

# 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 (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. 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 lncRNA 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 lncRNA homology search, synteny analysis, and lncRNA secondary structure prediction, uncovered some limited sequence similarity and sub-regions elucidating putative conserved secondary structures.

INDEX WORDS: lncRNA, 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

© 2013

YING SUN

All Rights Reserved

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

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

## LIST OF TABLES

	Page
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 <i>Zea mays</i> , <i>Sorghum bicolor</i> , <i>Setaria italica</i> and <i>Oryza sativa</i> .....	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 <i>Oryza sativa</i> Reveal by Infernal Search .....	65
Table 2.9: The Primers Designed to Confirm 19 Putative Transcripts in <i>Setaria italica</i> .....	66
Table 2.10: List of lncRNA Candidates in <i>Zea mays</i> 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 <i>Zea mays</i> , <i>Sorghum bicolor</i> , <i>Setaria italica</i> and <i>Oryza sativa</i> 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 <i>Zea mays</i> .....	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 Lodomery 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 spliceosome which removes introns from pre-mRNA (McKeown 1993); Small nucleolar RNA (snoRNA) conducts the 2'-O-ribose 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 latter 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

assemblies, Boerner and McGinnis (2012) reported 1,011 putative novel lncRNAs (Boerner and McGinnis 2012).

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 (lncRNA) 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 lncRNA 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 lncRNAs**

### **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 lncRNAs (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 lncRNAs 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 lncRNAs 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 lncRNAs, 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, Mirg, 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 (Hupaló 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 lincRNAs detector lincRScan (including ORF length cutoff, PhyloCSF coding-potential evaluation and Pfam homology search), and reported 308 novel mouse embryonic lincRNAs, among which, 13 lincRNAs 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 lincRNAs 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 lincRNAs

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

et al. 2007; Clemson, Hutchinson et al. 2009; Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Gupta, Shah et al. 2010; Huarte, Guttman et al. 2010; Orom, Derrien et al. 2010; Cabili, Trapnell et al. 2011; Baldassarre and Masotti 2012; Liu, Jung et al. 2012; Ulitsky, Shkumatava et al. 2012; Hangauer, Vaughn et al. 2013; Li, Feng et al. 2013; Moghe, Lehti-Shiu et al. 2013). 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.

2013). 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.

2012). 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 lncRNAs.*** Sense lncRNAs 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 lncRNA 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 lncRNA 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 lncRNAs which have been experimentally investigated (Amaral, Clark et al. 2011), demonstrates a large variety of levels at which lncRNAs 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 (Ebraldize, 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.** lncRNAs 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 lncRNAs, 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 lncRNA npc536 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 TPSI1/Mt4 lncRNA 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 TPSI1 in tomato (Liu, Muchhal et al. 1997). Liu et al. (1997) isolated the 474nt TPSI1 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 TPSI1, 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 TPSI1 (Burleigh and Harrison 1997). Besides the low conservatory ORFs, TPSI/Mt4 family lncRNAs 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) lncRNAs, 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 lncRNA 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 lncRNAs 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 lncRNA 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 lncRNAs 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, lncRNAs 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 lncRNA Sequences**

In contrast to protein-coding mRNAs, most of which are highly conserved in sequence to keep their protein products consistent in function, lncRNA 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, lncRNAs 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 lncRNA 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 lncRNA Structure Evolvability**

### **1.4.2.1 Secondary Structure and lncRNA 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 Lodomery 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-pairing (Elliott and Lodomery 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, Hennessey 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).

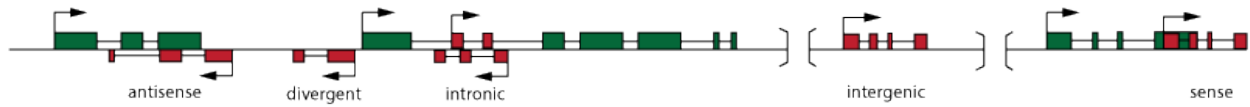
Briefly, lncRNAs might (1) only conserve in the function level within small stretches of their sequence which are closely related to their functions, and the utility of other parts of their sequences are left unknown; (2) primarily maintenance a particular base-paired structure (Eddy and Durbin 1994), 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

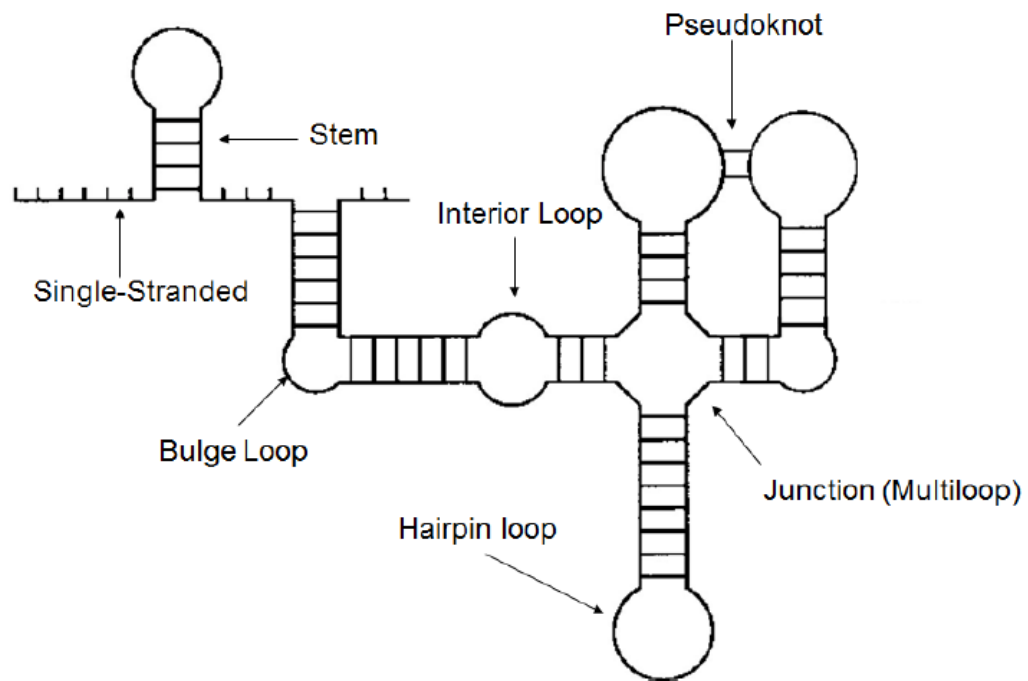
The abundance and multiple functions of long non-coding RNAs (lncRNA) in mammalian systems have been one of the most important discoveries in molecular biology in recent years. However, the identification and characterization of lncRNAs in plants, especially cereals, is in its early stages. This thesis is dedicated to identify lncRNAs in four economically important cereal genomes (*Zea mays*, *Sorghum bicolor*, *Setaria italica* 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.



## Figures



**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<sup>1</sup>

---

<sup>1</sup> 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 Lodomery 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 syntenic 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 aid in the lncRNA evolutionary conservation

analysis, considering that lncRNAs always lack known orthologs cross species at the sequence level (Cabili, Trapnell et al. 2011; Ulitsky, Shkumatava et al. 2012); however, even syntenic lncRNAs may exhibit distinct expression levels in phylogenetically close species (Kutter, Watt et al. 2012).

A handful of lncRNAs 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 lncRNA 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 italica*, 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 italica*, 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 italica* 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 italica* (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,  $P\text{-value} < 2.2\text{e-}16$ ).

Generally, the lncRNA candidates were significantly shorter than the annotated mRNAs (Table 2.2; Two-sample Z-test for mean,  $P\text{-value} < 2.2\text{e-}16$ ) and the average GC content were slightly lower than that of annotated mRNAs (Figure 2.5; Two-sample Z-test for mean,  $P\text{-value} < 2.2\text{e-}16$ , 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 italica*, 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 italica*, 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 italica*, 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 Delihis 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 italica*, 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 synteny 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 $\leq$ 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

lncRNA candidates in *Zea mays*, 14 in *Sorghum bicolor*, 18 in *Setaria italica*, and 6 in *Oryza sativa*; among which, only 1 pair of maize lncRNAs was 100 percent identical and another 6 lncRNAs from the 1,011 dataset were part of our lncRNAs candidates (Supplementary Table S8). To verify the unexpected lack of overlap between the 1,011 putative maize lncRNA dataset and our maize lncRNA candidates, we performed BLASTN again on the 1,011 putative maize lncRNAs and all the assembled transcripts from our dataset. It turned out that 684 out of the putative 1,011 maize lncRNAs 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 ( $\text{ORF} \leq 70$  amino acids) that led to the very limited similarity between the two candidate lncRNA 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 lncRNA candidates of *Zea mays*; 23 exhibited an average of 31nt alignment with our lncRNA candidates of *Sorghum bicolor*; 29 showed an average of 32nt alignment with our lncRNA candidates of *Setaria italica*; and 36 showed an average of 30nt alignment with our lncRNA candidates of *Oryza sativa* (Supplementary Table S9). Compared with the average length of alignment shared between our lncRNA candidates (Supplementary Table S7), the limited similarity at the sequence level between the 6,480 *Arabidopsis* lincRNAs and our lncRNA candidates strongly suggested the propensity of poor sequence conservation of lncRNAs cross-species, especially for those phylogenetically distant ones.



### 2.3.7 Structure-Based lncRNA Homology Search

Secondary structure is known to be associated with lncRNA 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 lncRNAs 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 lncRNA database (Table 2.2) for conserved RNA structures to any of the 225 lncRNA-family-alignment profiles from Rfam (Burge, Daub et al. 2013) (Supplementary Table S10).

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 Rfam lncRNA families, and only 2 Rfam lncRNA families had cross-species hits (Table 2.7). They were RF02189 family (human lncRNA ST7 overlapping transcript 4 conserved region 3 (Vincent, Petek et al. 2002)) homologous to lncRNA Maize\_TCONS\_00188946 and Oryza\_TCONS\_00032931, and RF02247 family (mouse lncRNA 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 lncRNA homologs were from syntenic regions based on information obtained from CoGe Platform (genomeevolution.org), the conserved structures of the 9 Rfam lncRNA families (Figure 2.9 and Supplementary Figure S1) gave us some insight into the secondary structures and potential functionality of their lncRNA homologs from the four cereals.

The lncRNA sequence homology analysis above uncovered 21 homologous triplets of *Zea mays*, *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 lncRNA triplets of *Zea mays*, *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* lncRNAs for RNA structure homolog by Infernal (Nawrocki, Kolbe et al. 2009) (Material and Methods). There was no homolog found among the 6,480 *Arabidopsis* lncRNAs, 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 lncRNA 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 lncRNA secondary structures by using the 7 distinct multiple alignments of orthologous lncRNA triplets obtained from sequence homology search and synteny analysis.

The RNAz program detected 2 groups of lncRNA 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 lncRNA identification and homology search, we would conduct PCR validation to investigate the expression of potential RNAs in *Setaria italica*, which were orthologous to lncRNAs 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 these 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 lncRNA 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 lncRNA 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 lncRNA controlling mRNA translation (Carrieri, Cimatti et al. 2012). Although the majority of our lncRNA candidates are not associated with TEs, the association between TEs and lncRNAs might shed a light on the functions of our lncRNA candidates.

Few of our lncRNA 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 lncRNAs, 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 lncRNAs are less likely to host small RNAs, given the potential large amount of lncRNAs uncovered.

#### **2.4.2 Evolutionary Conservation of lncRNAs 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 lncRNA candidates allow us to assess the evolutionary sequence conservation. At the sequence level, the lncRNA 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 lncRNAs 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 lncRNAs are too low to be captured in the RNA-Seq experiment. Correspondingly, we have designed PCR experiments to test the existence of potential lncRNAs in *Setaria italica* whose orthologous lncRNAs 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  $\geq 200$  nt; (2) ORF size  $\leq 70$ aa; (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 ( $\geq 80\%$ ) and high percentage of identity ( $\geq 90\%$ ).

#### **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](http://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 genomes 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 evaluate="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 evaluate="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-lncRNA-family-multiple-sequence-alignment profiles, and then cmsearch function was performed to search our candidate lncRNA database (Table 2.2) for RNAs homologous to any of the 225 profiles in the CM database.

The 21 homologous triplets of *Zea mays*, *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 lncRNA 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 dimers (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/6so9p4dqzfz70jp/YingSun\\_MS\\_Thesis\\_Supplementary\\_Material.rar](https://www.dropbox.com/s/6so9p4dqzfz70jp/YingSun_MS_Thesis_Supplementary_Material.rar)

## Tables

**Table 2.1** Summary of Transcriptome Assembly Results (Compared with Annotation)

Genome	Total # cDNAs (Assembly/compared with annotation)	# Strand-specific (Assembly/compared with annotation)	Multi-Exon (Assembly/compared with annotation)	Single-Exon (Assembly/compared with annotation)
<i>Zea mays</i>	245920/63540	202078/63540	185697/50531	60223/13009
<i>Sorghum bicolor</i>	71859/29448	70017/29448	63189/23665	8670/5783
<i>Setaria italica</i>	76711/40599	74904/40599	64667/31405	12044/9194
<i>Oryza sativa</i>	125208/49061	117117/49061	102507/53622	22701/12716
Genome	Len (Assembly/compared with annotation)			
	Mean	Sd	Min	Max
<i>Zea mays</i>	1974.2/1538.62	1573.14/873.91	32/69	63870/14668
<i>Sorghum bicolor</i>	1960.99/1489.92	1341.5/880.72	29/120	18659/14671
<i>Setaria italica</i>	1909.08/1427.52	1386.55/962.51	36/36	32036/15620
<i>Oryza sativa</i>	1968.01/1708.18	1523.48/1222.43	25/84	21871/16311
Categories				
Complete match	Other	Intronic	Antisense	Intergenic
63345	102804	45	11612	68114
29382	31257	7	2005	9208
40490	30001	2	2098	4120
49057	41308	14	4899	29930

**Table 2.2** Summary of lncRNA Length And Number of Exons

Genome	Total # lncRNAs	# lncRNAs with N exons (Assembly/compared with annotation)			
		Single-Exon	2 Exons	3 Exons	>3 Exons
<i>Zea mays</i>	7196	1575/13009	4401/9688	798/7263	422/33580
<i>Sorghum bicolor</i>	1974	277/5783	962/5144	330/3820	405/14701
<i>Setaria italica</i>	4236	2726/9194	984/7433	285/4859	241/19113
<i>Oryza sativa</i>	2542	502/10005	1606/9637	266/6362	168/23057
Length (Assembly/compared with annotation)					
		Mean	Sd	Min	Max
		790.30/1538.62	630.72/873.91	200/69	6889/14668
		985.74/1489.92	689.28/880.72	201/120	5687/14671
		639.35/1427.52	490.42/962.51	200/36	5089/15620
		835.79/1708.18	687.13/1222.43	200/84	6024/16311

**Table 2.3** Summary of lncRNA Locations Relative to Their Adjacent Annotated Genes

Genome	Total # lncRNAs	Categories (based on relative position to adjacent annotated genes)					
		Sense		Intronic		Antisense	
		Complete Match	Other	Sense	Antisense	>0b	>=100b
<i>Zea mays</i>	7196	1141	803	1	17	1836	1829
<i>Sorghum bicolor</i>	1974	14	200	0	1	488	486
<i>Setaria italica</i>	4236	2890	407	0	1	533	522
<i>Oryza sativa</i>	2542	155	491	0	5	1296	1289
Overlapping the sense strand of a annotated gene				Overlapping with the intron of annotated genes in either sense or antisense orientation		Overlapping with the antisense strand of an annotated genes	
Categories (based on relative position to adjacent annotated genes)							
Intergenic							
	>=0 kb	Divergent	>=1kb	>2kb	>=3kb	>=4kb	>=5kb
	3398	143	2950	2648	2410	2250	2135
	1271	39	1100	938	825	690	608
	405	32	286	186	139	109	91
	595	47	437	313	224	184	140
Oriented head to head with an annotated gene within 1 kb							

**Table 2.4** Transposable Element Components of lncRNA Candidates

Genome		<i>Zea mays</i>	<i>Sorghum bicolor</i>	<i>Setaria italica</i>	<i>Oryza sativa</i>
# of Seqs		7196	1974	4236	2542
Total length (bp)		5686969	1945846	2708269	2124587
GC%		46.11	45.74	44.9	44
Bases Masked	bp	1038581	198243	119618	78743
	% of total	18.26	10.19	4.42	3.71
Our TE database	# of elements	5419	960	790	270
	length occupied	1052709	172709	103548	31561
	percentage %	18.51	8.88	3.82	1.49
Simple repeats	# of elements	1480	756	667	1090
	length occupied	68596	36365	29117	51740
	percentage %	1.21	1.87	1.08	2.44

**Table 2.5** Categories of lncRNA Candidates with Transposable Elements

Genome	# of Seqs with TEs	Categories (based on relative position to adjacent annotated genes)							
		Sense			Antisense			Intergenic	
		Complete Match	Other Sense	Intronic	>0b	>=100b	>=0 kb	Divergent	>=5kb
<i>Zea mays</i>	2894	426	273	11	509	506	1675	102	1128
<i>Sorghum bicolor</i>	562	2	28	0	56	55	476	14	258
<i>Setaria italica</i>	515	297	81	0	49	48	88	5	27
<i>Oryza sativa</i>	195	7	36	1	79	79	72	18	24

**Table 2.6** Candidate lncRNA Homologous Pairs among *Zea mays*, *Sorghum bicolor*, *Setaria italica* and *Oryza sativa*

<i>Sorghum bicolor</i>	<i>Setaria italica</i>	<i>Oryza sativa</i>	Pair Triplet
147 (119; 6)	85 (51; 0)	46 (30; 1)	<i>Zea mays</i>
	63 (56; 2)	18 (15; 4)	<i>Sorghum bicolor</i>
# of Homologous pairs (# of Synteny Pairs; # of Pairs with at least one putative miRNA precursor )		29 (15; 6)	<i>Setaria italica</i>
			29 (21; 0)

**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).

#target name	query name	accession	mdl	mdl from	mdl to	seq from	seq to	strand	trunc	gc	bias	score	E-value	inc
Oryza_TCONS_00026565	TUG1_3	RF01891	cm	1	103	261	361	+	3'	0.45	0	17.5	0.0091	!
Setaria_TCONS_00015923	RMST_7	RF01968	cm	1	268	897	1112	+	no	0.39	3.1	24.2	0.0067	!
Setaria_TCONS_00015924	RMST_7	RF01968	cm	1	268	1373	1588	+	no	0.39	3.1	24.2	0.0067	!
Setaria_TCONS_00015931	RMST_7	RF01968	cm	1	268	1221	1436	+	no	0.39	3.1	24.2	0.0067	!
Setaria_TCONS_00015932	RMST_7	RF01968	cm	1	268	1483	1698	+	no	0.39	3.1	24.2	0.0067	!
Setaria_TCONS_00015930	RMST_7	RF01968	cm	1	268	1455	1670	+	no	0.39	3.1	24.2	0.0067	!
Oryza_TCONS_00068156	CDKN2B-AS_3	RF02045	cm	1	144	8	150	+	no	0.32	0.3	27.9	0.0069	!
Oryza_TCONS_00026046	DAOA-AS1_2	RF02091	cm	1	205	1590	1766	+	no	0.35	0.6	18.7	0.0089	!
Oryza_TCONS_00114045	PART1_2	RF02160	cm	1	249	960	1084	+	no	0.38	0	17.8	0.0078	!
Maize_TCONS_00188946	ST7-OT4_3	RF02189	cm	1	302	293	430	+	no	0.36	0	25.3	0.00022	!
Oryza_TCONS_00032931	ST7-OT4_3	RF02189	cm	1	302	328	425	+	no	0.3	0.3	18.1	0.0099	!
Oryza_TCONS_00016157	WT1-AS_7	RF02209	cm	1	294	616	751	+	no	0.32	0.2	17.3	0.0072	!
Oryza_TCONS_00017781	ZFAT-AS1_2	RF02212	cm	1	205	561	719	+	no	0.33	0.3	21.8	0.0059	!
Maize_TCONS_00199607	Six3os1_2	RF02247	cm	1	217	1098	1318	+	no	0.49	0.4	29.7	6.90E-06	!
Maize_TCONS_00012212	Six3os1_2	RF02247	cm	1	217	261	396	+	no	0.62	4.4	24.5	0.00022	!
Maize_TCONS_00067710	Six3os1_2	RF02247	cm	1	217	576	781	+	no	0.46	0.1	23.1	0.00059	!
Maize_TCONS_00151483	Six3os1_2	RF02247	cm	1	217	64	200	+	no	0.61	0	21.2	0.0022	!
Sorghum_TCONS_00046394	Six3os1_2	RF02247	cm	1	217	446	483	+	no	0.5	0	19.5	0.0024	!

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.

Sequence-level Homologs	Structue-based Homologs	mdl	mdl from	mdl to	seq from	seq to	strand	trunc	pass	gc	bias	score	E-value	inc
Maize_TCONS_00088624	Oryza_TCONS_00086492	hmm	6	104	194	295	+	-	6	0.6	0	46.8	4.10E-13	!
Maize_TCONS_00088626	Oryza_TCONS_00005598	hmm	32	114	187	272	+	-	6	0.6	0	30.9	3.10E-08	!
Maize_TCONS_00088627	Oryza_TCONS_00059479	hmm	34	105	373	447	+	-	6	0.57	0.1	23.4	6.00E-06	!
Maize_TCONS_00088628														
Sorghum_TCONS_00036916														
Setaria_TCONS_00044659														
Setaria_TCONS_00044658														
Setaria_TCONS_00044660														
Maize_TCONS_00212716	Oryza_TCONS_00086492	hmm	97	308	115	295	+	-	6	0.59	0	64.1	5.90E-18	!
Sorghum_TCONS_00036916	Oryza_TCONS_00005598	hmm	146	410	112	380	+	-	6	0.58	0	33.6	1.10E-08	!
Setaria_TCONS_00044659	Oryza_TCONS_00059479	hmm	75	386	206	607	+	-	6	0.56	0.1	32.3	2.70E-08	!
Setaria_TCONS_00044658														
Setaria_TCONS_00044660														
Maize_TCONS_00217063	Oryza_TCONS_00012226	hmm	141	416	233	515	+	-	6	0.6	5.6	26.8	2.20E-06	!
Sorghum_TCONS_00067754														
Setaria_TCONS_00024967														
Maize_TCONS_00222866	Oryza_TCONS_00005598	hmm	2	220	80	293	+	-	6	0.62	0.4	161.2	1.60E-47	!
Sorghum_TCONS_00030871	Oryza_TCONS_00059479	hmm	2	211	258	467	+	-	6	0.64	0.5	126.3	8.90E-37	!
Setaria_TCONS_00038354	Oryza_TCONS_00086492	hmm	79	186	192	291	+	-	6	0.62	0	27	3.50E-06	!

**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*.

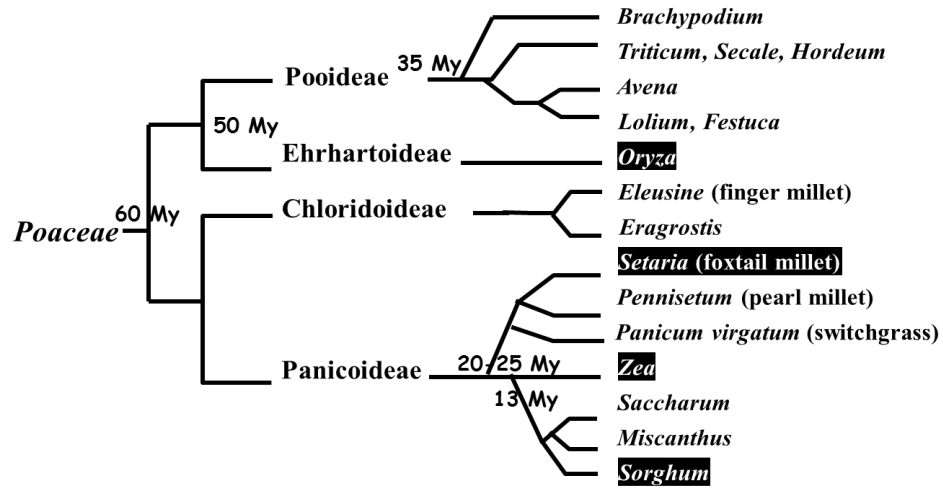
<i>Setaria italica</i>				Primer		Tm		GC %	
putative transcription loci				left primer (5'-3')	right primer (5'-3')	left	right	left	right
scaffold_9	+	50639465	50641593	TCTCTCACAGACGCAGATCC	GCCTTCGAACCCTAGTCACC	59.08	61.02	55	60
scaffold_9	-	4496385	4498702	GCTTTCTGAACGGTGAAGGA	CGTGCGCTAGTGAAGATCAG	60.38	59.76	50	55
scaffold_7	-	34130751	34133070	GGCGGAGTTGAGTCTCTTGA	ACAGCAAGACAGCATCATGG	60.53	59.86	55	50
scaffold_7	+	31391000	31393246	ATGGAACCTCCTCTTGCAA	TGCTGGCTACCACTGTTTTG	59.8	59.9	50	50
scaffold_5	+	40654932	40657598	TGTTTGGAATAGCAGCAAC	GTGATTGGGATTTGGTGGAG	59.88	60.17	45	50
scaffold_7	+	25026606	25029954	TTGCACCCCAAACCTGAACCT	GGCCATGTTTGCAGGTGTTT	60.03	59.89	50	50
scaffold_1	+	31724333	31726587	GAAGATTAGCATGGGCCAAA	GAACCTTTCTGCTCCCTTCA	60.04	59.41	45	50
scaffold_1	+	30709552	30711824	AGGGCGGGAGGTAGAACTT	GTCGTTCCCATCCTTCTTCC	60.08	60.83	57.89	55
scaffold_1	+	31724005	31726587	GAACCTTCTCTTGGGCATCA	CTTACGGTGTCACCCGATTC	60.2	60.38	50	55
scaffold_7	+	25026606	25029954	GAGCTTTCTCTTGGGCATCA	TCCTTACTGTGTCACCGGATT	60.48	59.46	50	47.62
scaffold_9	-	2125423	2128389	GGAACGTCCAAGCCTCTCAA	TCGACACCTTACCAAGCACC	59.97	59.97	55	55
scaffold_1	+	27484473	27486831	AGCACTCACATACTTCCCAGG	ACTGTGTTCTGCAGCTGACT	59.44	59.53	52.38	50
scaffold_3	+	22537214	22538660	GAGAGGAAGGGGAGGTAGGG	ATCTCCGGCTAGTGGCAATG	60.11	59.89	65	55
scaffold_3	-	10106810	10110221	GGTCGCAATAACCGATCCCT	CGCTGTCAACAATGCTGCTT	59.89	60.04	55	50
scaffold_2	-	3470141	3472699	TTCTCTGCCTCCTCAAGGGT	AGGGTTTAGTCCCGAGCTTG	60.18	59.39	55	55
scaffold_2	-	3470808	3473024	AAAGTTCTGGCCGAGGTTGT	TGGAGACATTTGCTGGTGCT	59.82	59.89	50	50
scaffold_2	-	37097972	37100165	TCGAATGGATGTGTTGAACC	AATTGATGGGATGGAACAGC	59.35	59.76	45	45
scaffold_2	-	38288352	38290574	AGTGTTAGCATACCACCACGA	GAGGGAGGCTCCACTTTGTC	59.1	60.04	47.62	60
scaffold_5	-	40778623	40782359	TACAAAACCTCCTCCCGATG	ACGAGGACTGATGCATGTGA	59.93	60.28	50	50

**Table 2.10** List of lncRNA Candidates in *Zea mays* and Their Corresponding UniformMu Insertions Chosen for Mutagenesis lncRNA

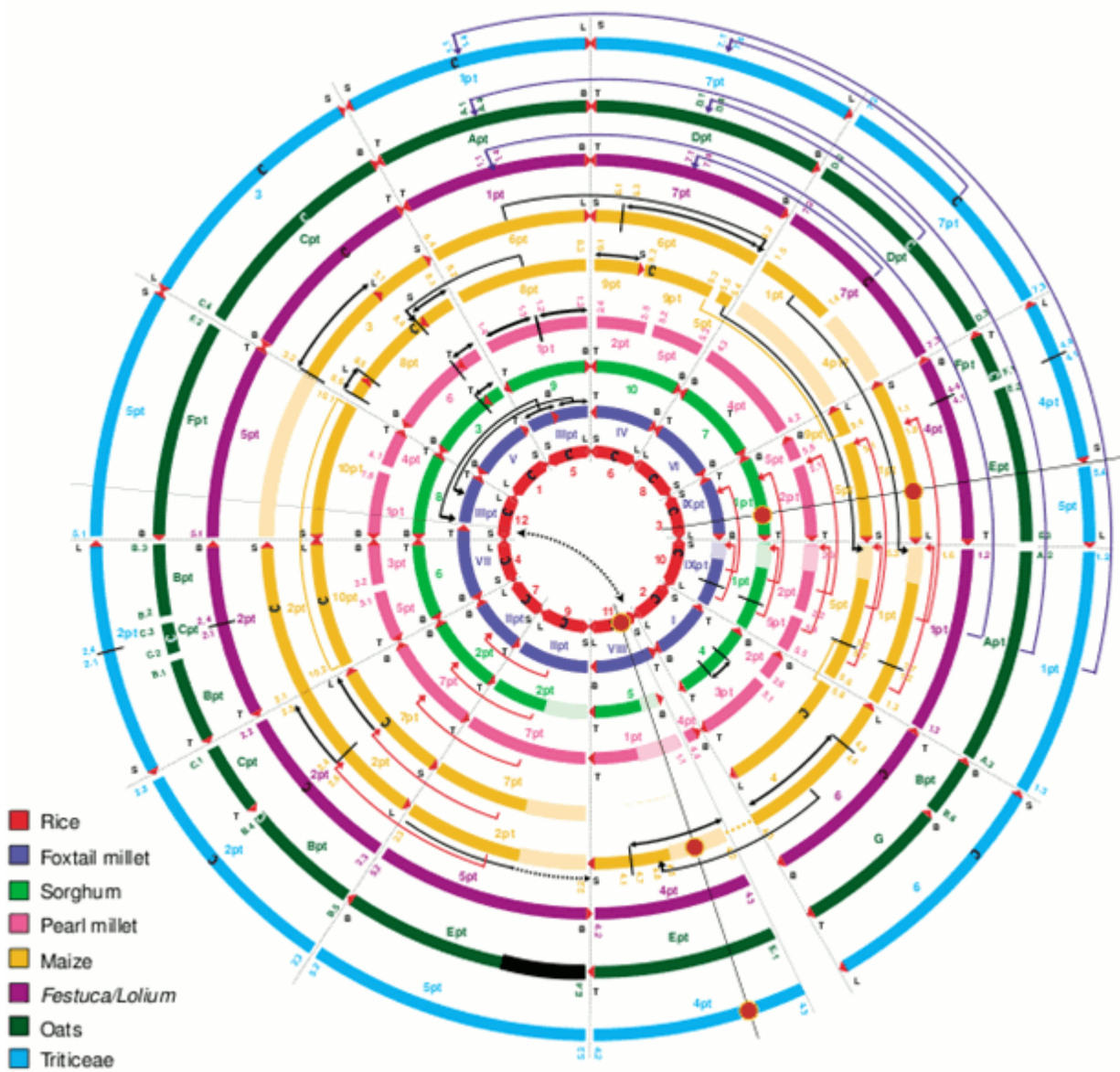
Function Analysis. These maize lncRNAs have orthologs in *Oryza sativa*, but are absent of orthologs in *Setaria italica*.

<i>Zea mays</i> lncRNA Candidates	Orthologs in <i>Oryza sativa</i>	Accession of UniformMu Insertion
Maize_TCONS_00049566	Oryza_TCONS_00071457	mu1046917
Maize_TCONS_00152318	Oryza_TCONS_00097993/Oryza_TCONS_00097994/Oryza_TCONS_00097996	mu1008548
Maize_TCONS_00156420	Oryza_TCONS_00043275	mu1053332
Maize_TCONS_00170480	Oryza_TCONS_00086542	mu1045554
Maize_TCONS_00172447/Maize_TCONS_00172448/Maize_TCONS_00172449/Maize_TCONS_00172450	Oryza_TCONS_00083534	mu1057331, mu1060504, mu1048972, mu1002948
Maize_TCONS_00197940	Oryza_TCONS_00123837	mu1049188, mu1030467

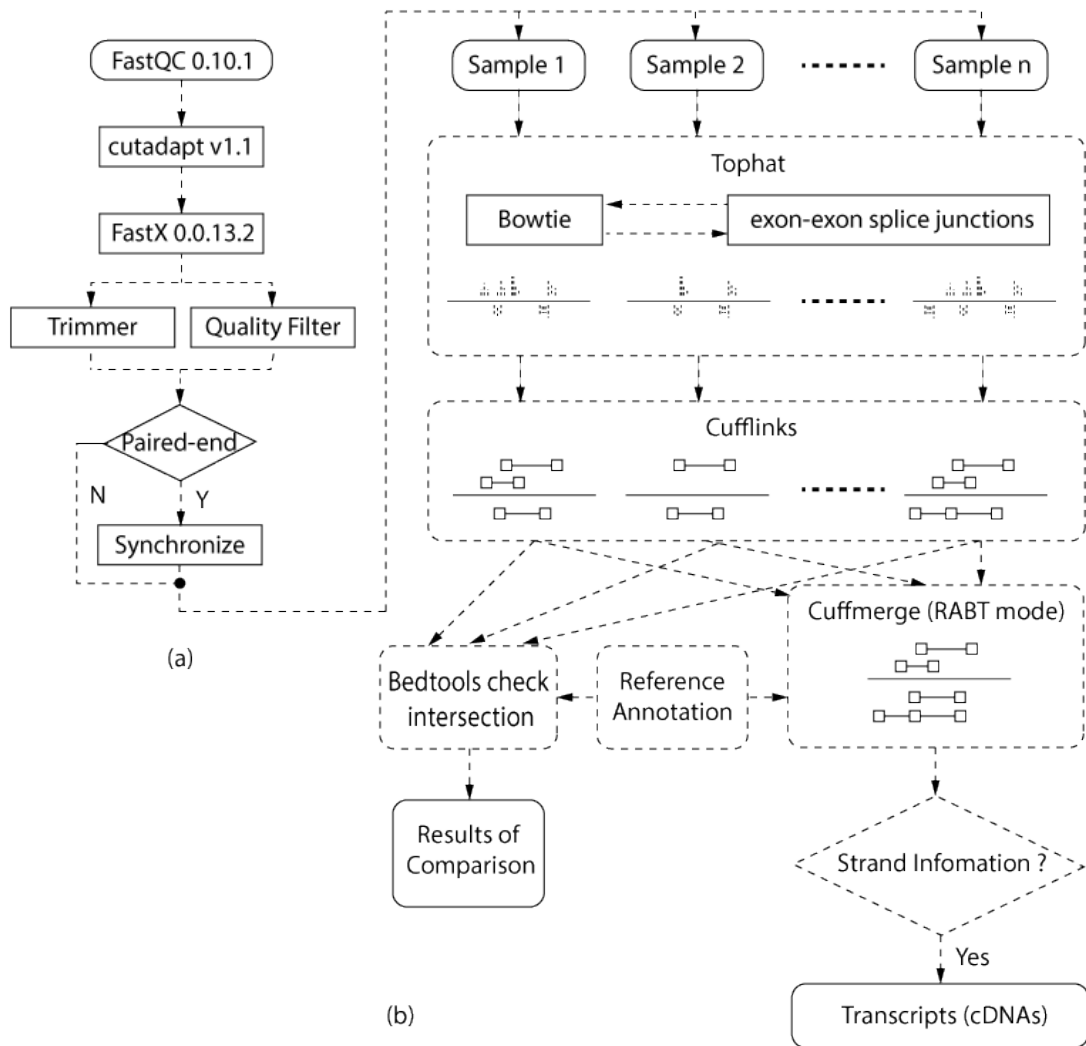
## Figures



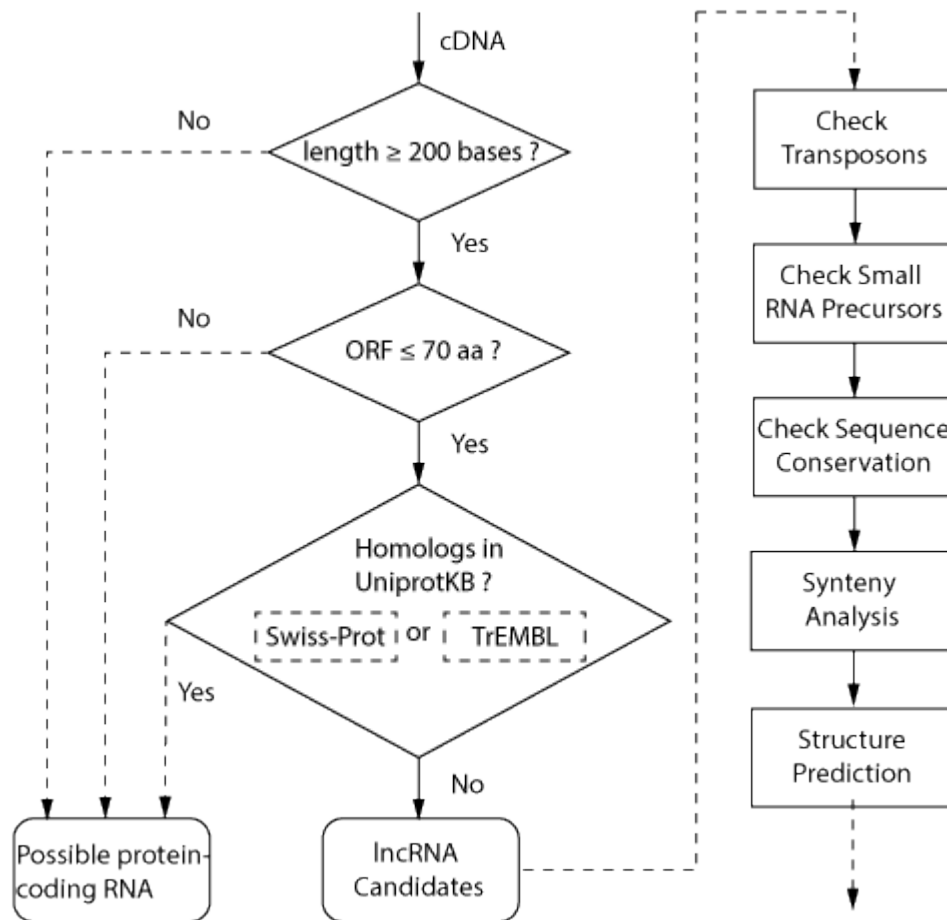
**Figure 2.1** Mini Phylogenetic Tree of Cereals (Devos 2005).



**Figure 2.2** Synteny of Cereal Genetic Maps (Devos 2005)

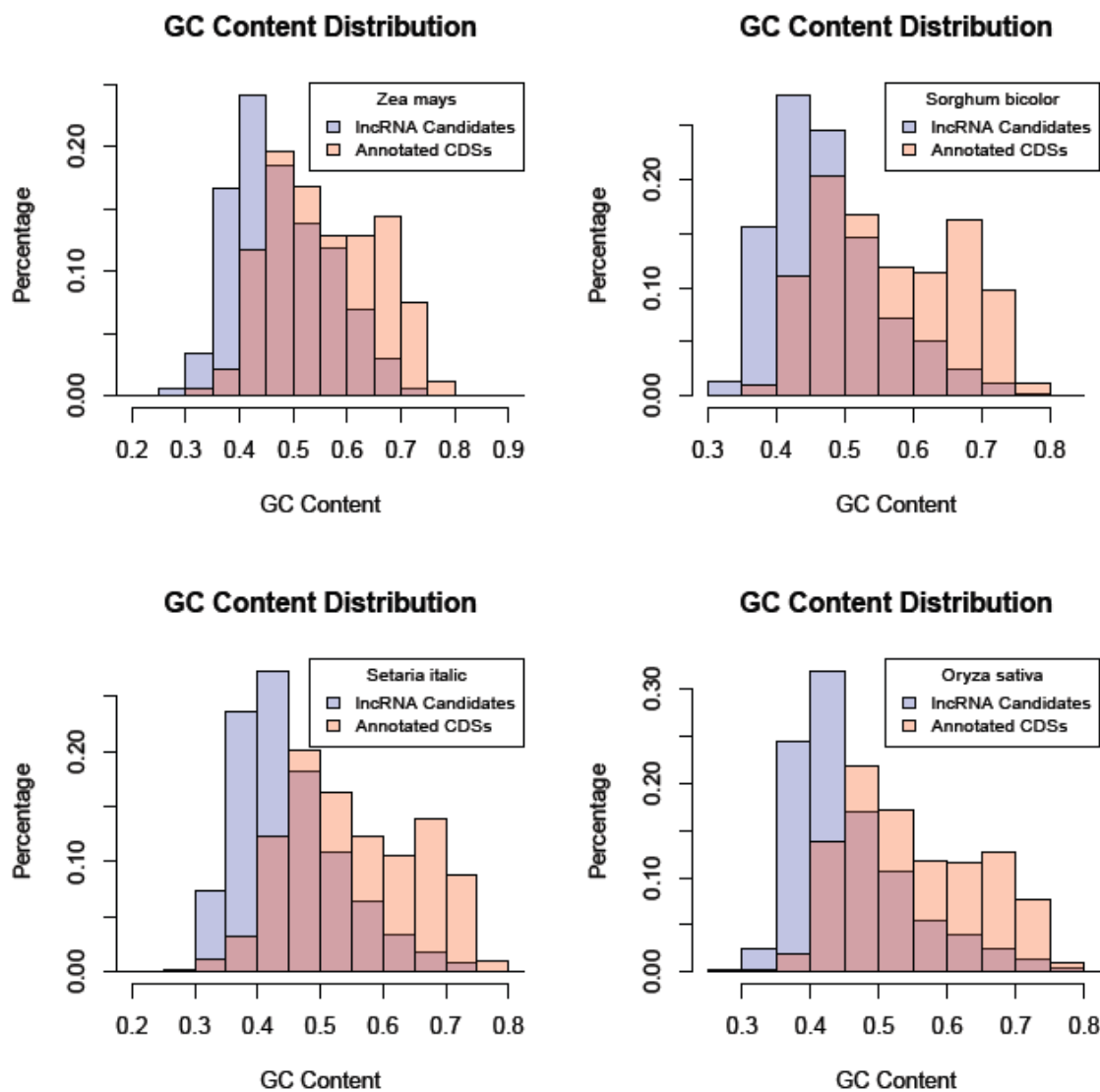


**Figure 2.3** Quality Control, Pre-processing of RNA-Seq Data (a) and "Tuxedo" Reference-Based Transcriptome Assembly (b).

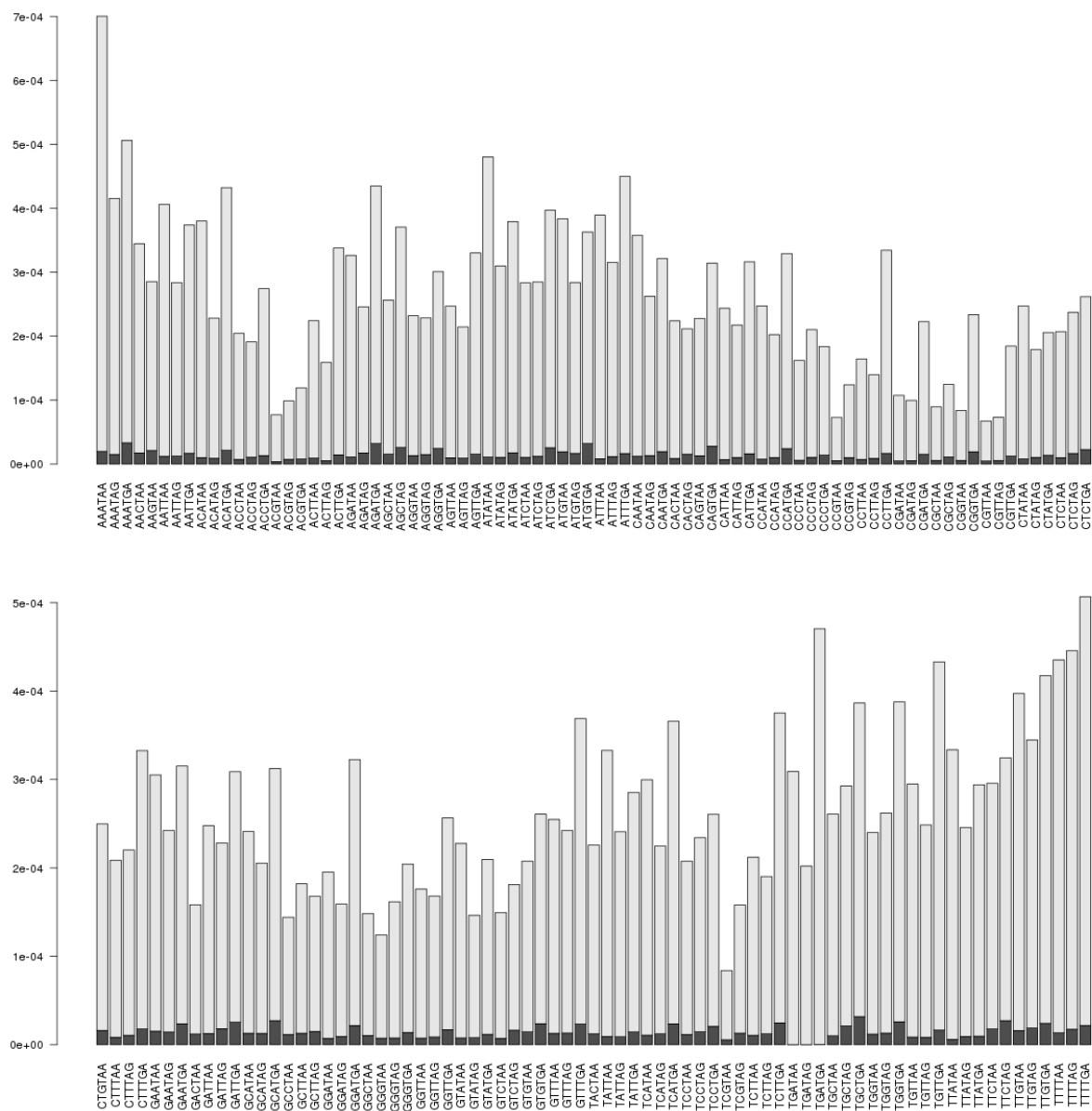


**Figure 2.4** Transcript Length, ORF Prediction and Homology Search IncRNA Identification Pipeline.

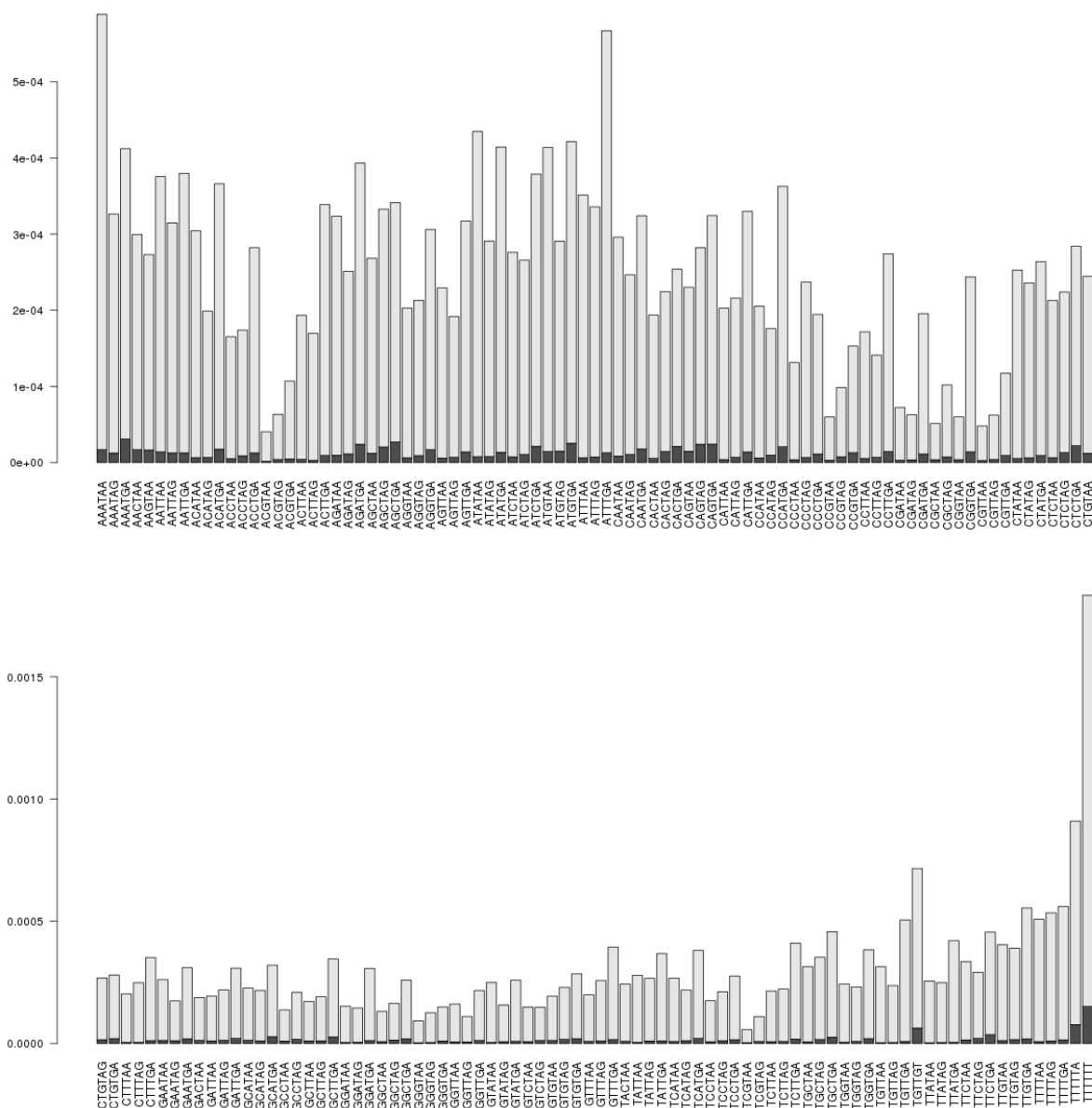




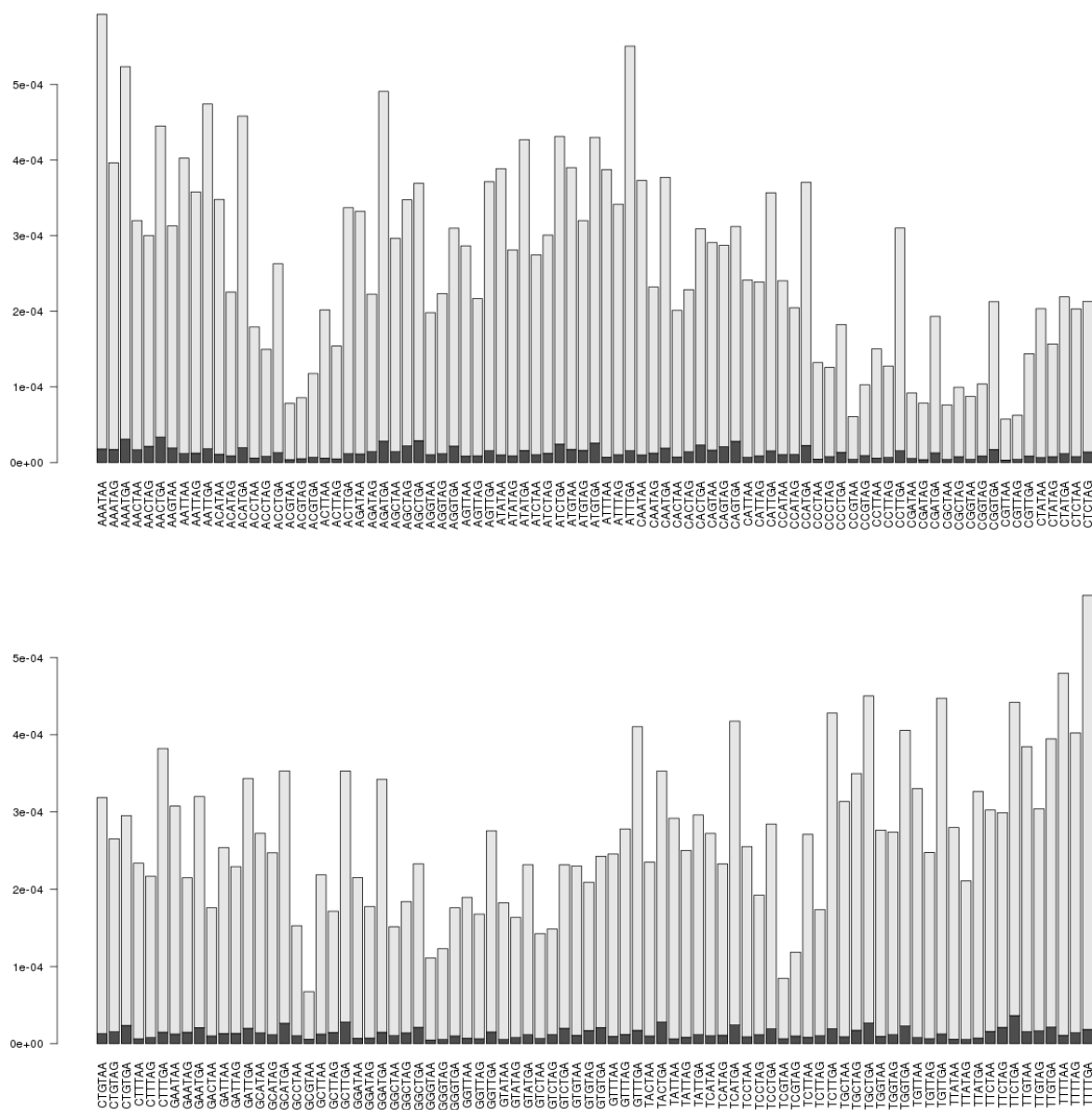
**Figure 2.5** Comparison of GC content of annotated genes (red) and our lncRNA candidates (blue). The y-axis shows the percentage of transcripts of all annotated coding sequence (CDS without UTR) or all lncRNAs; the x-axis indicates the GC content %.



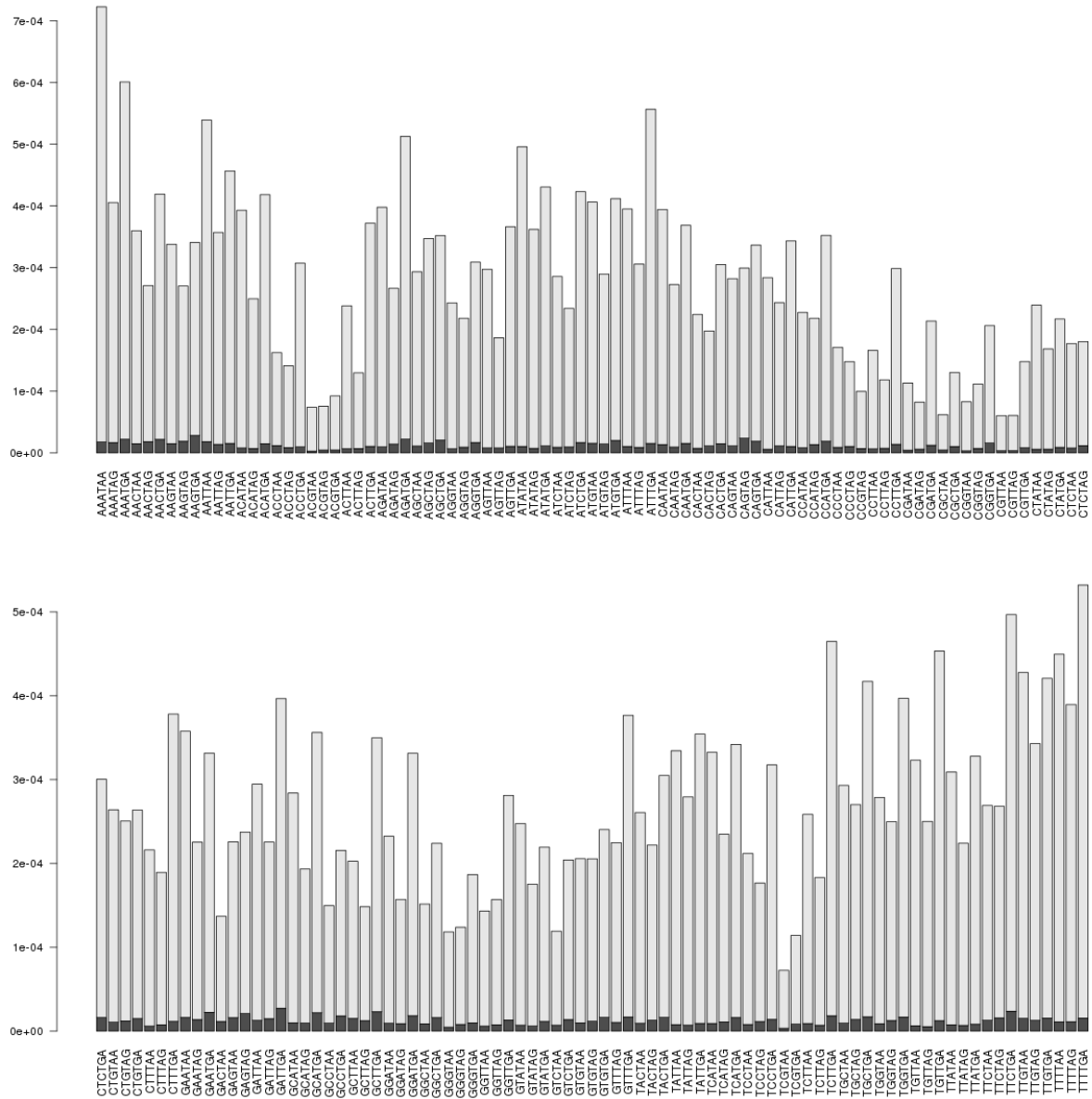
(A) *Zea mays*



(B) *Sorghum bicolor*



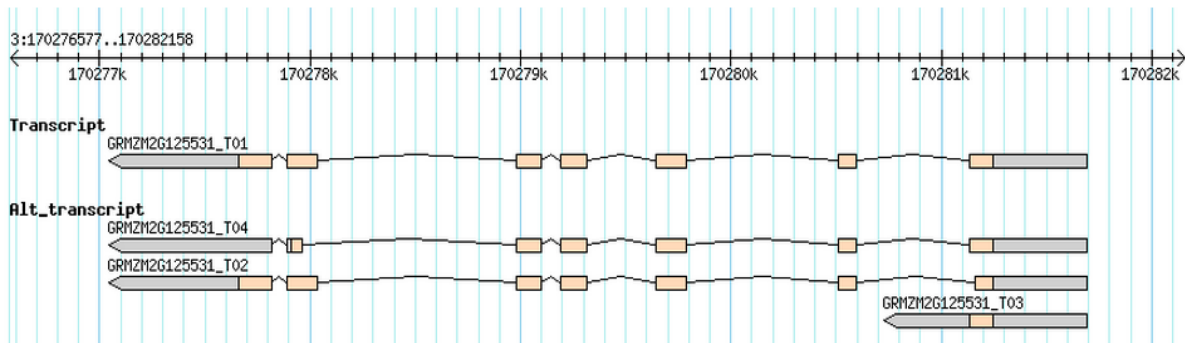
(C) *Setaria italica*



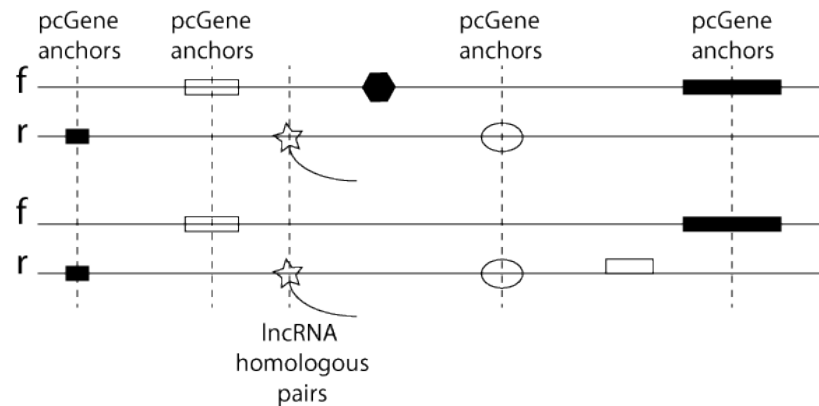
(D) *Oryza sativa*

**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*

*italica*, and 168 in *Oryza sativa*; and among these, 145 hexamers showed significant at least ten fold change between the two in all four genomes.

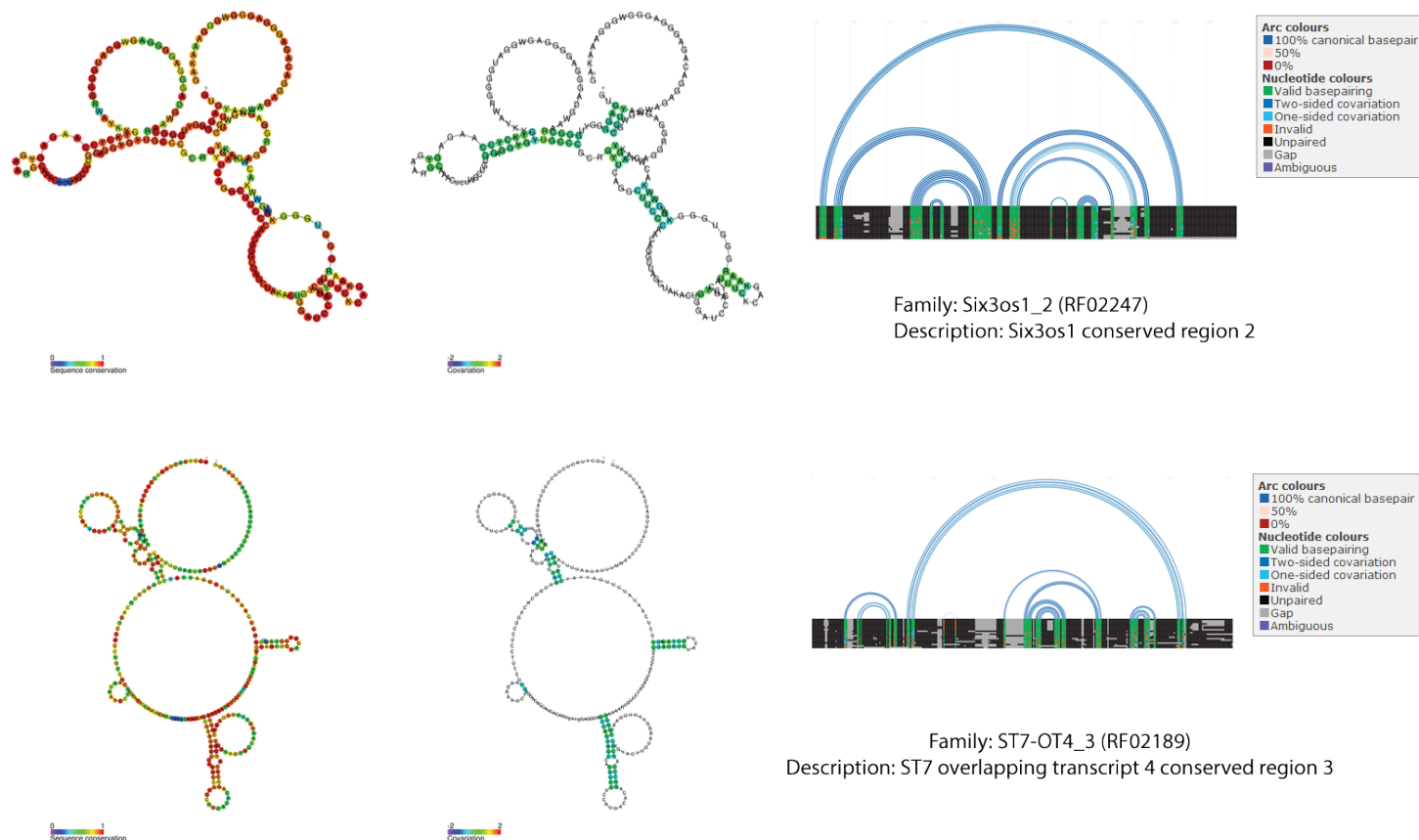


**Figure 2.7** An example of annotated genes which are identical to our lncRNA candidates. GRMZM2G125531\_T03 (right-bottom) is an alternative spliced isoform originated from a protein-coding locus in *Zea mays* genome and predicted to be a non-coding transcript by our pipeline.

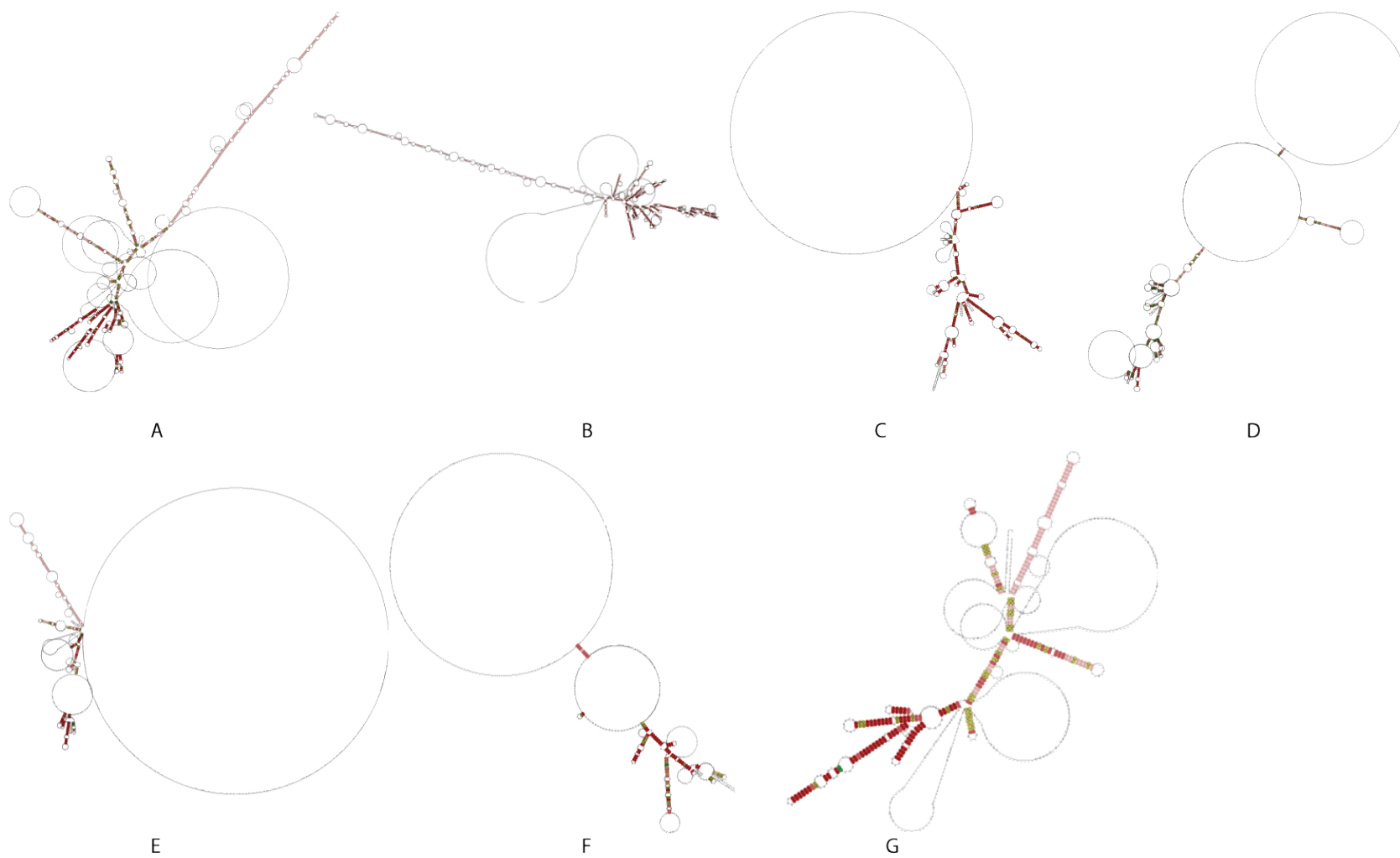


**Figure 2.8** LncRNA Orthologous Pairs (Sequence Homology and Synteny). We defined a lncRNA homologous pair as synteny pair if the lncRNA genes were located in the same synteny block between the two genomes with their adjacent upstream and downstream protein-coding gene forming un-disturbed protein-coding gene ortholog anchors. Different shapes represent different genes, and genes with the same shape are from syntenic regions. "pcGene" stands for protein-coding gene.



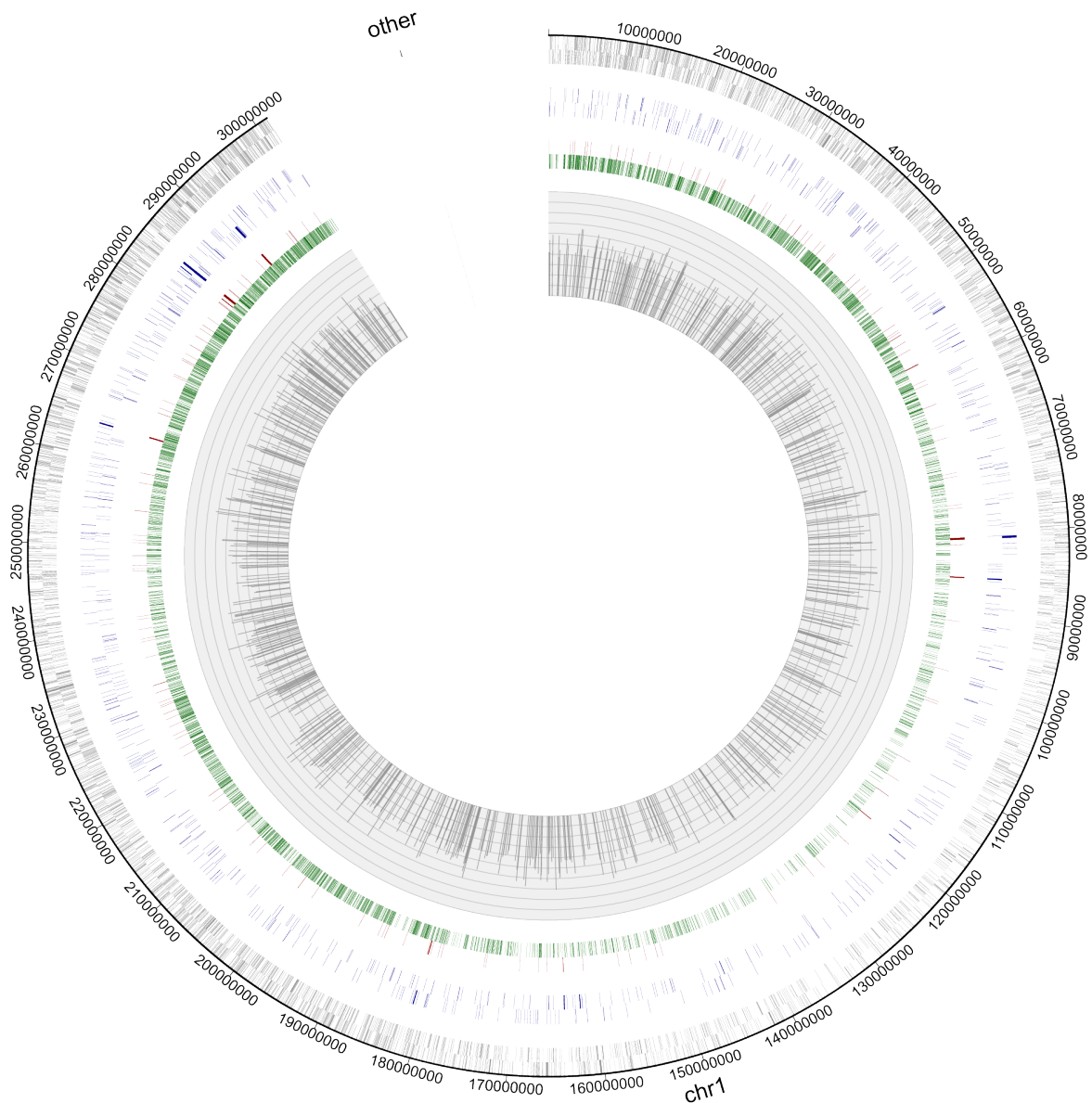


**Figure 2.9** Conserved Structures of 2 lncRNA families from Rfam, which have structure-based lncRNA homologs in more than one of *Zea mays*, *Sorghum bicolor*, *Setaria italica* and *Oryza sativa* genomes. The colors represent the sequence conservation, with red showing the most conserved sequences. 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 ( $>0.53$  and  $>0.79$  respectively) to have thermostatically stable conserved structures evaluated by RNAz global mode.



**Figure 2.11** 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 lncRNAs 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

- Alfano, G., C. Vitiello, et al. (2005). "Natural antisense transcripts associated with genes involved in eye development." Hum Mol Genet **14**(7): 913-923.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
- Amaral, P. P., M. B. Clark, et al. (2011). "lncRNADB: a reference database for long noncoding RNAs." Nucleic Acids Res **39**(Database issue): D146-151.
- Axtell, M. J., J. O. Westholm, et al. (2011). "Vive la difference: biogenesis and evolution of microRNAs in plants and animals." Genome Biol **12**(4): 221.
- Azzalin, C. M., P. Reichenbach, et al. (2007). "Telomeric repeat-containing RNA and RNA surveillance factors at mammalian chromosome ends." Science **318**(5851): 798-801.
- Badger, J. H. and G. J. Olsen (1999). "CRITICA: Coding region identification tool invoking comparative analysis." Mol Biol Evol **16**(4): 512-524.
- Baldassarre, A. and A. Masotti (2012). "Long Non-Coding RNAs and p53 Regulation." Int J Mol Sci **13**(12): 16708-16717.
- Ballabio, A. and H. F. Willard (1992). "Mammalian X-chromosome inactivation and the XIST gene." Curr Opin Genet Dev **2**(3): 439-447.
- Barciszewski, J. and V. A. Erdmann (2003). "Non-Coding RNAs: Molecular Biology and Molecular Medicine."
- Batista, P. J. and H. Y. Chang (2013). "Cytotopic localization by long noncoding RNAs." Curr Opin Cell Biol **25**(2): 195-199.
- Ben Amor, B., S. Wirth, et al. (2009). "Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses." Genome Res **19**(1): 57-69.
- Bennetzen, J. L., J. Schmutz, et al. (2012). "Reference genome sequence of the model plant *Setaria*." Nat Biotechnol **30**(6): 555-561.
- Bernstein, B. E., E. Birney, et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.



- Boerner, S. and K. M. McGinnis (2012). "Computational identification and functional predictions of long noncoding RNA in *Zea mays*." PLoS One **7**(8): e43047.
- Brannan, C. I., E. C. Dees, et al. (1990). "The product of the H19 gene may function as an RNA." Mol Cell Biol **10**(1): 28-36.
- Brockdorff, N., A. Ashworth, et al. (1991). "Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome." Nature **351**(6324): 329-331.
- Brockdorff, N., A. Ashworth, et al. (1992). "The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus." Cell **71**(3): 515-526.
- Brown, S. D. (1991). "XIST and the mapping of the X chromosome inactivation centre." Bioessays **13**(11): 607-612.
- Bu, D., K. Yu, et al. (2012). "NONCODE v3.0: integrative annotation of long noncoding RNAs." Nucleic Acids Res **40**(Database issue): D210-215.
- Burge, S. W., J. Daub, et al. (2013). "Rfam 11.0: 10 years of RNA families." Nucleic Acids Res **41**(Database issue): D226-232.
- Burleigh, S. H. and M. J. Harrison (1997). "A novel gene whose expression in *Medicago truncatula* roots is suppressed in response to colonization by vesicular-arbuscular mycorrhizal (VAM) fungi and to phosphate nutrition." Plant Mol Biol **34**(2): 199-208.
- Burleigh, S. H. and M. J. Harrison (1999). "The down-regulation of Mt4-like genes by phosphate fertilization occurs systemically and involves phosphate translocation to the shoots." Plant Physiol **119**(1): 241-248.
- Cabili, M. N., C. Trapnell, et al. (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes Dev **25**(18): 1915-1927.
- Campalans, A., A. Kondorosi, et al. (2004). "Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in *Medicago truncatula*." Plant Cell **16**(4): 1047-1059.
- Cantara, W. A., P. F. Crain, et al. (2011). "The RNA modification database, RNAMDB: 2011 update." Nucleic Acids Res **39**: D195-D201.
- Carninci, P., T. Kasukawa, et al. (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.

- Carrieri, C., L. Cimatti, et al. (2012). "Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat." Nature **491**(7424): 454-457.
- Cech, T. R. and J. Lingner (1997). "Telomerase and the chromosome end replication problem." Ciba Found Symp **211**: 20-28; discussion 28-34.
- Charon, C., C. Sousa, et al. (1999). "Alteration of enod40 expression modifies medicago truncatula root nodule development induced by sinorhizobium meliloti." Plant Cell **11**(10): 1953-1966.
- Chen, X. M. (2009). "Small RNAs and Their Roles in Plant Development." Annual Review of Cell and Developmental Biology **25**: 21-44.
- Cheng, J., P. Kapranov, et al. (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution." Science **308**(5725): 1149-1154.
- Cho, J. K., D. H. Koo, et al. (2005). "Isolation and characterization of cDNA clones expressed under male sex expression conditions in a monoecious cucumber plant (Cucumis sativus L. cv. Winter Long)." Euphytica **146**(3): 271-281.
- Chooniedass-Kothari, S., E. Emberley, et al. (2004). "The steroid receptor RNA activator is the first functional RNA encoding a protein." FEBS Lett **566**(1-3): 43-47.
- Chooniedass-Kothari, S., M. K. Hamedani, et al. (2010). "The steroid receptor RNA activator protein is recruited to promoter regions and acts as a transcriptional repressor." FEBS Lett **584**(11): 2218-2224.
- Chooniedass-Kothari, S., M. K. Hamedani, et al. (2006). "The steroid receptor RNA activator protein is expressed in breast tumor tissues." International Journal of Cancer **118**(4): 1054-1059.
- Church, D. M., L. Goodstadt, et al. (2009). "Lineage-specific biology revealed by a finished genome assembly of the mouse." PLoS Biol **7**(5): e1000112.
- Clark, M. B., R. L. Johnston, et al. (2012). "Genome-wide analysis of long noncoding RNA stability." Genome Res **22**(5): 885-898.
- Clemson, C. M., J. N. Hutchinson, et al. (2009). "An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles." Mol Cell **33**(6): 717-726.
- Core, L. J., J. J. Waterfall, et al. (2008). "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." Science **322**(5909): 1845-1848.

- Crespi, M. D., E. Jurkevitch, et al. (1994). "enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth." EMBO J **13**(21): 5099-5112.
- Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.
- Dai, X. Y., J. J. Yu, et al. (2007). "Overexpression of Zm401, an mRNA-like RNA, has distinct effects on pollen development in maize." Plant Growth Regulation **52**(3): 229-239.
- Das, S., U. Roymondal, et al. (2009). "Analyzing gene expression from relative codon usage bias in Yeast genome: a statistical significance and biological relevance." Gene **443**(1-2): 121-131.
- Dawe, R. K. (2004). "RNA interference on chromosomes." Nat Genet **36**(11): 1141-1142.
- De Lucia, F. and C. Dean (2011). "Long non-coding RNAs and chromatin regulation." Curr Opin Plant Biol **14**(2): 168-173.
- Derrien, T., R. Johnson, et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.
- Devos, K. M. (2005). "Updating the 'crop circle'." Curr Opin Plant Biol **8**(2): 155-162.
- Ding, J., Q. Lu, et al. (2012). "A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice." Proc Natl Acad Sci U S A **109**(7): 2654-2659.
- Ding, J. H., J. Q. Shen, et al. (2012). "RNA-Directed DNA Methylation Is Involved in Regulating Photoperiod-Sensitive Male Sterility in Rice." Molecular Plant **5**(6): 1210-1216.
- Dinger, M. E., K. C. Pang, et al. (2008). "Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities." PLoS Comput Biol **4**(11).
- Djebali, S., C. A. Davis, et al. (2012). "Landscape of transcription in human cells." Nature **489**(7414): 101-108.
- Du, Z., X. Zhou, et al. (2010). "agriGO: a GO analysis toolkit for the agricultural community." Nucleic Acids Res **38**(Web Server issue): W64-70.
- Ebralidze, A. K., F. C. Guibal, et al. (2008). "PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element." Genes Dev **22**(15): 2085-2092.

- Eddy, S. R. (2002). "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure." BMC Bioinformatics **3**: 18.
- Eddy, S. R. and R. Durbin (1994). "Rna Sequence-Analysis Using Covariance-Models." Nucleic Acids Res **22**(11): 2079-2088.
- Elbashir, S. M., J. Harborth, et al. (2001). "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells." Nature **411**(6836): 494-498.
- Elliott, D. and M. Lodomery (2010). "Molecular Biology of RNA." Oxford University Express.
- Fejes-Toth K, S. V., Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, et al. (2009). "Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs." Nature **457**(7232): 1028-1032.
- Feschotte, C., N. Jiang, et al. (2002). "Plant transposable elements: where genetics meets genomics." Nat Rev Genet **3**(5): 329-341.
- Finn, R. D., J. Mistry, et al. (2010). "The Pfam protein families database." Nucleic Acids Res **38**: D211-D222.
- Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*." Nature **391**(6669): 806-811.
- Forne, T., J. Oswald, et al. (1997). "Loss of the maternal H19 gene induces changes in Igf2 methylation in both cis and trans." Proc Natl Acad Sci U S A **94**(19): 10243-10248.
- Franco-Zorrilla, J. M., A. Valli, et al. (2007). "Target mimicry provides a new mechanism for regulation of microRNA activity." Nat Genet **39**(8): 1033-1037.
- Freyhult, E. K., J. P. Bollback, et al. (2007). "Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA." Genome Res **17**(1): 117-125.
- Furini, A. (2008). "CDT retroelement: The stratagem to survive extreme vegetative dehydration." Plant Signal Behav **3**(12): 1129-1131.
- Gent, J. I. and R. K. Dawe (2012). "RNA as a Structural and Regulatory Component of the Centromere." Annual Review of Genetics, Vol 46 **46**: 443-453.
- Georg, J., A. Honsel, et al. (2010). "A long antisense RNA in plant chloroplasts." New Phytologist **186**(3): 615-622.
- Gesteland, R. F. (2005). "The RNA World." Cold Spring Harbor Laboratory Press.

- Girard, G., A. Roussis, et al. (2003). "Structural motifs in the RNA encoded by the early nodulation gene enod40 of soybean." Nucleic Acids Res **31**(17): 5003-5015.
- Gish, W. and D. J. States (1993). "Identification of Protein Coding Regions by Database Similarity Search." Nat Genet **3**(3): 266-272.
- Gough, J., K. Karplus, et al. (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." J Mol Biol **313**(4): 903-919.
- Gruber, A. R., S. Findeiss, et al. (2010). "RNAz 2.0: improved noncoding RNA detection." Pac Symp Biocomput: 69-79.
- Gruber, A. R., R. Neuboeck, et al. (2007). "The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures." Nucleic Acids Res **35**: W335-W338.
- Gu, R., Z. Zhang, et al. (2009). "Gene regulation by sense-antisense overlap of polyadenylation signals." RNA **15**(6): 1154-1163.
- Guerrier-Takada, C., K. Gardiner, et al. (1983). "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." Cell **35**(3 Pt 2): 849-857.
- Guil, S., M. Soler, et al. (2012). "Intronic RNAs mediate EZH2 regulation of epigenetic targets." Nat Struct Mol Biol **19**(7): 664-670.
- Gulyaev, A. P. and A. Roussis (2007). "Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants." Nucleic Acids Res **35**(9): 3144-3152.
- Gupta, R. A., N. Shah, et al. (2010). "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." Nature **464**(7291): 1071-1076.
- Guttman, M., I. Amit, et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.
- Guttman, M., M. Garber, et al. (2010). "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nat Biotechnol **28**(5): 503-U166.
- Hadjiargyrou, M. and N. Delihas (2013). "The Intertwining of Transposable Elements and Non-Coding RNAs." Int J Mol Sci **14**(7): 13307-13328.

- Hammond, S. M. (2006). "RNAi, microRNAs, and human disease." Cancer Chemother Pharmacol **58 Suppl 1**: s63-68.
- Hangauer, M. J., I. W. Vaughn, et al. (2013). "Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs." PLoS Genet **9**(6): e1003569.
- Hansey, C. N., B. Vaillancourt, et al. (2012). "Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing." PLoS One **7**(3): e33071.
- Hao, Y., T. Crenshaw, et al. (1993). "Tumour-suppressor activity of H19 RNA." Nature **365**(6448): 764-767.
- Harrow, J., A. Frankish, et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774.
- Heo, J. B. and S. Sung (2011). "Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA." Science **331**(6013): 76-79.
- Hirsch, J., V. Lefort, et al. (2006). "Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts." Plant Physiol **140**(4): 1192-1204.
- Hu, W. Q., B. B. Yuan, et al. (2011). "Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation." Genes Dev **25**(24): 2573-2578.
- Huang, Y., J. L. Zhang, et al. (2013). "Molecular functions of small regulatory noncoding RNA." Biochemistry (Mosc) **78**(3): 221-230.
- Huang, Z., Malmberg, R., Mohebbi M., and Cai, L. (2010). "RNAv: Non-coding RNA secondary structure variation search via graph homomorphism." In CSB Conference Proceedings, CA, USA: 56-69
- Huarte, M., M. Guttman, et al. (2010). "A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response." Cell **142**(3): 409-419.
- Hung, T., Y. L. Wang, et al. (2011). "Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters." Nat Genet **43**(7): 621-U196.
- Hupaló, D. and A. D. Kern (2013). "Conservation and Functional Element Discovery in 20 Angiosperm Plant Genomes." Mol Biol Evol **30**(7): 1729-1744.
- Hurst, L. D., C. Pal, et al. (2004). "The evolutionary dynamics of eukaryotic gene order." Nat Rev Genet **5**(4): 299-310.

- Ilott, N. E. and C. P. Ponting (2013). "Predicting long non-coding RNAs using RNA sequencing." Methods **63**(1): 50-59.
- Jackson, A. L., J. Burchard, et al. (2006). "Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity." RNA **12**(7): 1179-1187.
- Jeffares, D. C., A. M. Poole, et al. (1998). "Relics from the RNA world." J Mol Evol **46**(1): 18-36.
- Jia, H., M. Osak, et al. (2010). "Genome-wide computational identification and manual annotation of human long noncoding RNA genes." RNA **16**(8): 1478-1487.
- Jouannet, V. and M. Crespi (2011). "Long Nonprotein-Coding RNAs in Plants." Prog Mol Subcell Biol **51**: 179-200.
- Kapranov, P., J. Cheng, et al. (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription." Science **316**(5830): 1484-1488.
- Kapranov, P. and G. St Laurent (2012). "Dark Matter RNA: Existence, Function, and Controversy." Front Genet **3**: 60.
- Katayama, S., Y. Tomaru, et al. (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-1566.
- Kelley, D. and J. Rinn (2012). "Transposable elements reveal a stem cell-specific class of long noncoding RNAs." Genome Biol **13**(11): R107.
- Khalil, A. M., M. Guttman, et al. (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." Proc Natl Acad Sci U S A **106**(28): 11667-11672.
- Kim, E. D. and S. Sung (2012). "Long noncoding RNA: unveiling hidden layer of gene regulatory networks." Trends Plant Sci **17**(1): 16-21.
- Kino, T., D. E. Hurt, et al. (2010). "Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor." Science Signaling **3**(107).
- Klattenhoff, C. and W. Theurkauf (2008). "Biogenesis and germline functions of piRNAs." Development **135**(1): 3-9.
- Knowles, D. G. and A. McLysaght (2009). "Recent de novo origin of human protein-coding genes." Genome Res **19**(10): 1752-1759.

- Kogo, R., T. Shimamura, et al. (2011). "Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers." Cancer Res **71**(20): 6320-6326.
- Kong, L., Y. Zhang, et al. (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Res **35**: W345-W349.
- Kotake, Y., T. Nakagawa, et al. (2011). "Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene." Oncogene **30**(16): 1956-1962.
- Kouchi, H., K. Takane, et al. (1999). "Rice ENOD40: isolation and expression analysis in rice and transgenic soybean root nodules." Plant J **18**(2): 121-129.
- Kung, J. T., D. Colognori, et al. (2013). "Long noncoding RNAs: past, present, and future." Genetics **193**(3): 651-669.
- Kurth, H. M. and K. Mochizuki (2009). "Non-coding RNA: a bridge between small RNA and DNA." RNA Biol **6**(2): 138-140.
- Kutter, C., S. Watt, et al. (2012). "Rapid turnover of long noncoding RNAs and the evolution of gene expression." PLoS Genet **8**(7): e1002841.
- Lanz, R. B., N. J. McKenna, et al. (1999). "A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex." Cell **97**(1): 17-27.
- Lee, J. T. (2009). "Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome." Genes Dev **23**(16): 1831-1842.
- Lee, J. T. (2012). "Epigenetic Regulation by Long Noncoding RNAs." Science **338**.
- Lee, J. T., L. S. Davidow, et al. (1999). "Tsix, a gene antisense to Xist at the X-inactivation centre." Nat Genet **21**(4): 400-404.
- Lee, T. H., H. Tang, et al. (2013). "PGDD: a database of gene and genome duplication in plants." Nucleic Acids Res **41**(Database issue): D1152-1158.
- Lercher, M. J., A. O. Urrutia, et al. (2003). "A unification of mosaic structures in the human genome." Hum Mol Genet **12**(19): 2411-2415.
- Li, D., J. Feng, et al. (2013). "Long intergenic noncoding RNA HOTAIR is overexpressed and regulates PTEN methylation in laryngeal squamous cell carcinoma." Am J Pathol **182**(1): 64-70.



- Li, L. and Y. Liu (2011). "Diverse small non-coding RNAs in RNA interference pathways." Methods Mol Biol **764**: 169-182.
- Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.
- Lin, M. F., I. Jungreis, et al. (2011). "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." Bioinformatics **27**(13): I275-I282.
- Lisch, D., P. Chomet, et al. (1995). "Genetic characterization of the Mutator system in maize: behavior and regulation of Mu transposons in a minimal line." Genetics **139**(4): 1777-1796.
- Liu, C. M., U. S. Muchhal, et al. (1997). "Differential expression of TPS11, a phosphate starvation-induced gene in tomato." Plant Mol Biol **33**(5): 867-874.
- Liu, J., C. Jung, et al. (2012). "Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis." Plant Cell **24**(11): 4333-4345.
- Liu, J. F., J. Gough, et al. (2006). "Distinguishing protein-coding from non-coding RNAs through support vector machines." PLoS Genet **2**(4): 529-536.
- Lorenz, R., S. H. Bernhart, et al. (2011). "ViennaRNA Package 2.0." Algorithms Mol Biol **6**: 26.
- Louro, R., A. S. Smirnova, et al. (2009). "Long intronic noncoding RNA transcription: expression noise or expression choice?" Genomics **93**(4): 291-298.
- Luke, B. and J. Lingner (2009). "TERRA: telomeric repeat-containing RNA." Embo Journal **28**(17): 2503-2510.
- Lyle, R., D. Watanabe, et al. (2000). "The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1." Nat Genet **25**(1): 19-21.
- Lyons E, F. M. (2008). "How to usefully compare homologous plant genes and chromosomes as DNA sequences." The Plant Journal (53): 661-673.
- Ma, J. X., B. X. Yan, et al. (2008). "Zm401, a short-open reading-frame mRNA or noncoding RNA, is essential for tapetum and microspore development and can regulate the floret formation in maize." J Cell Biochem **105**(1): 136-146.
- Ma L, Bajic VB, et al. (2013). "On the classification of long non-coding RNAs." RNA Biology **10**: 925 - 934.

- MacIntosh, G. C., C. Wilkerson, et al. (2001). "Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs." Plant Physiol **127**(3): 765-776.
- Martin, A. C., J. C. del Pozo, et al. (2000). "Influence of cytokinins on the expression of phosphate starvation responsive genes in Arabidopsis." Plant Journal **24**(5): 559-567.
- Matouk, I. J., N. DeGroot, et al. (2007). "The H19 non-coding RNA is essential for human tumor growth." PLoS One **2**(9): e845.
- Matsui, A., J. Ishida, et al. (2008). "Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array." Plant and Cell Physiology **49**(8): 1135-1149.
- Matsui, K., M. Nishizawa, et al. (2008). "Natural antisense transcript stabilizes inducible nitric oxide synthase messenger RNA in rat hepatocytes." Hepatology **47**(2): 686-697.
- Mayer, C., M. Neubert, et al. (2008). "The structure of NoRC-associated RNA is crucial for targeting the chromatin remodelling complex NoRC to the nucleolus." EMBO Rep **9**(8): 774-780.
- McCarty, D. R., A. M. Settles, et al. (2005). "Steady-state transposon mutagenesis in inbred maize." Plant J **44**(1): 52-61.
- McKeown, M. (1993). "The role of small nuclear RNAs in RNA splicing." Curr Opin Cell Biol **5**(3): 448-454.
- Meister, G. and T. Tuschl (2004). "Mechanisms of gene silencing by double-stranded RNA." Nature **431**(7006): 343-349.
- Meli, M., B. Albert-Fournier, et al. (2001). "Recent findings in the modern RNA world." Int Microbiol **4**(1): 5-11.
- Mercer, T. R., M. E. Dinger, et al. (2009). "Long non-coding RNAs: insights into functions." Nature Reviews Genetics **10**(3): 155-159.
- Mercer, T. R., D. J. Gerhardt, et al. (2012). "Targeted RNA sequencing reveals the deep complexity of the human transcriptome." Nat Biotechnol **30**(1): 99-104.
- Mercer, T. R. and J. S. Mattick (2013). "Structure and function of long noncoding RNAs in epigenetic regulation." Nat Struct Mol Biol **20**(3): 300-307.
- Moghe, G. D., M. D. Lehti-Shiu, et al. (2013). "Characteristics and significance of intergenic polyadenylated RNA transcription in Arabidopsis." Plant Physiol **161**(1): 210-224.

- Morozova, O., M. Hirst, et al. (2009). "Applications of new sequencing technologies for transcriptome analysis." Annu Rev Genomics Hum Genet **10**: 135-151.
- Murchison, E. P. and G. J. Hannon (2004). "miRNAs on the move: miRNA biogenesis and the RNAi machinery." Curr Opin Cell Biol **16**(3): 223-229.
- Nagano, T., J. A. Mitchell, et al. (2008). "The Air Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin." Science **322**(5908): 1717-1720.
- Nam, J. W. and D. P. Bartel (2012). "Long noncoding RNAs in *C. elegans*." Genome Res **22**(12): 2529-2540.
- Nawrocki, E. P., D. L. Kolbe, et al. (2009). "Infernal 1.0: inference of RNA alignments." Bioinformatics **25**(10): 1335-1337.
- Niazi, F. and S. Valadkhan (2012). "Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs." Rna-a Publication of the Rna Society **18**(4): 825-843.
- Novikova, I. V., S. P. Hennelly, et al. (2012). "Structural architecture of the human long non-coding RNA, steroid receptor RNA activator." Nucleic Acids Res **40**(11): 5034-5051.
- Ohno, S. (1972). "So much "junk" DNA in our genome." Brookhaven Symp Biol **23**: 366-370.
- Okamoto, M., K. Tatematsu, et al. (2010). "Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of *Arabidopsis* using tiling arrays." Plant Journal **62**(1): 39-51.
- Okonechnikov, K., O. Golosova, et al. (2012). "Unipro UGENE: a unified bioinformatics toolkit." Bioinformatics **28**(8): 1166-1167.
- Orom, U. A., T. Derrien, et al. (2010). "Long noncoding RNAs with enhancer-like function in human cells." Cell **143**(1): 46-58.
- Osato, N., Y. Suzuki, et al. (2007). "Transcriptional interferences in cis natural antisense transcripts of humans and mice." Genetics **176**(2): 1299-1306.
- Ouyang, S., W. Zhu, et al. (2007). "The TIGR Rice Genome Annotation Resource: improvements and new features." Nucleic Acids Res **35**(Database issue): D883-887.
- Pandey, R. R., T. Mondal, et al. (2008). "Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation." Mol Cell **32**(2): 232-246.

- Parker, B. J., I. Moltke, et al. (2011). "New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes." Genome Res **21**(11): 1929-1943.
- Paterson, A. H., J. E. Bowers, et al. (2009). "The Sorghum bicolor genome and the diversification of grasses." Nature **457**(7229): 551-556.
- Pedersen, J. S., G. Bejerano, et al. (2006). "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput Biol **2**(4): e33.
- Ponjavic, J., C. P. Ponting, et al. (2007). "Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs." Genome Res **17**(5): 556-565.
- Ponting, C. P., P. L. Oliver, et al. (2009). "Evolution and functions of long noncoding RNAs." Cell **136**(4): 629-641.
- Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.
- Rearick, D., A. Prakash, et al. (2011). "Critical association of ncRNA with introns." Nucleic Acids Res **39**(6): 2357-2366.
- Rinn, J. L. and H. Y. Chang (2012). "Genome regulation by long noncoding RNAs." Annu Rev Biochem **81**: 145-166.
- Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics **2**: 8.
- Roberts, A., H. Pimentel, et al. (2011). "Identification of novel transcripts in annotated genomes using RNA-Seq." Bioinformatics **27**(17): 2325-2329.
- Runge, S., F. C. Nielsen, et al. (2000). "H19 RNA binds four molecules of insulin-like growth factor II mRNA-binding protein." J Biol Chem **275**(38): 29562-29569.
- Rychlik, W. (2007). "OLIGO 7 primer analysis software." Methods Mol Biol **402**: 35-60.
- Salmena, L., L. Poliseno, et al. (2011). "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?" Cell **146**(3): 353-358.
- Sasaki, Y. T. F., T. Ideue, et al. (2009). "MEN epsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles." Proc Natl Acad Sci U S A **106**(8): 2525-2530.
- Schnable, P. S., D. Ware, et al. (2009). "The B73 maize genome: complexity, diversity, and dynamics." Science **326**(5956): 1112-1115.

- Schoeftner, S. and M. A. Blasco (2008). "Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II." Nat Cell Biol **10**(2): 228-U106.
- Seila, A. C., J. M. Calabrese, et al. (2008). "Divergent transcription from active promoters." Science **322**(5909): 1849-1851.
- Seitz, H. (2009). "Redefining microRNA targets." Curr Biol **19**(10): 870-873.
- Semon, M., D. Mouchiroud, et al. (2005). "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance." Hum Mol Genet **14**(3): 421-427.
- Shin, H., H. S. Shin, et al. (2006). "Loss of At4 function impacts phosphate distribution between the roots and the shoots during phosphate starvation." Plant Journal **45**(5): 712-726.
- Sievers, F., A. Wilm, et al. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**: 539.
- Sigova, A. A., A. C. Mullen, et al. (2013). "Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells." Proc Natl Acad Sci U S A **110**(8): 2876-2881.
- Smilnich, N. J., C. D. Day, et al. (1999). "A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome." Proc Natl Acad Sci U S A **96**(14): 8064-8069.
- Smit, A., Hubley, R & Green, P. (1996-2010). "RepeatMasker Open-3.0." <http://www.repeatmasker.org>.
- Smith, M. A., T. Gesell, et al. (2013). "Widespread purifying selection on RNA structure in mammals." Nucleic Acids Res **41**(17): 8220-8236.
- Sun, L., Z. H. Zhang, et al. (2012). "Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study." BMC Bioinformatics **13**.
- Swiezewski, S., F. Q. Liu, et al. (2009). "Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target." Nature **462**(7274): 799-U122.
- Taylor, C. B. and P. J. Green (1995). "Identification and Characterization of Genes with Unstable Transcripts (Guts) in Tobacco." Plant Mol Biol **28**(1): 27-38.
- Teramoto, H., T. Toyama, et al. (1996). "Noncoding RNA for CR20, a cytokinin-repressed gene of cucumber." Plant Mol Biol **32**(5): 797-808.

- Thomas, C. A., Jr. (1971). "The genetic organization of chromosomes." Annu Rev Genet **5**: 237-256.
- Tollervey, D. (1996). "Small nucleolar RNAs guide ribosomal RNA methylation." Science **273**(5278): 1056-1057.
- Topp, C. N., C. X. Zhong, et al. (2004). "Centromere-encoded RNAs are integral components of the maize kinetochore." Proc Natl Acad Sci U S A **101**(45): 15986-15991.
- Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.
- Trapnell, C., A. Roberts, et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat Protoc **7**(3): 562-578.
- Trapnell, C., B. A. Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.
- Tsai, M. C., O. Manor, et al. (2010). "Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes." Science **329**(5992): 689-693.
- Tsang, W. P., E. K. Ng, et al. (2010). "Oncofetal H19-derived miR-675 regulates tumor suppressor RB in human colorectal cancer." Carcinogenesis **31**(3): 350-358.
- Uchida, W., S. Matsunaga, et al. (2002). "Interstitial telomere-like repeats in the Arabidopsis thaliana genome." Genes Genet Syst **77**(1): 63-67.
- Ulitsky, I. and D. P. Bartel (2013). "lincRNAs: genomics, evolution, and mechanisms." Cell **154**(1): 26-46.
- Ulitsky, I., A. Shkumatava, et al. (2012). "Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution (vol 147, pg 1537, 2011)." Cell **151**(3): 684-686.
- van Hoof, A., Kastenmayer, J.P., Taylor, C.B. and Green, P.J. (1997). "GUT15 cDNAs from tobacco (Accession No. U84972) and Arabidopsis (Accession No. U84973) correspond to transcripts with unusual metabolism and a short conserved open reading frame (PGR97-048)." Plant Physiol.
- Vannier, J. B., A. Depeiges, et al. (2009). "ERCC1/XPF Protects Short Telomeres from Homologous Recombination in Arabidopsis thaliana." PLoS Genet **5**(2).

- Vincent, J. B., E. Petek, et al. (2002). "The RAY1/ST7 tumor-suppressor locus on chromosome 7q31 represents a complex multi-transcript system." Genomics **80**(3): 283-294.
- Vrbsky, J., S. Akimcheva, et al. (2010). "siRNA-mediated methylation of Arabidopsis telomeres." PLoS Genet **6**(6): e1000986.
- Wan, Y., M. Kertesz, et al. (2011). "Understanding the transcriptome through RNA structure." Nature Reviews Genetics **12**(9): 641-655.
- Wang, K. C., Y. W. Yang, et al. (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression." Nature **472**(7341): 120-U158.
- Wang, L., H. J. Park, et al. (2013). "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model." Nucleic Acids Res **41**(6).
- Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.
- Washietl, S., S. Findeiss, et al. (2011). "RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data." Rna-a Publication of the Rna Society **17**(4): 578-594.
- Washietl, S., S. Findeiss, et al. (2011). "RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data." RNA **17**(4): 578-594.
- Washietl, S. and I. L. Hofacker (2007). "Identifying structural noncoding RNAs using RNAz." Curr Protoc Bioinformatics **Chapter 12**: Unit 12 17.
- Washietl, S., I. L. Hofacker, et al. (2005). "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nat Biotechnol **23**(11): 1383-1390.
- Weinberg, Z. and W. L. Ruzzo (2004). "Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy." Bioinformatics **20 Suppl 1**: i334-341.
- Weinberg, Z. and W. L. Ruzzo (2006). "Sequence-based heuristics for faster annotation of non-coding RNA families." Bioinformatics **22**(1): 35-39.
- Willingham, A. T., A. P. Orth, et al. (2005). "A strategy for probing the function of noncoding RNAs finds a repressor of NFAT." Science **309**(5740): 1570-1573.
- Wilusz, J. E., H. Sunwoo, et al. (2009). "Long noncoding RNAs: functional surprises from the RNA world." Genes Dev **23**(13): 1494-1504.

- Wong, L. H., K. H. Brettingham-Moore, et al. (2007). "Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere." Genome Res **17**(8): 1146-1160.
- Wu, D. D., D. M. Irwin, et al. (2011). "De novo origin of human protein-coding genes." PLoS Genet **7**(11): e1002379.
- Wu, H. J., Z. M. Wang, et al. (2013). "Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants." Plant Physiol **161**(4): 1875-1884.
- Wuchty, S., W. Fontana, et al. (1999). "Complete suboptimal folding of RNA and the stability of secondary structures." Biopolymers **49**(2): 145-165.
- Xiao, S., F. Scott, et al. (2002). "Eukaryotic ribonuclease P: a plurality of ribonucleoprotein enzymes." Annu Rev Biochem **71**: 165-189.
- Xie, C., Y. E. Zhang, et al. (2012). "Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs." PLoS Genet **8**(9): e1002942.
- Xin, M. M., Y. Wang, et al. (2011). "Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing." BMC Plant Biol **11**.
- Yao, H., K. Brick, et al. (2010). "Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA." Genes Dev **24**(22): 2543-2555.
- Ye, J., G. Coulouris, et al. (2012). "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction." BMC Bioinformatics **13**: 134.
- Yoshimizu, T., A. Miroglio, et al. (2008). "The H19 locus acts in vivo as a tumor suppressor." Proc Natl Acad Sci U S A **105**(34): 12417-12422.
- Zhang, G., X. Liu, et al. (2012). "Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential." Nat Biotechnol **30**(6): 549-554.
- Zhang, M., H. Zhao, et al. (2011). "Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm." Proc Natl Acad Sci U S A **108**(50): 20042-20047.
- Zhang, Q. F., B. Z. Shen, et al. (1994). "Using Bulk Extremes and Recessive Class to Map Genes for Photoperiod-Sensitive Genic Male-Sterility in Rice." Proc Natl Acad Sci U S A **91**(18): 8675-8679.



- Zhang, X., K. Rice, et al. (2010). "Maternally Expressed Gene 3 (MEG3) Noncoding Ribonucleic Acid: Isoform Structure, Expression, and Functions." Endocrinology **151**(3): 939-947.
- Zhang, Y. (2012). "RNA-Seq Module 2: From QC to differential gene expression (workshop)."
- Zhang, Y. C. and Y. Q. Chen (2013). "Long noncoding RNAs: new regulators in plant development." Biochem Biophys Res Commun **436**(2): 111-114.
- Zhao, J., B. K. Sun, et al. (2008). "Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome." Science **322**(5902): 750-756.
- Zhu, Q.-H. and M.-B. Wang (2012). "Molecular Functions of Long Non-Coding RNAs in Plants." Genes **3**(1): 176-190.