis and rice genomes for the enrichments of NATs for small RNA and showed that siRNA were enriched in at least 84 and 119 of them, respectively (Zhang et

al., 2012). Some NAT-siRNAs are induced by abiotic and biotic stresses (Borsani et al., 2005; Katiyar-Agarwal et al., 2006) and thus play important roles in stress response. Sun et al. (2005) showed that overlapping pairs in human and mouse genomes have a significantly higher probability of having co-expression and inverse expression (i.e. characteristic of sense–antisense regulation) than do overlapping pairs in only one of the two species (Sun et al., 2005). They have not explored the small RNA production in their study but this conservation in gene expression may imply that the conservation of regulatory mechanisms may be due to NAT-siRNA production. Although NAT-siRNAs have been discovered in several plants including Arabidopsis and rice, it is not known whether they are mainly a transient phenomenon during evolution or they are well conserved. We will answer this question by comparing NAT-siRNAs from conserved cis-antisense overlapping genes in closely related grass species.

The grass family contains economically important crops such as rice, maize, sorghum, wheat, and barley, and is the major contributor to human nutrition and domestic animal feed. The grass genomes diverged greatly in genome size and gene number since they separated from a common ancestor 50-70 million years ago (mya), but gene order was well-maintained (Bennetzen and Ramakrishna, 2002). The availability of the whole genome sequences of four grasses (rice, maize, sorghum, and *Brachypodium*) gives us an opportunity to compare the type and frequency of overlapping genes, study the mechanisms of their birth and death, and compare the generation of nat-siRNAs in these closely related species. Because overlapping genes are not well annotated in the sorghum genome, we focused on the other three genomes. Our results indicate that number of

overlapping genes does not correlate well with genome size and the majority of overlapping genes are species-specific. Unlike rice and *Brachypodium*, the large maize genome is enriched with nested overlapping genes that occupy different strands. Translocation and gene creation and deletion are the main mechanisms for the origination and loss of overlapping genes. Although overlapping genes are a major source of generating nat-siRNAs in grass genomes, nat-siRNAs are rarely generated in the similar amount from conserved overlapping pairs.

2.3 Results

Number and types of overlapping genes

Based on current annotation, proportion of genes that overlap in three cereals differs significantly. After excluding transposon-related genes, out of 40,577 and 25,532 protein coding genes, 1494 genes (4% of total number of genes) in rice and 694 genes in *Brachypodium* (3% of total number of genes) are involved in overlapping (see Supp Tables 2.1; 2.2; 2.3 for the list of overlapping genes). Due to the preliminary nature of the annotation in the maize genome, we considered three maize gene datasets: a) Transposon-excluded gene set (32,540 genes), b) Pure computational prediction excluded gene set (30,339 genes), and c) cDNA supported gene set (20,480 Genes). We identified 3128, 2464 and 1614 genes involve in overlapping pattern in these three datasets, respectively, which account for 10%, 9% and 8% of the total number of genes, respectively. It shows that maize genome has much more overlapping genes than rice and *Brachypodium*, but the overall trend in the three species does not show any correlation

between the number of overlapping genes and size or total number of genes. In the rest of the study, we used the cDNA-supported gene set of in the maize genome.

As shown in Table 2.1, we categorized overlapping genes based on gene orientation. In three species considered, the different-strand overlaps outnumber the same-strand overlaps. In the same-strand overlaps, nested genes (where one gene resides completely in the other one) represent the primary arrangement. In rice and Arabidopsis the majority of overlaps (71% and 79%, respectively) have convergent orientation (3' to 3' overlap) whereas in *Brachypodium* the divergent orientation (5' to 5' overlap) is predominant (48%). In contrast, in larger maize genome, nested overlapping genes (46%) are apparently enriched.

We have classified the overlapping genes based on whether gene-coding sequences are involved in the overlapping region. Nearly half of the overlapping pairs involve coding sequence from at least one gene and in 34% to 46% of pairs, coding sequences from both genes are involved in overlapping. The percentage of overlapping pairs with the involvement of exon or CDS in maize is higher than the other two species. There are a few cases in which there is no exonic overlap and a gene is located inside the intronic region of the other one. Accurate annotation of UTRs is critical for determining the number and types of overlapping genes. We found that only 53% to 60% of the overlapping genes with annotated UTRs in the three cereal species considered (Table 2.2). In contrast, the well-annotated Arabidopsis genes 78% of the overlapping genes contain annotated UTRs. Therefore, we have probably underestimated the number of overlaps in the three cereals due to incomplete annotation.

Conservation and timing of overlaps in cereal genomes

In order to find which percentage of total overlapping genes in one specie has the corresponding overlapping pattern in the other two species and which percentage is species specific, we first found all the homologs of the two genes which were involved in the overlapping pattern in the other two species and then study whether any of these homologs are involved in an overlapping pattern. We didn't use the synteny data to find the orthologs at this step since we did not want to limit ourselves only to those overlaps, which fall in collinear regions. Table 2.3 shows that 80%, 64% and 89% of overlapping genes are unique in rice, *Brachypodium* and maize, respectively. So the majority of overlaps are species specific. They are the result of either gain of overlapping in that species or loss of overlapping in the orthologous genes of the other two species. We've examined these mechanisms in more details in the next step of our study. Among the three species considered, rice and *Brachypodium* are closer to each other in terms of evolutionary distance than maize. Based on the phylogenetic relationships among the three species, we use a parsimonious approach to determine whether the overlaps were present in their ancestor and the gain and loss of overlaps during evolution. For example, if an overlap is present in in maize and rice but not in *Brachypodium*, the most parsimonious explanation is that the overlap was present in the ancestor and has been lost in *Brachypodium* after its divergence from rice. If an overlap is present in rice and *Brachypodium* but not in maize, the most likely scenario is that it originated in the ancestor of rice and *Brachypodium* after its divergence from the maize lineage. If an

overlap is found in only one of the three species, most likely it emerged recently in that particular lineage.

Maize has gone through an additional round of whole genome duplication (WGD) after its separation from the rice lineage. It is therefore interesting to investigate the fate of overlapping genes after WGD. We started with the overlapping gene pairs in rice and found the two collinear regions in maize. 151 overlapping pairs in rice have two collinear regions in maize. Among them there was only one case in which the overlapping pattern has been kept in both segments. 30 of the pairs remain overlapping in one segment but not the other. 38 of the pairs showed neighboring pattern in one segment but not the other. 120 pairs do not show overlapping pattern in both segments. We also found that in 60% of the cases one of the genes of the overlapping pair has been deleted in the second segment and in 30% of cases both of genes involve in overlap were removed from the second segment.

Gain and loss of overlapping genes

Having the genome-scale information about the annotation of overlapping genes in three closely related species, we developed several hypotheses to explain the mechanisms, which give rise to the birth or death of different types of overlapping genes. To do so, we used the evolutionary relationship and also the orientation of overlaps' orthologs in the three cereal genomes. Based on the presence or absence of these orthologs we can place all the probable patterns into three major groups.

First group contains all those cases where both genes in an overlapping pair have orthologs in the other two species. We started with the set of overlapping genes in one species and used collinearity data to find orthologous pairs in the other two species and then categorized these orthologous pairs into overlapping (if they overlap), neighboring (if they were not overlapping but located one next to each other without any gene between them), and separated (if they were on different genome segments or were separated by other genes). The number of orthologous gene pairs in each category is shown in parenthesis in Figure 2.1, 2.2 and 2.3. For instance, out of 245 rice overlapping pairs (where an ortholog have been identified in *Brachypodium* and maize for both genes in the overlap pair), the orthologs of 44 pairs are overlapping, 148 are neighboring and 53 pairs are separated in *Brachypodium*. These numbers are 58, 60 and 127 pairs in maize, respectively. These numbers show that the relative orientation of rice overlaps is much more conserved in *Brachypodium* rather than maize. This could be due to dynamic maize genome and also it's farther evolutionary distance from rice. Based on the above organization of orthologous gene pairs in three species we considered 9 possible scenarios of overlapping gene evolution in each organism (Figure 2.1, 2.2 and 2.3). 4 of these scenarios are clear cases of gain or loss of overlaps. Because maize is an out-group to the rice/*Brachypodium* lineage, if two genes are overlapping in maize and in either rice or *Brachypodium*, but they do not overlap (neighboring or separated) in the other one, the most parsimonious explanation is that the overlap is lost in that organism either by exaptation (a process that gives rise to new genes or new variants from preexisting nucleotide sequences e.g. shortening of an existing gene) (Figure 2.1 a) or translocation

(Figure 2.1 b). As we see in Figure 2.1 a, 29 of rice overlaps have lost this pattern in *Brachypodium* due to exaptation. These are clear case of overlap loss in *Brachypodium*. The corresponding number for rice overlap death by exaptation is 15 as is clear in Figure 2.3 a. In Figure 2.1 b it's obvious that in 7 cases overlapping pattern is lost in *Brachypodium* due to translocation since the genes are overlap in the out-group. We didn't see any example for this scenario in loss of overlaps in rice (Figure 2.3 b). In another model we see that in 40 cases genes are neighboring in *Brachypodium* and maize but are overlapping in rice. Here the simplest explanation is that there is a gain of the overlap in rice by exaptation (Figure 2.1). The equivalent number in *Brachypodium* is 15 (Figure 2.3 c). If two genes do not overlap in maize and in either rice or *Brachypodium,* the simplest explanation is that there is a gain of the overlap by translocation (Figure 2.1 d, Figure 2.3 d). For the other 5 scenarios we cannot determine whether the current orientation of the overlaps in the three species is the support of overlaps birth in one of them or death in the other.

Among the second group of the overlapping genes one of the genes involved in overlap does not have ortholog in the collinear region in at least one of the closely related species that were compared. There can be 7 different orientations among the overlaps and the corresponding orthologs in this group. Pattern a' in rice overlaps which includes 3 different gene orientations in Figure 2.4 shows that in 163 cases the corresponding ortholog of only one of the genes involved in overlaps exists in either *Brachypodium* or maize or both species (Figure 2.4 a). Considering maize as the out-group, the most parsimonious explanation suggests that the overlapping phenomenon has happened after

divergence of maize. In those 36 cases that the ortholog exists in maize but not *Brachypodium* most probably gene deletion or translocation has happened in *Brachypodium* after maize divergence. Rice has gained the overlapping pattern either by translocation of the second gene or creation of the second gene. To determine exactly which mechanism is working here we searched *Brachypodium* genome for the homologs of the second gene. If there was a homolog we propose that translocation has caused the overlapping pattern in rice. Otherwise we propose that emergence of a new gene has created the overlapping pattern in rice (Table 2.4). This phenomena fits overprinting hypothesis in which the different frame of different genes are coded from the same locus on the genome and create a new gene. Overall 73 & 361 cases of overlapping genes have the similar pattern (pattern "a") in *Brachypodium* and maize, respectively (Figure 2.5 a, Figure 2.6 a). Comparable to what was explained for this pattern in rice, *Brachypodium* has gained the overlapping pattern either by translocation or gene creation (Table 2.4). But we can't precisely explain what has caused this pattern in maize since we don't have out-group for maize. So these 361 (220+75+66) observed overlaps, which are maize specific, could be either the result of overlaps birth in maize or overlaps death in rice & *Brachypodium*. These results also show that maize has more species-specific overlaps than rice and *Brachypodium*.

14 overlaps in rice represent pattern c' which is a clear case of overlap loss in *Brachypodium* by either translocation or gene deletion since the overlapping pattern is present in the out-group (Figure 2.4 c'). Number of cases in each scenario is summarized in Table 2.4. Accordingly we have 3 clear case of overlapping loss in rice (Figure 2.6 c').

The remaining 3 other orientations in this group (patterns d' & e') are less informative (Figure 2.4 d' & Figure 2.4 e'). In pattern d' the orthologs of both rice genes in the 38 overlapping pairs are present in the out-group but one is missed in *Brachypodium* collinear region due to either translocation or deletion. However we can't conclude anything about the timing of overlapping phenomena and it could be before or after maize divergence. Pattern e' is the least informative scenario since we cannot conclude anything about the timing of overlaps or mechanism that has caused these orientations (Figure 2.4 e'). Similar conclusion would be obtained from patterns d' & e' in *Brachypodium* overlapping genes (Figure 2.6 d' & Figure 2.6 e').

In the third group of overlapping genes both genes involved in the overlap did not have the ortholog in neither of the other two compared genomes (Figure 2.4 b', Figure 2.5 b', Figure 2.6 b'). Based on the most parsimonious explanation they are clear case of overlapping birth in rice and *Brachypodium* (Figure 2.4 b', Figure 2.6 b') but we cannot conclude whether it is birth or death in maize since we don't have out-group to study. Then we searched the two other genomes to find overlaps' homologs. If there are no homologs, we assumed that mechanism of overlapping is gene creation. But if the homologs do exist, the mechanism could be translocation (Table 2.4).

All the conclusions we've got from these scenarios are summarized in Table 2.4. In all the three plant species we had 390 cases of overlap's birth and 96 cases of overlap's death.

Our theoretical scenarios about overlaps evolution show that gene creation (40%) and translocation (55%) play the major role in the emergence of them and translocation

(35%) and exaptation (47%) are the predominant mechanism in disappearance of them. Based on all the explained patterns of birth and death of overlapping genes, by dividing number of gain and loss of overlapping genes by evolutionary time between species, we calculated the rates of gain and loss. Our calculations showed that the birth rate of overlapping genes is 6.2 overlaps/myr in rice and 2.8 overlaps/myr in *Brachypodium.* The death rate of overlapping genes is 0.6 overlaps/myr in rice and 1.2 overlaps/myr in *Brachypodium.*

Using cases in which two genes overlap in one species and but are only neighboring genes in another species, we studied the probable molecular mechanisms of the gain and loss of overlaps. All these mechanisms can be categorized under the general term of exaptation (Figure 2.1 c, Figure 2.2 c, Figure 2.3 c). As shown in Figure 2.7, UTR expansion is one of the major exaptation mechanisms, which cause neighboring genes to overlap and we found 28 cases in this category. There are 11 cases like Figure 2.8 in which an intergenic region has been eliminated and two genes have become probably first neighboring and then by using each other UTRs overlapping. In few other pairs the introns play the role in overlapping like intron elongation (Figure 2.9) and intron insertion in UTRs (Figure 2.10). And finally we did find three rare instances in which the adaptation of a new CDS has caused genes to overlap (Figure 2.11). All the results of this part have been summarized in Table 2.5.

Conservation of the generation of nat-siRNAs in three cereal genomes

When two overlapping genes are located on opposite strands (last three columns of Table 2.1), they could produce cis-natural antiscence transcripts (NATs) which have the potential to produce double stranded RNA, which can be further cleaved by Dicer like proteins to produce natural antisense small interfering RNAs (NAT-siRNAs). In order to compare the production of NAT-siRNAs in three cereals, we mapped 20.7 million clean small RNA sequences from 13 public libraries to 688 different-strand overlapping pairs in rice and 36.1 million clean small RNA sequences from 12 public libraries to 791 different-strand overlapping pairs in maize. We didn't include *Brachypodium* since there were just 3 public small RNA libraries available. After mapping we performed an enrichment test to see whether the overlapping region of the overlapping gene pairs is significantly enriched for small RNA production or not (see method). We found that 17%, and 15% of overlapping genes are enriched for small RNA hits in rice and maize, respectively.

Next, we studied the conservation of this enrichment pattern along these two species. 24% of rice overlaps that are conserved in maize can produce Nat-siRNA but only 15% of them can produce Nat-siRNA in mazie as well. An example of this enrichment in both species is shown in 2.12. In maize overlaps that are conserved in rice only 15% can produce Nat-siRNA; however, 26% of them can produce Nat-siRNA in rice.

2.4 Discussion

Recent studies show that overlapping genes are much more common in eukaryotic genomes than previously thought. Considering the lack of specific evolutionary force on the eukaryotic genome size, high frequency of overlaps is unexpected and probably they represent a hidden source of gene expression complexity in the genome. Although there have been a few single-specie studies of a particular type of overlapping genes (NATs) in Arabidopsis and rice (Jin et al., 2008; Zhou et al., 2009; Zhang et al., 2012), to the best of our knowledge, there has not been any exhaustive study on either all types of overlapping genes or their origin and probable transcriptional control mechanism in plants. We still don't know exactly how these genes have evolved among different species and does sharing the same genomic locus have any regulatory or functional meaning specifically. There are a few large-scale comparative studies on overlapping genes in vertebrates (Veeramachaneni et al., 2004; Makalowska et al., 2005; Makalowska et al., 2007; Sanna et al., 2008), which show they might be functionally important. Here we present a genome-wide comparative analysis of the overlapping genes phenomena and their evolution and origination in cereals.

Our results show that overlapping genes are frequent (3%-10% protein coding genes do overlap) in cereal genomes. These ratios are close to what we see in vertebrates (Makalowska et al., 2007) but are much less than the amount of overlaps in microbial genomes in which approximately one-third of all genes have this pattern (Johnson and Chisholm, 2004). There is no positive or negative correlation between these percentages and either total number of genes or genome size, therefore the driving force for the

creation and loss of overlapping genes is not saving genome space and increasing the information density. This is in contrast to what we see in bacteria and viruses. Overlapping genes have different types and orientation and there is no single major type that shows up in all studied genomes. Overall the majority of them belong to different-strand overlapping pairs, which is in agreement to what we see in other eukaryotes like human and mouse (Sanna et al., 2008; Solda et al., 2008). Solda et al. (2008) have claimed that generally overlapping genes are counter selected and the observed number of overlapping events is less than the expected number in case of neutrality, but some specific arrangement of overlaps which can provides selective advantages are kept in the organism more than what is expected (Solda et al., 2008). So probably in the case of genes on opposite strands the advantage could be represented by antisense regulation and therefore they are more abundant.

Unlike other plant genomes analyzed, the larger maize genome possesses a unique pattern of overlaps with significantly more nested overlapping genes. This property is similar to that of the human genome. Comparing to smaller genomes, larger genomes contain much longer introns, which make it easier to enclose a whole gene within another, creating nested overlapping genes. Taking into account the high number of transposable elements in maize one major formation mechanism for these nested genes could be retroposition in which a retro-gene has been inserted in the large introns of maize genes. These introns might provide an open chromatin environment for the external genes to enter. The same phenomenon has been reported in human (Yu et al., 2005).

Our study shows that the majority of overlapping genes are created recently and species-specific and there is low conservation of overlapping patterns in the closely-related cereal genomes, suggesting that overlapping is evolutionarily transient and there is a fast rate of gain and loss of overlaps in these species and since the birth rate is higher than death rate, we see more unique overlapping pattern. This is in contrast to what we see in microbial genomes in which overlapping genes have homologs in more microbes than do non-overlapping genes, and are therefore likely more conserved (Johnson and Chisholm, 2004). One major reason of the abundance of specie-specific overlaps could be that the poor conservation sometimes is a part of the functional role of a particular process as claimed by Solda et al (2008) (Solda et al., 2008). Providing an evidence for this hypothesis they mentioned the high amount of human specific alternatively spliced variants that are poorly conserved in other species but are very important in human gene expression regulation (Blencowe, 2006). Therefore we might propose the same hypothesis to justify the prevalence of species-specific overlaps in plants. The same species-specificity has been observed in mammals in another study (Sanna et al., 2008) but they claimed that this little number of shared overlapping relationship among mammals might be due to wrong assignment of orthologs. To test this hypothesis we examined the orthologs of both genes involved in overlaps in the other species and whenever we didn't see overlapping pattern between the orthologs, we looked at all possible homologs to see whether we can see overlapping relationship or not; And still we saw high frequency of species specific overlaps so wrong orthologs assignment does not play a major role here. Moreover, if gene annotation gets improved more, we might

discover that two neighboring genes become overlapping due to the identification of new UTRs boundaries. So we might have underestimated the number of overlaps and therefore conserved overlaps here.

Different orientations of overlaps within the species and different level of their conservation between the species imply that there is no unique evolutionary mechanism that leads to overlapping and different forces are involved. In all the three plant species we had more overlap's birth cases than death cases. Like mammals the birth rate of overlaps is higher than their death rate (Sanna et al., 2008) but both of these rates are higher in plants compared to mammals. We found that gene creation and translocation are major evolutionary forces that create overlaps. Translocation and exaptation play major roles in death of overlaps. Same mechanisms have been identified by Makalowska et al (2007) in birth of overlaps in vertebrates (Makalowska et al., 2007). However, Solda et al (2008) claimed that recent origin of overlapping genes is not the major reason of overlaps in Metazoa since they observed that most overlapping genes in one species had orthologs in the other species (Solda et al., 2008). In spite of this, we showed in our study that almost half of the overlaps don't have orthologs at least for one of their genes in the closely related species. In determining the overlap creation mechanisms we should take into account that we weren't able to find out the overlapping mechanisms in many cases since we didn't have the appropriate out-group to compare the current overlap relationship among species with. Therefore, if we consider all cases of gain and loss of overlaps (and not only clear cases) and also if we have a better UTR annotation of gene structures other mechanisms might become predominant.

And finally our study demonstrates that although overlapping genes are a rich source for the production of NAT-siRNAs in three cereals, the production of NAT-siRNAs is not well conserved; however we showed that the ratio of enrichment for small RNA in conserved rice overlaps is higher than the non-conserved ones. To generate NAT-siRNAs, both overlapping genes need to be expressed at a relatively high level within the same cell so that transcripts from the opposite strands can form double-stranded RNAs, which can be further cleaved by Dicer like proteins to produce siRNAs. However, in the course of evolution, the expression pattern of orthologous NATs may change significantly in closely-related species, which will certainly affect whether NAT-siRNAs can be produced or how much.

In summary, overlapping genes are frequently found in cereal genomes and were subject frequent gains and losses during evolution. Although antisense-overlapping genes can produce NAT-siRNAs that may play important roles in plant cellular function and stress response, the production of NAT-siRNAs appears to be also evolutionarily transient.

2.5 Materials and Methods

Sequence data

Genome sequences and annotation data were downloaded from the following websites: rice, the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu, version 6.1) (Goff et al., 2002); maize, http://www.maizesequence.org (version 4a)

(Schnable et al., 2009); and *Brachypodium distachyon*, http://www. *Brachypodium*.org (version 1.0) (2010).

Published small RNA libraries that were generated from various tissues and growth conditions were used to test for generation of nat-siRNAs from antisense overlapping gene pairs. All small RNA reads were downloaded from the NCBI Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo). We used 13, 12, and 3 small RNA libraries from rice, maize, and *Brachypodium*, respectively.

Identification of the overlapping genes

We defined a gene as a part of the genomic region from the start to the end of an annotated transcript. Any two genes whose coordinates overlap by the length more than 30bp are considered as overlapping.

Finding orthologous genes

In order to find the orthologs of the genes involved in overlaps, we used SynMap online program (Lyons and Freeling, 2008; Lyons et al., 2008) to determine the collinear regions between different species. We used the default parameters which are: minimum number of aligned pairs = 5, maximum distance between two matches = 20 and average distance expected between syntenic genes = 10. If homologs are found in the collinear regions, they are considered orthologs. If no ortholog was found in the collinear region, we searched the whole genome to find possible homologs using blastp with e-value <e^-

10. If we didn't find any homolog either by this search, we assumed that gene has been deleted or created after the divergence of the two organisms.

Identification of overlapping genes that produce NAT-siRNAs

We mapped clean small RNA reads from 13, 12 and 3 libraries from rice, maize and *Brachypodium*, respectively, to the annotated overlapping genes in the same species using SOAP2 (Li et al., 2009) Only perfect matches were considered. We used the same procedure as in (Zhou et al., 2009) to determine whether nat-siRNAs were generated from the overlapping region of a gene pair. We calculated the density of mapped small RNAs (number of small RNAs per 1kb excluding introns) in the overlapping region and over all exons of the gene pairs. An enrichment score was calculated as the ratio of small RNA density in the overlapping region and entire gene pairs. A cutoff score of 2 was used to decide whether nat-siRNAs were generated from an overlapping pair. To determine the significance of this cutoff, we used a simulation procedure. Briefly, we randomly chose n (n is the number of overlapping pairs which has been found in each organism) pairs of genes. Then we calculated the enrichment score for them by randomly choosing the start position of the length "L" (L is the average length of overlapping region of all the overlapping pairs in each organism) along these gene pairs. We considered this length "L" in the randomly chosen gene pairs as the counterpart of the overlapping region in the overlapping genes in order to be able to calculate the enrichment score. This sampling process was repeated 10,000 times and for each time we calculated the average enrichment score. A p-value was then estimated by calculating the

frequency that a sample had an enrichment score larger than 2. The p-values for rice, maize, and *Brachypodium* are 0.003, 0.0001, and 0.22, respectively.

Acknowledgements

This work was supported by the UCR Initial Complement Fund and A USDA hatch fund (CA-R*-BPS-7754 H) to RL.

2.6 Figures & Tables

Figure Legends

Figure 2.1. Nine possible scenarios of the evolution of overlapping gene pairs in rice where there are orthologs for both overlapping genes in the other two species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.2. Nine possible evolutionary scenarios of overlapping gene pairs in *Brachypodium* with orthologs for both overlapping genes in the other two species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.3. Nine evolutionary scenarios of overlapping genes in maize that have orthologs for both overlapping genes in the other two species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.4. Nine evolutionary scenarios of rice overlapping genes in which at least one of the overlapping genes does not have an ortholog in one of the other two species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.5. Nine evolutionary scenarios of *Brachypodium* overlapping genes in which at least one of the overlapping genes doesn't have an ortholog in one of the other two l species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.6. Nine evolutionary scenarios of maize overlapping genes in which at least one of the overlapping genes doesn't have an ortholog in one of the other two species. Numbers in parenthesis represent cases of overlaps in each category. The dashes represent genes. The bar between two genes indicates that the genes are located on different chromosomes.

Figure 2.7. UTR expansion has caused neighboring genes become overlapping.

Figure 2.8. Elimination of intergenic gap in neighbors has created overlaps.

Figure 2.9. Intron Elongation has caused neighboring genes become overlapping.

Figure 2.10. Intron insertion in UTR in neighbors has created overlaps.

Figure 2.11. Obtaining a new CDS has caused neighboring genes become overlapping.

Figure 2.12. Example of conserved overlapping pairs that can be the source of nat-siRNA in rice and maize.

Table 2.1. Classification of overlapping gene pairs (overlap length>=30bp) based on gene expression orientation.

Species		Same-strand overlaps		Different-strand overlaps		
		Not-nested	Nested	Not-nested		Nested
	Total # of overlaps (% in all genes)			Divergent	Convergent	
Rice	747 (2%)	27	32	54	532	102
Maize (all)	1564 (10%)	10	189	248	447	670
Maize (evidenceBased)	1232 (9%)	1	21	232	434	544
Maize (cdnaSupported)	807 (8%)	1	15	141	294	356
Arabidopsis	1226 (9%)	49	92	59	968	58

Table 2.2. Classification of overlapping genes based on gene segments in overlapping region.

Species	**Total number of overlaps** (% in all genes)	**UTR-UTR overlap (exclusive)**	**Exon-exon overlap**	**CDS involved**	**Not exonic**	**UTR annotation in overlapping region**
Rice	747 (2%)	42	259	298	17	460
Brachypodium	347 (3%)	4	109	134	35	205
Maize (all)	1564 (10%)	128	672	901	81	704
Maize (evidenceBased)	1232 (9%)	128	584	665	44	666
Maize (cdnaSupported)	807 (8%)	90	372	450	23	435
Arabidopsis	1226 (9%)	706	959	181	18	953

Table 2.3. Prediction of the status of overlapping genes in the cereals ancestors based on their conservation in current cereals genomes.

Maize	Y	Y	Y	N	N	Y	Y
Rice	Y	Y	N	Y	N	N	Y
Brachypodium	Y	N	Y	N	Y	N	Y
Ancestor	Y	Y	Y	N	N	?	?
Count	32	39	17	603	225	720	73

Table 2.4. Confirmed mechanisms of birth and death of OGs by comparative analysis.

Species		Gene Creation/loss		Translocation	Exaptation
	Total # of overlaps	One gene creation/loss	Two gene creation/loss		
Rice (OG birth)	747	16%	2.5%	17%	6%
Brachy (OG birth)	347	7%	0	24%	5%
Rice (OG death)	747	0.4%	0	0.5%	3%
Brachy (OG death)	347	3%	0	11%	6%

Table 2.5. Different exaptation mechanisms in generation of overlapping genes.

Total	UTR expansion	Del of intergenic region	Int insertion in UTR	Int elongation	Obtaining new cds	Not determined
56	28	11	3	1	3	10

Figure 2.1.

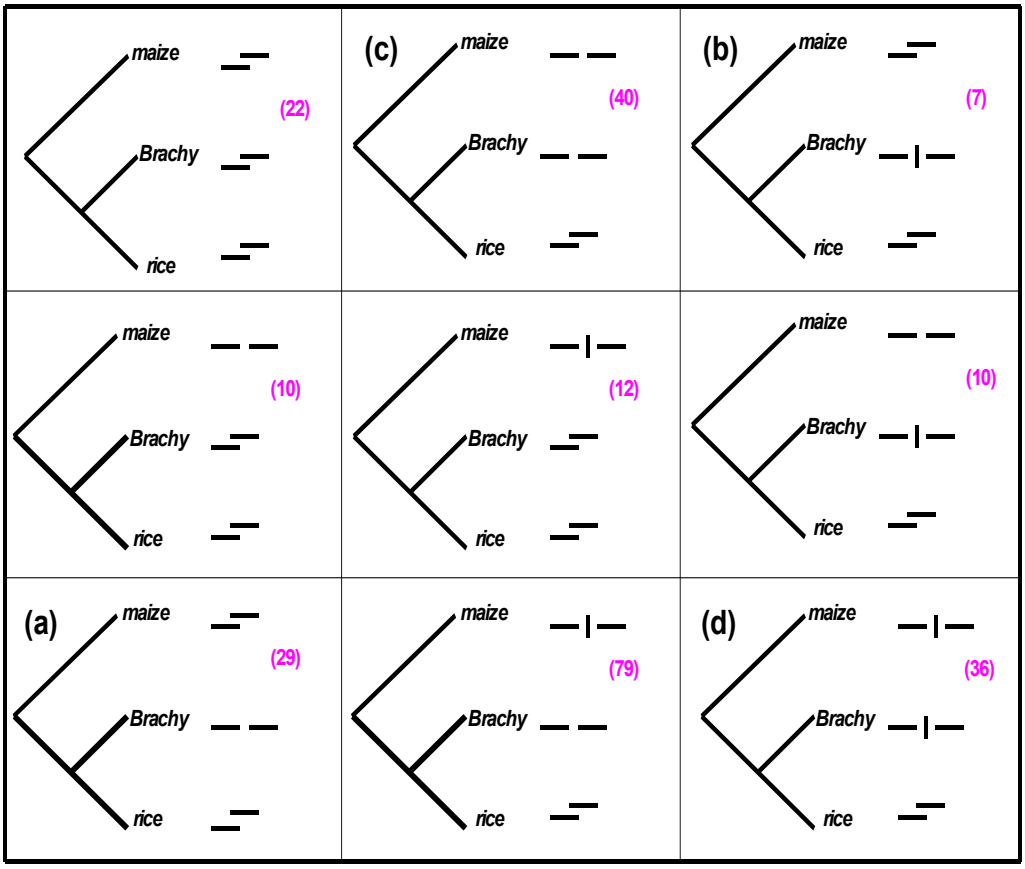

maize
Brachy
rice
(22)
(c)
maize
Brachy
rice
(40)
(b)
maize
Brachy
rice
(7)
maize
Brachy
rice
(10)
maize
Brachy
rice
(12)
maize
Brachy
rice
(10)
(a)
maize
Brachy
rice
(29)
maize
Brachy
rice
(79)
(d)
maize
Brachy
rice
(36)

Figure 2.2.

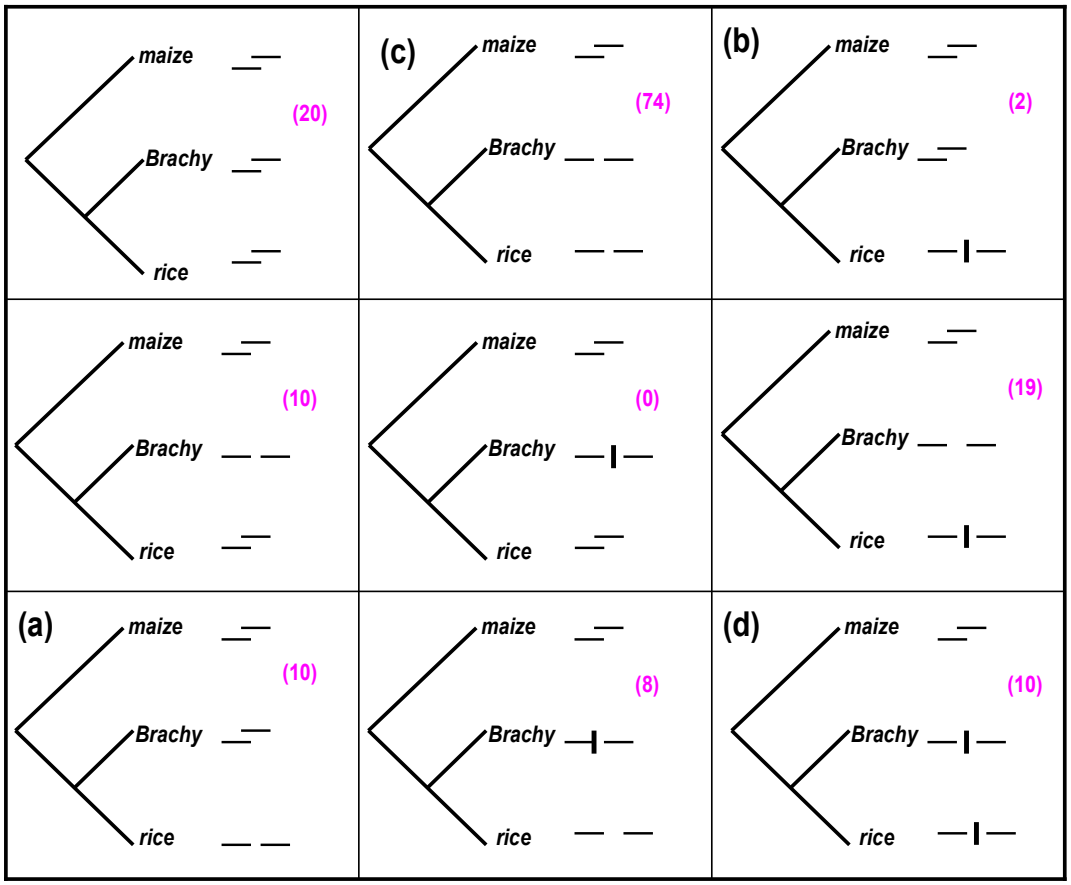

maize
Brachy
rice
(20)
(c)
maize
Brachy
rice
(74)
(b)
maize
Brachy
rice
(2)
maize
Brachy
rice
(10)
maize
Brachy
rice
(0)
maize
Brachy
rice
(19)
(a)
maize
Brachy
rice
(10)
maize
Brachy
rice
(8)
(d)
maize
Brachy
rice
(10)

Figure 2.3.

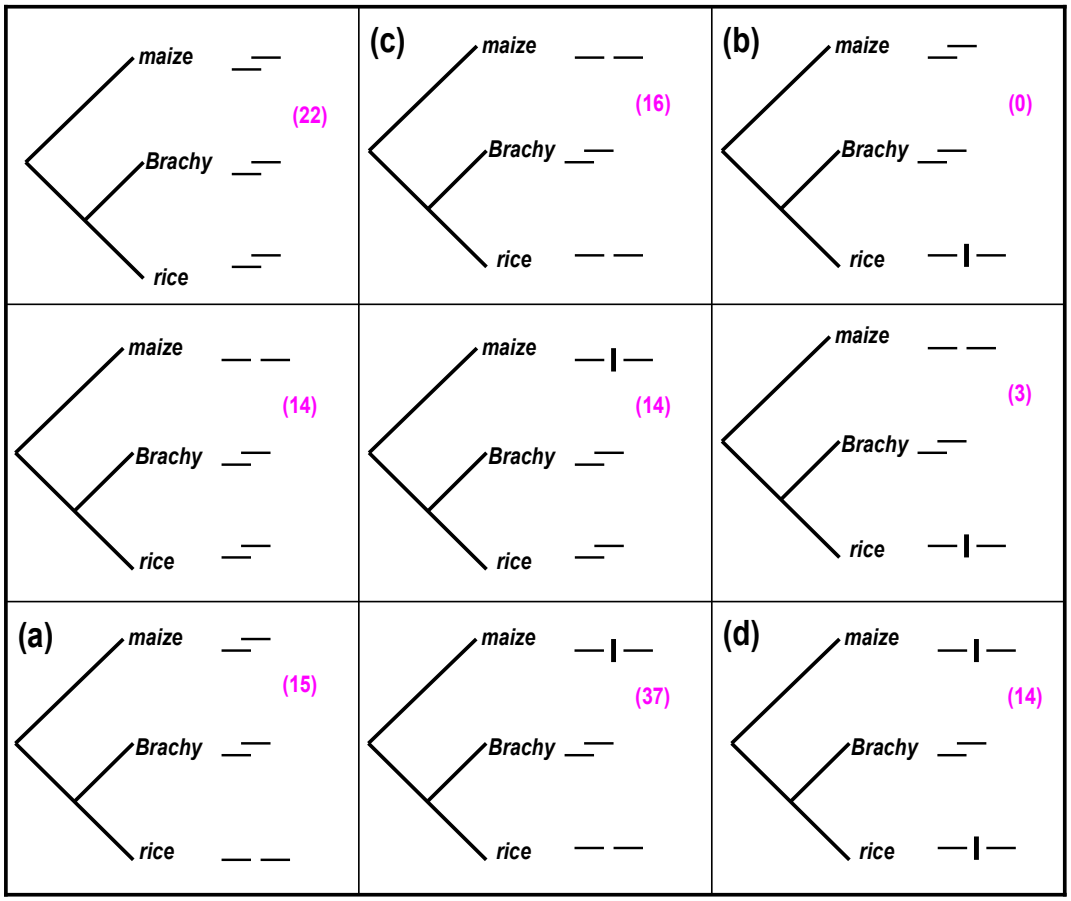

maize
Brachy
rice
(22)
(c)
maize
Brachy
rice
(16)
(b)
maize
Brachy
rice
(0)
maize
Brachy
rice
(14)
maize
Brachy
rice
(14)
maize
Brachy
rice
(3)
(a)
maize
Brachy
rice
(15)
maize
Brachy
rice
(37)
(d)
maize
Brachy
rice
(14)

Figure 2.4.

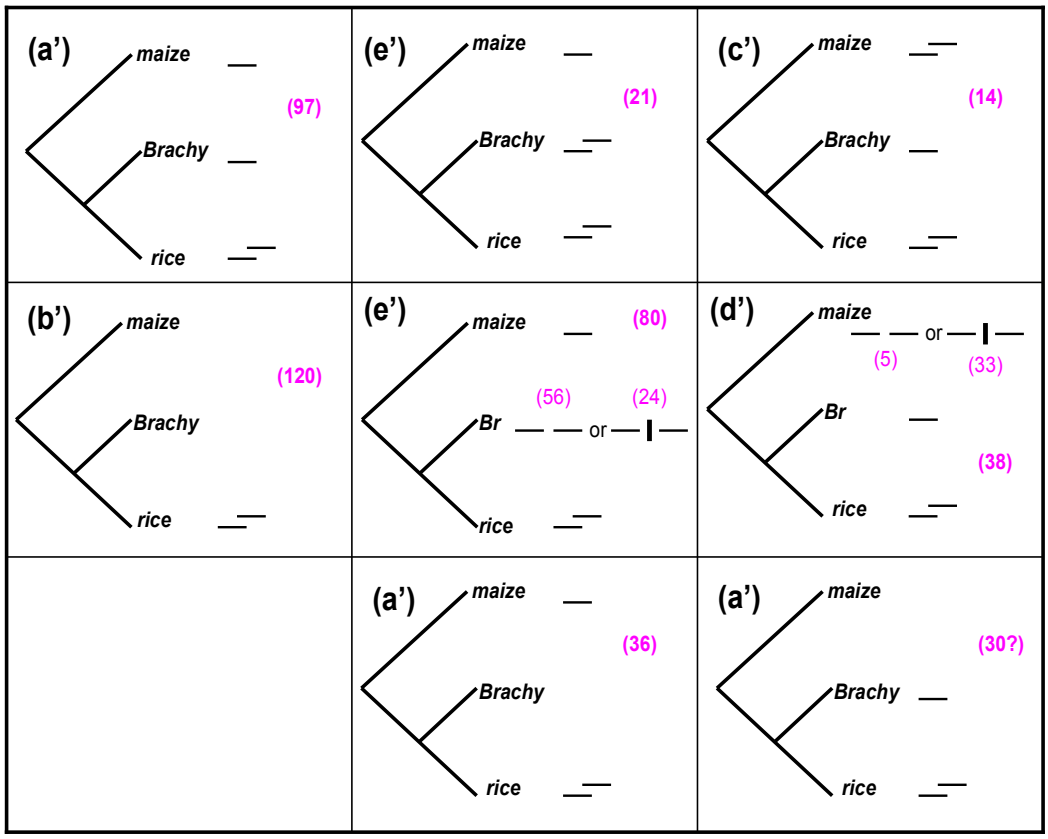

(a')
maize
(97)
Brachy
rice
(e')
maize
(21)
Brachy
rice
(c')
maize
(14)
Brachy
rice
(b')
maize
(120)
Brachy
rice
(e')
maize
(80)
(56)
(24)
Br
or
rice
(d')
maize
or
(5)
(33)
Br
(38)
rice
(a')
maize
(36)
Brachy
rice
(a')
maize
(30?)
Brachy
rice

Figure 2.5.

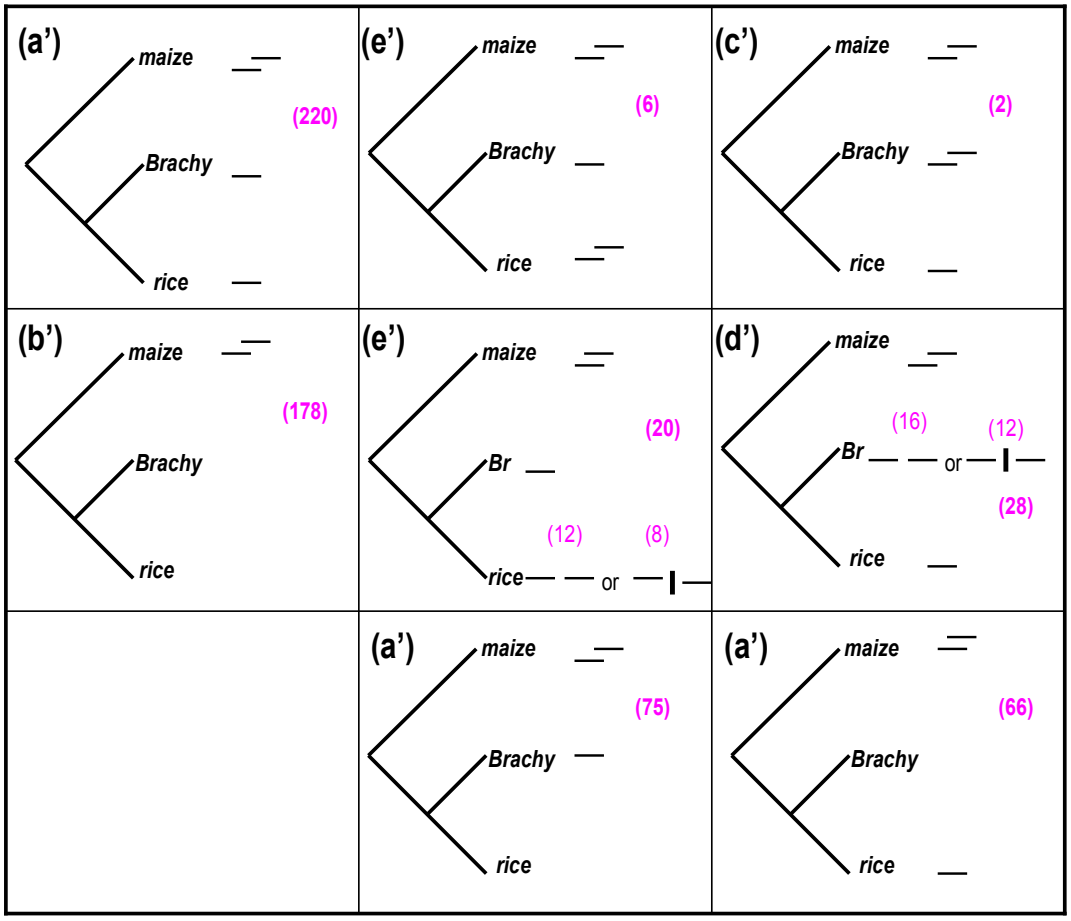

(a')
maize
(220)
Brachy
rice
(e')
maize
(6)
Brachy
rice
(c')
maize
(2)
Brachy
rice
(b')
maize
(178)
Brachy
rice
(e')
maize
(20)
Br
(12)
(8)
rice
or
(d')
maize
(16)
(12)
Br
or
(28)
rice
(a')
maize
(75)
Brachy
rice
(a')
maize
(66)
Brachy
rice

Figure 2.6.

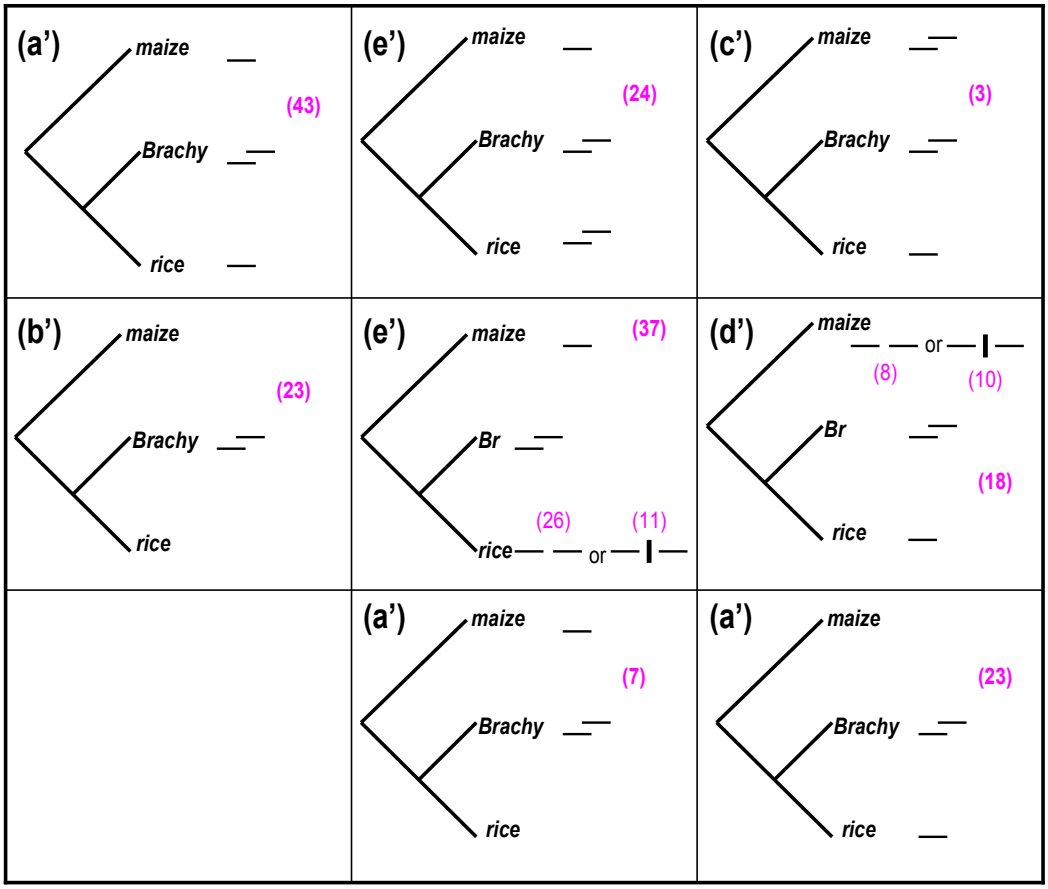

(a')
maize
(43)
Brachy
rice
(e')
maize
(24)
Brachy
rice
(c')
maize
(3)
Brachy
rice
(b')
maize
(23)
Brachy
rice
(e')
maize
(37)
Br
rice
(26)
or
(11)
(d')
maize
or
(8)
(10)
Br
(18)
rice
(a')
maize
(7)
Brachy
rice
(a')
maize
(23)
Brachy
rice

Figure 2.7.

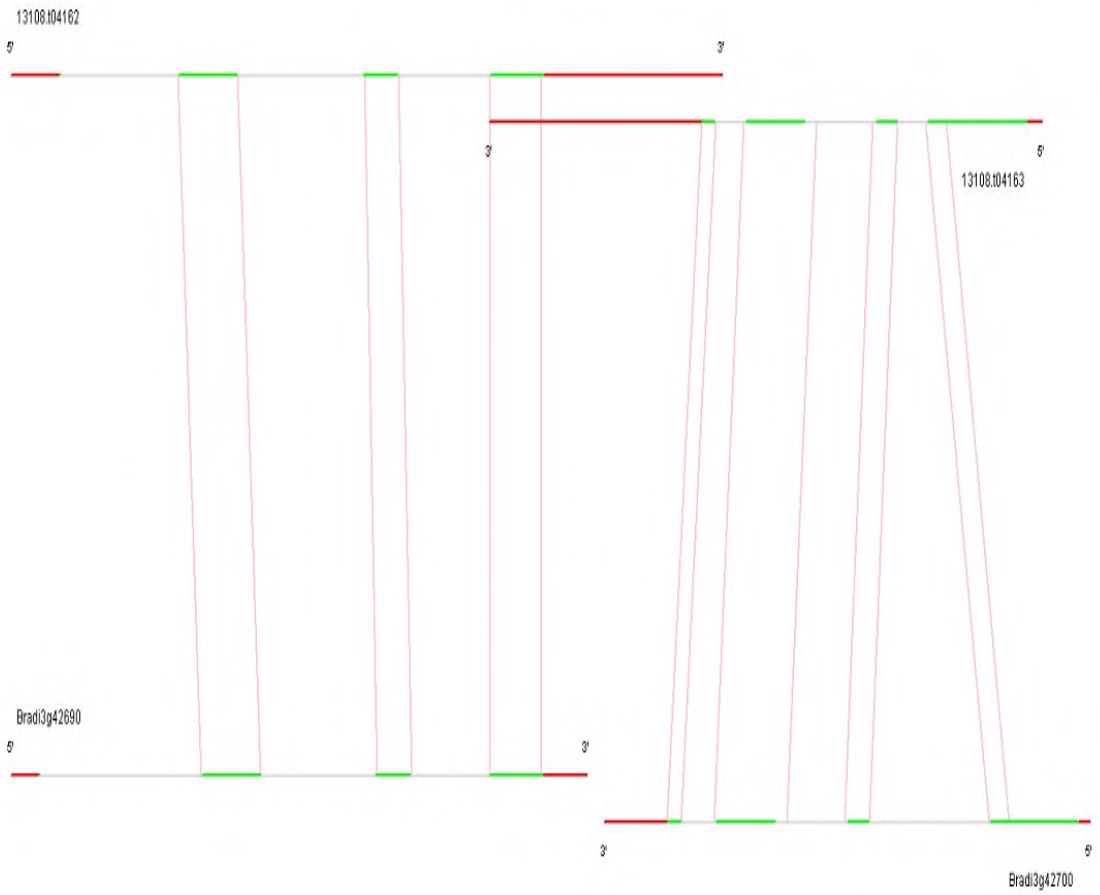

13108.t04162
5'
3'
3'
5'
13108.t04163
Bradi3g42690
5'
3'
3'
5'
Bradi3g42700

Figure 2.8.

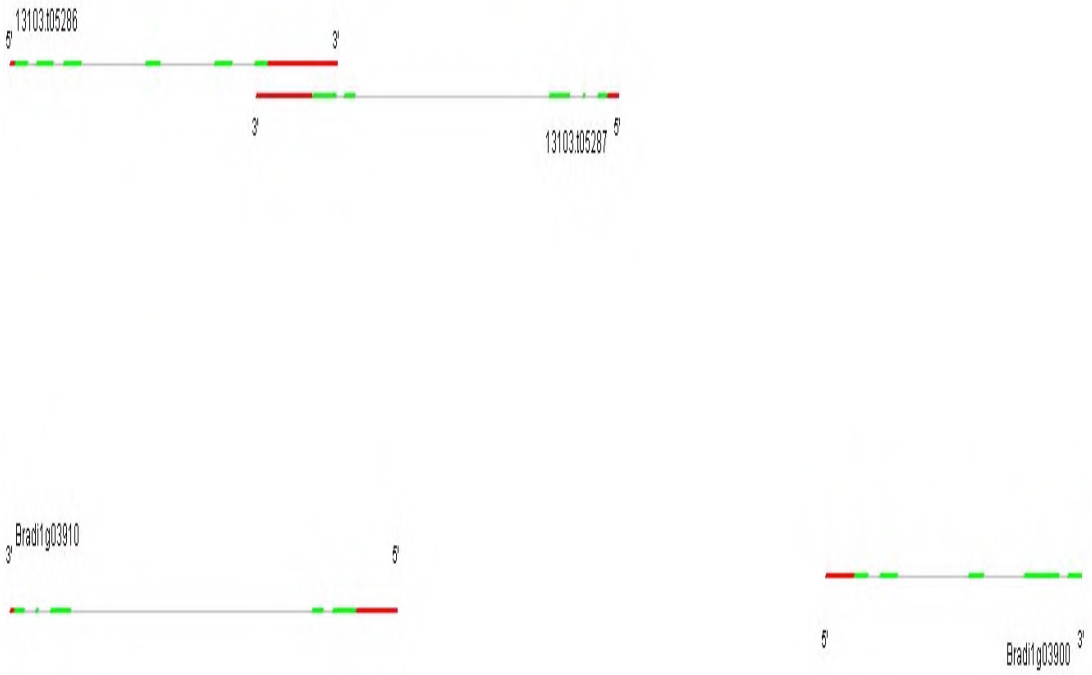

13103.t05286
5'
3'
3'
13103.t05287
5'
Bradi1g03910
3'
5'
5'
Bradi1g03900
3'

Figure 2.9.

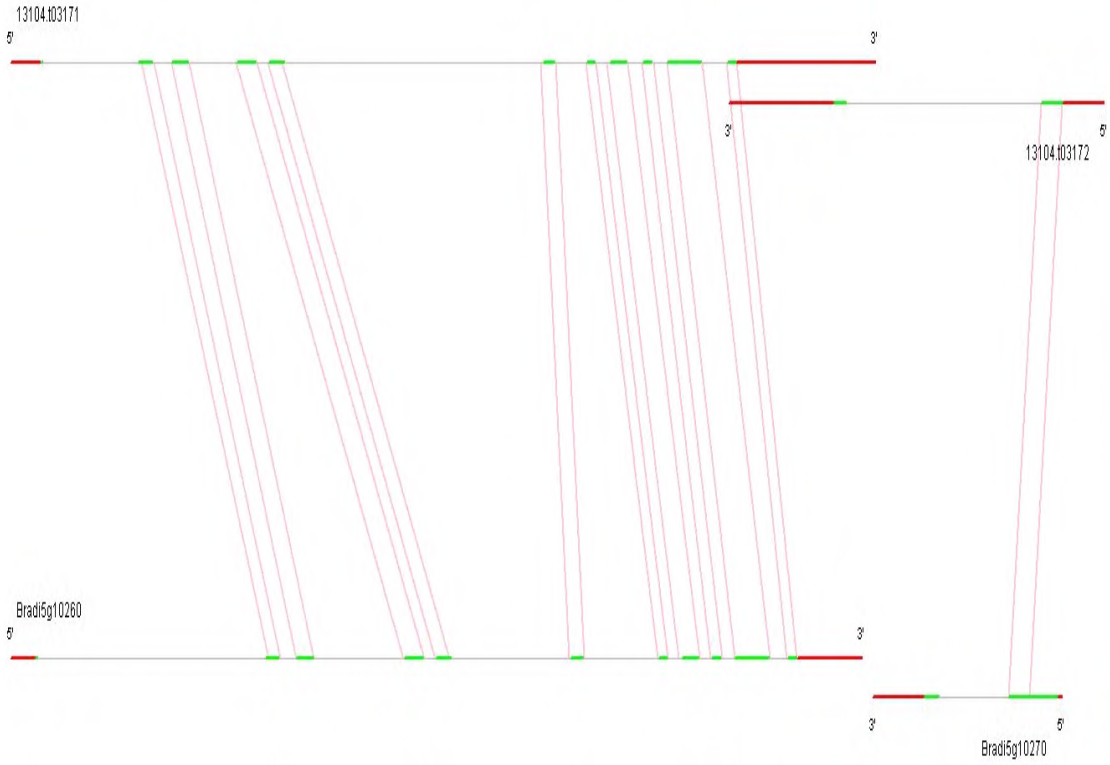

13104.t03171
5'
3'
3'
5'
13104.t03172
Bradi5g10260
5'
3'
3'
5'
Bradi5g10270

Figure 2.10.

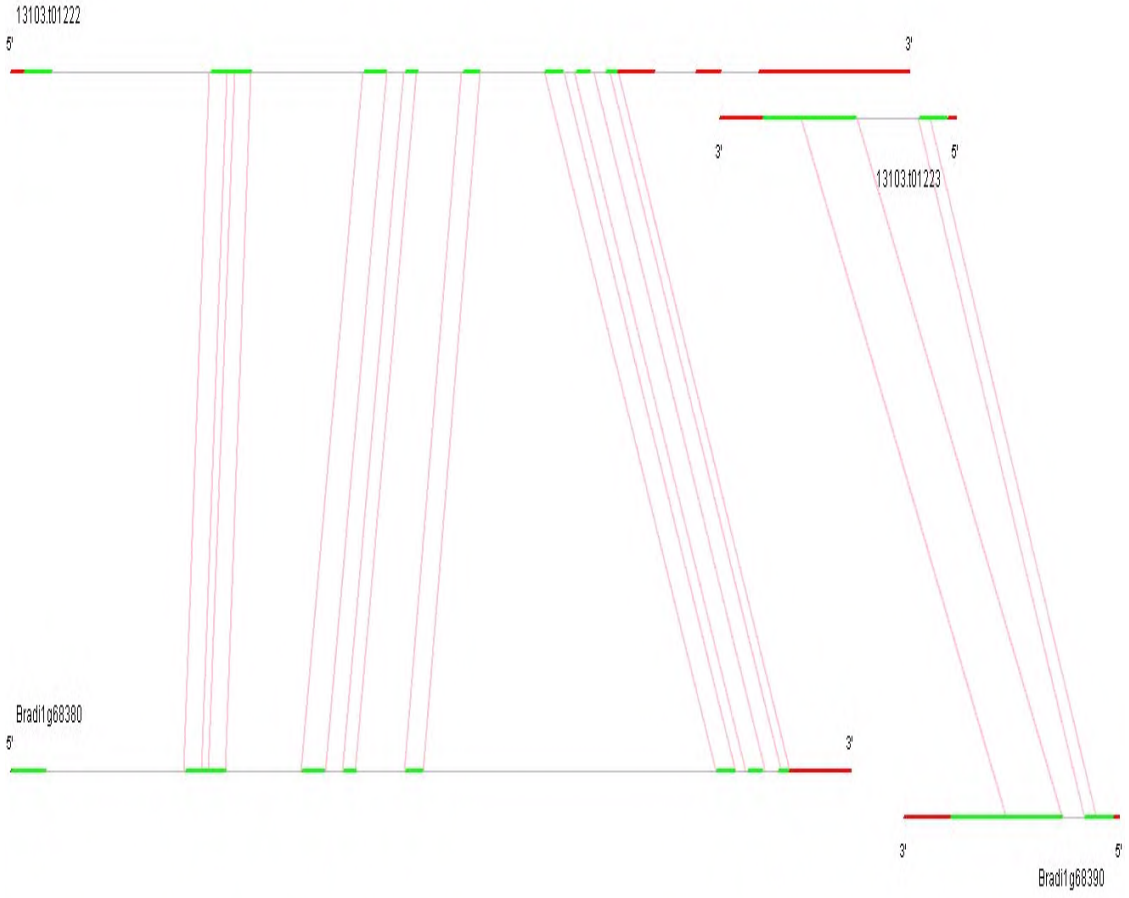

13103.t01222
5'
3'
3'
5'
13103.t01223
Bradi1g68380
5'
3'
3'
5'
Bradi1g68390

Figure 2.11.

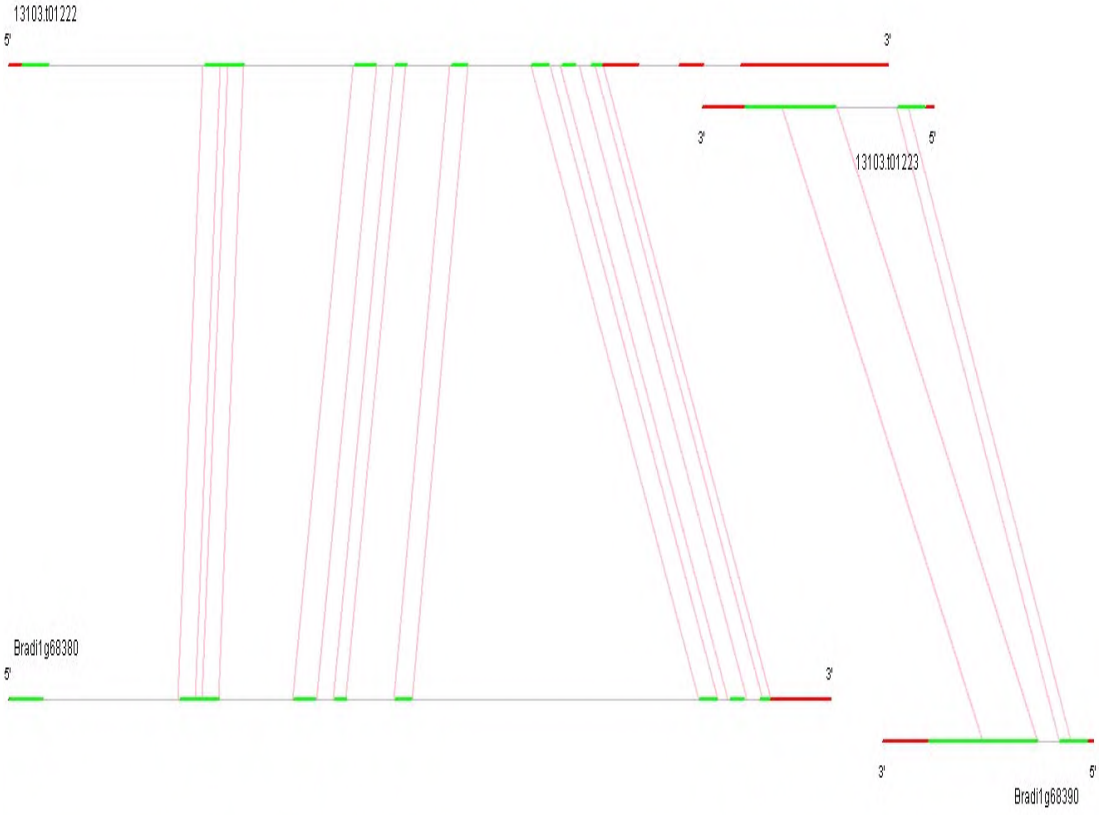
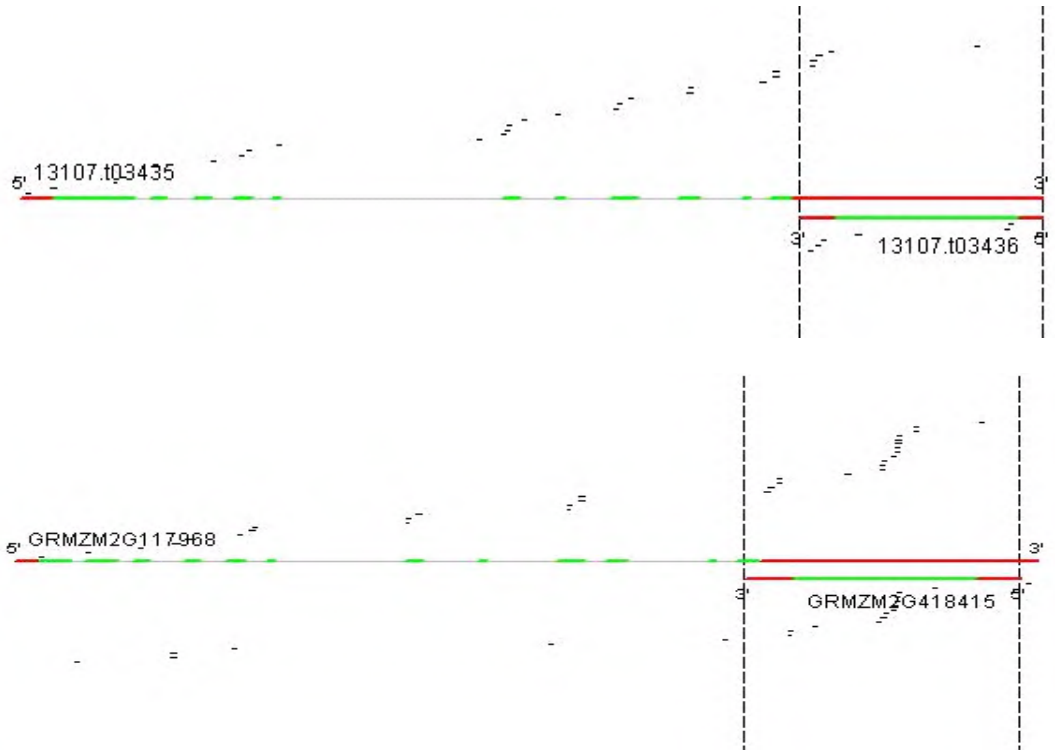

13103.t01222
5'
3'
3'
5'
13103.t01223
Bradi1g68380
5'
3'
3'
5'
Bradi1g68390

Figure 2.12.


5'
13107.t03435
3'
3'
13107.t03436
5'
5'
GRMZM2G117968
3'
3'
GRMZM2G418415
5'

2.7 References

(2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. In Nature , pp. 763-768.

Barrell, B.G., Air, G.M., and Hutchison, C.A., 3rd. (1976). Overlapping genes in bacteriophage phiX174. Nature **264,** 34-41.

Belshaw, R., Pybus, O.G., and Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. In Genome Res , pp. 1496-1504.

Bennetzen, J.L., and Ramakrishna, W. (2002). Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol **48,** 821-827.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell **126,** 37-47.

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. In Cell , pp. 1279-1291.

Dahary, D., Elroy-Stein, O., and Sorek, R. (2005). Naturally occurring antisense: transcriptional leakage or real overlap? In Genome Res , pp. 364-368.

Galante, P.A., Vidal, D.O., de Souza, J.E., Camargo, A.A., and de Souza, S.J. (2007). Sense-antisense pairs in mammals: functional and evolutionary considerations. In Genome Biol , pp. R40.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). In Science , pp. 92-100.

Henz, S.R., Cumbie, J.S., Kasschau, K.D., Lohmann, J.U., Carrington, J.C., Weigel, D., and Schmid, M. (2007). Distinct expression patterns of natural antisense transcripts in Arabidopsis. In Plant Physiol , pp. 1247-1255.

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. In BMC Mol Biol , pp. 6.

Johnson, Z.I., and Chisholm, S.W. (2004). Properties of overlapping genes are conserved across microbial genomes. In Genome Res , pp. 2268-2272.

Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Jr., Zhu, J.K., Staskawicz, B.J., and Jin, H. (2006). A pathogen-inducible endogenous siRNA in plant immunity. In Proc Natl Acad Sci U S A , pp. 18002-18007.

Keese, P.K., and Gibbs, A. (1992). Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A **89,** 9489-9493.

Lapidot, M., and Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. In EMBO Rep , pp. 1216-1222.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. In Bioinformatics , pp. 1966-1967.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. In Plant J , pp. 661-673.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., and Freeling, M. (2008). Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. In Plant Physiol , pp. 1772-1781.

Makalowska, I., Lin, C.F., and Makalowski, W. (2005). Overlapping genes in vertebrate genomes. In Comput Biol Chem , pp. 1-12.

Makalowska, I., Lin, C.F., and Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. In BMC Evol Biol , pp. 193.

Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P., and Olsson, O. (1983). Overlapping genes. Annu Rev Genet **17,** 499-525.

Osato, N., Yamada, H., Satoh, K., Ooka, H., Yamamoto, M., Suzuki, K., Kawai, J., Carninci, P., Ohtomo, Y., Murakami, K., Matsubara, K., Kikuchi, S., and Hayashizaki, Y. (2003). Antisense transcripts with rice full-length cDNAs. In Genome Biol , pp. R5.

Palleja, A., Harrington, E.D., and Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? In BMC Genomics , pp. 335.

Rancurel, C., Khosravi, M., Dunker, A.K., Romero, P.R., and Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. In J Virol , pp. 10719-10736.

Samuel, C.E. (1989). Polycistronic animal virus mRNAs. Prog Nucleic Acid Res Mol Biol **37,** 127-153.

Sanna, C.R., Li, W.H., and Zhang, L. (2008). Overlapping genes in the human and mouse genomes. In BMC Genomics , pp. 169.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke,

J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. In Science , pp. 1112-1115.

Shintani, S., O'HUigin, C., Toyosawa, S., Michalova, V., and Klein, J. (1999). Origin of gene overlap: the case of TCP1 and ACAT2. Genetics **152,** 743-754.

Solda, G., Suyama, M., Pelucchi, P., Boi, S., Guffanti, A., Rizzi, E., Bork, P., Tenchini, M.L., and Ciccarelli, F.D. (2008). Non-random retention of protein-coding overlapping genes in Metazoa. In BMC Genomics , pp. 174.

Steigele, S., and Nieselt, K. (2005). Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. In Nucleic Acids Res , pp. 5034-5044.

Sun, M., Hurst, L.D., Carmichael, G.G., and Chen, J. (2005). Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. In Nucleic Acids Res , pp. 5533-5543.

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., and Makalowska, I. (2004). Mammalian overlapping genes: the comparative perspective. In Genome Res , pp. 280-286.

Yu, P., Ma, D., and Xu, M. (2005). Nested genes in the human genome. In Genomics , pp. 414-422.

Zhang, X., Xia, J., Lii, Y.E., Barrera-Figueroa, B.E., Zhou, X., Gao, S., Lu, L., Niu, D., Chen, Z., Leung, C., Wong, T., Zhang, H., Guo, J., Li, Y., Liu, R., Liang, W., Zhu, J.K., Zhang, W., and Jin, H. (2012). Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. In Genome Biol , pp. R20.

Zhou, X., Sunkar, R., Jin, H., Zhu, J.K., and Zhang, W. (2009). Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa. In Genome Res , pp. 70-78.

Chapter 3

Refining microRNA Gene Boundaries in Cereals Using RNAseq Data

3.1 Abstract

RNAs are small RNA molecules that play important regulatory roles in plant development and stress response. MiRNA genes are transcribed into pri-miRNAs that form hairpin structures. Pri-miRNAs are cleaved by Dicer like proteins into pre-miRNAs, which are cleaved further into misran/mina* duplexes. Annotation of plant miRNAs is usually limited to the identification of mature miRNAs and pre-miRNAs without defining full-length miRNA gene boundaries. High-throughput RNA sequencing (RNAseq) has become a popular method for characterizing the whole transcriptome, including the transcripts of protein-coding genes and non-coding RNAs. Mapping RNAseq data to the genome can thus help define miRNA gene boundaries, which make it easier for identifying regulatory elements in the upstream regions of miRNA genes. We developed a methodology to use a large number of the recently available RNAseq data to define gene boundaries of miRNAs in three plant genomes and did a comparative analysis on these gene boundaries. By using the two known fully characterized miRNA gene datasets in Arabidopsis and maize, we optimized the parameters of our prediction procedure and evaluate its accuracy. We started the mapping from the pre-miRNA boundaries and then extended it towards pri-miRNA 5' boundaries using RNAseq data. In two training sets, in 75% of maize extended pre-miRNA, the average extension error rate (transcription start

site (TSS) prediction error rate) was only 20% of the total extended length. This average error rate was 15% in 82% of the extended *Arabidopsis* pre-miRNAs. Using this procedure, we predicted the upstream boundaries for 84 and 123 miRNA genes in maize and rice, respectively. This shows that our methodology could improve the annotation of considerable number of miRNAs with unknown gene boundaries. Mapping RNAseq data to the corresponding genomes is an effective approach for defining the upstream boundaries of miRNAs in plants, which provide a foundation for the identification of regulatory elements of miRNAs and the construction of miRNA-mediated regulatory networks in plants.

3.2 Introduction

MicroRNAs (miRNAs) are endogenous ~ 21 nucleotide small RNAs that regulate the expression of the target genes through sequence complementary (Bartel, 2004; Voinnet, 2009; Tang, 2010). They were found in plants for the first time in 2002 (Reinhart et al., 2002) and they play essential roles in plant growth, development, and stress response (Chen, 2004; Wang et al., 2005; Boualem et al., 2008; Ding et al., 2009).

Biogenesis of plant miRNAs is a multi-step process. MiRNA genes are first transcribed by RNA Polymerase II into 5' capped and poly (A)-tailed transcripts named pri-miRNAs (Lee et al., 2004). The pri-miRNAs are then cleaved by Dicer like proteins into imperfect fold back structures named precursor miRNAs (pre-miRNAs), which are further cleaved into ~21-nt miRNA/miRNA * duplexes (Kurihara et al., 2006; Voinnet, 2009; Meng et al., 2011). miRNA/miRNA* duplexes are methylated by HEN1 and exported to the cytoplasm by HASTY (Park et al., 2005; Yang et al., 2006). Mature miRNAs are loaded into miRNA-induced silencing complex (mi-RISC) and are used for silencing target transcripts through mRNA cleavage (Palatnik et al., 2003) or translational repression (Cai et al., 2009).

Little is known about the sequence and structure of the whole miRNA genes since it is difficult to identify full length miRNA transcripts *in vivo* (Meng et al., 2011). The majority of microRNAs are annotated with mature miRNA sequences and their precursors (pre-miRNAs) through computational and experimental methods (Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004). Pre-miRNAs are easy to define based on sequence homology of sequenced mature miRNAs and the secondary hairpin structure

of pre-miRNAs (Zhang et al., 2006; Fahlgren et al., 2010). However, the boundaries and structures of full-length miRNA genes are usually not defined.

It is important to know the exact miRNA gene boundaries because they help us understand the role of miRNAs in the regulatory networks (Meng et al., 2011). It has been shown that by integrating upstream sequence data of miRNAs, miRNA-target pair information and gene expression profile, the core regulatory modules of miRNAs which play an important role in our understanding of the gene expression regulations can be identified (Joung and Fei, 2009). The major player in miRNA gene regulations is their promoter region. In order to identify miRNA promoters, it is critical to identify the transcription start sites (TSS) of miRNA genes (Megraw et al., 2006). It has been shown in maize that TSS may be as little as <100 nucleotides and as much as >1000 nucleotides from the miRNA hairpin structure (this range is between 26 and 598 nucleotides in *Arabidopsis*), suggesting that promoter location can't be inferred directly from pre-miRNA coordinates alone (Xie et al., 2005a; Zhang et al., 2009). Due to the shortage of information about the exact promoter position, relatively little is known about the regulation of miRNA genes themselves, although much effort has been focused on elucidating the regulatory role of mature miRNAs.

In recent years several studies have been performed to characterize plant pri-microRNAs. Xie et al (2005) defined the boundaries of 99 *Arabidopsis* MIRNA loci by 5'RACE (Rapid Amplification of cDNA Ends) using locus specific primers (Xie et al., 2005b). The TSSs for 52 miRNA genes were mapped. The majority (86%) of transcripts were initiated with an adenosine, of which 93% were preceded by a pyrimidine. TATA

box-like sequences were detected in the upstream of 83% of these miRNAs. The identification of TSSs made it possible to discover promoter elements of *Arabidopsis* miRNA genes. They identified four new transcription binding motifs in these regions (Megraw et al., 2006). The structures of unprocessed primary miRNA transcripts were determined by 5' RACE and 3' RACE in maize as well (Zhang et al., 2009). Out of 89 tested miRNAs, they were able to capture the upstream-transcribed regions of 55 miRNA genes (TSS to the stem-loop). This low ratio is mainly because miRNA genes might be expressed in highly specific tissue/cell types, developmental stages, or environmental conditions. Only 18% of maize microRNA sequences in miRBase have a match in data for maize full-length complementary DNA due to the same reasons (Zhang et al., 2009). In another study via bioinformatics approaches the promoter of 212 rice miRNA genes were detected (Cui et al., 2009).

High-throughput RNA sequencing data have been used to resolve the transcription landscape in several studies (Mortazavi et al., 2008; Zhang et al., 2011; Wang et al., 2009; Garber et al., 2011) and has been applied to transcriptome analysis in several species such as Saccharomyces Cerevisiae, Saccharomyces Pombe, *Arabidopsis* and mouse (Cloonan et al., 2008; Lister et al., 2008; Nagalakshmi et al., 2008). RNAseq is an experimental procedure that generates sequence reads derived from the whole transcriptome using the recently developed deep sequencing technologies (Garber et al., 2011). The whole population of RNAs (total or fractionated using poly(A) tail) is converted to cDNAs using random primers (Cloonan et al., 2008; Mortazavi et al., 2008). The cDNAs are then sequenced in a high-throughput manner (with or without

amplification) to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). The reads are typically 30–400 bps, depending on the DNA sequencing technology (Ding et al., 2009; Wang et al., 2009). These reads can then be mapped to the reference genome for transcriptome reconstruction or expression quantification (Garber et al., 2011).

Compared to other methods for transcriptome profiling such as microarrays or cDNA/EST sequencing, RNAseq provides higher throughput and resolution with lower cost (Wang et al., 2009; Costa et al., 2010; Garber et al., 2011). Therefore, RNAseq is a preferred method for mapping gene and exon boundaries, discovery of novel or alternatively spliced transcripts, and measuring transcription level (Wang et al., 2009). RNAseq can reveal the precise location of transcription boundaries, to a single base resolution. Nagalakshmi et al. showed that 5′ and 3′ boundaries could be mapped to within 10–50 bases by a significant drop in mapping signal. 3′ boundaries can be precisely mapped by searching for poly (A) tags, and introns can be mapped by searching for tags that span GT–AG splicing consensus sites (Nagalakshmi et al., 2008). Furthermore, short reads from RNAseq that span known splice junctions give qualitative and quantitative information about how two exons are connected, whereas longer reads or pair-end short reads can reveal connectivity between multiple exons. RNAseq has been used to discover alternative splicing in 33% of rice genes and more than 200 chimeric transcripts (Zhang et al., 2010). Novel transcripts that are expressed at very low level and potential functional noncoding RNAs can be identified when poly(A)-enriched RNAseq reads are mapped to the unannotated regions in the genome. Using this strategy, more

than 7,000 transcriptional units, including ~1000 non-TE, long novel transcripts with known protein domains, were identified in rice (Zhang et al., 2010).

Several of studies have been done on the conservation of plant miRNA genes at hairpin level (Axtell and Bowman, 2008; Voinnet, 2009; Fahlgren et al., 2010; Tang, 2010); However little has been done on the evolution of miRNA gene structures. In both plants and animals, some MIRNA families are highly conserved through hundreds of millions of years. There is a large number of young miRNAs, too (Axtell and Bowman, 2008). It has been shown that evolutionary stability of different sub-regions of miRNA hairpins varies substantially (Fahlgren et al., 2010; Tang, 2010). Whereas mature miRNA and miRNA* are usually quite stable, other sub-regions are much more dynamic (Fahlgren et al., 2010; Tang, 2010). The recent determination of the genome sequences of closely related plant species and availability of lots of next generation sequencing data provide excellent opportunities for a more detailed study of the evolution of whole miRNA genes.

Here we set to use RNAseq data to define miRNA boundaries and to investigate the evolution of miRNA genes in three closely related cereal genomes. We developed a new methodology and used a subset of miRNA genes in *Arabidopsis* and maize with known boundaries as a training set. We reached a good prediction performance in our procedure. The average error rates for predicting 5' miRNA gene boundary for the majority of known miRNA genes in maize and Arabidopsis are 20% and 15%, respectively. Using our methodology, we are able to predict 5' gene boundary for 84 additional miRNA genes in maize and 123 miRNA genes in rice. Comparative analysis of

orthologous miRNA genes in maize and rice shows that gene structure is not well conserved in these two closelyrelated cereals.

3.3 Results

Extending miRNA boundaries using RNAseq data

We developed a new method for defining miRNA gene boundaries using RNAseq data. We mapped the RNAseq reads to the corresponding genome sequence and then focused on the RNAseq reads that were located in the upstream region of the annotated pre-miRNAs. We calculated the read density in a window located at the center of the pre-miRNA with a predefined widow size. We then used a sliding window with a predefined step length to extend miRNA gene structure towards the 5' end. The extension was stopped if the ratio of read density in the new window and the original window fell below a threshold.

We optimized the parameters of our algorithm (window size, step size and change in density) using a subset of miRNA genes in Arabidopsis and maize with known 5' gene boundaries. We systematically tested the performance of our method using parameters in the following range: window size: 50bp to 200bp, step size: 4bp to 25bp, and change in read density: 2 to 20 times. For each parameter set, we compared our predicted boundary and known boundary of known miRNA genes in the training set and calculated an error rate, which was defined as the ratio of the length difference between the predicted and real TSS to the total predicted extended length. The optimal parameters (window size,

step size and change in density) were 100, 20 and 7 for *Arabidopsis* and 100, 12 and 20 for maize, respectively.

It should be mentioned we filtered the training set in order to eliminate high error rates in two ways. First we only focused on those miRNA genes that were expanded below 520nt by our method. One reason to choose this number was that in maize the 5' regions (measured from the TSS to the stem-loop) have the mean length of 523nt for the experimentally confirmed pri-miRNAs. The miRNA gene boundary prediction becomes unreliable for those miRNA genes with very long extension on the 5' end. Second parameter that we picked to select the reliable extensions was the read density. We only focused on those miRNA genes with read density greater than 5 reads per 100 bp in the pre-miRNA region.

The performance of our method on two pri-miRNA training sets

We used the two available datasets on miRNA gene's TSSs in plants. In one dataset, 61 maize pri-miRNA 5' boundaries have been confirmed either by 5' RACE or full-length cDNA. A similar dataset has been obtained from *Arabidopsis* in which TSSs of 52 miRNA genes were mapped. Using the criteria we obtained in our methodology by maximizing the prediction performance, we extended the 5' region of 24 maize and 17 *Arabidopsis* pre-miRNAs with the known TSS. In 75% of known maize miRNA genes, the average extension error rate was 20% of the total extended length (Figure 3.1). This average error rate was 15% on 82% of the known miRNA genes in *Arabidopsis* (Figure 3.2).

Defining the new gene boundaries of miRNAs in rice and maize

Using the optimized parameters, we predicted the TSS of miRNA genes in rice and maize, which are without a known TSS. In maize we combined miRBase and the unpublished miRNAs from our laboratory and removed redundant miRNAs. Among 361 miRNAs in maize, 84 miRNAs had high enough read density for gene boundary prediction. The lengths of extension from the hairpin structure to TSS are shown in Figure 3.3.

In rice we used 546 miRNAs in the miRBase with known pre-miRNA location. We were able to define 5' boundary for 123 miRNA genes (Figure 3.4). The detailed results of these two sets are presented in Table 3.1 & 3.2.

Conservation of pri-miRNA 5' region between rice and maize

Based on the collinearity data and sequence homology, and after removing those miRNA with <1000bp distance to the upstream miRNA and those that were overlapping with a protein-coding genes, we found 74 rice-maize miRNA unique orthologous pairs. Among 18 orthologous pairs that had newly defined 5' boundary in both maize and rice, 73% of the miRNA orthologs had totally different 5' region extended length i.e. even if we consider the average error rate of our methodology, the extended length wouldn't be the same. In only 4 cases the difference between the extended 5' regions were less than 20% of the average extended lengths (Figure 3.5), suggesting that miRNA gene length is not well conserved between these two species.

3.4 Discussion

The availability of large number of RNAseq libraries from threeplants (rice, *Arabidopsis* and maize) makes it possible to computationally define miRNA gene boundaries. Figure 3.6 gives a good demonstration of how RNAseq reads can help pri-miRNA annotation. Because deep sequencing reads were generated randomly over the full length of RNA transcripts using random primers, they should be evenly distributed over the transcripts. Therefore, gene and exon boundaries can be precisely defined by monitoring mapped read density using a sliding window method (Sultan et al., 2008). We applied this approach to define miRNA gene boundaries and optimized parameters using two training sets of miRNAs with known gene boundaries. We were able to predict 5' boundaries for 84 and 123 miRNA genes in maize and rice, respectively. Comparative analysis of 5' boundaries of orthologous miRNA genes in maize and rice indicated that gene structure is not well conserved in these two closely related cereals.

Defining miRNA gene boundaries is important for studying the evolution of miRNA gene structure and finding cis-regulatory elements of miRNAs. Current computational approaches for identification of miRNA promoters were carried out by looking at the upstream regions of pre-miRNAs with arbitrarily chosen length (Zhou et al., 2007; Cui et al., 2009; Zhou et al., 2011). However, the length of miRNA genes varies greatly in plants. Experimental approaches such as 5' RACE have been used to define miRNA gene structures (Xie et al., 2005b; Zhang et al., 2009), but they are time-

and labor- intensive and are not high throughput. We showed here that mapping RNAseq reads is an effective computational approach for defining miRNA gene boundaries.

The effectiveness of our approach depends highly on the availability of sequence reads that are derived from full-length miRNA transcripts. The expression of miRNAs is usually tissue- and growth condition-specific. For lowly expressed miRNAs, it is hard to obtain high coverage on the miRNA genes. In addition, once miRNA genes are transcribed, they are quickly processed by Dicer like proteins to pre-miRNAs and then mature miRNAs. Therefore, reads from full-length miRNA transcripts are greatly reduced by this process. Performing RNAseq using RNA samples from mutant plants in which Dicer like protein is knocked out or knocked down will greatly improve the effectively of our method.

3.5 Materials and Methods

Sequence data

Genome sequences and annotation data were downloaded from the following websites: rice, the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu, version 6.1) (Goff et al., 2002); maize, http://www.maizesequence.org (version 5a) (Schnable et al., 2009) and *Arabidopsis Thaliana*, www.arabidopsis.org (TAIR9) (2000).

All RNAseq reads were downloaded from the NCBI Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo). We downloaded 23 paired-end sequencing SRA files and 2 single-end sequencing libraries for rice, 10 libraries for *Arabidopsis* and 12 paired-end sequencing libraries and 6 single-end sequencing for maize.

Trimming and Mapping RNAseq data to the genomic sequences

Published RNA_Seq libraries that were generated from various tissues and growth conditions were either clean reads (removed low quality, short reads and adapters) or raw reads. Raw reads were first converted into the fastq format using fastq_dump and then used to generate clean reads with the SolexaQA program (Cox et al., 2010). Clean reads were mapped to the 3kb upstream- and downstream-regions of annotated pri-miRNAs using SOAP2 (Li et al., 2009).

Annotated pre-miRNA datasets

Rice pre-miRNAs and their genomic coordinates were downloaded from the miRBase (version 18). Maize pre-miRNAs include the pre-miRNAs from the miRBase (version 18) and unpublished pre-miRNA from our lab after removing redundancy.

Identification of the orthologous miRNAs in rice and maize

Orthologous regions on the rice and maize chromosomes were identified using DAGchainer (Haas et al., 2004). Orthologous regions are regions with at least five anchor protein coding genes with no more than ten intervening genes between neighboring anchors. miRNAs of the same family that located in the collinear regions and flanked by the same anchor genes are considered orthologous miRNAs in the two species.

3.6 Figures & Tables

Figure Legends

Figure 3.1. Accuracy of miRNA TSS prediction using RNAseq data in maize maize (error bars show the real position of TSS)

Figure 3.2. Accuracy of miRNA TSS prediction using RNAseq data in *Arabidopsis* (error bars show the real position of TSS)

Figure 3.3. Distribution of extended length on 5' end of pre-miRNAs in maize

Figure 3.4. Distribution of extended length on 5' end of pre-miRNAs in rice

Figure 3.5. Comparison of the 5' end length of orthologous miRNA genes in rice and maize

Figure 3.6. An example of determining miRNA gene boundaries by RNAseq read mapping

Figure 3.1.

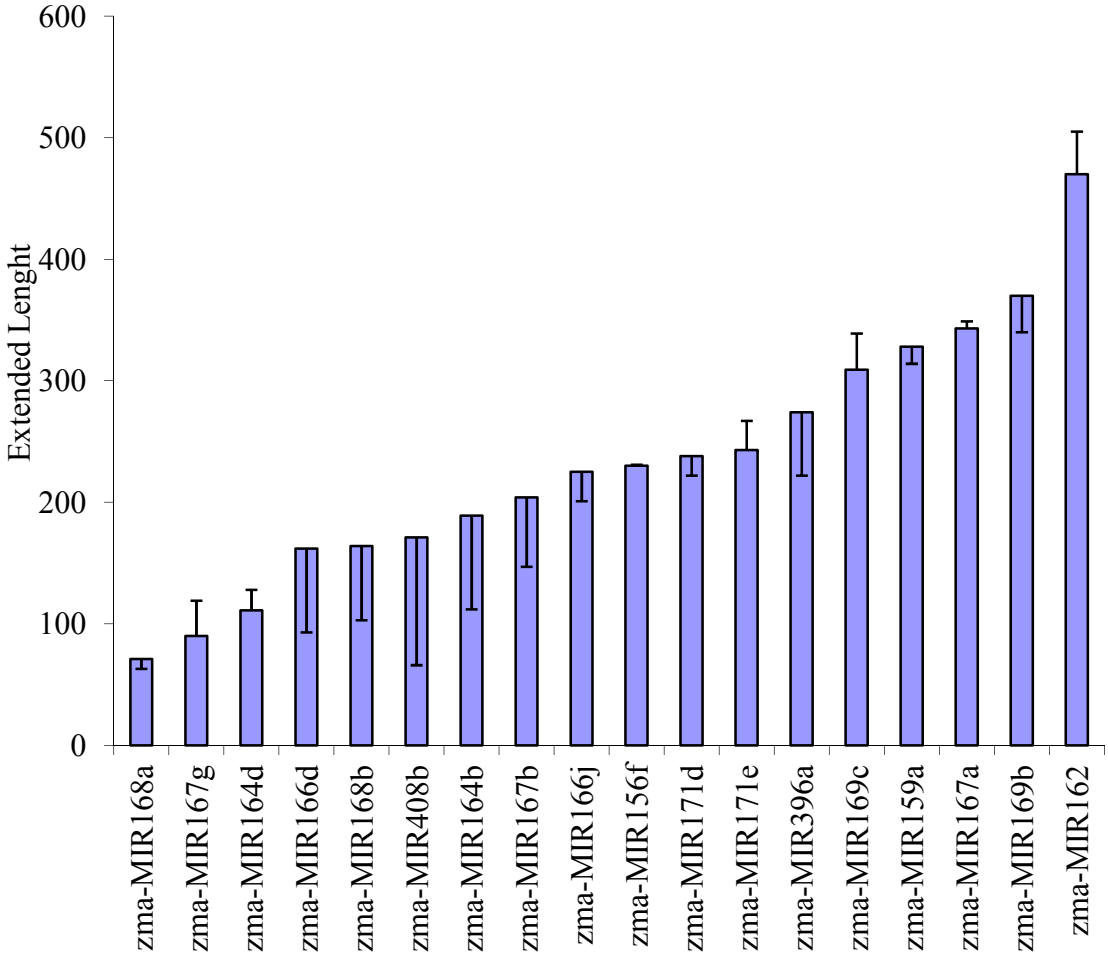

Extended Lenght
600
500
400
300
200
100
0
zma-MIR168a
zma-MIR167g
zma-MIR164d
zma-MIR166d
zma-MIR168b
zma-MIR408b
zma-MIR164b
zma-MIR167b
zma-MIR166j
zma-MIR156f
zma-MIR171d
zma-MIR171e
zma-MIR396a
zma-MIR169c
zma-MIR159a
zma-MIR167a
zma-MIR169b
zma-MIR162

Figure 3.2.

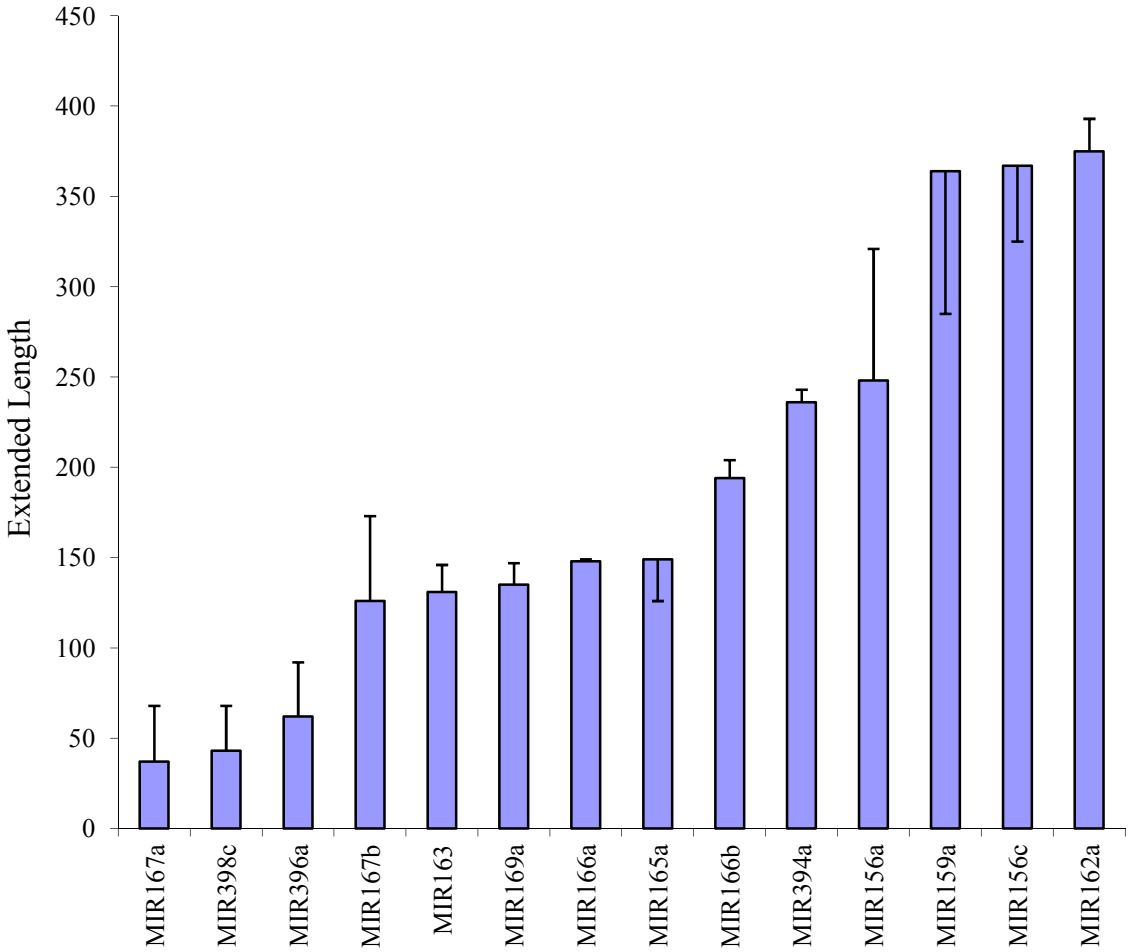

Extended Length
450
400
350
300
250
200
150
100
50
0
MIR167a
MIR398c
MIR396a
MIR167b
MIR163
MIR169a
MIR166a
MIR165a
MIR166b
MIR394a
MIR156a
MIR159a
MIR156c
MIR162a

Figure 3.3.

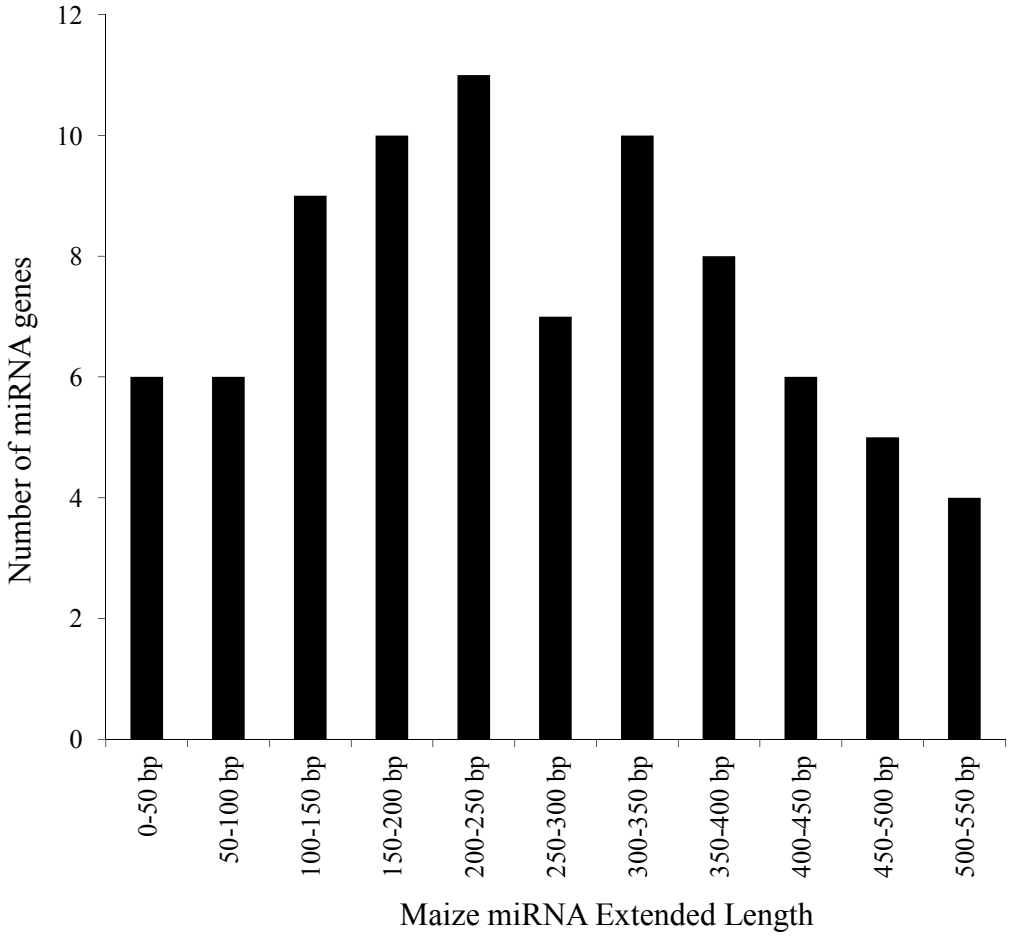

12
10
8
6
4
2
0
Number of miRNA genes
0-50 bp
50-100 bp
100-150 bp
150-200 bp
200-250 bp
250-300 bp
300-350 bp
350-400 bp
400-450 bp
450-500 bp
500-550 bp
Maize miRNA Extended Length

Figure 3.4.

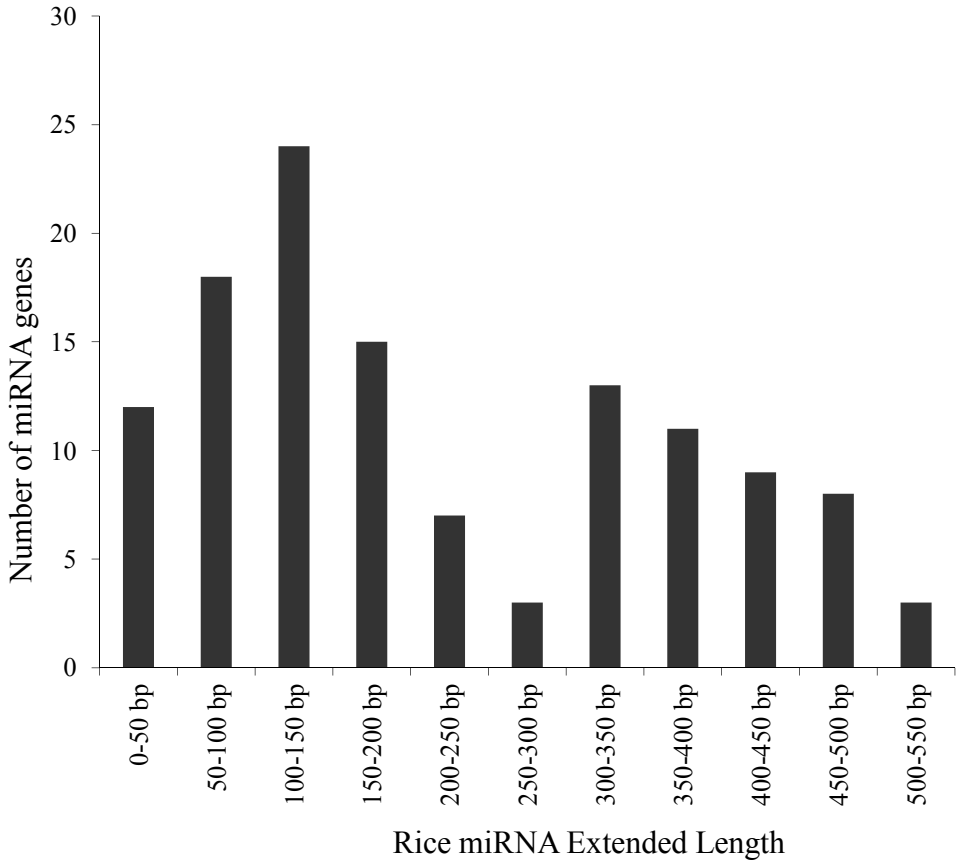

30
25
20
15
10
5
0
Number of miRNA genes
0-50 bp
50-100 bp
100-150 bp
150-200 bp
200-250 bp
250-300 bp
300-350 bp
350-400 bp
400-450 bp
450-500 bp
500-550 bp
Rice miRNA Extended Length

Figure 3.5.

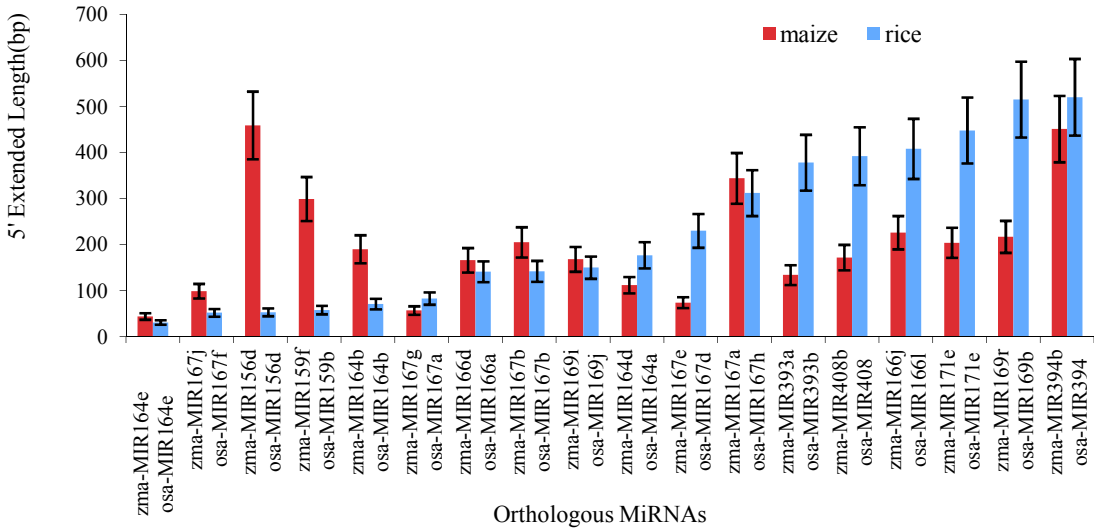

5' Extended Length(bp)
700
600
500
400
300
200
100
0
maize
rice
zma-MIR164e
osa-MIR164e
zma-MIR167j
osa-MIR167f
zma-MIR156d
osa-MIR156d
zma-MIR159f
osa-MIR159b
zma-MIR164b
osa-MIR164b
zma-MIR167g
osa-MIR167a
zma-MIR166d
osa-MIR166a
zma-MIR167b
osa-MIR167b
zma-MIR169i
osa-MIR169j
zma-MIR164d
osa-MIR164a
zma-MIR167e
osa-MIR167d
zma-MIR167a
osa-MIR167h
zma-MIR393a
osa-MIR393b
zma-MIR408b
osa-MIR408
zma-MIR166j
osa-MIR166l
zma-MIR171e
osa-MIR171e
zma-MIR169r
osa-MIR169b
zma-MIR394b
osa-MIR394
Orthologous MiRNAs

Figure 3.6.

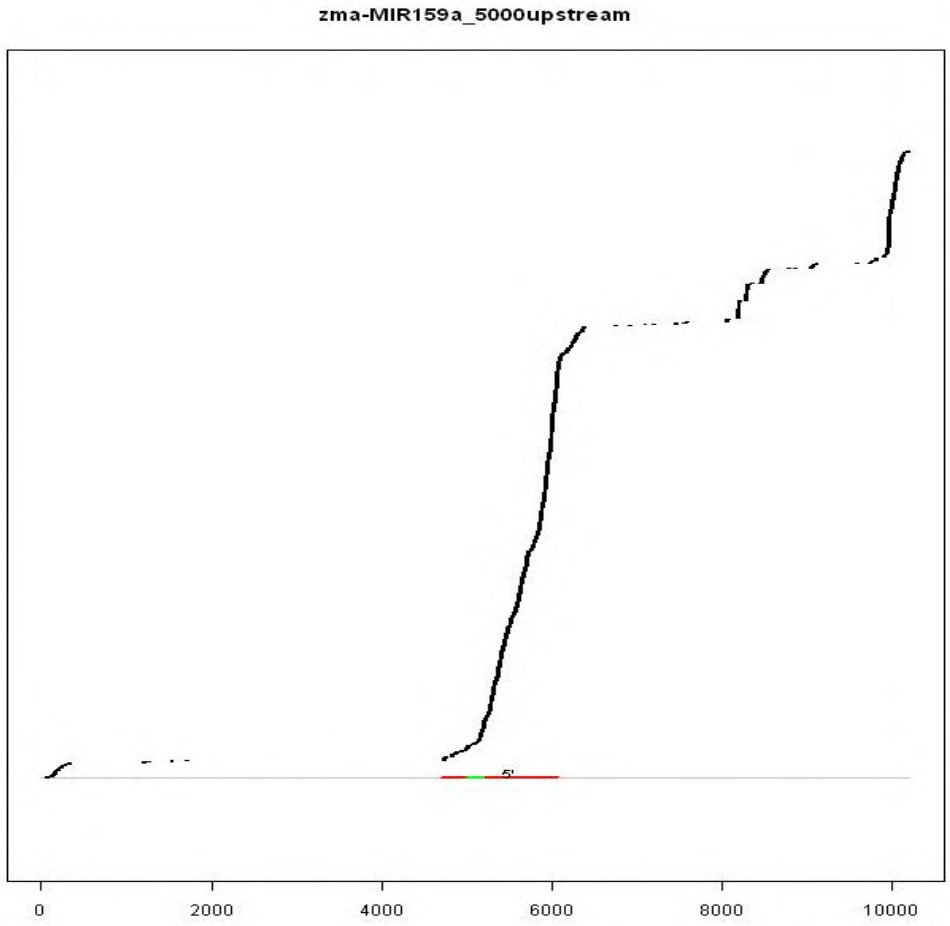


Table 3.1. Maize pre-miRRNA extended list by RNAseq mapping

miRNA genes	chr number	pre_miRNA start	pre_miRNA end	strand	extended_len
zma-MIR395o	chr10	144744547	144744772	+	218
zma-MIR528a	chr1	6415592	6415714	+	364
zma-MIR171e	chr1	7954980	7955105	+	204
zma-MIR167j	chr1	12595523	12595631	+	99
zma-MIR164e	chr1	46848652	46848767	+	44
zma-MIR390a	chr1	292578889	292579069	+	227
zma-MIR393a	chr2	751658	751783	+	134
zma-MIR169c	chr2	11299979	11300112	+	276
zma-MIR398a	chr2	169527758	169527897	+	328
zma-MIR156f	chr2	180193496	180193662	-	201
zma-MIR169j	chr2	192700339	192700489	+	429
zma-MIR159f	chr3	25490976	25491190	-	299
zma-MIR169a	chr3	37610280	37610409	+	504
zma-MIR169m	chr3	96704652	96704752	-	53
zma-MIR167a	chr3	119175685	119175874	+	344
zma-MIR167g	chr3	119177648	119177890	+	57
zma-MIR172e	chr3	144884289	144884462	+	44
zma-MIR397a	chr3	180667115	180667257	-	911
zma-MIR169n	chr3	229987641	229987744	+	207
zma-MIR169i	chr4	47241963	47242153	+	168
zma-MIR156i	chr4	137179087	137179203	-	132
zma-MIR394b	chr4	154166525	154166646	+	451
zma-MIR396a	chr4	173295127	173295263	+	275
zma-MIR396e	chr4	173300108	173300273	-	249
zma-MIR168b	chr4	239095553	239095656	-	165
zma-MIR167b	chr5	7688809	7688935	+	205
zma-MIR166d	chr5	21933694	21933797	-	166
zma-MIR168a	chr5	74340501	74340604	-	36
zma-MIR156d	chr5	92369224	92369342	-	459
zma-MIR169f	chr5	164711362	164711511	+	25
zma-MIR827	chr5	167348212	167348333	-	143
zma-MIR162	chr5	182040463	182040592	-	372
zma-MIR394a	chr5	193819088	193819213	+	456
zma-MIR396f	chr5	214340396	214340512	+	63
zma-MIR167h	chr6	93326220	93326362	-	22
zma-MIR156k	chr6	96127864	96127986	-	457
zma-MIR164b	chr6	141610040	141610167	+	190
zma-MIR167e	chr7	9830212	9830330	-	74
zma-MIR398b	chr7	38540171	38540278	+	357
zma-MIR171g	chr7	42481791	42481902	-	25

zma-MIR166j	chr7	124539680	124539833	+	226
zma-MIR156j	chr7	130652700	130652824	-	372
zma-MIR164d	chr7	172723300	172723515	-	112
zma-MIR169b	chr8	4791975	4792130	+	371
zma-MIR159a	chr8	10392580	10392825	+	291
zma-MIR159k	chr8	10585048	10585247	+	356
zma-MIR408b	chr8	38488497	38488643	-	172
zma-MIR169r	chr9	109138534	109138659	-	217
zma-MIR171d	chr9	126244364	126244484	-	201
Zma1_6388497_140	chr1	6415574	6415713	+	170
Zma10_128234646_140	chr10	128589865	128590004	-	180
Zma10_22176584_120	chr10	22126126	22126245	-	290
Zma2_24969704_120	chr2	25045695	25045814	+	310
Zma2_24987682_120	chr2	25045695	25045814	+	310
Zma2_24994237_120	chr2	25045694	25045813	+	309
Zma4_241922640_140	chr4	236310623	236310762	-	158
Zma1_102008498_160	chr1	103171107	103171266	+	303
Zma2_204027200_220	chr2	206954966	206955185	+	401
Zma4_241215956_160	chr4	235608739	235608898	-	343
Zma1_130204986_120	chr1	131335570	131335689	+	365
Zma9_50232939_120	chr9	50159779	50159898	+	312
Zma1_198071480_240	chr1	198218523	198218672	-	161
Zma1_233561788_120	chr1	234219221	234219340	+	352
Zma4_218793973_120	chr4	213174172	213174291	-	501
Zma1_240283455_140	chr1	240998447	240998586	+	301
Zma1_289217222_120	chr1	290272584	290272703	-	108
Zma1_289260479_120	chr1	290315841	290315960	-	139
Zma1_33794210_120	chr1	33722002	33722121	+	111
Zma1_4589306_240	chr1	4616456	4616669	-	269
Zma2_162544499_140	chr2	165389462	165389601	-	170
Zma6_94909262_120	chr6	94719365	94719484	-	330
Zma6_117789986_260	chr6	117659072	117659331	+	228
Zma2_9469755_160	chr2	9515506	9515665	+	82
Zma3_120369444_140	chr3	124186046	124186185	+	296
Zma3_20056264_120	chr3	20308986	20309105	+	501
Zma4_12883831_160	chr4	12944187	12944346	+	438
Zma5_208342376_140	chr5	209148061	209148200	-	122
Zma5_44465677_120	chr5	45128068	45128187	-	409
Zma6_135647868_160	chr6	135465761	135465920	+	431
Zma7_107563134_120	chr7	112983047	112983166	+	424
Zma7_151546650_140	chr7	157204603	157204742	-	499
Zma7_21627125_160	chr7	21639529	21639688	-	118
Zma8_160460280_260	chr8	161763754	161764013	-	511
Zma8_43034601_180	chr8	43032690	43032830	+	39

Table 3.2. Rice pre-miRRNA extended list by RNAseq mapping

miRNA genes	chr number	pre_miRNA start	pre_miRNA end	strand	extended_len
osa-MIR1860	Chr1	933420	933567	-	336
osa-MIR159b	Chr1	1214030	1214217	+	58
osa-MIR5521	Chr1	2333077	2333260	-	139
osa-MIR156c	Chr1	4664975	4665123	+	184
osa-MIR2096	Chr1	6442388	6442624	+	389
osa-MIR319b	Chr1	6676570	6676766	-	237
osa-MIR159f	Chr1	6692112	6692299	+	133
osa-MIR408	Chr1	12300635	12300847	+	392
osa-MIR159a	Chr1	17680877	17681148	+	442
osa-MIR1436	Chr1	20315541	20315701	-	184
osa-MIR1428b	Chr1	25389302	25389425	+	43
osa-MIR2925	Chr1	25800014	25800133	+	467
osa-MIR2862	Chr1	26802912	26803132	-	437
osa-MIR319a	Chr1	26822238	26822428	-	369
osa-MIR1846d	Chr1	39636022	39636137	+	220
osa-MIR172b	Chr1	42922692	42922929	-	299
osa-MIR806a	Chr1	42926954	42927205	-	42
osa-MIR397b	Chr2	3280779	3280896	-	140
osa-MIR818b	Chr2	4007187	4007299	+	33
osa-MIR2863b	Chr2	4195662	4195762	-	411
osa-MIR156d	Chr2	4512881	4513009	-	53
osa-MIR399i	Chr2	7650684	7650799	+	341
osa-MIR399g	Chr2	7675284	7675402	+	47
osa-MIR1884a	Chr2	10873351	10873564	+	480
osa-MIR437	Chr2	17044466	17044678	-	93
osa-MIR162a	Chr2	23599930	23600100	+	421
osa-MIR827a	Chr2	23895299	23895415	-	198
osa-MIR166d	Chr2	26124190	26124314	+	181
osa-MIR394	Chr2	27142285	27142394	+	520
osa-MIR168b	Chr2	27330559	27330664	+	378
osa-MIR2863c	Chr2	29863624	29863737	+	192
osa-MIR169e	Chr2	31199997	31200128	-	207
osa-MIR819c	Chr2	33750674	33750827	+	157
osa-MIR396f	Chr2	35630677	35630852	-	61
osa-MIR166e	Chr3	764722	764858	+	52
osa-MIR171e	Chr3	1969483	1969601	+	448
osa-MIR808	Chr3	8866921	8867072	+	279
osa-MIR164e	Chr3	10541073	10541204	+	31

osa-MIR435	Chr3	18208232	18208363	+	358
osa-MIR171i	Chr3	21190113	21190232	+	335
osa-MIR1428e	Chr3	23075358	23075467	+	78
osa-MIR1428d	Chr3	23075756	23075879	+	476
osa-MIR166i	Chr3	25293034	25293178	+	89
osa-MIR1867	Chr3	30507952	30508075	+	475
osa-MIR167b	Chr3	30539817	30539979	-	142
osa-MIR5497	Chr3	32498885	32499022	+	356
osa-MIR167c	Chr3	33123489	33123651	+	481
osa-MIR5513	Chr3	34000873	34001018	+	438
osa-MIR806c	Chr3	36187171	36187440	-	63
osa-MIR5511	Chr4	14643725	14643852	-	75
osa-MIR5499	Chr4	17069758	17069882	+	121
osa-MIR5515	Chr4	17072179	17072302	+	134
osa-MIR416	Chr4	17284208	17284316	+	116
osa-MIR1423	Chr4	19543152	19543287	+	439
osa-MIR2118a	Chr4	21470973	21471151	-	54
osa-MIR2118b	Chr4	21473335	21473513	-	179
osa-MIR2118e	Chr4	21475478	21475654	-	136
osa-MIR2118d	Chr4	21475739	21475914	-	453
osa-MIR2118f	Chr4	21478564	21478732	-	314
osa-MIR2118j	Chr4	21482644	21482814	-	483
osa-MIR162b	Chr4	24220760	24220893	+	479
osa-MIR156e	Chr4	24841188	24841291	-	79
osa-MIR5501	Chr4	28652731	28653096	+	419
osa-MIR399j	Chr4	28687978	28688083	+	50
osa-MIR442	Chr4	32183563	32183795	+	501
osa-MIR396d	Chr4	34251706	34251847	+	153
osa-MIR393b	Chr4	34746766	34746897	-	378
osa-MIR819f	Chr4	35104592	35104735	-	188
osa-MIR1425	Chr5	8862077	8862178	+	180
osa-MIR164b	Chr5	15838704	15838812	-	71
osa-MIR1850	Chr5	26212571	26212703	-	370
osa-MIR399k	Chr5	26248616	26248721	+	130
osa-MIR5502	Chr5	28917291	28917432	-	22
osa-MIR399d	Chr6	1560790	1561075	+	168
osa-MIR819h	Chr6	10052973	10053127	-	136
osa-MIR811a	Chr6	13901553	13901742	+	125
osa-MIR399f	Chr6	20887634	20887750	-	24
osa-MIR5517	Chr6	24617624	24617726	-	319
osa-MIR156h	Chr6	26553831	26553931	+	314
osa-MIR1861h	Chr6	27237193	27237296	-	30
osa-MIR169b	Chr6	27294911	27295038	-	515
osa-MIR160b	Chr6	28080022	28080160	+	239

osa-MIR166b	Chr6	30326086	30326291	-	308
osa-MIR160e	Chr7	2215762	2215875	-	372
osa-MIR167d	Chr7	4165296	4165405	-	230
osa-MIR5518	Chr7	9525310	9525472	+	53
osa-MIR809d	Chr7	14712001	14712167	+	27
osa-MIR1424	Chr7	26661361	26661660	-	350
osa-MIR819i	Chr7	27792077	27792274	+	177
osa-MIR164a	Chr7	28522349	28522504	-	177
osa-MIR1861j	Chr8	15130013	15130122	-	27
osa-MIR812j	Chr8	19803620	19803916	+	117
osa-MIR169i	Chr8	26800402	26800577	+	110
osa-MIR169h	Chr8	26804635	26804756	+	130
osa-MIR806f	Chr8	27962767	27963007	+	342
osa-MIR5491	Chr9	6895342	6895458	+	150
osa-MIR156g	Chr9	15064252	15064367	+	148
osa-MIR166l	Chr9	16950816	16950932	+	408
osa-MIR1875	Chr9	17566946	17567197	-	331
osa-MIR169j	Chr9	19788380	19788504	+	150
osa-MIR1846a	Chr10	528250	528361	+	123
osa-MIR398a	Chr10	9145071	9145185	-	295
osa-MIR167f	Chr10	14651810	14651922	-	52
osa-MIR166a	Chr10	19915634	19915778	+	141
osa-MIR171d	Chr10	21165791	21165925	+	250
osa-MIR806g	Chr10	22846562	22846801	+	135
osa-MIR2872	Chr11	2520898	2521257	+	197
osa-MIR2118p	Chr11	7803178	7803351	-	381
osa-MIR2118q	Chr11	7806454	7806630	-	136
osa-MIR440	Chr11	9159384	9159503	-	239
osa-MIR1880	Chr11	13496198	13496326	+	54
osa-MIR1846b	Chr11	27244208	27244319	-	119
osa-MIR419	Chr12	8234925	8235027	-	303
osa-MIR413	Chr12	10133792	10133904	+	381
osa-MIR5520	Chr12	10267043	10267373	-	39
osa-MIR5074	Chr12	12671855	12671990	+	82
osa-MIR5505	Chr12	13386344	13386551	+	302
osa-MIR5506	Chr12	15162073	15162183	-	197
osa-MIR5488	Chr12	18010567	18010779	+	130
osa-MIR166g	Chr12	18300803	18300947	+	76
osa-MIR1441	Chr12	19457211	19457390	-	119
osa-MIR167a	Chr12	25443203	25443343	+	83
osa-MIR167h	Chr12	25447013	25447132	+	312

3.7 References

(2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408,** 796-815.

Axtell, M.J., and Bowman, J.L. (2008). Evolution of plant microRNAs and their targets. In Trends Plant Sci , pp. 343-349.

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

Boualem, A., Laporte, P., Jovanovic, M., Laffont, C., Plet, J., Combier, J.P., Niebel, A., Crespi, M., and Frugier, F. (2008). MicroRNA166 controls root and nodule development in Medicago truncatula. In Plant J , pp. 876-887.

Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A brief review on the mechanisms of miRNA regulation. In Genomics Proteomics Bioinformatics (China: 2009 Beijing Genomics Institute. Published by Elsevier Ltd), pp. 147-154.

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

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J., and Grimmond, S.M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. In Nat Methods , pp. 613-619.

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol **2010,** 853916.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. In BMC Bioinformatics , pp. 485.

Cui, X., Xu, S.M., Mu, D.S., and Yang, Z.M. (2009). Genomic analysis of rice microRNA promoters and clusters. In Gene (Netherlands), pp. 61-66.

Ding, D., Zhang, L., Wang, H., Liu, Z., Zhang, Z., and Zheng, Y. (2009). Differential expression of miRNAs in response to salt stress in maize roots. In Ann Bot , pp. 29-38.

Fahlgren, N., Jogdeo, S., Kasschau, K.D., Sullivan, C.M., Chapman, E.J., Laubinger, S., Smith, L.M., Dasenko, M., Givan, S.A., Weigel, D., and Carrington, J.C. (2010). MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. In Plant Cell , pp. 1074-1089.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. In Nat Methods , pp. 469-477.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C.,

Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). In Science , pp. 92-100.

Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. In Bioinformatics , pp. 3643-3646.

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

Joung, J.G., and Fei, Z. (2009). Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model. In Bioinformatics , pp. 387-393.

Kurihara, Y., Takashi, Y., and Watanabe, Y. (2006). The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. In RNA , pp. 206-212.

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. In EMBO J , pp. 4051-4060.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. In Bioinformatics , pp. 1966-1967.

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. In Cell , pp. 523-536.

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

Meng, Y., Shao, C., Gou, L., Jin, Y., and Chen, M. (2011). Construction of microRNA- and microRNA*-mediated regulatory networks in plants. In RNA Biol , pp. 1124-1148.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. In Nat Methods , pp. 621-628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. In Science , pp. 1344-1349.

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

Park, M.Y., Wu, G., Gonzalez-Sulser, A., Vaucheret, H., and Poethig, R.S. (2005). Nuclear processing and export of microRNAs in Arabidopsis. In Proc Natl Acad Sci U S A , pp. 3691-3696.

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

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. In Science , pp. 1112-1115.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. In Science , pp. 956-960.

Sunkar, R., and Zhu, J.K. (2004). Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. In Plant Cell , pp. 2001-2019.

Tang, G. (2010). Plant microRNAs: an insight into their gene structures and evolution. In Semin Cell Dev Biol, pp. 782-789.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. In Cell , pp. 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. In Plant Cell , pp. 2204-2216.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. In Nat Rev Genet , pp. 57-63.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005a). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. In Nature , pp. 338-345.

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

Yang, Z., Ebright, Y.W., Yu, B., and Chen, X. (2006). HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. In Nucleic Acids Res , pp. 667-675.

Zhang, B., Pan, X., Wang, Q., Cobb, G.P., and Anderson, T.A. (2006). Computational identification of microRNAs and their targets. In Comput Biol Chem , pp. 395-407.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. In Genome Res , pp. 646-654.

Zhang, L., Chia, J.M., Kumari, S., Stein, J.C., Liu, Z., Narechania, A., Maher, C.A., Guill, K., McMullen, M.D., and Ware, D. (2009). A genome-wide characterization of microRNA genes in maize. PLoS Genet **5,** e1000716.

Zhang, Y., Stupka, E., Henkel, C.V., Jansen, H.J., Spaink, H.P., and Verbeek, F.J. (2011). Identification of common carp innate immune genes with whole-genome sequencing and RNA-Seq data. In J Integr Bioinform (Germany), pp. 169.

Zhou, M., Sun, J., Wang, Q.H., Song, L.Q., Zhao, G., Wang, H.Z., Yang, H.X., and Li, X. (2011). Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA genes. FEBS J **278,** 929-940.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. In PLoS Comput Biol , pp. e37.

Chapter 4

Comparative Analysis of Regulatory Elements of microRNAs in Cereal Genomes

4.1 Abstract

Studies concerning microRNA transcription regulation in plants have mostly focused on locating the miRNA gene promoters. Little is known about the regulatory elements that control miRNA gene expression. Identification of these elements helps us to have a better understanding of the role of miRNAs in the regulatory networks. Here we have studied the existence and abundance of putative miRNA specific regulatory motifs that are conserved in cereal genomes. We initially found the orthologous miRNA genes among maize, sorghum and rice using both sequence homology and synteny data. Next we searched for motifs occurrence in the area 1000 bp upstream of the maize, sorghum and rice orthologous miRNA precursors. We showed that generally when the copy number of motifs is high in maize (>10), it is present in the upstream region of sorghum or rice miRNA genes while low copy number motifs are mostly species specific. We demonstrate that the most significant motif found by MEME, is also the most conserved one with highest copy number in the three species miRNA upstream region. This motif matches significantly to E2F-variant in AGRIS motif database. E2F-variant motifs have been reported before to play a major role in animal miRNAs and also their corresponding transcription factor is regulated by miRNAs. To conclude, our results show that some

miRNA regulatory elements are conserved in different cereal genomes and might have a critical role in miRNA regulatory networks. Also these data provides several good motif candidates for experimental verification.

4.2 Introduction

MicroRNAs (miRNAs) are endogenous, noncoding 20-24 nt RNA molecules that act as regulators of gene expression. They are generated from one arm of hairpin precursors that are derived from pri-miRNA transcripts produced by RNA polymerase II (Lee et al., 2004).

MRNAs are play a major role in regulation of the expression of protein-coding genes (Bartel, 2004; Lim et al., 2005) and are thus important for development, stress response and other biological processes in plants (Shukla et al., 2008; Voinnet, 2009). One miRNA can repress several different mRNAs (Lim et al., 2005) and sometimes they are compared with transcription factors in terms of the effect that they can impose on gene regulation (Hobert, 2004). They operate by binding to mRNA sequences with base pairing and cleave the target mRNAs or repressing the translation of mRNAs or attaching to DNA (Voinnet, 2009). For example, Li et al. found 160 experimentally confirmed target genes that are cleaved by 53 rice miRNA families using the degradome sequencing approach (Li et al., 2010).

The expression regulation of miRNA genes is not fully understood and little is known about the possible regulatory elements in the upstream region of these genes. It has been shown that most miRNA genes are transcribed from their own promoters (Bartel, 2004). Some miRNAs are also organized into clusters and transcribed as multicistronic units (Lau et al., 2001). One major approach to study the miRNA transcription is to inspect the upstream regions of these genes to find regulatory motifs

(Ohler et al., 2004; Zhou et al., 2007). Identification and analyzing the promoters of different miRNA genes could provide clues to understanding regulatory networks.

A few studies have used the alignment of the miRNA upstream sequences of orthologous genes to identify miRNA-specific regulatory elements (Inouchi et al., 2007; Zhou et al., 2007; Heikkinen et al., 2008). For example, the alignment of the upstream regions of the orthologous miRNAs in two nematode worms revealed a significant conserved regulatory motif around 500bp upstream of the miRNA hairpin start site (Heikkinen et al., 2008). This motif was conserved in phylogenetically distant species including human and mouse. In another study, the authors compared the upstream sequences of 242 human and 290 mouse miRNAs and found some significantly conserved motifs that are good candidates for experimentally testing of miRNA expression as well as possible interaction with regulatory factors (Inouchi et al., 2007). Several attempts have been made to study the regulatory motifs of miRNA genes in plants. A recent study investigated the flanking regions of four plant miRNA genes and several conserved and non-conserved motif elements were discovered (Zhou et al., 2011). They characterized the core promoters and regulatory elements of miRNA genes in four species and found that most miRNAs have the same type of promoters as protein-coding genes (Zhou et al., 2007). In another study, Cui et al. detected the promoters of 212 rice miRNAs and studied their clustering patterns (Cui et al., 2009).

Several motif-finding programs are available for transcriptional regulatory element analysis. These programs employ different algorithms such as exhaustive search and bootstrapping (Kankainen and Holm, 2005), exhaustive enumeration (Pavesi et al.,

2001), probabilistic local alignment based method (Bailey et al., 2006) and Gibbs sampler (Lawrence et al., 1993). One approach to reduce the false positive rate of motif finding is to focus on those motifs that are conserved in several species. Some comparative studies have been performed to identify the phylogenetically conserved sequence motifs in animals (Ohler et al., 2004; Inouchi et al., 2007; Heikkinen et al., 2008) but little is known in plants (Zhou et al., 2007).

Here we set to study the miRNA-specific regulatory elements and their conservation in three cereal genomes (maize, sorghum and rice). We find that some motifs are frequently found in the miRNA upstream regions in different species while others are either species-or miRNA-specific.

4.3 Results

Clusters of microRNAs in cereal genomes

Vast majority of microRNA genes are located in intergenic regions in plant genomes (Cui et al., 2009; Zhou et al., 2011). Whereas the majority of microRNAs exist as single copy on the chromosomes, some miRNAs are close to each other to form miRNA clusters, presumably originated through tandem duplications (Zhou et al., 2011). MiRNAs from a single cluster are usually transcribed together and their expression is regulated by the same regulatory elements in the upstream region of the first miRNA (Lau et al., 2001). In order to correctly identify miRNA regulatory motifs, it is necessary to first find miRNA clusters in the genomes. As described in methods, for miRNA clustering, both upstream and downstream pre-miRNAs with pair wise distance less than

1 kb were considered as clustered miRNAs. When the clustered were organized, we only used the 5'-end sequence of the first miRNAs in the upstream region of the cluster for motif analysis. Based on this strategy, we removed 41, 14 & 28 of maize, sorghum and rice miRNAs (those that were organized in downstream regions of the clusters) from the miRNA lists. Only the upstream regions of the first miRNA in each cluster were used for discovery of regulatory motifs.

Finding the orthologous miRNA precursors in maize, sorghum and rice

We first found orthologous miRNA gene pairs between maize and sorghum using sequence homology and synteny data. We started with 543, 373, and 447 miRNA in rice, maize, and sorghum, respectively, that were either deposited in miRBase or identified in our lab after redundancy has been removed. We found 68 orthologous miRNA pairs between maize and sorghum, and 74 orthologous miRNA pairs between maize and rice.

Motif analysis and conservation study in maize, sorghum and rice

The overall process flowchart of our motif discovery method is displayed in Figure 4.1. We used MEME program (Bailey et al., 2009) to find a set of 20 most overrepresented motifs in maize miRNA upstream region which have orthologs in sorghum. Then we searched for these motifs in the upstream region of the otrhologous sorghum miRNA. Next we used rice orthologous miRNAs as the out-group for motif conservation study. The reason that we chose maize among these three species to start our work is first we have the genome data of a closely related species to it which is

sorghum, and this helps us to do the motif comparative analysis more precisely and second we were able to extend 84 maize miRNA upstream region in last chapter and can evaluate the effect of that in motif discovery. We ran the motif enrichment analysis test (adj.P.Val<0.01) and found that eleven of the motifs are not miRNA specific and are found frequently in upstream region of protein-coding genes (we searched the upstream region of 1000 random prompter of maize protein coding genes). These include a motif that matches the T-box, which is a common regulatory element in protein coding genes, and participate in cell development. Excluding these non-significant motifs, we searched for the existence and copy number of the other motifs in sorghum and rice. Significance and sequence of these motifs have been summarized in Table 4.1.

First we counted the overall copy number of the significant motifs in maize, sorghum and rice upstream regions of the orthologous miRNA list to study their conservation (Figure 4.2). Interestingly, the motif copy numbers in not similar even in maize and sorghum, which are evolutionarily close. There is only one motif with a high copy number (> 50) in all the three species (motifs 1), which is the most significant one, as well. There is another motif (motif_d), which was present in rice and maize but not sorghum. So we can conclude that this motif (or the related miRNA ortholog) has been removed from sorghum genome. Also 5 of the motifs are maize specific and haven't been detected in the other two species therefore they are good candidates to study how miRNA get specialized in a particular organism.

Second, we studied which percentages of the total conserved miRNAs in the three plant species possess the significant motifs (Figure 4.3). As it is shown in this figure,

there are some motifs which present in the upstream region of the majority of miRNAs while some others are present in only few ones most of which belong to the same family. We also see that some of motifs are limited to specific miRNA families in different species.

At the end, based on these figures we decided to analyze some of the motifs with particular characteristics in more details.

4.4 Discussion

Characteristics of the interesting MEME motifs in the three species

-Motif_a (the most significant MEME motif).

The most interesting and significant MEME motif found, motif_a, is also the most conserved one with highest copy number in the three species miRNA upstream region. The logo of this motif is presented in Figure 4.4. It is present in the majority of miRNAs in sorghum and rice with a very high copy number (>350). Interestingly, in most of the cases this motif has either a very high or very low copy number in different miRNA genes in the three species. This explains how different miRNAs family members can get specialized independently (Figure 4.4). This motif is not limited to one or few miRNA families and exists in several of them. It matches well to the E2F-variant, which is known to participate in cell cycle, transcription, stress and defense or signaling (Ramirez-Parra et al., 2003). Entry into the S phase of the cell cycle is controlled by E2F transcription factors that induce the transcription of genes required for cell cycle progression and DNA replication. Although the E2F pathway is highly conserved in higher eukaryotes, only a

few E2F target genes have been experimentally validated in plants. They found more 200 E2F target genes in Arabidopsis and rice, which were expressed almost exclusively during G1 and S phases (Vandepoele et al., 2005). In mammals it has been shown that E2F-induced miRNAs play some role in response to mitogenic signaling (Bueno et al., 2010). It has also been shown that E2F activity can be regulated by micro-RNAs (miRNAs) and in turn, miRNAs themselves are targets of E2F family proteins establishing negative feedback loops (Emmrich and Putzer, 2010). Just recently it has been revealed that there is crosstalk between E2Fs and microRNAs participate in the regulation hypoxia response in animals (Biyashev and Qin, 2011). Similar crosstalk could probably exist in plants. This motif also has a significant match to and Sorlips, "Sequences Over-Represented in Light-Induced Promoters". They have been reported to be present in promoters that are induced by phytochrome A in Arabidopsis (Hudson and Quail, 2003).

-Motif_b (absent in rice)

This motif is not present in rice but is frequent in maize and sorghum miRNA genes upstream region (Figure 4.5). It has a good match to RY-repeat, which is conserved in the promoters of seed-specific genes in both monocot and dicot species and has been reported before in the miRNA core promoters of four model species including rice and Arabidopsis (Zhou et al., 2007).

-Motf_d (absent in sorghum)

This motif is the only case that is absent in all sorghum miRNA upstreams but present in several rice and maize miRNA upstream sequences. It is also the only motif that showed up exclusively in the maize refined miRNA upstream regions. It matches significantly to CArG3 (Figure 4.6). It has been shown that CArG3 is the binding site for a negatively acting factor during early floral stages in Arabidopsis (Tilly et al., 1998).

-Motif_f (maize specific)

This motif (along with 4 more motifs) is only present in maize miRNA upstream region. One characteristic of all of them is that they have a low copy number and present in just few miRNAs (Figure 4.7). It is significantly similar to OBF4_5binding site, which has been reported to participate in flowering regulation (Song et al., 2008).

Study the changes in maize miRNAs motif content after refining the gene boundaries using RNAseq

We had tried to refine the maize miRNA gene boundaries and getting closer to transcript Start Site by RNAseq mapping before (See chapter 2). Here we compared the significant motif list before and after refining. Using the improved gene boundaries coordinates, in most of the cases only changed the significance level of the identified motifs and also the significance level of their matches to the known motifs. However in one case, one motif (motif_d which matches significantly to CArG3 which is a known motif in AGRIS database) was identified in the refined miRNA upstream set which didn't have any counterpart in the non-refined upstream set (we didn't see any opposite pattern).

This shows the importance of refining the miRNA boundaries as much as we can to get a more accurate list of significant motifs.

In conclusion, we have identified over-represented and conserved motifs in the upstream regions of three monocots miRNA stem-loop sequences. By applying the conservation analysis on the identified motifs, we found several motifs that most probably are true positive and are good candidate for experimental verification of possible regulatory functions. Also by studying the existence of maize miRNA motifs in the closely related genomes, we report five explicit maize-specific miRNA regulatory elements.

4.5 Materials and Methods

Sequence data

Genome sequences and annotation data were downloaded from the following websites: for rice, the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu, version 6.1) (Goff et al., 2002); for maize, http://www.maizesequence.org (version 5a) (Schnable et al., 2009), and for *Sorghum Bicolor*, www.phytozome.net (Paterson et al., 2009).

Annotated miRNAs in three cereals

Rice miRNAs were downloaded from the miRBase (version 18). Maize miRNAs included the miRNAs from the miRBase (version 18) and unpublished miRNAs that were

identified in our lab after removing redundancy. *Sorghum* miRNAs were identified in our lab.

MiRNA precursor clustering and extraction of their upstream region

To identify miRNA clusters in the studied genomes, we defined 1 kb as the maximum inter-miRNA distance for two miRNA genes to be considered as clustered. Then we only extracted the 1 kb upstream regions of the miRNAs that were located at the 5' end of the clusters and other single miRNAs. We also removed those miRNAs which are overlapping with protein-coding genes or too close to the upstream protein-coding genes (<1 kb if both are on the same strand and <2 kb if they are on opposite strands).

Identification of the maize orthologous miRNAs in sorghum and rice

Maize collinear regions in sorghum and rice genomes were identified using DAGchainer (Haas 2004), requiring at least five anchor genes with no more than ten intervening genes between neighboring anchor genes. MiRNA precursors that were flanked by the same anchor genes and had the lowest alignment e-value using blastall were considered orthologous miRNAs.

Motif extraction and analysis

The summary of our methodology has been demonstrated in Figure 4.1. After finding miRNA orthologs, we searched for the motifs occurring in the area 1000 bp upstream from maize miRNA using MEME (used parameters: -mod anr -minw 5 -maxw

15 –nmotifs 20). To test the statistical significance of the found motifs, we used the promoter region of 1000 randomly chosen maize protein coding genes as the background. We ran motif enrichment test using Bioconductor Biostrings library to find enrichment/depletion of motifs in our sample sequence set relative to the background set. To do so, we started with PFM (positional frequency matrix) in MEME output, convert it to PWM (positional weight matrix) using a Biostrings function. We chose these parameters for enrichment test: revcomp=TRUE, cutoff=0.9, occurrence=1. The cutoff (minimum score of matching) is the percentage of the maximum possible score, here.

Next we mapped the significant maize miRNA motifs to the 1000 bp upstream region of sorghum miRNA precursor which are orthologs of the corresponding miRNAs in maize using FIMO in MEME Suite package (q-value<0.01). Then we did the same thing in rice. The idea here is to find whether the enriched miRNA motifs in maize are conserved in sorghum and rice or not. We calculated the frequency of each motif in the orthologous miRNAs in maize, sorghum and rice. In the last step we mapped the identified motifs against AGRIS database which contains *Arabidopsis* known gene regulator elements using STAMP (Mahony and Benos, 2007).

4.6 Figures & Tables

Figure Legends

Figure 4.1. The workflow of the discovery and study of conserved miRNA motifs in maize, sorghum and rice.

Figure 4.2. Comparison of the total number of identified miRNA motifs in the three genomes.

Figure 4.3. Comparison of the number of conserved miRNA which have the specific motifs

Figure 4.4. Motif_a frequency in orthologous miRNAs in the three cereals and its logo.

Figure 4.5. Motif_b frequency in orthologous miRNAs in the three cereals.

Figure 4.6. Motif_d frequency in orthologous miRNAs in the three cereals.

Figure 4.7. Motif_f frequency in orthologous miRNAs in the three cereals.

Figure 4.1.

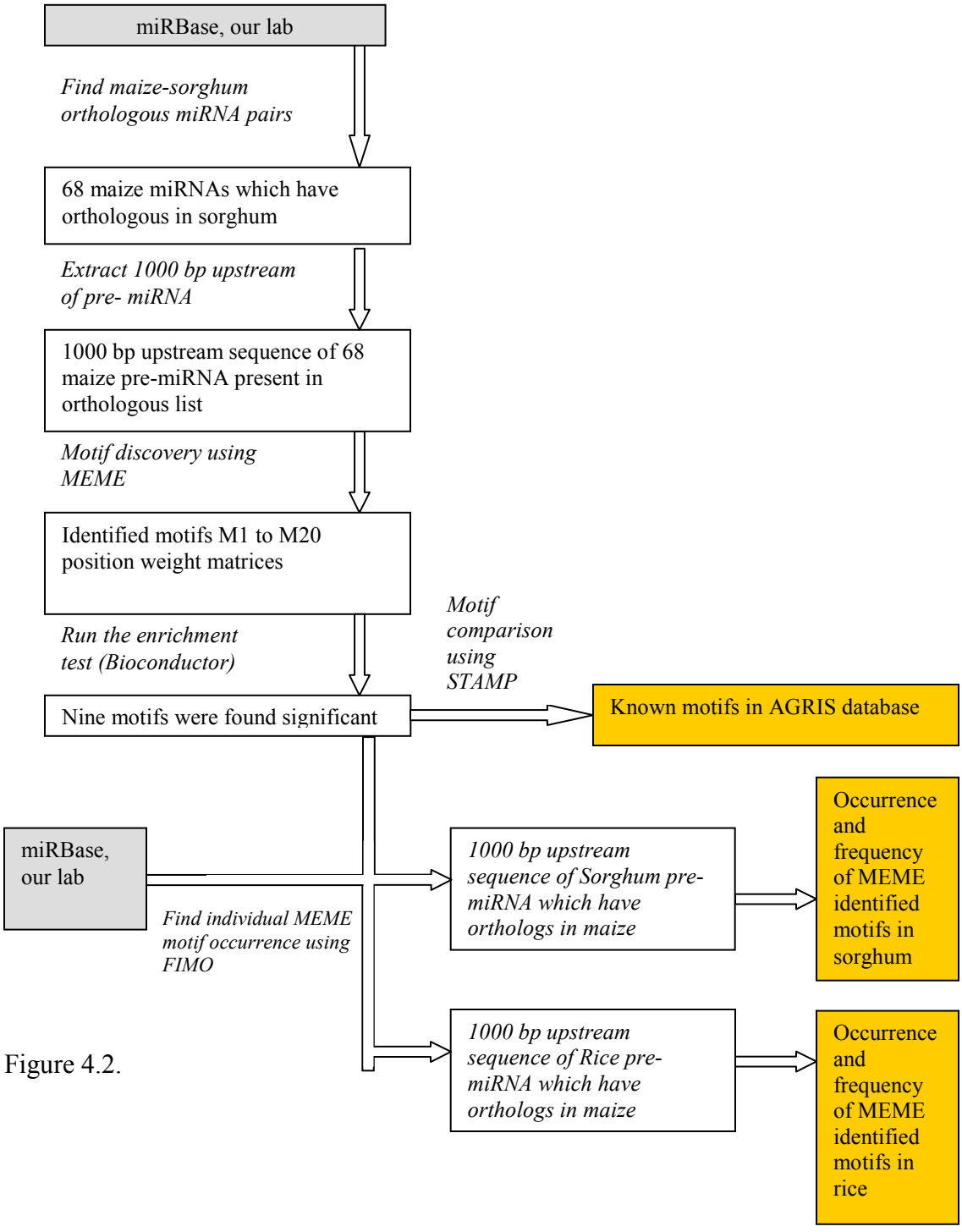

miRBase, our lab
Find maize-sorghum orthologous miRNA pairs
68 maize miRNAs which have orthologous in sorghum
Extract 1000 bp upstream of pre- miRNA
1000 bp upstream sequence of 68 maize pre-miRNA present in orthologous list
Motif discovery using MEME
Identified motifs M1 to M20 position weight matrices
Run the enrichment test (Bioconductor)
Motif comparison using STAMP
Nine motifs were found significant
Known motifs in AGRIS database
miRBase, our lab
Find individual MEME motif occurrence using FIMO
1000 bp upstream sequence of Sorghum pre-miRNA which have orthologs in maize
Occurrence and frequency of MEME identified motifs in sorghum
1000 bp upstream sequence of Rice pre-miRNA which have orthologs in maize
Occurrence and frequency of MEME identified motifs in rice

Figure 4.2.

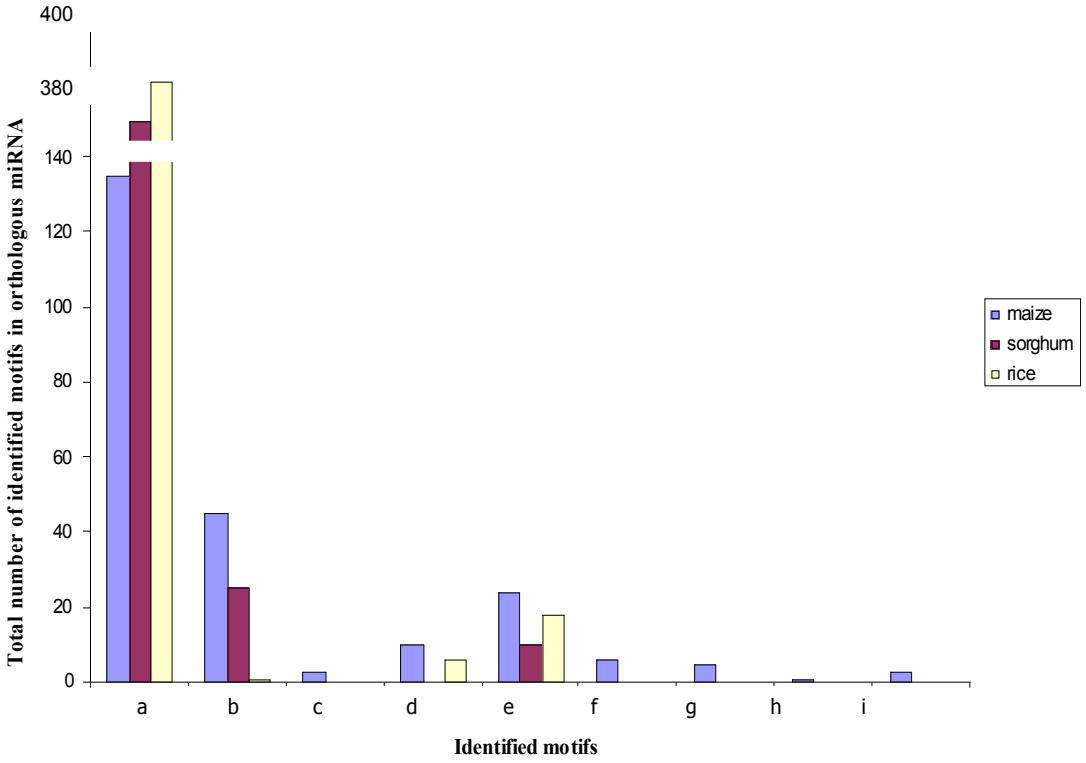

400
380
140
120
100
80
60
40
20
0
Total number of identified motifs in orthologous miRNA
a
b
c
d
e
f
g
h
i
Identified motifs
maize
sorghum
rice

Figure 4.3.

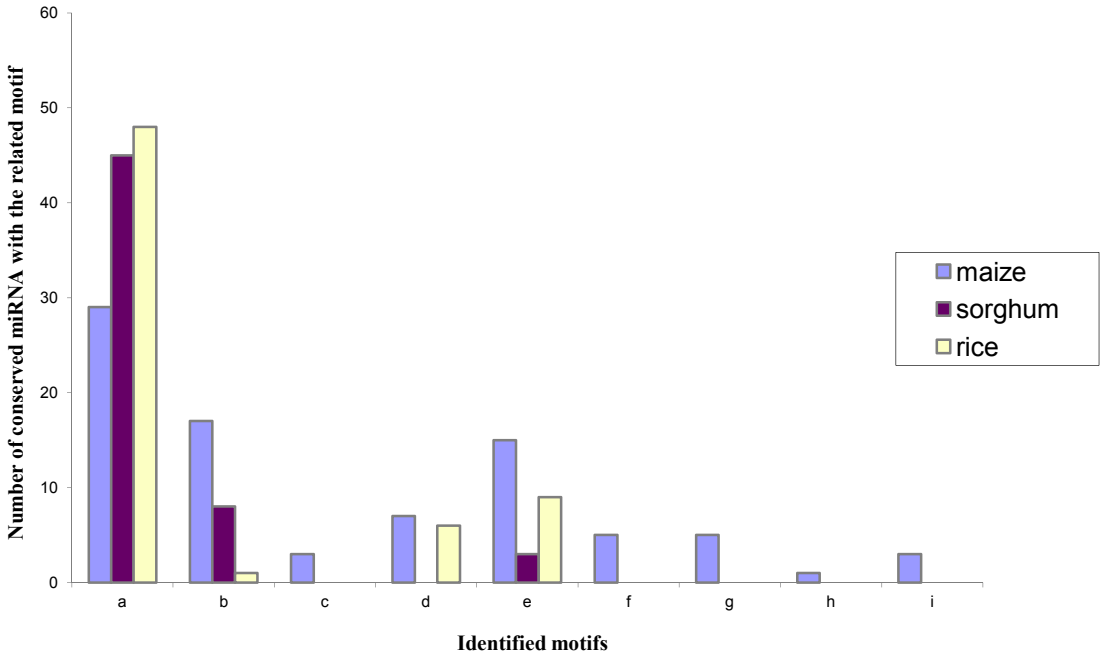
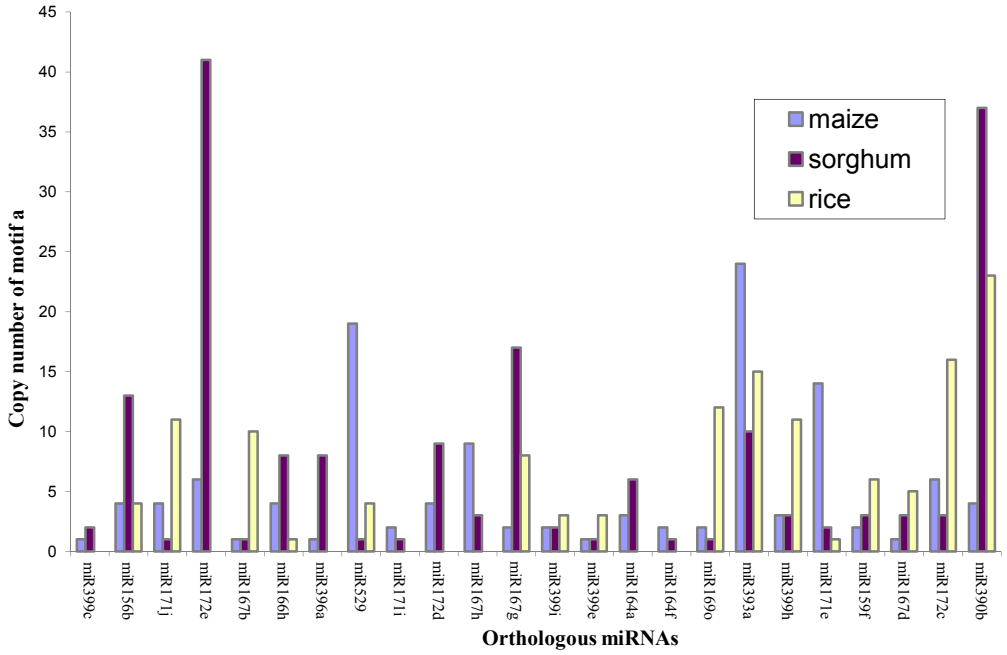

Number of conserved miRNA with the related motif
60
50
40
30
20
10
0
a
b
c
d
e
f
g
h
i
Identified motifs
maize
sorghum
rice

Figure 4.4.


Copy number of motif a
45
40
35
30
25
20
15
10
5
0
maize
sorghum
rice
miR399c
miR156b
miR171j
miR172e
miR167b
miR166h
miR396a
miR529
miR171i
miR172d
miR167h
miR167g
miR399i
miR399e
miR164a
miR164f
miR169o
miR393a
miR399h
miR171e
miR159f
miR167d
miR172c
miR390b
Orthologous miRNAs

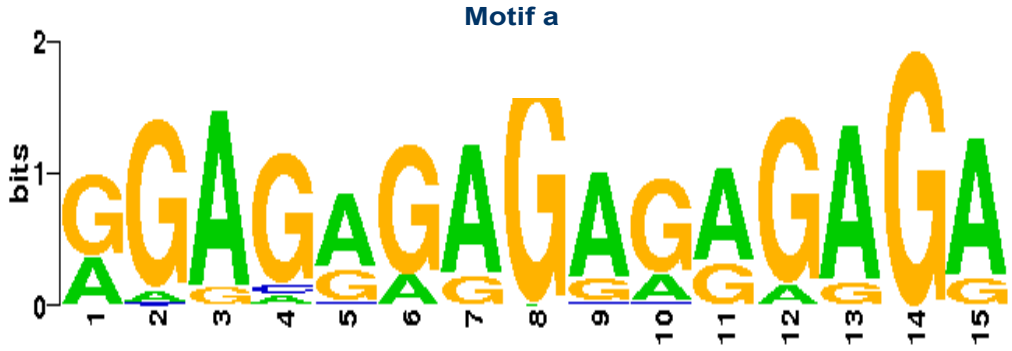

Motif a
bits
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 4.5.

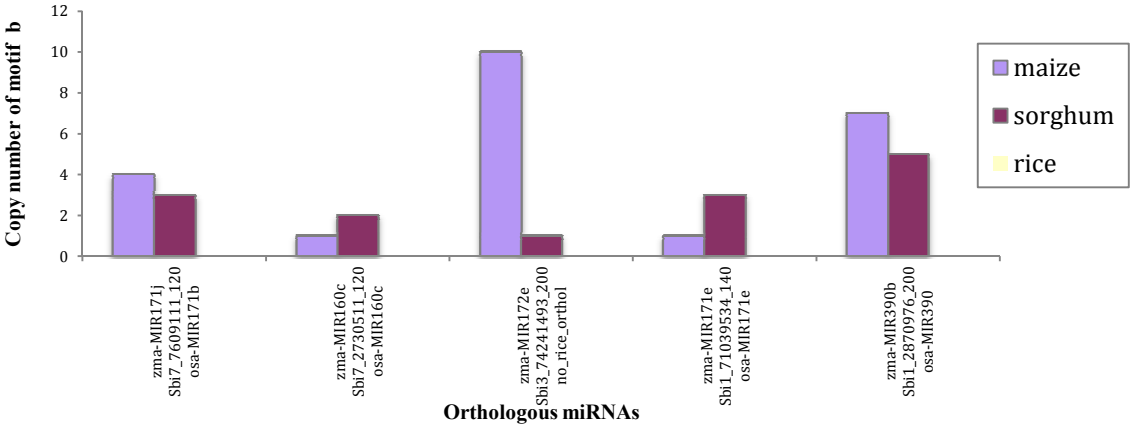

Copy number of motif b
12
10
8
6
4
2
0
maize
sorghum
rice
zma-MIR171j
Sbi7_7609111_120
osa-MIR171b
zma-MIR160c
Sbi7_2730511_120
osa-MIR160c
zma-MIR172e
Sbi3_74241493_200
no_rice_orthol
zma-MIR171e
Sbi1_71039534_140
osa-MIR171e
zma-MIR390b
Sbi1_2870976_200
osa-MIR390
Orthologous miRNAs

Figure 4.6.

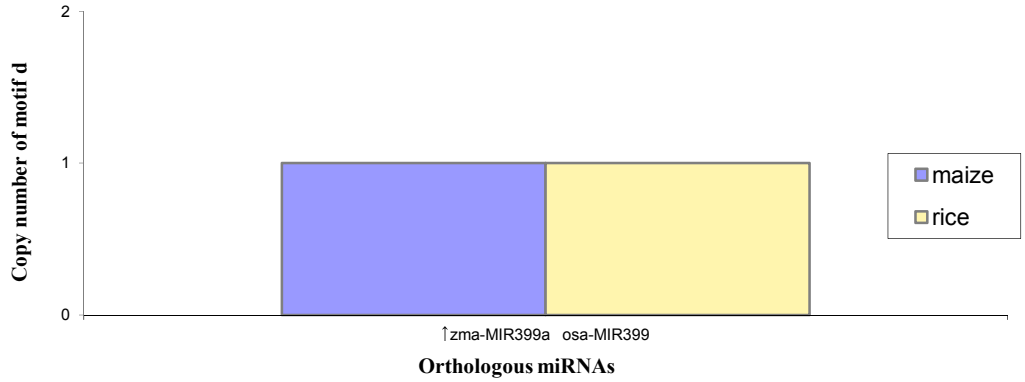

2
1
0
Copy number of motif d
maize
rice
zma-MIR399a
osa-MIR399
Orthologous miRNAs

Figure 4.7.

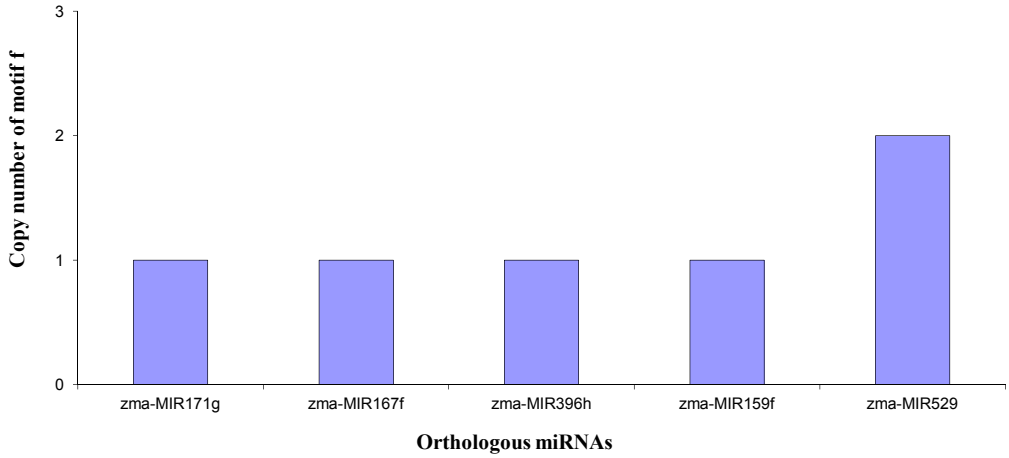

Copy number of motif f
3
2
1
0
zma-MIR171g
zma-MIR167f
zma-MIR396h
zma-MIR159f
zma-MIR529
Orthologous miRNAs

Table 4.1. The identified sequence motifs in the upstream regions of maize miRNA stem-loop sequences which have orthologs in sorghum and rice

Motif names	adj-P-Val (enrichment test)	Motifs regular expression	E-val of matches to known AGRIS motifs	Known motifs
a	7.40E-04	[GA]GAG[AG][GA][AG]GA[GA][AG]GAGA	9.40E-05	E2F-varian/Solis
b	2.10E-04	[CT]A[GCT]ATGCATGCATG[CGA]	1.20E-11	PY-repeat
c	1.40E-07	[AG]TG[AG][AT]A[CT]A[TA][CG][CT]ATTG	5.10E-04	Evening Element
d	1.40E-03	[TAC][AG]AA[GA]A[TA][GA]AAA[CTA][AC]A[AT]	5.80E-05	CArG3
e	0	[GC][CT][TAG][GT]CTGCTGCTG[GAC][TC]	4.30E-02	RAV1-A
f	7.10E-07	AC[TA][AC]T[GT][AC][TAC]ATATCA[TA]	1.40E-05	OBF4_5
g	0	AA[TG]A[GA]TG[AT][ATC][GC]TGA[AC]A	4.20E-04	SORLIP3
h	9.70E-07	GAGT[AT][GT]CATG[GT]GA[AT]G	6.8E-06	AtMYC2
i	3.50E-04	GC[GCT][AG][ACT][TAG]GGCAGGCAG	6.60E-03	CBF1

4.7 References

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. In Nucleic Acids Res, pp. W369-373.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. In Nucleic Acids Res, pp. W202-208.

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

Biyashev, D., and Qin, G. (2011). E2F and microRNA regulation of angiogenesis. Am J Cardiovasc Dis **1,** 110-118.

Bueno, M.J., Gomez de Cedron, M., Laresgoiti, U., Fernandez-Piqueras, J., Zubiaga, A.M., and Malumbres, M. (2010). Multiple E2F-induced microRNAs prevent replicative stress in response to mitogenic signaling. In Mol Cell Biol, pp. 2983-2995.

Cui, X., Xu, S.M., Mu, D.S., and Yang, Z.M. (2009). Genomic analysis of rice microRNA promoters and clusters. In Gene, pp. 61-66.

Emmrich, S., and Putzer, B.M. (2010). Checks and balances: E2F-microRNA crosstalk in cancer control. In Cell Cycle, pp. 2555-2567.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). In Science, pp. 92-100.

Heikkinen, L., Asikainen, S., and Wong, G. (2008). Identification of phylogenetically conserved sequence motifs in microRNA 5' flanking sites from C. elegans and C. briggsae. In BMC Mol Biol, pp. 105.

Hobert, O. (2004). Common logic of transcription factor and microRNA action. In Trends Biochem Sci , pp. 462-468.

Hudson, M.E., and Quail, P.H. (2003). Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. In Plant Physiol, pp. 1605-1616.

Inouchi, A., Shinohara, S., Inoue, H., Kita, K., and Itakura, M. (2007). Identification of specific sequence motifs in the upstream region of 242 human miRNA genes. In Comput Biol Chem, pp. 207-214.

Kankainen, M., and Holm, L. (2005). POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. In Nucleic Acids Res, pp. W427-431.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. In Science, pp. 858-862.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. In Science **262,** 208-214.

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. In EMBO J , pp. 4051-4060.

Li, Y.F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M.J., Zhang, W., and Sunkar, R. (2010). Transcriptome-wide identification of microRNA targets in rice. In Plant J , pp. 742-759.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. In Nature , pp. 769-773.

Mahony, S., and Benos, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. In Nucleic Acids Res , pp. W253-258.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. In RNA , pp. 1309-1322.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F., Messing, J., and Rokhsar, D.S. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457,** 551-556.

Pavesi, G., Mauri, G., and Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics **17 Suppl 1,** S207-214.

Ramirez-Parra, E., Frundt, C., and Gutierrez, C. (2003). A genome-wide identification of E2F-regulated genes in Arabidopsis. In Plant J , pp. 801-811.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E.,

Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. In Science , pp. 1112-1115.

Shukla, L.I., Chinnusamy, V., and Sunkar, R. (2008). The role of microRNAs and other endogenous small RNAs in plant stress responses. In Biochim Biophys Acta (Netherlands), pp. 743-748.

Song, Y.H., Song, N.Y., Shin, S.Y., Kim, H.J., Yun, D.J., Lim, C.O., Lee, S.Y., Kang, K.Y., and Hong, J.C. (2008). Isolation of CONSTANS as a TGA4/OBF4 interacting protein. In Mol Cells (Korea South), pp. 559-565.

Tilly, J.J., Allen, D.W., and Jack, T. (1998). The CArG boxes in the promoter of the Arabidopsis floral organ identity gene APETALA3 mediate diverse regulatory effects. Development **125,** 1647-1657.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inze, D., and De Veylder, L. (2005). Genome-wide identification of potential plant E2F target genes. In Plant Physiol , pp. 316-328.

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

Zhou, M., Sun, J., Wang, Q.H., Song, L.Q., Zhao, G., Wang, H.Z., Yang, H.X., and Li, X. (2011). Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA genes. FEBS J **278,** 929-940.

Zhou, X., Ruan, J., Wang, G., and Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. In PLoS Comput Biol , pp. e37.

Chapter 5

Conclusions

5-1 Summary of our findings and how they relate to other works

In the last few years, many large-scale studies and technological advances have helped to pave the way toward greater understanding of the genome organization and function in different organisms. One of the first steps that has been done to a great amount recently is the generation of large scale sequencing data. There are two major challenges dealing with this huge amount of data: (1) handling and managing these data using the available computational processing power, and (2) developing innovative genomic informatics to process the raw data and extract meaningful information.

This thesis has tried to address the second challenge. Utilizing publicly available genome and NGS data, comparative genomics approach, and common and also novel computational methodologies, we were able to develop and test some biological hypothesis and improve the annotation of current data for further use.

The overall purpose of the current study was to assess three genomic features in cereal genomes - genomic source of two different types of small RNAs (nat-siRNAs and miRNAs) and miRNA-specific regulatory elements- and also determine how they have evolved. This can provide a foundation for the community to study small RNA expression and function more precisely in plants. Also the bioinformatic predictions that we made about the sources of nat-siRNA and miRNA regulatory elements could generate

a number of candidates for experimental analysis and save the initial testing of thousands of potential genomic fragments.

To be more specific, one of the most significant outcomes emerged from the first part of this thesis is providing a genome-wide and also a comparative view of overlapping gene content and orientations in several plant genomes for the first time. We showed that like other eukaryotic genomes, overlapping genes are frequent in cereals (3% to 10% of cereal protein-coding genes have this arrangement in different species), which is more common than previously thought. The frequencies of cereal overlaps are close to what we see in vertebrates (Makalowska et al., 2007) but are much less than the amount of overlaps in microbial genomes (Johnson and Chisholm, 2004). There is no positive or negative correlation between the percentages of overlaps in the genome and either number of genes or genome size. Therefore saving genome space and increasing the information density is not a major force in creation of overlaps. In terms of their orientation, the majority of overlaps belong to different-strand pairs, which is in agreement to what we see in other eukaryotes like human and mouse.

The second major finding of the first part is that the majority of overlapping genes are recent and species-specific. There is not much conservation of overlapping pattern in closely related cereal genomes. This suggests a fast rate of gain and loss of overlaps in these species. Our results also show that maize has more species-specific overlaps than rice and *Brachypodium.* The same species-specificity have been observed in mammals (Sanna et al., 2008; Solda et al., 2008) but they proposed that this little number of shared overlap relationship among mammals might be due to wrong assignment of orthologs. To

test this hypothesis in plants we examined the overlapping relationship not only between orthologs but also between homologs in different species. Still we found high frequency of species-specific overlaps, so this specificity should not be the result of wrong ortholog assignment. Also we showed that maize has a unique pattern of overlapping genes. Like what exists in the human genome it has lots of nested genes. The major formation mechanism for these nested genes could be retroposition in which a retro-gene has been inserted in the large introns of maize genes. These introns might provide an open chromatin environment for the insertion of the external genes.

The third major finding of the second chapter is that gene creation and translocation are major evolutionary forces that cause overlaps. Same mechanisms have been identified in birth of overlaps in vertebrates (Makalowska et al, 2007). In the three species considered we found 390 clear cases of overlap's birth and 96 clear cases of overlap's death. Similar to what has been found in mammals, the birth rate of overlaps is higher than their death rate but both of these rates are higher in plants compared to mammals (Sanna et al. 2008).

The last finding of the first part of our study shows that overlapping genes can be a rich source of nat-siRNAs. The expression levels of the genes involved in overlapping pairs have been shown to have negative correlation in Arabidopsis (Henz et al., 2007) which can be explained through the regulation of nat-siRNAs though the RNA interference pathway . This shows that overlapping pattern could impose a great regulatory effect on the gene expression and probably enable the fine-tuning of gene expression. Our comparative analysis of nat-siRNAs in cereals showed that the

generation of nat-siRNAs are not well conserved, suggesting that the fine tuning of expression by nat-siRNAs is species-specific and may be a cause for the unique phenotypes of each species.

In the third chapter we showed for the first time that by adopting a novel computational methodology, publicly available NGS data could be used to improve the annotation of miRNA genes. This is particularly useful for those model organisms with a plenty of NGS data. Our analysis is also one of the first ones in its type, which uses this number of public NGS data to help the improvement of the biological information in plants. The other very helpful piece of information that we had was two sets of miRNA genes with known gene boundaries. They were used as the training set to optimize the parameters of our methodology. The procedure we developed here was encouraging. It could be a reliable predictor of the upstream regions of the poorly annotated miRNA genes although there is still room for improvement. Accurate identification of miRNA transcription start site is useful because it helps to analyze the promoter and regulatory element more precisely and provides some foundations about the regulation of miRNA genes and their role in regulatory modules. By getting closer to the real TSS location it is possible to look at a narrower range for different motifs and regulatory elements. Also knowing the full sequence of miRNA transcripts gives us the opportunity to do a more accurate evolutionary analysis on them in closely related species.

And finally in the last chapter, we have identified over-represented and conserved motifs in the upstream regions of three cereals miRNA stem-loop sequences. By applying the conservation analysis on the identified motifs, we found several motifs that most

probably are true positive regulators of miRNAs and are good candidates for experimental verification. Also by studying the existence of maize miRNA motifs in the closely related genomes, we provide an explicit list of maize-specific miRNA regulatory elements. We not only studied the presence and absence of the motifs but also we compared their copy numbers in the three genomes. It was shown that a few motifs have a very high copy number in the studied miRNAs and species but this is not true for the majority of them. . Moreover, we showed that refine the upstream boundaries of miRNA genes changed the significance level of the identified motifs and the level of their matches to the known ones.

5.2 Evaluation

Now if we stand back and do a self-analysis, we can see some limitations in our study, which could have affected the results. These limitations include the following.

In the study of overlapping gene content we might have over- or underestimated the number of overlaps and therefore conserved ones due to imperfect gene annotation. We tried to resolve this issue by mapping the RNAseq reads to the gaps between neighbouring genes to examine whether they are actual overlapping or not, but we are not able to significantly improve the annotation of overlapping genes.

When we investigated the causes of birth and death of overlaps, we were unable to determine the mechanisms in many cases since we didn't have the appropriate out-group. Once we have additional genomes being sequenced and annotated from cereals species

that are more closely related to the species we have studied, this situation will be improved.

We didn't see high rate of conservation of nat-siRNAs among the conserved overlaps that might be due to not having enough small RNA data available yet. The expression of overlapped genes is dependent on plant tissues and growth conditions from which total RNA was extracted. As a consequence, nat-siRNAs might be generated from one sample, but not another. More small RNA data from various tissues and growth conditions will provide more evidence for the generation of nat-siRNAs.

We used small training sets to optimize the parameters of our methodology in refining the miRNA gene boundaries. If there were more experimentally confirmed pri-miRNAs, we had a larger dataset to train the computational procedure and also we could have kept some of the known data to test the performance of the procedure. Also known dataset might help us to find some sequence signatures in specific locations around TSSs that could be used to predict unknown TSSs more precisely. Furthermore, pri-miRNAs are processed into pre-miRNAs and mature miRNAs very soon after they are transcribed. RNAseq data from miRNA biogenesis pathway should significantly increase the chance of obtaining reads from intact pri-miRNAs and improve our ability to refine pri-miRNA boundaries.

Choosing of the1000 bp upstream region of miRNA precursors was an almost arbitrary decision. Probably we could have picked a more precise upstream range for each organisms based on the average length of known gene upstream regions that are enriched for regulatory elements.

5.3 Future Works

It is recommended that further research be undertaken in the following areas.

For the second chapter, further detailed studies with more lineages and better gene structure annotations are needed to confirm the mechanisms of overlapping emergence and loss. This gives us an idea how genomes evolve in this aspect. Also analyzing more lineages helps us to find those patterns that have been conserved throughout evolution and might have an important functional role.

Enrichment of overlaps for small RNA reads shows that overlapping pattern can impose a great regulatory effect on the gene expression. We provided some conserved and non-conserved enriched overlapping genes candidates and believe that they could be a good start for experimental functional analysis of their probable small RNA product in different species.

Not being able to predict 100% of miRNA gene boundaries does not suggest that our method for refining gene boundaries is not efficient or cannot be generalized. What it does highlight is that without enough transcriptome data (particularly RNAseq) at different developmental stages and sections, finding miRNA gene start sites can be very challenging. In theory, if we had enough RNAseq data, we should be able to exactly predict all the TSSs (enough sequencing depth will result in full coverage). At the time of conducting this research, we did our best to use as many public RNAseq libraries as possible. Our methodology could be optimized more whenever more RNAseq data and also pri-miRNA annotation become available. The next step for improving our method is

to integrate promoter motif location data with RNAseq read mapping data to predict TSS more accurately. We tried TATA-box but it didn't help the prediction that much since its location varies in different miRNA genes and sometimes it is absent. It is also helpful to study the performance of our methodology in animal miRNA genes with a known TSS and optimize it if possible.

5.4 References

Henz, S.R., Cumbie, J.S., Kasschau, K.D., Lohmann, J.U., Carrington, J.C., Weigel, D., and Schmid, M. (2007). Distinct expression patterns of natural antisense transcripts in Arabidopsis. In Plant Physiol, pp. 1247-1255.

Johnson, Z.I., and Chisholm, S.W. (2004). Properties of overlapping genes are conserved across microbial genomes. In Genome Res, pp. 2268-2272.

Makalowska, I., Lin, C.F., and Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. In BMC Evol Biol, pp. 193.

Sanna, C.R., Li, W.H., and Zhang, L. (2008). Overlapping genes in the human and mouse genomes. In BMC Genomics, pp. 169.

Solda, G., Suyama, M., Pelucchi, P., Boi, S., Guffanti, A., Rizzi, E., Bork, P., Tenchini, M.L., and Ciccarelli, F.D. (2008). Non-random retention of protein-coding overlapping genes in Metazoa. In BMC Genomics, pp. 174.