GENOME-WIDE IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF LONG NON-CODING RNAs IN CEREALS

by

YING SUN

(Under the Direction of Russell L. Malmberg)

ABSTRACT

The abundance and multiple functions of long non-coding RNAs (IncRNA) in mammalian systems have been one of the most important discoveries in molecular biology in recent years. However, the identification and characterization of IncRNAs in plants, especially cereals, is in its early stages. We conducted a reference-guided transcriptome assembly with RNA-Seq data from four economically important cereals, and screened for RNAs that were at least 200 bases in length, at most 70 amino acids in open reading frames and lack of homology in Uniprot database. We identified 7,196 IncRNA candidates in Zea mays, 1,974 in Sorghum bicolor, 4,236 in Setaria italica and 2,542 in Oryza sativa, and conducted sequence composition analysis, transposable elements detection and miRNA precursor screen. Further, a cross-species comparison, including sequence- and structure-based IncRNA homology search, synteny analysis, and IncRNA secondary structure prediction, uncovered some limited sequence similarity and sub-regions elucidating putative conserved secondary structures.

INDEX WORDS: IncRNA, Comparative Genomics, Evolutionary Conservation, Cereals
GENOME-WIDE IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF LONG NON-CODING RNAs IN CEREALS

by

YING SUN

B.S., Beijing University of Posts and Telecommunications, China, 2007
M.S., Beijing University of Posts and Telecommunications, China, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013
GENOME-WIDE IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF LONG NON-CODING RNAs IN CEREALS

by

YING SUN

Major Professor: Russell L. Malmberg
Committee: Liming Cai
Katrien Devos

Electronic Version Approved:
Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2013
ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Malmberg for his inspiring ideas of this project, insightful critiques, immense knowledge and patience. His guidance helped and encouraged me in those moments when I was in frustration with the research. Besides my advisor, I would also like to thank my committee members, Dr. Cai and Dr. Devos, for their valuable comments and perceptive questions. Finally, I wish to thank my parents for their generous support throughout my life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION AND LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Prologue: The Small Non-Coding RNA World</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Long Non-Coding RNA Emerging</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Long Non-Coding RNA Genomic Context and Regulatory Networks</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Evolution of Long Non-Coding RNAs</td>
<td>18</td>
</tr>
<tr>
<td>1.5 My Objectives</td>
<td>23</td>
</tr>
<tr>
<td>2 GENOME-WIDE IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF</td>
<td>26</td>
</tr>
<tr>
<td>LONG NON-CODING RNAs IN CEREALS</td>
<td></td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>27</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>27</td>
</tr>
<tr>
<td>2.3 Results</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>43</td>
</tr>
<tr>
<td>2.5 Material and Methods</td>
<td>49</td>
</tr>
<tr>
<td>3 CONCLUSION AND FUTURE PERSPECTIVES</td>
<td>85</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>87</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 2.1: Summary of Transcriptome Assembly Results (Compared with Annotation) ........57
Table 2.2: Summary of lncRNA Length And Number of Exons ........................................58
Table 2.3: Summary of lncRNA Locations Relative to Their Adjacent Annotated Genes ..........59
Table 2.4: Transposable Element Components of lncRNA Candidates .............................60
Table 2.5: Categories of lncRNA Candidates with Transposable Elements ......................61
Table 2.6: Candidate lncRNA Homologous Pairs among Zea mays, Sorghum bicolor, Setaria
     italica and Oryza sativa ..................................................................................................62
Table 2.7: Significant Structure-based lncRNA Homologs among the lncRNA Candidates and
     the 225 lncRNA Families from Rfam ............................................................................63
Table 2.8: Putative Homologs in Oryza sativa Reveal by Infernal Search .........................65
Table 2.9: The Primers Designed to Confirm 19 Putative Transcripts in Setaria italica ........66
Table 2.10: List of lncRNA Candidates in Zea mays and Their Corresponding UniformMu
     Insertions Chosen for Mutagenesis lncRNA Function Analysis ...................................67
LIST OF FIGURES

Page

Figure 1.1: LncRNA (Red) Categories Based on Their Locations Relative to Nearby Protein-Coding Genes (Green)........................................................................................................24

Figure 1.2: RNA Secondary Structure Motifs (Wuchty, Fontana et al. 1999) ........................................25

Figure 2.1: Mini Phylogenetic Tree of Cereals (Devos 2005)................................................................68

Figure 2.2: Synteny of Cereal Genetic Maps (Devos 2005)................................................................69

Figure 2.3: Quality Control, Pre-processing of RNA-Seq Data (a) and "Tuxedo" Reference-Based Transcriptome Assembly (b)........................................................................70

Figure 2.4: Transcript Length, ORF Prediction and Homology Search LncRNA Identification Pipeline...........................................................................................................................71

Figure 2.5: Comparison of GC Content of Annotated Genes (Red) and the LncRNA Candidates (Blue)...........72

Figure 2.6: Hexamers That Showed at Least Ten Fold Changes (Either Under- or Over-Representation) in LncRNA Candidates and Annotated Protein-Coding CDS Sequences.73

Figure 2.7: An Example of Annotated Genes Which Are Identical to Our LncRNA Candidates..78

Figure 2.8: LncRNA Orthologous Pairs (Sequence Homology and Synteny). .........................79

Figure 2.9: Conserved Structures of 2 LncRNA Families from Rfam, Which Have Structure-based LncRNA Homologs in More than One of Zea mays, Sorghum bicolor, Setaria italica and Oryza sativa Transcriptomes. .................................................................80
Figure 2.10: RNAalifold RNA Structure Prediction Based on TE-Masked lncRNA Multiple-Sequence Alignment .................................................................81

Figure 2.11: Distribution of UniformMu Insertions along the Chromosome 1 of Zea mays ........83
CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

Since the breakthrough discovery of RNase P catalytic activity (Guerrier-Takada, Gardiner et al. 1983) led us to the "RNA world" (Gesteland 2005), surprises keep coming. RNAs are no longer confined in the "Central Dogma" (mRNA, rRNA, tRNA) (Crick 1970), and might serve as the cradle of modern life by carrying both genetic information and catalytic capacity (Gesteland 2005; Elliott and Ladomery 2010). Recent studies have revealed a large variety of non-coding RNAs (ncRNA) with distinct biological features and cellular functions dramatically expanding the RNA family (Kurth and Mochizuki 2009; Louro, Smirnova et al. 2009; Mercer, Dinger et al. 2009; Ponting, Oliver et al. 2009; Wilusz, Sunwoo et al. 2009; De Lucia and Dean 2011; Jouannet and Crespi 2011; Li and Liu 2011; Kim and Sung 2012; Rinn and Chang 2012; Huang, Zhang et al. 2013; Kung, Colognori et al. 2013; Ma L, Bajic VB et al. 2013).

1.1 Prologue: The Small Non-Coding RNA World

Cellular small non-coding RNAs perform crucial functions in cellular metabolism; some of them are involved in co-transcriptional or post-transcriptional modifications. Several examples are: small nuclear RNA (snRNA) and its corresponding protein (snRNP) make up the splicesome which removes introns from pre-mRNA (McKeown 1993); Small nucleolar RNA (snoRNA) conducts the 2'-O-ribosie methylation for rRNA maturation (Tollervey 1996); RNase P catalyzes the cleavage of primary tRNA's 5' leader sequence (Xiao, Scott et al. 2002), otherwise the later is unable to achieve the highly conserved cloverleaf secondary structure for functioning; Telomerase extends chromosomal DNA ends by using the RNA moiety of its
ribonucleoprotein subunit as a primer for replication (Cech and Lingner 1997). All these ncRNAs mentioned above are considered to be housekeeping non-coding RNAs (Ponting, Oliver et al. 2009).

In parallel, small regulatory ncRNAs play an essential role in gene silencing. siRNA mediates RNA interference (RNAi) by mRNA cleavage, translational repression and chromatin modification (Fire, Xu et al. 1998; Meister and Tuschl 2004), and the successful induction of RNAi in cultured mammalian cells by synthetic siRNAs (Elbashir, Harborth et al. 2001) suggested a new method for achieving rapid gene knockdown in reverse genetic studies. In particular, piRNA is one special siRNA involved in transposon defense. It addresses RNA-PIWI protein complexes onto retrotransposon RNAs, resulting in target RNA degradation (Klattenhoff and Theurkauf 2008).

The miRNA is another gene-silencing trigger, and shares similar structure and RNAi machinery with siRNA (Murchison and Hannon 2004). A major difference between siRNA and miRNA, especially those in animals, falls within their target recognition. Typically, siRNA shares full sequence complementarity with their intended targets, despite unintended off-target effects (Jackson, Burchard et al. 2006). Comparable with siRNA, miRNA in plants requires near-perfect complementarity to induce direct target mRNA cleavage; However, in contrast, miRNA in animals tends to induce translation repression with incomplete hybridization to its target (Axtell, Westholm et al. 2011). Furthermore, miRNA is encoded by the genome and regarded as a "functioning monitor" of normal cells. Clinical studies have proved miRNA dysregulation is associated with various human diseases (Hammond 2006).

Although small ncRNAs function in quite distinct ways to long non-coding RNAs (lncRNAs), their diversity of functioning gives us a hint of how complicated the RNA regulatory
networks could be. In addition, considering that their ubiquity and central place in cellular metabolism, these "relics from the RNA world" (Jeffares, Poole et al. 1998) are great models of RNA evolution. The fact that small ncRNAs derived from a common ancestry may share short conserved sequence domains and also well-conserved secondary structure elements, rather than strong sequence homologies (Meli, Albert-Fournier et al. 2001), stresses the importance of RNA secondary structure in functional continuity during evolution.

1.2 Long Non-Coding RNA Emerging

1.2.1 From Junk DNA to Pervasive Transcription

In 1971, Thomas CA Jr. coined the term "C-value paradox" to state the discrepancy between the developmental complexity and the amount of DNA per haploid genome (Thomas 1971) among eukaryotes, and the large chunks of non-coding space were referred to as "junk DNA" for years (Ohno 1972). Later, during 1970s to the early 1980s, the identification of rRNA, tRNA, snoRNA, snRNA and RNase P demonstrated the existence of functional elements in the non-coding regions; however, these were far from genome-wide. Since the late 1990s, the advent of genomic tiling arrays and high-throughput transcriptome sequencing technologies dramatically speed up the whole-genome search of functional RNAs (Morozova, Hirst et al. 2009). The ENCODE project assigned biochemical functions for 80% of the human genome, the majority of which (approximately 62%) belong to different RNA types (Bernstein, Birney et al. 2012); the interpretation of this result is still controversial. Although Arabidopsis thaliana might be the most well-annotated plant genome, transcriptome sequencing still keeps on revealing the presence of novel transcripts. Moghe et al. (2013) found 6,545 intergenic transcribed fragments occupying 3.6% of its presumed intergenic space, of which, 32.1% showed evidence of translation (Moghe, Lehti-Shiu et al. 2013). As to cereals, by analyzing Zea mays EST

Coming with the accumulation of novel transcripts is an increasing number of non-coding RNAs, some of which are much longer than previous characterized small ncRNAs. Typically, long non-coding RNA (IncRNA) refers to "non-protein coding transcripts longer than 200 nucleotides", and may either be polyadenylated or not (Carninci, Kasukawa et al. 2005; Cheng, Kapranov et al. 2005). They may be tissue- and cell-type specific (Xin, Wang et al. 2011; Liu, Jung et al. 2012; Ulitsky, Shkumatava et al. 2012), and observed within both nuclear domains and cytoplasm (Batista and Chang 2013). As mentioned above, while studies of small ncRNAs have accumulated enormous knowledge about their functionality, the field of IncRNA has just recently gained momentum and they may form functionally diverse ncRNA categories (Mercer, Dinger et al. 2009; Wilusz, Sunwoo et al. 2009; Kim and Sung 2012; Lee 2012; Rinn and Chang 2012; Kung, Colognori et al. 2013; Mercer and Mattick 2013).

1.2.2 Transcriptome Analysis and Extensive Identification of IncRNAs

1.2.2.1 Benchwork: De Novo Identification

The analysis of transcriptomes has been experiencing a great technology jump in recent years, from candidate gene-based detection of RNA (northern blotting) to high-throughput next-generation sequencing of entire transcriptome (RNA-Seq). Currently, tiling microarrays, RNA-Seq and genome-wide chromatin signature-based approaches have been widely used for the identification of IncRNAs (Jouannet and Crespi 2011; Rinn and Chang 2012).

Tiling Microarrays. Probes in tiling microarrays are designed to cover the entire genome or contiguous segments of interest, thus are able to detect novel transcripts. Using the Arabidopsis Tiling Array-based Detection of Exons (ARTADE)-based method, Matsui et al.
(2008) performed *Arabidopsis* whole-genome expression profiling studies and identified 7,719 stress-responsive transcription units (non- *Arabidopsis* Genome Initiative (non-AGI) TUs) from un-annotated regions, with length varying from 50 bases to 12,800 bases (Matsui, Ishida et al. 2008). Later, another 6,105 novel genes were discovered from un-annotated regions under ABA treatment using whole-genome Affymetrix tiling arrays (Okamoto, Tatematsu et al. 2010). A large majority of these non-AGI TUs possibly encoded hypothetical non-coding RNAs and included a large number of antisense RNAs (Matsui, Ishida et al. 2008; Okamoto, Tatematsu et al. 2010). However, some limitations of tiling arrays include the potential of cross hybridization, the lack of strand-specific information, limited resolution due to high background noise and the requirement of a reference genome for probe design (Rinn and Chang 2012).

**RNA-Seq**, also called "whole transcriptome shotgun sequencing", can tackle the transcriptome profiling in an high-throughput and unbiased manner (Wang, Gerstein et al. 2009). It is able to directly identify transcripts by sequence instead of measuring them by hybridization intensities, and to reach an ultra sequencing depth, which may help to reveal more accurate exonic structures of novel transcripts and detect certain functional RNAs with low expression level. What's more, it also allows for accurate quantification of expression levels based on the number of reads. Through an RNA-Seq analysis, Hu et al (2011) found 428 IncRNAs that were dynamically regulated during erythropoiesis (Hu, Yuan et al. 2011). Introducing a targeted RNA sequencing strategy, Mattick and Rinn et al. (2012) detected rare non-coding variant isoforms of known primary mRNAs, an additional 163 neighboring and antisense IncRNAs and mature spliced transcripts within intronic regions from the human transcriptome (Mercer, Gerhardt et al. 2012). Beyond the GENCODE version 7 (Derrien, Johnson et al. 2012; Harrow, Frankish et al. 2012) annotated elements, Djebali et al. (2012) extended over all 73,325 intergenic and antisense
transcripts from 41,204 novel genes (Djebali, Davis et al. 2012); So far, GENCODE keeps a record of 13,562 lncRNA genes producing 23,105 lncRNA transcripts (GENCODE version 18). Hangauer et al. (2013) assembled 53,864 distinct putative lincRNAs from 127 human RNA-Seq datasets; of which, over 4,000 are highly expressed, rivaling with protein-coding genes (Hangauer, Vaughn et al. 2013). By using computational analysis and experimental approaches, Xin et al. (2011) identified 125 putative wheat stress responsive lincRNAs, of which 2 are signal recognition particle (SRP) 7S RNA variants and 3 are U3 snoRNAs (Xin, Wang et al. 2011). The astonishing amount of lncRNAs identified through RNA-Seq analysis has demonstrated the power of the next-generation sequencing technologies in transcriptome studies.

**Chromatin signature-based approaches** are designed to detect actively transcribed regions (K4-K35 domain) marked by H3K4me3 and H3K36me3. The combination of chromatin signature-based approaches and massively parallel sequencing of targeted DNA sites is able to identify unknown genomic architectures (Rinn and Chang 2012). Surveying the entire mouse genome by chromatin marks, Guttman et al. (2009) identified 1,600 large multi-exonic large intergenic non-coding RNAs (lincRNAs) across four mouse cell types with greater than 95% showing clear evolutionary conservation (Guttman, Amit et al. 2009). Further, by analyzing chromatin-state maps of various human cell types, Khalil and Guttman et al. (2009) expanded the number of human lincRNAs to as many as 3,300, and observed that about 20% of these lincRNAs regulate gene expression by recruiting PRC2 to specific genomic loci (Khalil, Guttman et al. 2009).

**1.2.2.2 Bioinformatics: In Silico Identification**

The rationale for computational identification of lncRNAs is to distinguish lncRNAs from protein-coding mRNAs on the basis of biological features, such as the length of open
reading frame (ORF) and homology search against known protein database or protein-coding mRNA database. Diverse bioinformatics tools have been developed and successfully revealed a myriad of novel lncRNAs.

The most traditional strategy to separate lncRNAs from the rest is probably by estimating their ORF length. Typically, a 100 amino acid threshold is used for this purpose; and various ORF prediction programs have been developed; for instance, GenScan, ESTScan2, ANGLE (Jouannet and Crespi 2011). However, debate against this arbitrary threshold has never stopped, given the fact that well-known lncRNAs, such as H19, Xist, Mirtg, Gtl2, and KcnqOT1, all contain putative long ORFs (Dinger, Pang et al. 2008).

Sequence homology search against mRNA or protein databases could be performed by bioinformatics tools, such as BLAST (Gish and States 1993), Pfam (Finn, Mistry et al. 2010) and SUPERFAMILY (Gough, Karplus et al. 2001), assuming that homology provides indirect evidence of function as an mRNA (Dinger, Pang et al. 2008). Applying an ORF-Predictor/BLASTP parsing pipeline with manual annotation, Jia et al. (2010) derived 6,736 human lncRNA genes by combining 5,446 lncRNA genes from their own lncRNA identification pipeline and 4 public databases; they suggested that 62% of the "hypothetical protein" genes are classified as non-coding by their protein-coding capacity evaluation pipeline (Jia, Osak et al. 2010).

Besides direct comparison of sequence similarity, the evolutionary conservation of protein-coding ORFs can also be inferred by the tendency for protein-coding sequences to favor synonymous base changes over non-synonymous ones. CRITICA (Badger and Olsen 1999), RNAcode (Washietl, Findeiss et al. 2011) and PhyloCSF (Lin, Jungreis et al. 2011) are all designed to calculate phylogenetic conservation scores. The main drawback of these statistical
phylogenetic algorithms are their dependence on the genome-wide multiple sequence alignment and known coding and non-coding information. This would not be a problem for vertebrates given the UCSC 45-vertebrate-genome alignment with reference to human genome, but is not very realistic for plant genomes considering the complexity of their genome architecture. For instance, the most recent conservation study on 20 angiosperm plant genomes only cover around 26% of *Sorghum bicolor*, *Oryza sativa* and *Zea mays* genomes (Hupalo and Kern 2013).

The presence of conserved RNA secondary structures may be another property to infer lncRNA homology (Wan, Kertesz et al. 2011). Accordingly, a number of programs have been designed, such as QRNA (Rivas and Eddy 2001), RNAz (Gruber, Neuboeck et al. 2007), EvoFOLD (Pedersen, Bejerano et al. 2006) and Infernal (Nawrocki, Kolbe et al. 2009) which search for RNAs based upon structure.

Distinguishing protein-coding from non-coding RNAs through support vector machines (SVM), a machine learning approach, has emerged recently. CPC (Kong, Zhang et al. 2007) and CONC (Liu, Gough et al. 2006) comprehensively evaluated the weight of each biological features in the model, including peptide length, amino acid composition, protein homologs, secondary structure, and protein alignment information, based on large amount of training data; and thus are supposed to show high level of accuracy (Dinger, Pang et al. 2008). Nevertheless, the performance of their SVM might be confined by the quality and specificity of the training data; practically, CPC works pretty well for human transcripts, but would not be an ideal choice for cereals. A properly trained SVM would represent a promising discrimination method which might lead to a dramatic increase in the number of lncRNAs identified in plants. Actually, CPC was implemented in a maize lncRNA identification pipeline and provided 1,913 putative ncRNA candidates in their initial step (Boerner and McGinnis 2012).
Another novel coding-potential assessment tool, CPAT, implements a logistic regression model built with four sequence features: ORF length, ORF coverage, Fickett TESTCODE statistic and hexamer usage bias (Wang, Park et al. 2013). The speed of this method is really impressive; however, given its unstable performance (intensively influenced by the input reference data), it seems not practical to our cereal lncRNA studies.

Programs listed above can be combined together to get more confident predictions. Nam et al. (2012) exploited the coding potential of 262 lincRNA transcripts from 170 loci in *C. elegans* by a combination of CPC and RNAcode (Nam and Bartel 2012). Cabili et al. (2011) removed any putative ORFs with a positive phylogenetic codon substitution frequency (PhyloCSF) metric, which was calculated for each locus across 29 mammals, and scanned the rest through Pfam for homologs (Cabili, Trapnell et al. 2011). This led them to present more than 8,000 human lincRNAs across 24 tissues and cell types; of which, 993 (12%) had expressed orthologous transcripts in another vertebrate species (Cabili, Trapnell et al. 2011). Sun et al. (2012) proposed a lncRNAs detector lncRScan (including ORF length cutoff, PhyloCSF coding-potential evaluation and Pfam homology search), and reported 308 novel mouse embryonic lncRNAs, among which, 13 lncRNAs may be regulated by Klf1 and involved in erythroid development (Sun, Zhang et al. 2012).

1.3 Long Non-Coding RNA Genomic Context and Regulatory Networks

The study of lncRNAs can be traced back to 1990s when placental X chromosome inhibitor Xist (Brockdorff, Ashworth et al. 1991; Brown 1991; Ballabio and Willard 1992; Brockdorff, Ashworth et al. 1992; Lee 2009) and cancer inducer H19 (Brannan, Dees et al. 1990) were characterized. As mentioned above, the development of technology, from traditional gene mapping to high-throughput transcriptome sequencing, has identified thousands of lncRNAs.
consequently (Amaral, Clark et al. 2011; Bu, Yu et al. 2012; Derrien, Johnson et al. 2012); however, the functions of the vast majority of them are still elusive. NONCODE v3.0 contained 73,327 lncRNAs, among which only 1,635 have been annotated with potential functions (Bu, Yu et al. 2012), compared with 197 detailed records in lncRNAdb (Amaral, Clark et al. 2011). To better understand their functional significance, scientists attempted to classify them into different categories according to their distinct features, such as genomic context or mechanism of functioning, such as epigenetic chromatin modification, cis-/trans-transcriptional regulation, post-transcriptional regulation and sub-cellular structure construction (Mercer, Dinger et al. 2009; Ponting, Oliver et al. 2009; Rinn and Chang 2012; Kung, Colognori et al. 2013; Ma L, Bajic VB et al. 2013).

1.3.1 Genomic Location and Context

According to their location relative to nearby protein-coding genes, lncRNAs may be roughly classified into 5 categories: (1) sense, (2) antisense, (3) intronic, (4) divergent and (5) intergenic (Ponting, Oliver et al. 2009; Rinn and Chang 2012; Kung, Colognori et al. 2013; Ma L, Bajic VB et al. 2013) (Figure 1.1).

*Intergenic lncRNAs (lincRNAs).* lincRNAs are those transcribed from intergenic regions. They are typically shorter than protein-coding genes, and have fewer exons, typically 2-3, and mostly polyadenylated (Ulitsky and Bartel 2013). Considering the large non-coding regions in eukaryotic genomes (for example, only 1~2% of human genome has been specified as protein-coding), lincRNAs are prevalently existing in eukaryotic transcriptomes and account for the lion’s share of currently identified lncRNA population (Ballabio and Willard 1992; Burleigh and Harrison 1997; Liu, Muchhal et al. 1997; Burleigh and Harrison 1999; Lee, Davidow et al. 1999; Martin, del Pozo et al. 2000; Topp, Zhong et al. 2004; Shin, Shin et al. 2006; Matouk, DeGroot
lincRNAs function through various mechanisms, in epigenetics (recruiters, tethers and scaffolds for chromatin modification), in transcription (decoys, co-regulators, and PolII inhibitors), post-transcriptional regulation (translational control, splicing regulation), etc (Kung, Colognori et al. 2013; Ma L, Bajic VB et al. 2013; Ulitsky and Bartel 2013).

lincRNAs are more conserved than introns and antisense transcripts, although less conserved than mRNAs (Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Guttman, Garber et al. 2010; Orom, Derrien et al. 2010; Ma L, Bajic VB et al. 2013); they are commonly expressed in a tissue-specific pattern (Guttman, Amit et al. 2009; Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011; Derrien, Johnson et al. 2012); and are more stable than intronic lncRNAs (Clark, Johnston et al. 2012). However, unexpectedly, only a small fraction of the vertebrate lincRNAs contain conserved secondary structures (Ulitsky and Bartel 2013), which is considered to be closely related to RNA functions.

**Antisense lncRNAs.** Antisense lncRNAs are transcribed from the antisense strand of protein-coding genes. Abundant antisense lncRNAs have been found in both mammalian and plant transcriptome studies (Smilinich, Day et al. 1999; Lyle, Watanabe et al. 2000; Matsui, Ishida et al. 2008; Pandey, Mondal et al. 2008; Swiezewski, Liu et al. 2009; Georg, Honsel et al. 2010; Okamoto, Tatematsu et al. 2010); as reported, 87% of mouse protein-coding genes have antisense counterparts (Katayama, Tomaru et al. 2005) and approximate 32% of the human lncRNAs are antisense to coding genes (Derrien, Johnson et al. 2012; Ma L, Bajic VB et al.
The overlap between the sense protein-coding gene and the antisense lncRNA can be complete (nested) or partially, while natural antisense transcripts (NATs) tend to accumulate around promoter and terminator regions (Kung, Colognori et al. 2013).

Many sense and antisense (SAS) pairs have been found in imprinted regions. Examples are, the dual lncRNA SAS pair Xist/Tsix (Lee, Davidow et al. 1999) (also considered to be lincRNAs according to their relative position to protein-coding genes), and lncRNA and protein-coding pairs, such as Igf2r/Air (Lyle, Watanabe et al. 2000), Kcnq1/Kcnq1ot1 (Pandey, Mondal et al. 2008) and FLC/COOLAIR (Swiezewski, Liu et al. 2009). The expression of SAS pairs is more correlated (either positive or negative) than by chance (Kung, Colognori et al. 2013). Another interesting correlation between SAS pair is the negative correlation between the length of their overlap and the level of their expression, suggesting a transcriptional collision model caused by convergent transcription (Osato, Suzuki et al. 2007).

NATs function not only via transcriptionally influencing the expression of the antisense gene, but also by masking splicing sites through base complementarity that consequently affects the alternative splicing of their overlapping transcripts (Gu, Zhang et al. 2009). NATs can also serve as scaffolds by virtual of recruiting stabilizing factors to increase the stability of its sense counterpart sites, such as the recruitment of HuRNA to ARE-containing transcripts by iNOS (Matsui, Nishizawa et al. 2008).

**Intronic lncRNAs.** Introns produce not only small ncRNAs but also lncRNAs (Louro, Smirnova et al. 2009; Rearick, Prakash et al. 2011; Kung, Colognori et al. 2013). In the human genome, a role for intronic lncRNAs in guiding PRC2 to specific chromatin regions has been proposed, and their over-expression patterns are treated as hallmarks of cancer (Guil, Soler et al. 2013).
In plants, COLDAIR is located in the first intron of FLC and required for triggering vernalization-mediated epigenetic silencing of FLC (Heo and Sung 2011).

**Sense IncRNAs.** Sense IncRNAs are those which overlap with part or cover the entire sequence of a protein-coding gene. A well-known example is steroid receptor RNA activator (SRA), initially isolated and functional characterized in 1999 (Lanz, McKenna et al. 1999). It may be the first example of an IncRNA which can play dual roles as either a transcript or a protein (SRAP) in cellular process (Lanz, McKenna et al. 1999; Chooniedass-Kothari, Emberley et al. 2004): its IncRNA acts as a transcriptional repressor in specific promoter regions (Chooniedass-Kothari, Hamedani et al. 2010), whereas its protein SRAP is a biomarker in breast tumor tissues (Chooniedass-Kothari, Hamedani et al. 2006).

1.3.2 From A Functional Perspective:

1.3.2.1 The Big Picture

Genome-wide expression and evolutionary analysis (Bu, Yu et al. 2012; Derrien, Johnson et al. 2012), in conjunction with a handful of IncRNAs which have been experimentally investigated (Amaral, Clark et al. 2011), demonstrates a large variety of levels at which IncRNAs are regulating gene expression (Kung, Colognori et al. 2013): (1) epigenetic chromatin modification, including epigenetic silencing (Ballabio and Willard 1992; Nagano, Mitchell et al. 2008; Pandey, Mondal et al. 2008; Zhao, Sun et al. 2008; Heo and Sung 2011; Kogo, Shimamura et al. 2011), facilitation of the histone exchange reaction (Dawe 2004; Topp, Zhong et al. 2004; Gent and Dawe 2012), enhancer-like long-range gene activation via chromosomal looping (Wang, Yang et al. 2011) and etc.; (2) cis-/trans-transcriptional regulation, including acting as decoys for TFs to inhibit gene expression (Kino, Hurt et al. 2010; Hung, Wang et al. 2011), sequestering miRNAs away from their mRNA target (Franco-Zorrilla, Valli et al. 2007; Seitz
2009; Salmena, Poliseno et al. 2011), transcriptional co-regulation (Yao, Brick et al. 2010) (Kung, Colognori et al. 2013) and etc.; (3) post-transcriptional regulation, including mRNA splicing (Gu, Zhang et al. 2009), stability control (Matsui, Nishizawa et al. 2008), translational regulation (Ebralidze, Guibal et al. 2008) and etc.; (4) Others, including sub-cellular structural organization (Azzalin, Reichenbach et al. 2007; Clemson, Hutchinson et al. 2009; Sasaki, Ideue et al. 2009), nuclear trafficking (Willingham, Orth et al. 2005; Wong, Brettingham-Moore et al. 2007; Wilusz, Sunwoo et al. 2009), and etc.

Many lncRNAs employ more than one mechanisms of action. H19, the first lncRNA described in mammalian genome (Brannan, Dees et al. 1990), is imprinted from the maternal allele at the Igf2 locus and influences the expression of the major fetal growth factor Igf2 via both transcriptional control (Forne, Oswald et al. 1997) and post-transcriptional binding of Igf2 mRNA binding-protein (IMP) family members (Runge, Nielsen et al. 2000); it is also a bi-functional RNA recruited in tumor development as either a miRNA precursor (oncogene) (Tsang, Ng et al. 2010) or a lncRNA (tumor suppressor) (Hao, Crenshaw et al. 1993; Yoshimizu, Miroglio et al. 2008). So does the steroid receptor RNA activator (SRA), as mentioned above.

1.3.2.2 Long Non-Coding RNAs in Plant Development

There have been some attempts made to systematically identify and annotate lncRNAs in plants (Hirsch, Lefort et al. 2006; Ben Amor, Wirth et al. 2009; Georg, Honsel et al. 2010; Xin, Wang et al. 2011; Zhang, Zhao et al. 2011; Boerner and McGinnis 2012; Liu, Jung et al. 2012; Wu, Wang et al. 2013). Meanwhile, several plant lncRNAs have been experimentally demonstrated to be recruited in crucial plant developmental processes, such as stress response and reproduction.
**Stress Response.** IncRNAs take part into responsive mechanisms of various stresses in plants. By using computational analysis and experimental approaches, Xin et al. (2011) identified 125 putative wheat stress responsive IncRNAs, of which 2 are signal recognition particle (SRP) 7S RNA variants and 3 are U3 snoRNAs (Xin, Wang et al. 2011). Hirsch et al. (2006) proposed 43 ncRNA transcripts in *Arabidopsis*, and 11 of them contain potential functional secondary structures suggested by strong nucleotide strand asymmetries and a biased GC content (Hirsch, Lefort et al. 2006). The npc48 transcript is one of these 11 ncRNAs with 983nt in length, which works in a similar way as miRNA MIR168; the over-expression of both will reduce levels of MIR168’s target mRNA AGO1 and lead to strong leaf serration (Hirsch, Lefort et al. 2006). Another IncRNA ncp536 is antisense to the Golgi-transport-complex-protein-related gene AT1G67930, suggesting cis-regulatory roles. It is up-regulated by phosphate starvation and salt stress, and its over-expression induced by salt stress significantly promotes root growth (Ben Amor, Wirth et al. 2009).

The TPS11/Mt4 IncRNA family, responding to phosphate deprivation, includes At4 (Burleigh and Harrison 1999; Shin, Shin et al. 2006) and AtIPS1 (Martin, del Pozo et al. 2000) paralogs in *Arabidopsis*, Mt4 in *Medicago truncatula* (Burleigh and Harrison 1997), Mt4-like in soybean (Burleigh and Harrison 1999), and TPS11 in tomato (Liu, Muchhal et al. 1997). Liu et al. (1997) isolated the 474nt TPS11 transcript specifically induced by phosphate starvation in tomato roots and leaves, and found that its promoter region shared sequence conservation with yeast phosphate-starvation-induced genes (Liu, Muchhal et al. 1997). Like TPS11, Mt4 in *Medicago truncatula* (Burleigh and Harrison 1997) has multiple small ORFs (longest 51aa), but only a small region of one ORF shares identity in sequence with TPS11 (Burleigh and Harrison 1997). Besides the low conservatory ORFs, TPS1/Mt4 family IncRNAs contain a highly conserved
motif intensively complementary to miR-399, a phosphate starvation-induced miRNA (Franco-Zorrilla, Valli et al. 2007). This complementarity helps AtIPS1 to sequester miRNAs away from their mRNA targets, while the cleavage of itself is prevented by a mismatch loop in the motif (Franco-Zorrilla, Valli et al. 2007). This miRNA pseudotarget activity is the first evidence of "competing endogenous RNA" (ceRNA) (Seitz 2009; Salmena, Poliseno et al. 2011), which reveals a potential bona fide genome-wide regulatory network. In addition, the fact that transcripts of this family are regulated by cytokinins and decrease rapidly after Pi re-supplement does suggest their functions in regulatory process.

The GUT15/CR20 family is another group of hormone/stress-induced lncRNAs, including GUT15 (gene with unstable transcript 15) in tobacco (Taylor and Green 1995), AtGUT15 and AtCR20-1 in Arabidopsis (Teramoto, Toyama et al. 1996; van Hoof 1997), and CR20 in cucumber (Teramoto, Toyama et al. 1996). Their transcripts appear to be unstable, compatible with its cytokinin-repressed characteristics, suggesting their regulatory roles (MacIntosh, Wilkerson et al. 2001).

A few other plant lncRNAs induced by biotic signals are CDT-1 in Craterostigma species induced by ABA or dehydration rendering callus desiccation tolerant (Furini 2008), CsM10 under male sex expression conditions in cucumber (potentially biotic-stress-related) (Cho, Koo et al. 2005), Enod40 induced during nodule development in leguminous plants (Crespi, Jurkevitch et al. 1994). Enod40 was initially characterized to play a role in root nodule organogenesis commonly in leguminous plants (Crespi, Jurkevitch et al. 1994; Charon, Sousa et al. 1999; Girard, Roussis et al. 2003; Campalans, Kondorosi et al. 2004), and later was also found in rice (Kouchi, Takane et al. 1999). It is characterized with two short peptide products as well as a highly conserved secondary structure (Gultyaev and Roussis 2007). Enod40 functions
mainly through the highly conserved secondary structure by protein re-localization (Barciszewski and Erdmann 2003), rather than its protein products; one evidence is the transmission of MtRBP1 from nuclear speckles into cytoplasmic granules (Campalans, Kondorosi et al. 2004; Zhu and Wang 2012).

**Reproduction.** Cold-induced COOLAIR and COLDAIR are respectively antisense-intragenic and sense-intronic (from the first intron) IncRNAs, associated with regulation of the FLC flowering locus in *Arabidopsis* (Heo and Sung 2011). COLDAIR is required for triggering vernalization-mediated epigenetic silencing of floral repressors FLC, by targeting PRC2 to FLC chromatin and resulting in H3K27me3 methylation (Heo and Sung 2011). COOLAIR, distinct from COLDAIR, is suggested to function in early cold induced FLC silencing, based on the observation of its accumulation to a peak around 10-14 days after cold treatment and declination with prolonged cold (Swiezewski, Liu et al. 2009). The detailed mechanism of its functions is unknown, but suggested through transcriptional interference (Swiezewski, Liu et al. 2009).

A photoperiod-sensitive male sterility (PSMS) is a spontaneous mutant in rice, and possesses a number of desirable characteristics that may facilitated the development of hybrids (Zhang, Shen et al. 1994). Ding et al. (2012) identified the sufficient expression of a IncRNA with 1,236 bases in length, referred to as long-day-specific male-fertility-associated RNA (LDMAR), regulates PSMS and is crucial for normal pollen development under long-day conditions in rice (Ding, Lu et al. 2012). By altering the secondary structure of LDMAR with a SNP, they observed a reduction of LDMAR expression under long-day condition and the consequential premature programmed cell death in developing anthers followed by PSMS (Ding, Lu et al. 2012). Later, this research group detected a novel siRNA from LDMAR promoter region, which suppressed LSMAR by DNA methylation in the promoter (Ding, Shen et al. 2012).
Zm401 is a pollen-specific gene in maize. Dai et al. (2007) proved the importance of Zm401 in maize pollen development, based on the observation that the over-expression of Zm401 resulted in abnormal tassels, degenerate anthers and degradation of tapetum, asynchronous fusion of pollen sacs, and aborted pollen grain development (Dai, Yu et al. 2007). Further knockdown studies revealed significant regulation effect on the expression of critical genes for pollen development, including ZmMADS2, MZm3-3, and ZmC5, in conjunction with aberrant development of the microspore and tapetum as well as finally male-sterility (Ma, Yan et al. 2008).

**Other Interesting IncRNAs in Plants.** Topp et al. (2004) reported that maize centromeric retrotransposons (CRMs), satellite repeats (CentC) and other centromeric RNAs may facilitate the histone exchange reaction either by opening chromatin during transcription, or by remaining attached to the centromere after transcription and serving as scaffolds to recruit centromeric histone H3 (CENH3) to centromere (Dawe 2004; Topp, Zhong et al. 2004; Gent and Dawe 2012). Deduction of centromeric RNA size in vivo from known chromatin-modifying IncRNA Xist suggested large centromeric RNAs with small functional segments (Topp, Zhong et al. 2004). In addition, evidence showed that a portion of these large centromeric RNAs were protected from RNAi machinery and maintain single-stranded, which is crucial for their role as scaffolds (Topp, Zhong et al. 2004).

1.4 Evolution of Long Non-Coding RNAs

Given "the RNA world" assumption and structural conservation shared by central small ncRNAs (Jeffares, Poole et al. 1998; Meli, Albert-Fournier et al. 2001), we expect that the natural selection pressure has been reflected on IncRNAs as conserved elements and structures. Here, we emphasize and rewrite two points of Daniel C. Jeffares’s assumptions (Jeffares, Poole
et al. 1998) as (1) Ubiquitous. RNAs that occur in all phylogenetically close species are more likely to be conserved functional RNA; (2) Continuity of function. "Under a Darwinian mechanism any complex structure cannot arise by chance de novo"; thus, IncRNAs from a common ancestor are ought to share either sequence similarity or conserved structure, to keep their functionality, even if the conserved domain is small.

1.4.1 Rapid Evolutionary Turnover of IncRNA Sequences

In contrast to protein-coding mRNAs, most of which are highly conserved in sequence to keep their protein products consistent in function, IncRNA always lack known orthologs cross species. Ulitskey et al. (2011) reported that only 6.7% of zebrafish lincRNAs showed sequence conservation to another zebrafish lincRNA, less than 6% of them had detectable homology with human or mouse orthologs (Ulitsky, Shkumatava et al. 2012), and approximately 12% of human lincRNAs had conserved orthologs in the other species (Church, Goodstadt et al. 2009; Cabili, Trapnell et al. 2011; Ulitsky and Bartel 2013). Kutter et al. (2012) observed that despite of the conserved syntenic sequences, the transcription of the lincRNAs might vary a lot between phylogenetically close rodent species (Kutter, Watt et al. 2012). Only 59.7% of lincRNAs expressed in Mus musculus liver are also expressed in Mus castaneus, compared with 91.7% of mRNA (Kutter, Watt et al. 2012). Complemented with the evidence that human lincRNA expression is strikingly tissue-specific compared with coding genes (Cabili, Trapnell et al. 2011), the high specificity of lincRNAs enable ideal indicators of distinct subpopulation of cells or organisms.

In spite of their limited sequence conservation, lincRNA functionality claims the support of imprinted purifying selection. Ponjavic et al. (2007) observed that, in contrary to the neutralist explanation, IncRNAs from mouse, especially their promoter regions, exhibited suppressed rates
of nucleotide substitutions, insertions and deletions, compared with neighboring ancestral repeats (Ponjavic, Ponting et al. 2007).

Actually, there is one interesting exception of IncRNA rapid evolution. TERRA, previously found in human and yeast (Azzalin, Reichenbach et al. 2007; Schoeftner and Blasco 2008; Luke and Lingner 2009), recently was also identified in Arabidopsis (Vrbsky, Akimcheva et al. 2010). Unlike the localization preference to telomeres in mammalian genome (Azzalin, Reichenbach et al. 2007), Arabidopsis TERRA arises either from telomeres or in the proximity of centromeres (Uchida, Matsunaga et al. 2002; Vannier, Depeiges et al. 2009). What’s more, a subset of these Arabidopsis TERRA transcripts tends to form partially double stranded structure and be processed into siRNAs. These TERRA-produced siRNA are involved in the RNA-dependent DNA methylation pathway and contribute to the maintenance of telomeric chromatin (Vrbsky, Akimcheva et al. 2010).

1.4.2 IncRNA Structure Evolvability

1.4.2.1 Secondary Structure and IncRNA Functions

Primary (sequence of nucleotides), secondary (double-stranded helices) and tertiary (compact and highly-organized) structure contribute to the 3 hierarchical levels of RNA structure organization (Elliott and Ladomery 2010), by virtual of hydrogen bonds on the Watson-Crick face and the Hoogsteen and ribose face (Mercer and Mattick 2013). There are five common secondary structure motifs, including helices, hairpin loops, bulges and pseudoknots (Figure 1.2) (Wuchty, Fontana et al. 1999), while they are connected in the RNA tertiary structure by non-Watson-Crick base-paring (Elliott and Ladomery 2010). Folding patterns of RNAs can be very complex, given that even the 4 core nucleotides have over 100 chemically distinct modified forms (Cantara, Crain et al. 2011).
Secondary structures of non-coding RNAs are very important for their functionality. For instance, the functionally important portions of the "cloverleaf" model of tRNA are the anticodon loop that pairs with a mRNA specifying a certain amino acid, and the amino acid binding stem attached with an amino acid (Elliott and Ladomery 2010). As to lncRNAs, more cases have verified the structure-function relationship for lncRNAs. MEG3 isoforms in human were demonstrated to share three common secondary structural motifs, two of which are important for p53 activation, by bioinformatics prediction (Mfold) and experimental deletion analysis (Zhang, Rice et al. 2010). Furthermore, the replacement of the p53-related motifs with dissimilar sequences that formed the same structures showed no disruption of MEG functioning (Zhang, Rice et al. 2010), which enhanced the conclusion that MEG3 mainly functions dependently on its secondary structure rather than its primary sequence. Mammalian NoRC-associated RNAs (pRNA), required for nucleolar localization of NoRC for rDNA silencing, are poorly conserved in sequence but able to fold into similar stem-loop secondary structures (Mayer, Neubert et al. 2008). The conserved stem-loop is crucial for TIP5 (the large subunit of NoRC) binding, and mutations of this stem-loop impaired the targeting of NoRC to rDNA locus (Mayer, Neubert et al. 2008). Similarly, Xist, HOTAIR and ANRIL recruit polycomb complex for gene silencing by their double stem-loop and other structural motifs (Zhao, Sun et al. 2008; Tsai, Manor et al. 2010; Kotake, Nakagawa et al. 2011). Complementing the enzymatic and chemical probing technologies with covariance analysis, Novikova et al. (2012) experimentally confirmed four functional domains with a variety of secondary structure elements of human SRA1; correspondingly, deletions study and site-directed mutagenesis both proved disruption of function (Novikova, Hennelly et al. 2012).
1.4.2.2 lncRNA Structure Evolvability

An RNA sequence is malleable to fold into a number of combinations of thermodynamically stable helices; vice versa, a similar structure may adopt quite distinct genotypes (Wan, Kertesz et al. 2011). For example, AT-AC and GT-AG snRNAs share well-conserved secondary structure elements critical for activity even if there is no strong homology shown at the sequence level (Meli, Albert-Fournier et al. 2001). As mentioned before, in contrast to the iconic tRNA cloverleaf structure, its nucleotide sequence may vary up to over 90% (Meli, Albert-Fournier et al. 2001; Mercer and Mattick 2013). Thus, indeed, it would not be a surprise to notice the low sequence conservation of lincRNAs cross different species, as mentioned above (Church, Goodstadt et al. 2009; Cabili, Trapnell et al. 2011; Ulitsky, Shkumatava et al. 2012; Ulitsky and Bartel 2013), even in the case of lncRNAs in the same family. Pi-starvation-responsive lncRNAs Mt4 and TPSI1 in plants only have a small region of one of their short ORFs sharing identity in sequence with each other (Burleigh and Harrison 1997), but contain a highly conserved motif intensively complementary to miR-399, which helps them sequester miRNAs away from their mRNA targets (Franco-Zorrilla, Valli et al. 2007).

Briefly, lncRNAs might (1) only conserve in the function level within small stretches of their sequence which are closely related to their functions, and the utility of other parts of their sequences are left unknown; (2) primarily maintenance a particular base-paired structure (Eddy and Durbin 1994), while allow accumulation of mutations at sequence level; (3) vice versa, the allowance of accumulating mutants benefits lncRNAs to re-fold into novel structures to achieve environment adaptation, which could explain their rapid evolvability (Mercer and Mattick 2013).
1.5 My Objectives

The abundance and multiple functions of long non-coding RNAs (lncRNA) in mammalian systems have been one of the most important discoveries in molecular biology in recent years. However, the identification and characterization of lncRNAs in plants, especially cereals, is in its early stages. This thesis is dedicated to identify lncRNAs in four economically important cereal genomes (Zea mays, Sorghum bicolor, Setaria italic and Oryza sativa), evaluate their evolutionary conservation and predict their secondary structures. Characterizing lncRNAs in cereals may reveal previously hidden regulatory networks of crucial cereal developmental processes, such as stress response and reproduction, and facilitate the development of new biotechnological applications for stress response and adaptation, growth control, and yield increment.
Figure 1.1 LncRNA (Red) Categories Based on Their Locations Relative to Nearby Protein-Coding Genes (Green). Intergenic lncRNAs are those transcribed from intergenic regions; Antisense lncRNAs are transcribed from the antisense strand of protein-coding genes; Intronic lncRNAs are produced from introns; Sense lncRNAs are those which overlap with part or cover the entire sequence of a protein-coding gene; Divergent (Bidirectional) lncRNAs are transcripts that initiate in a divergent fashion from the promoter region of a protein-coding gene (typically within a few hundred base pairs from the transcription start site) (Kung, Colognori et al. 2013).
Figure 1.2 RNA Secondary Structure Motifs (Wuchty, Fontana et al. 1999).
CHAPTER 2
GENOME-WIDE IDENTIFICATION AND EVOLUTIONARY ANALYSIS OF LONG NON-CODING RNAs IN CEREALS

1 Ying Sun, Katrien Devos, Liming Cai, and Russell L. Malmberg. To be submitted to The Plant Cell.
2.1 Abstract

We identified 7,196 lncRNA candidates in the cereal *Zea mays*, 1,974 in *Sorghum bicolor*, 4,236 in *Setaria italica* and 2,542 in *Oryza sativa*, using computational methods, then compared these across the species. Our approach involved a reference-guided transcriptome assembly with RNA-Seq data from these four economically important cereals, and screened for RNAs that were at least 200 bases in length, at most 70 amino acids in open reading frames and lack of homology in Uniprot database. A sequence composition analysis of the lncRNA candidates, in comparison to protein-coding transcripts, highlighted specific distinctive features, including low GC content, paucity of introns and hexamer usage bias, which were consistent with what was found in mammalian genomes. RepeatMasker identified from 1% (rice) to 19% (maize) of the candidate lncRNAs as being from transposable elements, based on our dataset with 3,853 transposable elements. We compared the candidate lncRNAs with 25,141 miRNAs from miRBase, and found that less than 1% of them would be potential miRNA precursors. The cross-species comparison, including sequence- and structure-based lncRNA homology search, synteny analysis, and lncRNA secondary structure prediction, uncovered some limited sequence similarity; in sub-regions we predicted conserved secondary structures using covariation analysis. Our results are consistent with a model of rapid evolution of lncRNAs.

2.2 Introduction

Since the breakthrough discovery of RNase P catalytic activity (Guerrier-Takada, Gardiner et al. 1983) led us to the "RNA world" (Gesteland 2005), surprises keep coming. RNAs are no longer confined in the "Central Dogma" (mRNA, rRNA, tRNA) (Crick 1970), and might serve as the cradle of modern life by carrying both genetic information and catalytic capacity (Gesteland 2005; Elliott and Ladomery 2010). Recent studies have revealed a large variety of
non-coding RNAs (ncRNA) with distinct biological features and cellular functions, dramatically expanding the RNA family (Kurth and Mochizuki 2009; Louro, Smirnova et al. 2009; Mercer, Dinger et al. 2009; Ponting, Oliver et al. 2009; Wilusz, Sunwoo et al. 2009; De Lucia and Dean 2011; Jouannet and Crespi 2011; Li and Liu 2011; Kim and Sung 2012; Rinn and Chang 2012; Huang, Zhang et al. 2013; Kung, Colognori et al. 2013; Ma L, Bajic VB et al. 2013). A simple classification based on their sequence length leads to two main categories, including the small ncRNAs and long ncRNAs (lncRNA) with at least 200 bases in length. While the former has been widely studied and proved to play diverse functional roles in the cells (Chen 2009; Kurth and Mochizuki 2009; Li and Liu 2011; Huang, Zhang et al. 2013), the lncRNAs just recently began to shed light on previously hidden layers of gene regulatory networks (Ponjavic, Ponting et al. 2007; Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011; Liu, Jung et al. 2012).

The recent advent of high-throughput transcriptome sequencing technologies, in cooperation with computational analysis, dramatically sped up the search for lncRNAs (Morozova, Hirst et al. 2009). Currently, the GENCODE (Derrien, Johnson et al. 2012; Harrow, Frankish et al. 2012), a sub-project of the human genome encyclopedia ENCODE (Bernstein, Birney et al. 2012), contains 13,562 lncRNA genes producing 23,105 lncRNA transcripts (GENCODE version 18). NONCODE (Bu, Yu et al. 2012) currently consists of 147,444 lncRNAs from various genomes, of which, 82,324 and 43,532 are respectively for human and mouse (NONCODE version 4.0).

In contrast to the intensive studies of lncRNAs in mammalian genomes, the identification and characterization of lncRNAs in plants, especially cereals, is in its early stages. There have been some attempts made to systematically identify and annotate lncRNAs in plants (Hirsch,
Lefort et al. 2006; Ben Amor, Wirth et al. 2009; Georg, Honsel et al. 2010; Xin, Wang et al. 2011; Zhang, Zhao et al. 2011; Boerner and McGinnis 2012; Liu, Jung et al. 2012; Wu, Wang et al. 2013). Xin et al. (2011) identified 125 putative wheat stress responsive lncRNAs, of which 2 are signal recognition particle (SRP) 7S RNA variants and 3 are U3 snoRNAs (Xin, Wang et al. 2011). Liu et al. (2012) characterized 6,480 lincRNAs in *Arabidopsis* from 200 public tiling array datasets using bioinformatics methods (Liu, Jung et al. 2012). These lncRNAs showed limited evolutionary conservation among plant species (Xin, Wang et al. 2011; Liu, Jung et al. 2012), and had tissue-dependent expression patterns similar to what was found in vertebrates (Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011; Ulitsky, Shkumatava et al. 2012). In addition, computational analysis revealed 1,011 putative novel lncRNAs in maize (Boerner and McGinnis 2012). However, few of these analyses approached a genome-wide scale, and even fewer focused on cereals as a group.

The cereals (*Poaceae*) are ideal subjects for comparative molecular evolution studies, given their relatively short speciation time (Figure 2.1 (Devos 2005)), morphological diversity and economic importance. Efforts to generate sequences from *Zea mays*, *Sorghum bicolor*, *Setaria italica* and *Oryza sativa* have produced extensive genomic and transcriptomic sequence resources for these four major cereal crops (Ouyang, Zhu et al. 2007; Paterson, Bowers et al. 2009; Schnable, Ware et al. 2009; Bennetzen, Schmutz et al. 2012; Zhang, Liu et al. 2012). Previous marker-based macro-colinearity studies of grass genomes revealed large synteny blocks among cereals in the form of concentric "crop circles" (Figure 2.2 (Devos 2005)). Meanwhile, current advances in computational comparative genomics (Lyons E 2008; Lee, Tang et al. 2013) allow the detailed characterization of genome-wide rearrangements during evolution besides colinearity. Synteny should be a considerable aide in the lncRNA evolutionary conservation
analysis, considering that IncRNAs always lack known orthologs cross species at the sequence level (Cabili, Trapnell et al. 2011; Ulitsky, Shkumatava et al. 2012); however, even syntenic IncRNAs may exhibit distinct expression levels in phylogenetically close species (Kutter, Watt et al. 2012).

A handful of IncRNAs have been experimentally demonstrated to be recruited in crucial plant developmental processes, such as stress response and reproduction (Zhu and Wang 2012; Zhang and Chen 2013). The TPSI1/Mt4 IncRNA family, including At4 (Burleigh and Harrison 1999; Shin, Shin et al. 2006) and AtIPS1 (Martin, del Pozo et al. 2000) paralogs in Arabidopsis, Mt4 in Medicago truncatula (Burleigh and Harrison 1997), Mt4-like in soybean (Burleigh and Harrison 1999), and TPSI1 in tomato (Liu, Muchhal et al. 1997), respond to phosphate deprivation and function through a highly conserved sequence motif to sequester miRNAs away from their mRNA targets, while the cleavage of itself is prevented by a mismatch loop in the motif (Franco-Zorrilla, Valli et al. 2007). This miRNA pseudotarget activity is the first evidence of "competing endogenous RNA" (ceRNA) (Seitz 2009; Salmena, Poliseno et al. 2011), which reveals a potential bona fide genome-wide regulatory network. Enod40, induced during nodule development (Crespi, Jurkevitch et al. 1994), functions through its highly conserved secondary structure for protein re-localization in leguminous plants (Girard, Roussis et al. 2003; Campalans, Kondorosi et al. 2004). Cold-induced COOLAIR and COLDAIR are associated with regulation of the FLC flowering locus in Arabidopsis (Heo and Sung 2011). COLDAIR is required for triggering vernalization-mediated epigenetic silencing of floral repressor FLC, by targeting PRC2 to FLC chromatin and resulting in H3K27me3 methylation (Heo and Sung 2011).

COOLAIR, distinct from COLDAIR, is suggested to function in early cold induced FLC silencing, based on the observation of its accumulation to a peak around 10-14 days after cold
treatment and declination with prolonged cold (Swiezewski, Liu et al. 2009). Ding et al. (2012) identified expression of a lncRNA with 1,236 bases in length, referred to as long-day-specific male-fertility-associated RNA (LDMAR), regulates photoperiod-sensitive male sterility (PSMS) and is crucial for normal pollen development under long-day conditions in rice (Ding, Lu et al. 2012). Zm401 is a pollen-specific gene in maize. Dai et al. (2007) proved the importance of Zm401 in maize pollen development, based on the observation that the over-expression of Zm401 resulted in abnormal tassels, degenerate anthers and degradation of tapetum, asynchronous fusion of pollen sacs, and aborted pollen grain development (Dai, Yu et al. 2007).

Here, we conducted a reference-guided transcriptome assembly with RNA-Seq data from four economically important cereals, and screened for long ncRNAs. We performed cross-species comparisons to characterize them and to study their evolution, including sequence- and structure-based lncRNA homology search, synteny analysis, and lncRNA secondary structure prediction.

2.3 Results

2.3.1 Transcriptome Assembly and Comprehensive Identification of lncRNAs in Zea mays, Sorghum bicolor, Setaria italica and Oryza sativa

We collected over 4G RNA-Seq reads of Zea mays, 0.5G of Sorghum bicolor, 0.8G of Setaria italica, and 0.9G of Oryza sativa from NCBI Sequence Read Archive (SRA) (Supplementary Table S1). All reads were pre-processed through our RNA-Seq data quality-control pipeline (Figure 2.3 (a)) and passed through the genome-guided transcriptome assembly pipeline based on the "Tuxedo" protocol (Trapnell, Roberts et al. 2012) (Figure 2.3 (b) and Material and Methods). Over 84% of reads from Zea mays, over 83% of reads from Sorghum bicolor, over 75% of reads from Setaria italica, and over 75% of reads from Oryza sativa were
mapped entirely or partially to the reference genome. Expression was detected for 84.8% of annotated exons in *Zea mays*, 83.0% of *Sorghum bicolor*, 81.3% of *Setaria italic*, and 76.8% of *Oryza sativa* (Trapnell, Roberts et al. 2012), while a substantial number of reads fell in unannotated regions.

To comprehensively identify transcripts including previous annotations, we implemented reference annotation based transcript (RABT) assembly (Roberts, Pimentel et al. 2011), and obtained 245,920 cDNA sequences for *Zea mays*, 71,859 for *Sorghum bicolor*, 76,711 for *Setaria italic*, 125,208 for *Oryza sativa*, compared with 63,540, 29,448, 40,599 and 49,061 mRNAs annotated in each genome respectively (Table 2.1). Table 2.1 also shows the number of intronic cDNAs, antisense cDNAs, and intergenic cDNAs found for each species.

To identify lncRNAs, all assembled cDNA sequences were filtered to remove (1) those cDNAs with length shorter than 200 bases, (2) those whose ORFs were longer than 70 amino acids, which was a lower and more stringent value than a previous lncRNA study in maize (Boerner and McGinnis 2012), and (3) those which had homologs in either UniprotKB/Swiss-prot or TrEMBL datasets (Figure 2.4). Among all the assembled cDNAs, 7,196 *Zea mays*, 1,974 *Sorghum bicolor*, 4,236 *Setaria italic* and 2,542 *Oryza sativa* lncRNA candidates were remained after the filtering (Table 2.2 and Table 2.3).

### 2.3.2 Sequence Composition Analysis of lncRNA Candidates

Except for the lncRNA candidates in *Setaria italic* (35.6%), most lncRNA candidates in the other three genomes were spliced (78.1% for *Zea mays*, 86.0% for *Sorghum bicolor*, and 80.2% for *Oryza sativa*) (Table 2.2), similar to what was stated in human genome that most (approximately 98%) lncRNAs are spliced (Derrien, Johnson et al. 2012). Since the number of lncRNAs identified by RNA-Seq are largely influenced by the sequencing depth and sample
variety (Ilott and Ponting 2013), our prediction might underestimate the abundance and diversity of lncRNAs in the four cereal transcriptome, especially for Setaria italica, and introduce the bias as shown above. In addition, 83.0% (5,976 out of 7,196) of Zea mays lncRNA candidates possessed single or two exons, 62.8% (1,239 out of 1,974) of Sorghum bicolor, 87.6% (3,710 out of 4,236) of Setaria italica, and 82.9% (2,108 out of 2,542) of Oryza sativa (Table 2.2), compared with the estimation that only 35.7% of Zea mays annotated mRNAs having less than three exons, and correspondingly 37.1% of Sorghum bicolor, 41.0% of Setaria italica, and 40.0% of Oryza sativa mRNAs, suggesting the paucity of introns of the lncRNA candidates (Two-sample Z-test for proportion, P-value< 2.2e-16).

Generally, the lncRNA candidates were significantly shorter than the annotated mRNAs (Table 2.2; Two-sample Z-test for mean, P-value< 2.2e-16) and the average GC content were slightly lower than that of annotated mRNAs (Figure 2.5; Two-sample Z-test for mean, P-value<2.2e-16, Material and Methods, Supplementary Table S2). Although the relatively low GC content was unexpected (considering higher GC content implies more stable secondary structures with more GC hydrogen bonds), these two features of our lncRNA candidates were also observed in mammalian lncRNAs (Supplementary Table S2) (Niazi and Valadkhan 2012).

We analyzed the hexamer composition of the lncRNA candidates and annotated protein-coding region of mRNAs (CDSs without UTRs) by a sliding window advanced at three-nucleotide steps for CDS and single-nucleotide steps for lncRNA candidates. Out of 4,096 types of hexamers, 33.4% (1368) of Zea mays, 36.7% (1,503) of Sorghum bicolor, 34.4% (1,411) of Setaria italica and 32.3% (1,323) of Oryza sativa showed more than twofold differential representation (either under- or over-representation) in lncRNA candidates and annotated protein-coding CDS sequences (Supplementary Table S3; Figure 2.6). When we did the analysis
with the same single-nucleotide window steps for both CDS and lncRNA candidates, the percentage, out of 4,096 types of hexamers, indicating two or more fold change between CDS sequences and lncRNA candidates would fall to 19.1% of *Zea mays*, 23.2% of *Sorghum bicolor*, 22.9% of *Setaria italic*, and 21.9% of *Oryza sativa* (Supplementary Table S3). However, these percentage values were still much higher than the corresponding percentages in the case of comparison between 3'UTR sequences and lncRNA candidates hexamer usages, with 1.4% for *Zea mays*, 8.4% for *Sorghum bicolor*, 2.0% for *Setaria italic*, and 3.8% for *Oryza sativa* (Supplementary Table S3). As to the 5'UTRs, the percentage of hexamers having over two fold changes were as high as the values obtained in the comparison of CDS sequences and lncRNA candidates, with 14.0% for *Zea mays*, 46.1% for *Sorghum bicolor*, 26.5% for *Setaria italic*, and 22.9% for *Oryza sativa* (Supplementary Table S3).

### 2.3.3 Categorization of lncRNA Candidates Based on Genomic Location and Context

Analyzing the genomic location and context of lncRNAs can help us to predict their functional roles and regulatory relationships with nearby genes (Kung, Colognori et al. 2013; Ulitsky and Bartel 2013). According to their relative positions to adjacent annotated genes, we classified the lncRNA candidates into sense-overlap, intronic, antisense and intergenic, resulting the following distribution as shown in (Table 2.3). Considering that there might be un-annotated alternative spliced transcripts in the transcriptome, it is difficult to predict the exact boundaries of each protein-coding locus; and given that changes of annotated genes (protein-coding or not) might be made with the progressing annotation projects for each genome, we currently just took all the annotated genes in the four genomes as putative protein-coding anchors for our classification and defined the boundaries of each gene from the start of 5'UTR to the end of 3'UTR.
Among the lncRNA candidates that overlap an annotated gene in the same strand, 1,141 of *Zea mays*, 14 of *Sorghum bicolor*, 2,890 of *Setaria italica* and 155 of *Oryza sativa* were found to be intersecting with exons of an annotated genes. We checked the annotation of these genes and found that: First, among these annotated genes, 1,125 (out of 1,141) of *Zea mays*, 13 (out of 14) of *Sorghum*, 2,886 (out of 2,890) of *Setaria italica* and 155 (out of 155) for *Oryza sativa* were marked as "there are no functional annotations for this locus", which implies their potential as non-coding RNAs. As to the other 16 (=1141-1125) annotated genes in *Zea mays*, 12 were alternative-spliced isoforms with small ORF and the functional annotations were for the primary transcript with larger ORFs (Figure 2.7 and Supplementary Table S4); 1 were annotated as "Protein of unknown function", 3 were putative peptides without start codons. Similarly, 1 (=14-13) annotated gene in *Sorghum bicolor* was a putative peptide without a start codon but which showed homology to the 4F5 protein family. 1 out of 4 (=2,890-2,886) annotated gene in *Setaria italica* was an alternative spliced isoform, 1 was a putative peptide without start codon, and 2 were short peptides (one belonged to small peptide DVL family and the other was identified as a member of Ctr copper transporter family). Secondly, 92.8% (1,059 out of 1,141) of *Zea mays* annotated genes, which overlap with the lncRNA candidates, only have single or two exons, 78.6% (11 out of 14) of *Sorghum bicolor*, 95.8% (2,770 out of 2,890) of *Setaria italica* and 78.1% (121 out of 155) of *Oryza sativa*. Although the bias towards fewer exons was not a sufficient condition for identifying lncRNAs, the features of these annotated genes showed some additional clues of the non-coding potential.

The majority of the lncRNA candidates in *Zea mays* and *Sorghum bicolor* were intergenic (3398 out of 7196 and 1271 out of 1974, respectively); while antisense lncRNAs account for a large amount of *Setaria italica* and *Oryza sativa* lncRNA candidates. The
intergenic lncRNA candidates were further categorized into divergent or 1-5kb apart from their adjacent annotated genes. There were 143 of *Zea mays*, 39 of *Sorghum bicolor*, 32 of *Setaria italica* and 47 of *Oryza sativa* lncRNA candidates were identified as intergenic divergent transcripts (If we counted the ones overlapping with annotated genes, there would be more divergent lncRNAs candidates transcribed head to head with an annotated gene within 1 kb).

There were 1,829 *Zea mays* lncRNA candidates, 486 *Sorghum bicolor*, 522 *Setaria italica* and 1,289 *Oryza sativa*, respectively, with at least 100nt of antisense overlapping with annotated genes. Previous studies exhibited that NATs can function either via transcriptionally influencing the expression of the antisense gene, or by masking splicing sites through base complementarity that consequently affects the alternative splicing of their overlapping transcripts (Gu, Zhang et al. 2009). Further expression level analysis may help to understand the correlation between the lncRNA candidates and their antisense protein-coding counterparts.

**2.3.4 The Majority of lncRNA Candidates Are Not Associated with Transposable Elements**

Transposable elements (TEs) are DNA fragments that can change their positions within the genome, duplicating themselves, and have been demonstrated to be "the single largest component of the genetic material of most eukaryotes" (Feschotte, Jiang et al. 2002). The first evidence of TE in maize was discovered by Barbara McClintock decades ago, and currently it has been reported that 85% of the maize genome are made of TEs (Schnable, Ware et al. 2009). Recently, accumulating evidence indicates a close association of TEs and ncRNAs, with the fact that a good many of small ncRNAs derived from TEs, including those gene-silencing mediators; in addition, Alu and LTR were found embedded in lncRNAs and were crucial for their functioning (Hadjiargyrou and Delihas 2013).
We used RepeatMasker (Smit 1996-2010) together with a combined maize transposable element database to determine if any of the lncRNA candidates were from transposons. We compiled 3,948 transposable elements from two public sources (1,526 from maizetdb.org and 2,422 from RepBase18.01 grasrep.ref) using CD-HIT-EST (Li and Godzik 2006) with 100% identity to obtain a non-redundant dataset with 3,853 transposable elements. RepeatMasker identified from 1% (Oryza sativa) to 19% (Zea mays) of the candidate lncRNA sequences as being from transposable elements (Table 2.4), while there are 40.2% (2,894 out of 7,196) lncRNA candidates of Zea mays, 28.5% (562 out of 1,974) of Sorghum bicolor, 12.2% (515 out of 4,236) of Setaria italic, and 7.7% (195 out of 2,542) of Oryza sativa were detected to have partial TE segments embedded in their sequences (Table 2.5 and Supplementary Table S5).

2.3.5 A Small Number of The lncRNAs May Be Small ncRNA Precursors

Small regulatory ncRNAs play an essential role in gene silencing, either by guiding heterochromatin formation at homologous loci, or by post-transcriptional mRNA degradation or translational inhibition (Chen 2009). Many lncRNAs may serve as precursors for small ncRNAs. LncRNAs in this group generally have no intrinsic functions of their own and will experience degradation after being processed by Dicers (Kapranov, Cheng et al. 2007; Fejes-Toth K 2009; Wilusz, Sunwoo et al. 2009).

To evaluate the miRNA precursor potential of our lncRNA dataset, we compared candidate lncRNAs with 25,141 miRNAs from miRBase (http://www.mirbase.org/ Release 19) to see if any were concatenated miRNAs. BLAST search showed that less than 1% of the candidate lncRNAs we detected contained miRNA with the cutoff that the coverage of miRNA sequence should be over 80% and the percentage of identity over 90%. Actually, unlike being the target of miRNA in plants, to be a precursor required exact matching between query
sequence and the pre-miRNA hairpin sequence. Only 26 (out of 7,196) of Zea mays lncRNA candidates, 8 (out of 1,974) of Sorghum bicolor, 3 (out of 4,236) of Setaria italica, and 19 (out of 2,542) of Oryza sativa showed potential to be miRNA precursors (Supplementary Table S6).

**2.3.6 Some lncRNA Candidates Showed Limited Cross-Species Sequence Homology**

We started by investigating the cross-species sequence similarity of the lncRNA candidates among Zea mays, Sorghum bicolor, Setaria italica and Oryza sativa. The reciprocal BLASTN (Altschul, Madden et al. 1997) search over TE-masked lncRNA candidates initially discovered 147 homologous pairs of Zea mays and Sorghum bicolor, 85 of Zea mays and Setaria italica, 46 of Zea mays and Oryza sativa, 63 of Sorghum bicolor and Setaria italica, 18 of Sorghum bicolor and Oryza sativa, and 29 of Setaria italica and Oryza sativa (Table 2.6 and Supplementary Table S7). Further, there were 29 homologous triplets of Zea mays, Sorghum bicolor and Setaria italica, and there was no quadruplets was found among four species. Synteny Analysis (Figure 2.8; Material and Methods) proved 119 (out of 147) pairs, 51 (out of 85), 30 (out of 46), 56 (out of 63), 15 (out of 18), and 15 (out of 29) were syntenic respectively for the six species pairs listed above in the same order, as well as 21 (out of 29) syntenic triplets of Zea may, Sorghum bicolor and Setaria italica (Table 2.6 and Supplementary Table S7), which suggested lncRNA genes underlying orthologous traits. The homologous pairs showed limited similarity with an average 18% ~42% of lncRNA sequence being aligned (Supplementary Table S7).

Boerner et al. (2012) identified 1,011 putative lncRNAs from Zea mays EST dataset using bioinformatics pipelines (Boerner and McGinnis 2012). Since they used relatively loose constraints in their pipeline (ORF≤120 amino acids and only search homologs in Swissprot database) and CPC as a complement to their pipeline, we checked whether there was any overlap between these two datasets and found that 122 out of the 1,011 showed limited homology to our
IncRNA candidates in *Zea mays*, 14 in *Sorghum bicolor*, 18 in *Setaria italica*, and 6 in *Oryza sativa*; among which, only 1 pair of maize IncRNAs was 100 percent identical and another 6 IncRNAs from the 1,011 dataset were part of our IncRNAs candidates (Supplementary Table S8). To verify the unexpected lack of overlap between the 1,011 putative maize IncRNA dataset and our maize IncRNA candidates, we performed BLASTN again on the 1,011 putative maize IncRNAs and all the assembled transcripts from our dataset. It turned out that 684 out of the putative 1,011 maize IncRNAs were included in our assembly, of which, 35 found identical transcripts in our assembly dataset and the rest seemed to be novel isoforms or incomplete EST assemblies; another 249 (out of 1,011) partially overlap with our assembled transcripts (Supplementary Table S8). Thus, it might be our more stringent threshold (ORF≤70 amino acids) that led to the very limited similarity between the two candidate IncRNA datasets.

Another sequence homolog search against the 6,480 *Arabidopsis* lincRNA (Liu, Jung et al. 2012) dataset revealed that 37 (out of 6,480) *Arabidopsis* lincRNA exhibited an average of 45nt alignment with our IncRNA candidates of *Zea mays*; 23 exhibited an average of 31nt alignment with our IncRNA candidates of *Sorghum bicolor*; 29 showed an average of 32nt alignment with our IncRNA candidates of *Setaria italica*; and 36 showed an average of 30nt alignment with our IncRNA candidates of *Oryza sativa* (Supplementary Table S9). Compared with the average length of alignment shared between our IncRNA candidates (Supplementary Table S7), the limited similarity at the sequence level between the 6,480 Arabidopsis lincRNAs and our IncRNA candidates strongly suggested the propensity of poor sequence conservation of IncRNAs cross-species, especially for those phylogenetically distant ones.
2.3.7 Structure-Based IncRNA Homology Search

Secondary structure is known to be associated with IncRNA functionality (Mayer, Neubert et al. 2008; Zhao, Sun et al. 2008; Tsai, Manor et al. 2010; Zhang, Rice et al. 2010; Kotake, Nakagawa et al. 2011). Given the structural conservation shared by central small ncRNAs (Jeffares, Poole et al. 1998; Meli, Albert-Fournier et al. 2001), we expect that the natural selection pressure has been reflected on IncRNAs as conserved elements and structures to maintain their continuity of functions, despite the lack of sequence homology.

Covariance models (CM) (Eddy and Durbin 1994) have been widely used to implement genome-wide computational screens and annotation for conserved RNA secondary structures (Eddy 2002; Weinberg and Ruzzo 2004; Weinberg and Ruzzo 2006; Freyhult, Bollback et al. 2007; Nawrocki, Kolbe et al. 2009; Huang 2010). CMs describe a particular capacity of RNA that strongly correlated base pairs in sequence tend to change specifically to maintain the secondary structure.

We implemented Infernal (Nawrocki, Kolbe et al. 2009), a CM-based RNA homology search program, for searching our candidate IncRNA database (Table 2.2) for conserved RNA structures to any of the 225 IncRNA-family-alignment profiles from Rfam (Burge, Daub et al. 2013) (Supplementary Table S10).

Only 18 IncRNA candidates (5 from Zea mays, 1 from Sorghum bicolor, 5 from Setaria italica and 7 from Oryza sativa) exhibited significant homology to 9 Rfam IncRNA families, and only 2 Rfam IncRNA families had cross-species hits (Table 2.7). They were RF02189 family (human IncRNA ST7 overlapping transcript 4 conserved region 3 (Vincent, Petek et al. 2002)) homologous to IncRNA Maize_TCONS_00188946 and Oryza_TCONS_00032931, and RF02247 family (mouse IncRNA Six3os1 conserved region 2 (Alfano, Vitiello et al. 2005))
homologous to Maize_TCONS_00199607, Maize_TCONS_00012212, Maize_TCONS-_00067710, Maize_TCONS_00151483 and Sorghum_TCONS_00046394. Neither the Maize_TCONS_00188946 and Oryza_TCONS_00032931 pair, nor the Maize_TCONS_-
00199607/00012212/00067710/00151483 and Sorghum_TCONS_00046394 pair were identified in the cross-species sequence similarity search as shown above. This suggested that structure-based RNA homology search could be a good complement of sequence similarity comparison. Although none of these cross-species structure-based IncRNA homologs were from syntenic regions based on information obtained from CoGe Platform (genomevolution.org), the conserved structures of the 9 Rfam IncRNA families (Figure 2.9 and Supplementary Figure S1) gave us some insight into the secondary structures and potential functionality of their IncRNA homologs from the four cereals.

The IncRNA sequence homology analysis above uncovered 21 homologous triplets of Zea may, Sorghum bicolor and Setaria italica from syntenic regions of these three genomes, suggesting orthologous traits (Table 2.6 and Supplementary Table S7). We selected the longest isoform as the representative sequence of a single transcription locus, and obtained 7 distinct orthologous IncRNA triplets of Zea may, Sorghum bicolor and Setaria italica. Accordingly, another 7 covariance models were built based on the orthologous triplets and then used to search all the assembled cDNA sequences of Oryza sativa and the 6,480 Arabidopsis lincRNAs for RNA structure homolog by Infernal (Nawrocki, Kolbe et al. 2009) (Material and Methods). There was no homolog found among the 6,480 Arabidopsis lincRNAs, but several putative homologs were found in Oryza sativa (Table 2.8). As mentioned above, no quadruplets was found among four species based on sequence similarity comparison. However, the model field in the results were all marked as “hmm”, because for models with zero base pairs, the cmsearch
function of infernal uses a profile hidden markov model (HMM) instead of a CM for final hit scoring. Thus, these results did not provide any conserved structural information.

### 2.3.8 IncRNA Secondary Structure Prediction

Computational de novo prediction of structural ncRNAs from multiple alignments has been successfully implemented for genome-wide screen of functional elements (Washietl, Hofacker et al. 2005; Pedersen, Bejerano et al. 2006; Parker, Moltke et al. 2011; Smith, Gesell et al. 2013), inspiring the development of structure search programs, such as RNAz (Washietl and Hofacker 2007; Gruber, Findeiss et al. 2010), EvoFold (Pedersen, Bejerano et al. 2006), RNAcode (Washietl, Findeiss et al. 2011). However, the usage of these programs on cereal genomes has been hampered by the lack of multiple alignment of whole-genome sequences. Instead, we predicted IncRNA secondary structures by using the 7 distinct multiple alignments of orthologous IncRNA triplets obtained from sequence homology search and synteny analysis.

The RNAz program detected 2 groups of IncRNA triplet alignment exhibiting a high probability (>0.53 and >0.79 respectively) to have stable conserved structures, when we used the RNAz algorithm with parameters set to work globally and to score the given alignment as a whole (Figure 2.10 by RNAalifold (Lorenz, Bernhart et al. 2011)). However, under the recommended "sliding-window" mode with a window size of 120 and a step-size of 40, no structural conservation was detected.

### 2.3.9 PCR Experimental Verification

To minimize the influence of the depth of sequencing and the variety of sample on IncRNA identification and homology search, we would conduct PCR validation to investigate the expression of potential RNAs in Setaria italica, which were orthologous to IncRNAs in both Zea mays and Oryza sativa but missing in the current dataset. Primers were designed to test 19
transcripts which were computationally predicted to express in *Setaria italica* (Table 2.9), based on the 30 lncRNA homologs from syntenic regions of *Zea mays* and *Oryza sativa* as shown above. The experiments are in progress.

**2.3.10 lncRNA Functionality Prediction by UniformMu Insertion**

UniformMu (McCarty, Settles et al. 2005) is a unique maize population of mutator (Mu) transposable elements (Lisch, Chomet et al. 1995) developed for experimental analysis of maize gene functions. To understand the potential functionality of the lncRNA candidates in *Zea mays* and their homologs in *Sorghum bicolor, Setaria italica* and *Oryza sativa*, we searched for maize lncRNA candidates with UniformMu mutant insertion locus in their transcription boundaries (from 150bp upstream to 150bp downstream of lncRNA transcription locus), and found 1,494 (out of 7,196) lncRNAs containing potential Mu insertion loci, and 945 of them consisted of Mu insertion loci in their exons (Figure 2.11, Supplementary Table S12 and Supplementary Figure S2).

We focused on maize lncRNAs which had orthologs in *Oryza sativa*. Based on the sequence homology and synteny analysis as shown above, 9 of these conserved maize lncRNA candidates, with a total of 10 UniformMu insertions were selected (Table 2.10) for lncRNA function analysis, and the maize seeds with specific mutations were ordered and ready for use. The phenotypic consequence of mutations might provide us some clues of the lncRNA functionality in *Zea mays*.

**2.4 Discussion**

**2.4.1 There Are Thousands of lncRNAs in These Species**

Analysis of these cereal lncRNAs and their comparison to protein-coding annotations uncovered some significant differences of sequence composition between the two groups.
LncRNAs tended to be spliced but have less exons in comparison with protein-coding RNAs, which is consistent with previous report in human transcriptome that lncRNAs display a striking bias toward two-exon transcript (42% in contrary to 6% of protein-coding genes) (Derrien, Johnson et al. 2012). LncRNAs also showed Hexamer usage biases similar to 3'UTRs of protein-coding RNAs, but distinct from 5'UTR and CDS regions. These variations of cDNA base composition were consistent with sequence features of mammalian lncRNAs (Niazi and Valadkhan 2012). Niazi et al. (2012) revealed that lncRNAs and the 3'UTR sequences shared great similarities both in structural features and sequence composition, and identified that the difference in hexamer composition between ORFs and lncRNAs was about twice of that in 5'UTRs versus lncRNAs, and much higher than that in 3'UTRs versus lncRNAs, suggesting distinct sequence composition of coding and non-coding RNAs (Niazi and Valadkhan 2012). Further, a statistically significant lower GC content of lncRNAs than protein-coding RNAs was found. These could give us some clues about their biological characteristics, such as low expression level. A positive correlation between the GC-content of genes and their expression level has been observed in several studies (Lercher, Urrutia et al. 2003; Semon, Mouchiroud et al. 2005; Das, Roymondal et al. 2009).

Jia et al. (2010) found that hundreds of known hypothetical-protein genes in humans, which are in the vicinity of lncRNAs, are lack of coding potential, based on their ORF-prediction and protein homology search pipeline, and then suggested that 62% of the "hypothetical protein" genes are potentially non-coding (Jia, Osak et al. 2010). If this also applies to plants, our sense lncRNAs, especially those completely overlapped ones, suggested that some of their overlapping annotated genes might also be non-coding rather than coding. Thus, the identification of lncRNA is important to improve the annotation of coding regions of genomes.
Divergent transcription over short distances was suggested to be crucial for granting access of TFs to control elements that reside upstream of core promoters and may help promoter regions maintain a state poised for subsequent regulation (Core, Waterfall et al. 2008; Seila, Calabrese et al. 2008). Sigova et al. (2013) showed that over 60% of IncRNA generated from human and murine embryonic stem cells originated from divergent transcription at promoters of active protein-coding genes, and exhibited coordinated changes in transcription with their protein-coding counterpart when embryonic stem cells are differentiated into endoderm (Sigova, Mullen et al. 2013). Thus, it would not be surprising if our divergent IncRNA candidates also play a role in regulating the transcription of their nearby protein-coding genes in cis.

Kelley et al. (2012) revealed that in sharp contrast to protein coding genes, 83% of human lincRNAs are associated with TEs and TEs comprise 42% of human lincRNA sequences (Kelley and Rinn 2012). The majority of these lincRNAs are associated with human endogenous retrovirus (HERV) LTRs and the rest overlap LINE or SINE elements (Kelley and Rinn 2012). They also observed HERVH transcriptional regulatory signals correlates strongly with stem cell-specific expression of lincRNAs (Kelley and Rinn 2012). There was also example of the importance of SINE/Alu repeats in the process of antisense IncRNA controlling mRNA translation (Carrieri, Cimatti et al. 2012). Although the majority of our IncRNA candidates are not associated with TEs, the association between TEs and IncRNAs might shed a light on the functions of our IncRNA candidates.

Few of our IncRNA candidates had the potential to be post-processed into small RNAs. Similarly, in human genome, Derrien et. al. (2012) found only 5% of small RNAs (rRNAs, miRNAs, snRNAs, and snoRNAs) fell into the boundaries of 4% of IncRNAs, compared with the observation that 27% of small RNAs were mapped within the genetic region of 7% protein-
coding genes (Derrien, Johnson et al. 2012). And only 2.5% of the 6,480 lincRNA identified in Arabidopsis were associated with small RNA (Liu, Jung et al. 2012). However, it is still far from the conclusion that IncRNAs are less likely to host small RNAs, given the potential large amount of IncRNAs uncovered.

### 2.4.2 Evolutionary Conservation of IncRNAs in Closely Related Cereal Genomes

Previous studies have indicated that only 6.7% of zebrafish lincRNAs showed sequence conservation to another zebrafish lincRNA, less than 6% of them had detectable homology with human or mouse orthologs (Ulitsky, Shkumatava et al. 2012), and only approximately 12% of human lincRNAs had conserved orthologs in the other species (Church, Goodstadt et al. 2009; Cabili, Trapnell et al. 2011; Ulitsky and Bartel 2013). Kutter et al. (2012) observed that despite of the conserved syntenic sequences, the transcription of the lincRNAs might vary a great deal between phylogenetically close rodent species (Kutter, Watt et al. 2012).

The reconstructed cereal IncRNA candidates allow us to assess the evolutionary sequence conservation. At the sequence level, the IncRNA candidates showed some but very limited sequence similarity between Zea mays, Sorghum bicolor, Setaria italica and Oryza sativa, and shared no common with the Arabidopsis lincRNAs. This result could be explained either by the extreme rapid evolution of IncRNAs even between phylogenetically close-related species (Church, Goodstadt et al. 2009; Cabili, Trapnell et al. 2011; Ulitsky and Bartel 2013), or by the possibility that the expression level of some IncRNAs are too low to be captured in the RNA-Seq experiment. Correspondingly, we have designed PCR experiments to test the existence of potential IncRNAs in Setaria italica whose orthologous IncRNAs had been detected in Zea mays and Oryza sativa, and these experiments are in progress. Furthermore, more future work will be
directed toward evaluating selective pressure upon these lncRNAs and their syntenic regions in close related species, such as substitution rate (Ponjavic, Ponting et al. 2007).

Given "the RNA world" assumption and structural conservation shared by central small ncRNAs (Jeffares, Poole et al. 1998; Meli, Albert-Fournier et al. 2001), we expect that the natural selection pressure has been reflected on lncRNAs as conserved elements and structures. The structure-based RNA homology search yielded a handful of the lncRNA candidates, of which, sub-regions showed highly conserved secondary structure with known vertebrate lncRNA families. However, no obvious covariance model has been detected based on the multiple alignment of orthologous lncRNAs from *Zea mays*, *Sorghum bicolor* and *Setaria italica*.

Only two orthologous triplets of the lncRNA candidates in *Zea mays*, *Sorghum bicolor* and *Setaria italica* were considered to have putative stable conserved structures by RNAz. An example of bioinformatics structure prediction combined with experimental testing is the MEG3 in human. MEG3 isoforms were demonstrated to share three common secondary structural motifs, two of which are important for p53 activation, by bioinformatics prediction (Mfold) and experimental deletion analysis (Zhang, Rice et al. 2010). Furthermore, the replacement of the p53-related motifs with dissimilar sequences that formed the same structures showed no disruption of MEG3 functioning (Zhang, Rice et al. 2010). This reinforces the conclusion that MEG3 mainly functions dependently on its secondary structure rather than its primary sequence.

Many known RNAs maintain a particular base-paired structure (Eddy and Durbin 1994) and simultaneously accumulate mutations at sequence level. For lncRNAs, we might expect a more complicated situation in which the accumulated mutants permit lncRNAs to re-fold into novel structures which have some environmental adaptation (Mercer and Mattick 2013).
2.4.3 Functionality of lncRNAs

Although thousands of lncRNAs have been detected in human and mouse genomes, the functions of the vast majority of them are still elusive. In our study, we designed mutagenesis experiments using UniformMu insertions to check the phenotypic consequence of mutations at the lncRNA transcription sites and hope this will provide us some clues of their functionality and the continuity of function during evolution with the aid of homology search.

In the view of the evidence that co-expressed and/or co-functional genes tend to be clustered along the genome (Hurst, Pal et al. 2004), we performed gene ontology enrichment analysis of protein-coding genes adjacent to the 1,494 lncRNA candidates (10 genes upstream and 10 genes downstream) using agriGO (Du, Zhou et al. 2010); but no enriched GO terms was identified in the neighborhood.

2.4.4 Summary

We identified 7,196 lncRNA candidates in *Zea mays*, 1,974 in *Sorghum bicolor*, 4,236 in *Setaria italica* and 2,542 in *Oryza sativa*. A small subset of these showed sequence or syntenic conservation, allowing us to suggest they are orthologs. Our current data provides a good resource for future studies of cereal lncRNA evolution and function. Accordingly, characterizing lncRNAs in cereals may reveal previously hidden regulatory networks of crucial cereal developmental processes, such as stress response and reproduction, and facilitate the development of new biotechnological applications for stress response and adaptation, growth control, and yield increment.
2.5 Material and Methods

2.5.1 Data Collection, Quality Control and Transcriptome Assembly

We initially collected 4795.9M *Zea mays* RNA-Seq reads of 106 samples, 761.3M *Sorghum bicolor* RNA-Seq reads of 25 samples, 1068.2M *Setaria italica* RNA-Seq reads of 17 samples and 1731.4M *Oryza sativa* RNA-Seq reads of 96 samples (Supplementary Table S1) from NCBI Sequence Read Archive (SRA), which had been sequenced using either Illumina Genome Analyzer (II/IIx) or Illumina HiSeq 2000. The quality evaluation of the high throughput sequencing data was performed using FASTQC (Version 0.10.1). We removed the adapter and PCR primer remnants using CutAdapt (v1.1), and filtered low-quality reads via the Fastx-Toolkit (Version 0.0.13). For paired-end reads, we synchronized the left and right reads before assembly, and removed those whose corresponding mate did not pass the quality control. Finally, we obtained 4157.9M high-quality reads for *Zea mays*, 555.4M for *Sorghum bicolor*, 836.4M for *Setaria italica* and 911.9M for *Oryza sativa*.

All sequenced reads were mapped to their reference genome (downloaded from Phytozome version 9.0, Supplementary Material and Methods), using the spliced read mapper TopHat (v2.0.6) (Trapnell, Pachter et al. 2009). We implemented the "Discovery Mode" (Zhang 2012) of RNA-Seq read alignment with two iterations of TopHat. The main purpose of the first TopHat implementation is to predict the potential splice junctions. And the concatenation of the splice junction.bed files from all the samples is used as the input of the second TopHat re-alignment for each sample. This specific mode provides more comprehensive possible intron information for reads mapping. The parameters for TopHat command line were modified based on the properties of each sample, since the default sets are for the good of mammalian transcriptome assembly. Commonly, "microexon-search" was used to identify alignments.
incident to microexons; "b2-very-sensitive" was used to increase the alignment sensitivity; "min-intron-length" requires 5 (Hansey, Vaillancourt et al. 2012); the default values of "segment-length" as 15 and "segment-mismatches" as 1 were taken in most cases, except for short reads less than 45 bp (suggested by TopHat 1.3.1 release announcement that half the read length for "segment-length" and 0 for "segment-mismatches"); "j" junction files was referred for the second iteration. As to paired-end reads, "mate-std-dev" and "mate-inner-dist" were obtained from the library information provided by SRA for each sample; "no-discordant" required paired-end reads to be mapped concordantly; and "library-type fr-unstranded" indicated the type of the library.

We applied Cufflinks (version 2.0.2) (Trapnell, Williams et al. 2010) without reference annotations (also in the "Discovery Mode") for transcriptome assembly. Parameters were used as default, except for "min-intron-length", "mate-std-dev" and "mate-inner-dist", adjusted in accordance as appropriate for each samples. Next, Bedtools (Quinlan and Hall 2010) was use to compare the Cufflinks gtf files and the reference genome annotation, in order to check the assembly coverage of each annotated gene. We introduced the reference-annotation-based-transcript assembly (RABT) (Roberts, Pimentel et al. 2011) method into the Cuffmerge step of our pipeline, given that around 25% annotated exons of Oryza sativa were missing in the Cufflinks gtf files. The RABT method generates faux-reads from reference-annotated transcripts, and tiles them along the reference genome as RNA-Seq reads. RABT can help to identify features that are missing due to read coverage gaps. Finally, Cuffmerge merges the resulting assemblies of Cufflinks for each sample into a whole.

2.5.2 Identification of lncRNAs by ORF-Predictor/BLASTX Homology Search Pipeline

The standards for our lncRNA pipeline were: (1) transcript length ≥ 200 nt; (2) ORF size ≤ 70aa; (3) no homologs in UniprotKB database (both Swiss-Prot and TrEMBL datasets). Our
pipeline worked in this way: After excluding all the transcripts with length less than 200nt, we used the cross-platform bioinformatics software UGENE (Okonechnikov, Golosova et al. 2012) for ORF prediction. Only transcripts whose maximum ORF length were no longer than 70aa are maintained. The 70aa threshold for ORF length was chosen to reduce potential false positives, based on a statistical analysis we performed of the Swissprot dataset that approximately 95% sequences in this manually-annotated-and-reviewed dataset are longer than 70aa. Later, in the homology search part, BLASTX was performed twice respectively on UniprotKB/ Swiss-Prot and UniprotKB/TrEMBL, and only transcripts which did not have homologs in either dataset were kept, with the parameters: strand="plus" (we only needed to check the translation on plus strand), max_target_seqs=1 (if any homolog was detected, then the transcript would be removed), evalue=0.001 (significance level for the blast search).

2.5.3 Sequence Composition Analysis

To reduce the bias caused by the discrepancy between the numbers of annotated mRNAs and our lncRNA candidates, we calculated the proportion instead of the frequency of transcripts with a certain GC content. Only annotated protein-coding sequences (only CDS without UTR) were used for calculating the GC content of annotated mRNAs, in order to reduce the influence of non-coding UTR sequences. In addition, a two-sample z-test was performed to compare the mean of GC content between annotated mRNAs and lncRNA candidates.

The hexamer usage of annotated protein-coding sequences (only CDS without UTR) and lncRNA candidates were calculated by a sliding window advanced at three-nucleotide steps (one codon) or single-nucleotide steps for CDS sequences and single-nucleotide steps for lncRNAs. R was used to compute the hexamer usage fold changes.
2.5.4 Preparation of Transposable Element Dataset and Perform of RepeatMasker

CD-HIT-EST (Li and Godzik 2006) clusters similar proteins (DNAs) into clusters that meet a user-defined similarity threshold. We compiled 3948 transposable elements from two public sources (1526 from maizetdb.org and 2422 from RepBase18.01 grasrep.ref) using CD-HIT-EST (Li and Godzik 2006) with 100% identity, to obtain a non-redundant dataset with 3853 transposable elements. RepeatMasker was run with default parameters on the merged TE dataset.

2.5.5 miRNA Precursor Detection

25,141 miRNAs as well as pre-miRNA hairpin sequences were downloaded from miRBase (http://www.mirbase.org/ Release 19). The hairpin RNA sequences were transformed into cDNAs before BLAST comparison. A perl script was written to parse the BLAST result for hits with high coverage (>=80%) and high percentage of identity (>=90%).

2.5.6 Sequence Homologous Analysis of lncRNA Candidates

2.5.6.1 Reciprocal BLASTN Search and Synteny Analysis among Candidates

All lncRNA candidates were pre-processed by RepeatMasker and transposable elements were masked based on our TE dataset. Reciprocal BLASTN search were performed on each pair of the four genomes (Zea mays versus Sorghum bicolor, Zea mays versus Setaria italica, Zea mays versus Oryza sativa, Sorghum bicolor versus Setaria italica, Sorghum bicolor versus Oryza sativa, Setaria italica versus Oryza sativa, and vice versa), with the parameters: strand="plus" and evalue="0.001". The initial purpose of Reciprocal BLASTN search, instead of single-orientation BLASTN, was to discover the best BLAST hit pairs between every two genomes. However, considering the existence of homologs caused by gene duplication after speciation (which is common in cereal genomes) as well as transcript isoforms, we kept all the multiple-hits pairs, if (1) transcript a1 in genome A obtained multiple top hits in genome B (b1, b2 and b3),
(b1, b2 and b3) were isoforms transcribed from the same lncRNA gene locus, and a1 was the best hit of all (b1, b2 and b3) in the reverse BLASTN search; (2) transcript a1 had multiple top hits in genome B (b1 and b2), the matched region of both hits were similar and a1 was the best hit of all (b1 and b2) in the reverse BLASTN search.

Whole genome synteny information of the four genomes was extracted from SynMap of the CoGe (Comparative Genomics) Platform (genomevolution.org) with all default settings (Supplementary Material and Methods). We defined a lncRNA homologous pair as synteny pair if the lncRNA genes were located in the same synteny block between the two gnomes with their adjacent upstream and downstream protein-coding gene forming un-disturbed protein-coding gene ortholog anchors (Figure 2.8).

2.5.6.2 Sequence Homolog Search in Other Datasets

*Homologs of the 1011 lncRNAs in Zea mays* (Boerner and McGinnis 2012). We performed BLASTN on the 1011 maize lncRNAs identified in (Boerner and McGinnis 2012) against the lncRNA candidates from the four cereal genomes with parameters: task="dc-megablast", strand="plus" and evalue="0.001", max_target_seqs="1".

*Homologs of the 6480 Arabidopsis lincRNAs* (Liu, Jung et al. 2012). We performed BLASTN on the 6480 Arabidopsis lincRNAs (Liu, Jung et al. 2012) against the lncRNA candidates from the four cereal genomes with parameters: task="blastn", strand="plus" and evalue="0.001", max_target_seqs="1". Considering the phylogenetic distance between Arabidopsis and the cereals, "blastn" was used instead of "dc-megablast".

2.5.7 Infernal Search for Conserved RNA Structure and Sequence similarities

The Stockholm alignment files of the “Seed” sequences (the alignment of the representative sequences) in 225 lncRNA families were downloaded from Rfam v11.0 via
BioMart (Burge, Daub et al. 2013) (Supplementary Table S10). The cmbuild and cmcalibrate functions of Infernal 1.1rc2 were used to build a CM (covariance model) database containing 225 Rfam-IncRNA-family-multiple-sequence-alignment profiles, and then cmsearch function was performed to search our candidate IncRNA database (Table 2.2) for RNAs homologous to any of the 225 profiles in the CM database.

The 21 homologous triplets of *Zea may*, *Sorghum bicolor* and *Setaria italica* from syntenic regions of these three genomes (Table 2.6 and Supplementary Table S7) were analyzed to only keep the longest isoform as the representative sequence of a single transcription locus. Next, 7 distinct orthologous IncRNA triplets were selected, substituting T with U to make cDNA sequences into RNA sequences, aligned by Clustal Omega with the "RNA" option (Sievers, Wilm et al. 2011) and built into another 7 covariance models using cmbuild and cmcalibrate functions of Infernal (Nawrocki, Kolbe et al. 2009). They were then used to search all the assembled cDNA sequences of *Oryza sativa* and the 6480 *Arabidopsis* lincRNAs for RNA structure homologs by Infernal cmsearch function (Nawrocki, Kolbe et al. 2009).

2.5.8 RNA Secondary Structure Prediction with RNAz and RNAalifold

RNAz (Gruber, Findeiss et al. 2010) detects functional RNA secondary structures in multiple sequence alignment based on thermodynamic stability and structural conservation. We tried two runs of RNAz structure search: (1) global mode, in which the RNAz algorithm attempted to detect base pairs globally and scored the given alignment as a whole; (2) sliding-window mode with a window size of 120 and a step size of 40, which is recommended for common usage by its developer. RNAalifold is part of the ViennaRNA package (Lorenz, Bernhart et al. 2011), and used by RNAz for initial conserved structure prediction. We used
RNAalifold to obtain a schematic diagram of the putative lncRNA structures based on the orthologous lncRNA multiple alignments.

2.5.9 Primer Design for RNA Validation in *Setaria italica*

30 lncRNA homologs from syntenic regions of *Zea mays* and *Oryza sativa* were collected, as shown above in the sequence homology search section, and those without lncRNA orthologs in *Setaria italica* were selected. A computational prediction based on both sequence conservation (BLASTN and nhmmer function of HMMER v3.1 (http://hmmer.janelia.org/) ) and synteny analysis (genomevolution.org) yielded 19 putative orthologous loci in *Setaria italica*. The candidate primers were designed using Primer3, and only those passed the examination of hairpins, self/cross dimmers (Oligo 7 (Rychlik 2007)) and specificity checking (Primer-Blast (Ye, Coulouris et al. 2012)) were kept.

2.5.10 lncRNA Function Prediction by UniformMu Mutation

We collected 51,827 UniformMu insertion loci in total from MaizeGDB database UniformMu insertion release 2 and release 5, and mapped them to the maize genome based on their location information (Supplementary Figure S2). lncRNAs with UniformMu insertions within their transcription boundaries (from 150bp upstream to 150bp downstream) were selected (Supplementary Table S12). The agriGO (Du, Zhou et al. 2010), the GO Analysis Toolkit and Database for Agricultural Community, was used for gene ontology enrichment analysis of the protein-coding genes adjacent to lncRNA candidates. Based on the sequence homology search and synteny analysis as shown above, 10 UniformMu insertions were selected for mutagenesis lncRNA function analysis. And maize seeds with specific mutations were ordered from Maize Genetics Cooperation Stock Center.
2.5.11 Supplementary Materials

https://www.dropbox.com/s/6so9p4dqvzf70ip/YingSun_MS_Thesis_Supplementary_Material.rar
### Table 2.1 Summary of Transcriptome Assembly Results (Compared with Annotation)

<table>
<thead>
<tr>
<th>Genome</th>
<th>Total # cDNAs (Assembly/compared with annotation)</th>
<th># Strand-specific (Assembly/compared with annotation)</th>
<th>Multi-Exon (Assembly/compared with annotation)</th>
<th>Single-Exon (Assembly/compared with annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Zea mays</em></td>
<td>245920/63540</td>
<td>202078/63540</td>
<td>185697/50531</td>
<td>60223/13009</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>71859/29448</td>
<td>70017/29448</td>
<td>63189/23665</td>
<td>8670/5783</td>
</tr>
<tr>
<td><em>Setaria italic</em></td>
<td>76711/40599</td>
<td>74904/40599</td>
<td>64667/31405</td>
<td>12044/9194</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>125208/49061</td>
<td>117117/49061</td>
<td>102507/53622</td>
<td>22701/12716</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome</th>
<th>Len (Assembly/compared with annotation)</th>
<th>Mean</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Zea mays</em></td>
<td>1974.2/1538.62</td>
<td>1573.14/873.91</td>
<td>32/69</td>
<td>63870/14668</td>
<td></td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>1960.99/1489.92</td>
<td>1341.5/880.72</td>
<td>29/120</td>
<td>18659/14671</td>
<td></td>
</tr>
<tr>
<td><em>Setaria italic</em></td>
<td>1909.08/1427.52</td>
<td>1368.55/962.51</td>
<td>36/36</td>
<td>32036/15620</td>
<td></td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>1968.01/1708.18</td>
<td>1523.48/1222.43</td>
<td>25/84</td>
<td>21871/16311</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Categories</th>
<th>Complete match</th>
<th>Other</th>
<th>Intronic</th>
<th>Antisense</th>
<th>Intergenic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63345</td>
<td>102804</td>
<td>45</td>
<td>11612</td>
<td>68114</td>
</tr>
<tr>
<td></td>
<td>29382</td>
<td>31257</td>
<td>7</td>
<td>2005</td>
<td>9208</td>
</tr>
<tr>
<td></td>
<td>40490</td>
<td>30001</td>
<td>2</td>
<td>2098</td>
<td>4120</td>
</tr>
<tr>
<td></td>
<td>49057</td>
<td>41308</td>
<td>14</td>
<td>4899</td>
<td>29930</td>
</tr>
</tbody>
</table>

57
Table 2.2 Summary of lncRNA Length And Number of Exons

<table>
<thead>
<tr>
<th>Genome</th>
<th>Total # lncRNAs</th>
<th># lncRNAs with N exons (Assembly/compared with annotation)</th>
<th>2 Exons</th>
<th>3 Exons</th>
<th>&gt;3 Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zea mays</td>
<td>7196</td>
<td>1575/13009</td>
<td>4401/9688</td>
<td>798/7263</td>
<td>422/33580</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>1974</td>
<td>277/5783</td>
<td>962/5144</td>
<td>330/3820</td>
<td>405/14701</td>
</tr>
<tr>
<td>Setaria italic</td>
<td>4236</td>
<td>2726/9194</td>
<td>984/7433</td>
<td>285/4859</td>
<td>241/19113</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>2542</td>
<td>502/10005</td>
<td>1606/9637</td>
<td>266/6362</td>
<td>168/23057</td>
</tr>
</tbody>
</table>

Length (Assembly/compared with annotation)

<table>
<thead>
<tr>
<th>Genome</th>
<th>Mean</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zea mays</td>
<td>790.30/1538.62</td>
<td>630.72/873.91</td>
<td>200/69</td>
<td>6889/14668</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>985.74/1489.92</td>
<td>689.28/880.72</td>
<td>201/120</td>
<td>5687/14671</td>
</tr>
<tr>
<td>Setaria italic</td>
<td>639.35/1427.52</td>
<td>490.42/962.51</td>
<td>200/36</td>
<td>5089/15620</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>835.79/1708.18</td>
<td>687.13/1222.43</td>
<td>200/84</td>
<td>6024/16311</td>
</tr>
<tr>
<td>Genome</td>
<td>Total # lncRNAs</td>
<td>Categories (based on relative position to adjacent annotated genes)</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------</td>
<td>-------------------------------------------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sense</td>
<td>&gt;=0 kb</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complete Match</td>
<td>Divergent</td>
<td></td>
</tr>
<tr>
<td>Zea mays</td>
<td>7196</td>
<td>1141</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>1974</td>
<td>14</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Setaria italic</td>
<td>4236</td>
<td>2890</td>
<td>405</td>
<td></td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>2542</td>
<td>155</td>
<td>595</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>&gt;=1kb</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sense</td>
<td>2950</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antisense</td>
<td>1100</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;2kb</td>
<td>286</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;3kb</td>
<td>2648</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;4kb</td>
<td>938</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;5kb</td>
<td>2410</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>825</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>139</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>224</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>437</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3 Summary of lncRNA Locations Relative to Their Adjacent Annotated Genes
Table 2.4 Transposable Element Components of lncRNA Candidates

<table>
<thead>
<tr>
<th>Genome</th>
<th>Zea mays</th>
<th>Sorghum bicolor</th>
<th>Setaria italica</th>
<th>Oryza sativa</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Seqs</td>
<td>7196</td>
<td>1974</td>
<td>4236</td>
<td>2542</td>
</tr>
<tr>
<td>Total length (bp)</td>
<td>5686969</td>
<td>1945846</td>
<td>2708269</td>
<td>2124587</td>
</tr>
<tr>
<td>GC%</td>
<td>46.11</td>
<td>45.74</td>
<td>44.9</td>
<td>44</td>
</tr>
<tr>
<td>Bases Masked bp</td>
<td>1038581</td>
<td>198243</td>
<td>119618</td>
<td>78743</td>
</tr>
<tr>
<td>% of total</td>
<td>18.26</td>
<td>10.19</td>
<td>4.42</td>
<td>3.71</td>
</tr>
<tr>
<td># of elements</td>
<td>5419</td>
<td>960</td>
<td>790</td>
<td>270</td>
</tr>
<tr>
<td>length occupied</td>
<td>1052709</td>
<td>172709</td>
<td>103548</td>
<td>31561</td>
</tr>
<tr>
<td>percentage %</td>
<td>18.51</td>
<td>8.88</td>
<td>3.82</td>
<td>1.49</td>
</tr>
<tr>
<td># of elements</td>
<td>1480</td>
<td>756</td>
<td>667</td>
<td>1090</td>
</tr>
<tr>
<td>length occupied</td>
<td>68596</td>
<td>36365</td>
<td>29117</td>
<td>51740</td>
</tr>
<tr>
<td>percentage %</td>
<td>1.21</td>
<td>1.87</td>
<td>1.08</td>
<td>2.44</td>
</tr>
</tbody>
</table>
Table 2.5 Categories of IncRNA Candidates with Transposable Elements

<table>
<thead>
<tr>
<th>Genome</th>
<th># of Seqs with TEs</th>
<th>Sense Complete Match</th>
<th>Sense Other Sense</th>
<th>Intronic &gt;0b</th>
<th>Antisense &gt;=100b</th>
<th>Antisense &gt;=0 kb</th>
<th>Intergenic Divergent</th>
<th>Intergenic &gt;=5kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zea mays</td>
<td>2894</td>
<td>426</td>
<td>273</td>
<td>11</td>
<td>509</td>
<td>506</td>
<td>1675</td>
<td>102</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>562</td>
<td>2</td>
<td>28</td>
<td>0</td>
<td>56</td>
<td>55</td>
<td>476</td>
<td>14</td>
</tr>
<tr>
<td>Setaria italic</td>
<td>515</td>
<td>297</td>
<td>81</td>
<td>0</td>
<td>49</td>
<td>48</td>
<td>88</td>
<td>5</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>195</td>
<td>7</td>
<td>36</td>
<td>1</td>
<td>79</td>
<td>79</td>
<td>72</td>
<td>18</td>
</tr>
</tbody>
</table>
### Table 2.6 Candidate lncRNA Homologous Pairs among *Zea mays*, *Sorghum bicolor*, *Setaria italica* and *Oryza sativa*

<table>
<thead>
<tr>
<th>Pair</th>
<th>Triplet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zea mays</strong></td>
<td>29 (21; 0)</td>
</tr>
<tr>
<td><strong>Sorghum bicolor</strong></td>
<td>29 (15; 6)</td>
</tr>
<tr>
<td><strong>Setaria italica</strong></td>
<td>63 (56; 2)</td>
</tr>
<tr>
<td><strong>Oryza sativa</strong></td>
<td>46 (30; 1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pair</th>
<th>Triplet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zea mays</strong></td>
<td>29 (21; 0)</td>
</tr>
<tr>
<td><strong>Sorghum bicolor</strong></td>
<td>29 (15; 6)</td>
</tr>
<tr>
<td><strong>Setaria italica</strong></td>
<td>63 (56; 2)</td>
</tr>
<tr>
<td><strong>Oryza sativa</strong></td>
<td>46 (30; 1)</td>
</tr>
</tbody>
</table>

# of Homologous pairs (# of Syntenic Pairs; # of Pairs with at least one putative miRNA precursor)
Table 2.7 Significant Structure-based lncRNA Homologs among the lncRNA Candidates and the 225 lncRNA Families from Rfam.

Only 18 lncRNA candidates (5 from *Zea mays*, 1 from *Sorghum bicolor*, 5 from *Setaria italica* and 7 from *Oryza sativa*) exhibited significant homology to 9 lncRNA families. And only two lncRNA families had cross-species hits (light purple).

<table>
<thead>
<tr>
<th>#target name</th>
<th>query name</th>
<th>accession</th>
<th>mdl</th>
<th>mdl from</th>
<th>mdl to</th>
<th>seq from</th>
<th>seq to</th>
<th>strand</th>
<th>trunc</th>
<th>gc</th>
<th>bias</th>
<th>score</th>
<th>E-value</th>
<th>inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oryza_TCONS_00026565</td>
<td>TUG1_3</td>
<td>RF01891</td>
<td>cm</td>
<td>1</td>
<td>103</td>
<td>261</td>
<td>361</td>
<td>+</td>
<td>3'</td>
<td>0.45</td>
<td>0</td>
<td>17.5</td>
<td>0.0091</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00015923</td>
<td>RMST_7</td>
<td>RF01968</td>
<td>cm</td>
<td>1</td>
<td>268</td>
<td>897</td>
<td>1112</td>
<td>+</td>
<td>no</td>
<td>0.39</td>
<td>3.1</td>
<td>24.2</td>
<td>0.0067</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00015924</td>
<td>RMST_7</td>
<td>RF01968</td>
<td>cm</td>
<td>1</td>
<td>268</td>
<td>1373</td>
<td>1588</td>
<td>+</td>
<td>no</td>
<td>0.39</td>
<td>3.1</td>
<td>24.2</td>
<td>0.0067</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00015931</td>
<td>RMST_7</td>
<td>RF01968</td>
<td>cm</td>
<td>1</td>
<td>268</td>
<td>1221</td>
<td>1436</td>
<td>+</td>
<td>no</td>
<td>0.39</td>
<td>3.1</td>
<td>24.2</td>
<td>0.0067</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00015932</td>
<td>RMST_7</td>
<td>RF01968</td>
<td>cm</td>
<td>1</td>
<td>268</td>
<td>1483</td>
<td>1698</td>
<td>+</td>
<td>no</td>
<td>0.39</td>
<td>3.1</td>
<td>24.2</td>
<td>0.0067</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00015930</td>
<td>RMST_7</td>
<td>RF01968</td>
<td>cm</td>
<td>1</td>
<td>268</td>
<td>1455</td>
<td>1670</td>
<td>+</td>
<td>no</td>
<td>0.39</td>
<td>3.1</td>
<td>24.2</td>
<td>0.0067</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_00068156</td>
<td>CDKN2B-AS_3</td>
<td>RF02045</td>
<td>cm</td>
<td>1</td>
<td>144</td>
<td>8</td>
<td>150</td>
<td>+</td>
<td>no</td>
<td>0.32</td>
<td>0.3</td>
<td>27.9</td>
<td>0.0069</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_00026046</td>
<td>DAOA-AS1_2</td>
<td>RF02091</td>
<td>cm</td>
<td>1</td>
<td>205</td>
<td>1590</td>
<td>1766</td>
<td>+</td>
<td>no</td>
<td>0.35</td>
<td>0.6</td>
<td>18.7</td>
<td>0.0089</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_000114045</td>
<td>PART1_2</td>
<td>RF02160</td>
<td>cm</td>
<td>1</td>
<td>249</td>
<td>960</td>
<td>1084</td>
<td>+</td>
<td>no</td>
<td>0.38</td>
<td>0</td>
<td>17.8</td>
<td>0.0078</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00188946</td>
<td>ST7-OT4_3</td>
<td>RF02189</td>
<td>cm</td>
<td>1</td>
<td>302</td>
<td>293</td>
<td>430</td>
<td>+</td>
<td>no</td>
<td>0.36</td>
<td>0</td>
<td>25.3</td>
<td>0.00022</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_00032931</td>
<td>ST7-OT4_3</td>
<td>RF02189</td>
<td>cm</td>
<td>1</td>
<td>302</td>
<td>328</td>
<td>425</td>
<td>+</td>
<td>no</td>
<td>0.3</td>
<td>0.3</td>
<td>18.1</td>
<td>0.0099</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_00016157</td>
<td>WT1-AS_7</td>
<td>RF02209</td>
<td>cm</td>
<td>1</td>
<td>294</td>
<td>616</td>
<td>751</td>
<td>+</td>
<td>no</td>
<td>0.32</td>
<td>0.2</td>
<td>17.3</td>
<td>0.0072</td>
<td>!</td>
</tr>
<tr>
<td>Oryza_TCONS_00017781</td>
<td>ZFAT-AS1_2</td>
<td>RF02212</td>
<td>cm</td>
<td>1</td>
<td>205</td>
<td>561</td>
<td>719</td>
<td>+</td>
<td>no</td>
<td>0.33</td>
<td>0.3</td>
<td>21.8</td>
<td>0.0059</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_000199607</td>
<td>Six3os1_2</td>
<td>RF02247</td>
<td>cm</td>
<td>1</td>
<td>217</td>
<td>1098</td>
<td>1318</td>
<td>+</td>
<td>no</td>
<td>0.49</td>
<td>0.4</td>
<td>29.7</td>
<td>6.90E-06</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00012212</td>
<td>Six3os1_2</td>
<td>RF02247</td>
<td>cm</td>
<td>1</td>
<td>217</td>
<td>261</td>
<td>396</td>
<td>+</td>
<td>no</td>
<td>0.62</td>
<td>4.4</td>
<td>24.5</td>
<td>0.00022</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00067710</td>
<td>Six3os1_2</td>
<td>RF02247</td>
<td>cm</td>
<td>1</td>
<td>217</td>
<td>576</td>
<td>781</td>
<td>+</td>
<td>no</td>
<td>0.46</td>
<td>0.1</td>
<td>23.1</td>
<td>0.00059</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_000151483</td>
<td>Six3os1_2</td>
<td>RF02247</td>
<td>cm</td>
<td>1</td>
<td>217</td>
<td>64</td>
<td>200</td>
<td>+</td>
<td>no</td>
<td>0.61</td>
<td>0</td>
<td>21.2</td>
<td>0.0022</td>
<td>!</td>
</tr>
<tr>
<td>Sorghum_TCONS_00046394</td>
<td>Six3os1_2</td>
<td>RF02247</td>
<td>cm</td>
<td>1</td>
<td>217</td>
<td>446</td>
<td>483</td>
<td>+</td>
<td>no</td>
<td>0.5</td>
<td>0</td>
<td>19.5</td>
<td>0.0024</td>
<td>!</td>
</tr>
</tbody>
</table>

Abbreviations:
Model "mdl" (Hidden Markov Model or Covariance Model); Boundaries of the alignment with respect to the query model ("mdl from" and "mdl to") and the target sequence ("seq from" and "seq to"); Following the “seq to” column is a + or - symbol indicating whether the hit is on the top (+) or bottom (-) strand; A truncated hit "trunc" is defined as a hit that is missing one or more nucleotides at the 5’ and/or 3’ end; "gc" GC content in the hit; "bias" biased-composition correction that a high bias scores may be a red flag for a false positive; "score" the score (in bits) for this target/query comparison; "E-value" the expectation value (statistical significance) of the target; "inc" indicates whether or not this hit achieves the inclusion threshold: ’!’ if it does, ’?’ if it does not (and rather only achieves the reporting threshold). By default, the inclusion threshold is an E-value of 0.01 and the reporting threshold is an E-value of 10.0, but these can be changed with command line options as described in the manual pages (Nawrocki, Kolbe et al. 2009).
Table 2.8 Putative Homologs in *Oryza sativa* Reveal by Infernal Search. But the model field is marked as "hmm", because for models with zero basepairs, cmsearch of infernal uses a profile HMM instead of a CM for final hit scoring. Thus, these results actually provided us no conserved structural information.

<table>
<thead>
<tr>
<th>Sequence-level Homologs</th>
<th>Structure-based Homologs</th>
<th>mdl from</th>
<th>mdl to</th>
<th>seq from</th>
<th>seq to</th>
<th>strand</th>
<th>trunc</th>
<th>pass</th>
<th>gc</th>
<th>bias</th>
<th>score</th>
<th>E-value</th>
<th>inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maize_TCONS_00088624</td>
<td>Oryza_TCONS_00086492</td>
<td>hmm</td>
<td>6</td>
<td>104</td>
<td>194</td>
<td>+</td>
<td>-</td>
<td>6</td>
<td>0.6</td>
<td>0</td>
<td>46.8</td>
<td>4.10E-13</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00088626</td>
<td>Oryza_TCONS_00005598</td>
<td>hmm</td>
<td>32</td>
<td>114</td>
<td>187</td>
<td>+</td>
<td>-</td>
<td>6</td>
<td>0.6</td>
<td>0</td>
<td>30.9</td>
<td>3.10E-08</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00088627</td>
<td>Oryza_TCONS_00059479</td>
<td>hmm</td>
<td>34</td>
<td>105</td>
<td>373</td>
<td>+</td>
<td>-</td>
<td>6</td>
<td>0.57</td>
<td>0.1</td>
<td>23.4</td>
<td>6.00E-06</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00088628</td>
<td>Sorghum_TCONS_00036916</td>
<td>hmm</td>
<td>97</td>
<td>308</td>
<td>115</td>
<td>295</td>
<td>+</td>
<td>-</td>
<td>0.59</td>
<td>0</td>
<td>64.1</td>
<td>5.90E-18</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00044659</td>
<td>Sorghum_TCONS_00036916</td>
<td>hmm</td>
<td>146</td>
<td>410</td>
<td>112</td>
<td>380</td>
<td>+</td>
<td>-</td>
<td>0.58</td>
<td>0</td>
<td>33.6</td>
<td>1.10E-08</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00044658</td>
<td>Sorghum_TCONS_00036916</td>
<td>hmm</td>
<td>75</td>
<td>386</td>
<td>206</td>
<td>607</td>
<td>+</td>
<td>-</td>
<td>0.56</td>
<td>0.1</td>
<td>32.3</td>
<td>2.70E-08</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00044660</td>
<td>Sorghum_TCONS_00036916</td>
<td>hmm</td>
<td>141</td>
<td>416</td>
<td>233</td>
<td>515</td>
<td>+</td>
<td>-</td>
<td>0.6</td>
<td>5.6</td>
<td>26.8</td>
<td>2.20E-06</td>
<td>!</td>
</tr>
<tr>
<td>Maize_TCONS_00217063</td>
<td>Sorghum_TCONS_000067754</td>
<td>hmm</td>
<td>2</td>
<td>220</td>
<td>80</td>
<td>293</td>
<td>+</td>
<td>-</td>
<td>0.62</td>
<td>0.4</td>
<td>161.2</td>
<td>1.60E-47</td>
<td>!</td>
</tr>
<tr>
<td>Sorghum_TCONS_000067754</td>
<td>Maize_TCONS_00217063</td>
<td>hmm</td>
<td>2</td>
<td>211</td>
<td>258</td>
<td>467</td>
<td>+</td>
<td>-</td>
<td>0.64</td>
<td>0.5</td>
<td>126.3</td>
<td>8.90E-37</td>
<td>!</td>
</tr>
<tr>
<td>Setaria_TCONS_00038354</td>
<td>Sorghum_TCONS_000067754</td>
<td>hmm</td>
<td>79</td>
<td>186</td>
<td>192</td>
<td>291</td>
<td>+</td>
<td>-</td>
<td>0.62</td>
<td>0</td>
<td>27</td>
<td>3.50E-06</td>
<td>!</td>
</tr>
</tbody>
</table>
Table 2.9 The Primers Designed to Confirm 19 Putative Transcripts in *Setaria italica*. The transcripts selected were computationally predicted based on the 30 lncRNA homologs from syntenic regions of *Zea mays* and *Oryza sativa.*

<table>
<thead>
<tr>
<th>Setaria italica</th>
<th>Primer</th>
<th>Tm</th>
<th>GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>putative transcription loci</strong></td>
<td><strong>left primer (5'-3')</strong></td>
<td><strong>right primer (5'-3')</strong></td>
<td><strong>left</strong></td>
</tr>
<tr>
<td>scaffold_9 +</td>
<td>50639465</td>
<td>50641593</td>
<td>TCTTCACAGACGCAGATCC</td>
</tr>
<tr>
<td>scaffold_9 -</td>
<td>4496385</td>
<td>4498702</td>
<td>GCTTTCGAACCCTAGTCACC</td>
</tr>
<tr>
<td>scaffold_7 -</td>
<td>34130751</td>
<td>34133070</td>
<td>GGCAGATGGTGCTTCTGGA</td>
</tr>
<tr>
<td>scaffold_7 +</td>
<td>31391000</td>
<td>31393246</td>
<td>ATGGAGATCCCTCGCTGCAA</td>
</tr>
<tr>
<td>scaffold_5 +</td>
<td>40654932</td>
<td>40657598</td>
<td>GGCGAATCATGGGCCAAA</td>
</tr>
<tr>
<td>scaffold_7 +</td>
<td>25026606</td>
<td>25029954</td>
<td>GGCGAATCATGGGCCAAA</td>
</tr>
<tr>
<td>scaffold_1 +</td>
<td>31724333</td>
<td>31726587</td>
<td>GGAAGATCTCTTGCTGCAA</td>
</tr>
<tr>
<td>scaffold_1 -</td>
<td>25026606</td>
<td>25029954</td>
<td>GAGCTTTCTCTTGGGCATCA</td>
</tr>
<tr>
<td>scaffold_1 +</td>
<td>31724005</td>
<td>31726587</td>
<td>GGAAGATCTCTTGCTGCAA</td>
</tr>
<tr>
<td>scaffold_7 +</td>
<td>40654932</td>
<td>40657598</td>
<td>GAGCTTTCTCTTGGGCATCA</td>
</tr>
<tr>
<td>scaffold_9 -</td>
<td>2152423</td>
<td>2128389</td>
<td>GAGATGCTACTCTTGCCAGCATCA</td>
</tr>
<tr>
<td>scaffold_1 +</td>
<td>27484473</td>
<td>27486831</td>
<td>ATCGATGCTACTCTTGCCAGCATCA</td>
</tr>
<tr>
<td>scaffold_3 +</td>
<td>22537214</td>
<td>22538660</td>
<td>ATCGATGCTACTCTTGCCAGCATCA</td>
</tr>
<tr>
<td>scaffold_3 -</td>
<td>10106810</td>
<td>10110221</td>
<td>GACTGCTACATCCTCCTGGCATCA</td>
</tr>
<tr>
<td>scaffold_2 -</td>
<td>3470141</td>
<td>3472699</td>
<td>TTCTCTGCTCCTCACAAGGTT</td>
</tr>
<tr>
<td>scaffold_2 -</td>
<td>3470808</td>
<td>3473024</td>
<td>AAAGTTCTGGCCAGGTTTG</td>
</tr>
<tr>
<td>scaffold_2 -</td>
<td>37097972</td>
<td>37100165</td>
<td>TCAGATGCTACTCTTGCCAGCATCA</td>
</tr>
<tr>
<td>scaffold_2 -</td>
<td>38288352</td>
<td>38290574</td>
<td>AGTGCTACATCCTCCTGGCATCA</td>
</tr>
<tr>
<td>scaffold_5 -</td>
<td>40778623</td>
<td>40782359</td>
<td>TACAAACACCTCCTCCCGATG</td>
</tr>
</tbody>
</table>
Table 2.10 List of IncRNA Candidates in *Zea mays* and Their Corresponding UniformMu Insertions Chosen for Mutagenesis IncRNA Function Analysis. These maize IncRNAs have orthologs in *Oryza sativa*, but are absent of orthologs in *Setaria italica*.

<table>
<thead>
<tr>
<th><em>Zea mays</em> IncRNA Candidates</th>
<th>Orthologs in <em>Oryza sativa</em></th>
<th>Accession of UniformMu Insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maize_TCONS_00049566</td>
<td>Oryza_TCONS_00071457</td>
<td>mu1046917</td>
</tr>
<tr>
<td>Maize_TCONS_00152318</td>
<td>Oryza_TCONS_00097993/Oryza_TCONS_00097994/Oryza_TCONS_00097996</td>
<td>mu1008548</td>
</tr>
<tr>
<td>Maize_TCONS_00156420</td>
<td>Oryza_TCONS_00043275</td>
<td>mu1053332</td>
</tr>
<tr>
<td>Maize_TCONS_00170480</td>
<td>Oryza_TCONS_00086542</td>
<td>mu1045554</td>
</tr>
<tr>
<td>Maize_TCONS_00172447/Maize_TCONS_00172448/Maize_TCONS_00172449/Maize_TCONS_00172450</td>
<td>Oryza_TCONS_00083534</td>
<td>mu1057331, mu1060504, mu1048972, mu1002948</td>
</tr>
<tr>
<td>Maize_TCONS_00197940</td>
<td>Oryza_TCONS_00123837</td>
<td>mu1049188, mu1030467</td>
</tr>
</tbody>
</table>
Figure 2.1 Mini Phylogenetic Tree of Cereals (Devos 2005).
Figure 2.2 Synteny of Cereal Genetic Maps (Devos 2005)
Figure 2.3 Quality Control, Pre-processing of RNA-Seq Data (a) and "Tuxedo" Reference-Based Transcriptome Assembly (b).
Figure 2.4 Transcript Length, ORF Prediction and Homology Search IncRNA Identification Pipeline.
Figure 2.5 Comparison of GC content of annotated genes (red) and our lncRNA candidates (blue). The y-axis shows the percentage of transcripts of all annotated coding sequence (CDS without UTR) or all lncRNAs; the x-axis indicates the GC content %.
(A) *Zea mays*
(B) Sorghum bicolor
(C) *Setaria italic*
Figure 2.6 Hexamers that showed at least ten fold changes (either under- or over-representation) in lncRNA candidates (Grey bars) and annotated protein-coding CDS sequences (Black bars). The y axis represents the usage proportion of certain hexamer (x axis) in all the 4096 hexamer combinations. Out of 4096, there were 155 in Zea mays, 167 in Sorghum bicolor, 167 in Setaria
*italic*, and 168 in *Oryza sativa*; and among these, 145 hexamers showed significant at least ten fold change between the two in all four genomes.
**Figure 2.7** An example of annotated genes which are identical to our lncRNA candidates. GRMZM2G125531_T03 (right-bottom) is an alternative spliced isoform originated from a protein-coding locus in *Zea mays* genome and predicted to be a non-coding transcript by our pipeline.
Figure 2.8 LncRNA Orthologous Pairs (Sequence Homology and Synteny). We defined a lncRNA homologous pair as synteny pair if the lncRNA genes were located in the same synteny block between the two genomes with their adjacent upstream and downstream protein-coding gene forming un-disturbed protein-coding gene ortholog anchors. Different shapes represent different genes, and genes with the same shape are from syntenic regions. "pcGene" stands for protein-coding gene.
Figure 2.9 Conserved Structures of 2 lncRNA families from Rfam, which have structure-based lncRNA homologs in more than one of *Zea mays*, *Sorghum bicolor*, *Setaria italica* and *Oryza sativa* genomes. The colors represent the sequence conservation, with red showing the most conserved sequences. The graphs in the middle display the covariance model sub-regions.
Figure 2.10 RNAalifold RNA structure prediction based on TE-masked lncRNA multiple-sequence alignment. (A)
Maize_TCONS_00005329, Sorghum_TCONS_00009140 and Setaria_TCONS_00073644; (B) Maize_TCONS_00007457, Sorghum_TCONS_00062992 and Setaria_TCONS_00022801; (C) Maize_TCONS_00067076, Sorghum_TCONS_00052860 and Setaria_TCONS_00019134; (D) Maize_TCONS_00088627, Sorghum_TCONS_00036916 and Setaria_TCONS_00044659; (E) Maize_TCONS_00212716, Sorghum_TCONS_00036916 and Setaria_TCONS_00044659; (F) Maize_TCONS_00217063, Sorghum_TCONS_00067754 and Setaria_TCONS_00024967; (G) Maize_TCONS_00222866, Sorghum_TCONS_00030871 and Setaria_TCONS_00038354. All the sequences were masked using RepeatMasker based on our TE dataset. The "Red" color marked the highly conserved folds and the bulges might be caused by the masked TE bases. (A) and (B) represent high probability (>0.53 amd >0.79 respectively) to have thermostatically stable conserved structures evaluated by RNAz global mode.
Figure 2.1 Distribution of UniformMu insertions along the chromosome 1 of Zea mays. The outmost grey circle represents all the annotated genes (two rings represent for genes on the forward and reverse strands respectively); the next blue circle marks the IncRNAs identified (two rings represent for genes on the forward and reverse strands respectively); the green circle shows all the UniformMu insertions along the chromosome 1, and the red bars marks those overlapping
with lncRNAs; the inner grey histogram exhibits the GC-content of annotated genes on the outmost grey circle. Other chromosomes (Supplementary Figure S2).
Chapter 3

Conclusions and Future Perspectives

The advent of high-throughput transcriptome sequencing technologies has provided a myriad of evidence for pervasive transcription (Kapranov and St Laurent 2012), and consequently revealed a more extensive, complex and dynamic biological world. Many of the novel transcripts uncovered recently are long functional RNAs with little coding capacity (Guttman, Amit et al. 2009; Cabili, Trapnell et al. 2011; Derrien, Johnson et al. 2012; Hangauer, Vaughn et al. 2013), showing obvious temporal- and spatial-specific expression pattern (Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011; Kutter, Watt et al. 2012; Ulitsky, Shkumatava et al. 2012). And a small portion of them have highly conserved secondary structures associated with their functions (Mayer, Neubert et al. 2008; Zhao, Sun et al. 2008; Tsai, Manor et al. 2010; Zhang, Rice et al. 2010; Kotake, Nakagawa et al. 2011).

The abundance and multiple functions of long non-coding RNAs (lncRNA) in mammalian systems has been one of the most important discoveries in molecular biology in recent years. However, the identification and characterization of lncRNAs in plants, especially cereals, is in its early stages. We conducted a reference-guided transcriptome assembly with RNA-Seq data from four economically important cereal genomes, and screened for RNAs that were at least 200 bases in length, at most 70 amino acids in open reading frames and lack of homology in Uniprot database. We identified 7,196 Zea mays, 1,974 Sorghum bicolor, 4,236 Setaria italica and 2,542 Oryza sativa lncRNA candidates, and conducted sequence composition analysis, transposable elements detection and miRNA precursor screen. Further, a cross-species
comparison, including sequence- and structure-based lncRNA homology search, synteny analysis, and lncRNA secondary structure prediction, uncovered some limited sequence similarity and sub-regions elucidating putative conserved secondary structures. Experiments to verify our results are in progress. Our current data provides a good resource for future studies of cereal lncRNA evolution and function. Accordingly, characterizing lncRNAs in cereals may reveal previously hidden regulatory networks of crucial cereal developmental processes, such as stress response and reproduction, and facilitate the development of new biotechnological applications for stress response and adaptation, growth control, and yield increment.

Another interesting topic which may be addressed in our following studies is the birth of de novo protein-coding genes from the lncRNA loci. There is a plenty of evidence proving that de novo protein-coding genes may derive from not only exon shuffling, genome duplication and recombination (fusion and fission), but also from non-coding regions in the genome. To gain the protein-coding capability, this has been proposed to progress from non-coding DNAs to ncRNAs, and then to evolve into protein-coding ones. Knowles et al. (2009) first reported 3 de novo hominoid-specific protein-coding genes that originate from ancestral non-coding DNAs by ORF frame-shifting caused by “disabler”, which have no homologs in other primate genomes (Knowles and McLysaght 2009). Wu et al. (2011) identified another 60 new protein-coding genes gained by human after the human-chimpanzee speciation event (Wu, Irwin et al. 2011). Xie et al. (2012) proposed that lncRNA could serve as the “birth pool” of de novo protein-coding genes, based on their identification of 24 de novo protein-coding genes whose orthologous regions in chimp and macaque encode lncRNAs (Xie, Zhang et al. 2012). Our search for potential de novo protein-coding genes from the lncRNA candidates in the four cereal transcriptomes are in progress.
REFERENCES


