# Evolution of Duplicated Non-Coding RNAs in Plants

by

Sishuo Wang

B.Sc., China Agricultural University, 2012

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2017

**Abstract**

Non-coding RNAs (ncRNAs) consist of microRNAs, lincRNAs (long intergenic non-coding RNA), rRNAs, tRNAs and the RNAs from other types of genes that do not have the potential to be protein-coding. Non-coding RNAs play various roles in cellular processes. Gene duplication is a major force in gene evolution and the evolution of duplicated protein-coding genes has been studied extensively. Whether the same evolutionary principles hold true for ncRNAs, especially lincRNAs, is still poorly understood particularly in plants. I characterized the effects of the change in microRNA binding sites on the divergence of multiple types of duplicated genes in *Arabidopsis thaliana* and *Brassica rapa* (Chapter 2). I found that the vast majority of duplicated genes showed divergence in their microRNA binding sites that could be associated with their expression and functional divergence. To better understand the evolutionary dynamics of lincRNAs in plants, I analyzed the sequence evolution of lincRNAs from five species (*Arabidopsis thaliana*, *Oryza sativa ssp. japonica*, *Zea mays*, *Medicago truncatula* and *Solanum lycopersicum*) across 55 plant genomes (Chapter 3). My analyses revealed that lincRNAs show more rapid sequence divergence compared with protein-coding genes and microRNAs. I also analyzed the expression conservation of lincRNAs between closely related species and showed rapid expression evolution of lincRNAs. I also identified a considerable number of conserved regions in the sequence of lincRNAs that are under stronger selection constraints than surrounding regions. To investigate the role of gene duplication in the evolution of plant lincRNAs, I identified duplicated lincRNAs

in several plant species (Chapter 4). I compared the expression patterns between duplicated lincRNAs using RNA-seq data from multiple tissue types and developmental stages, revealing extensive expression divergence of lincRNAs. Finally, I studied the effects of polyploidy and abiotic stress on the expression of lincRNAs in diploid and polyploid *Brassica* species (Chapter 5). My results showed extensive divergence of the expression of lincRNAs after polyploidy and in response to different stresses. This thesis provides new insights into lincRNA evolution and fates of lincRNAs after duplication in flowering plants.

## Lay Summary

Gene duplication is a major process by which new genes are created and evolve. Two major types of genes include those for proteins and those for non-coding RNAs (ncRNAs). To better understand the evolution of ncRNAs, I investigated the evolution and duplication of ncRNAs in plants. I analyzed the conservation of ncRNAs in a broad selection of plants and found rapid changes of ncRNAs in both sequence and expression. I identified short conserved regions in ncRNAs across species that might be associated with functions of ncRNAs. Also, I explored the evolution and expression of duplicated ncRNAs and found rapid divergence between ncRNAs after duplication. Additionally, I compared the evolution of ncRNAs and proteins, and proposed potential factors underlying their differences. This research considerably advances our knowledge of the evolutionary properties of ncRNAs and how they evolve after gene duplication.

**Preface**

Chapter 2 has been published. **Wang SS**, Adams KL (2015) Duplicate Gene Divergence by Changes in MicroRNA Binding Sites in Arabidopsis and Brassica. *Genome Biology And Evolution* 7: 646-655. I conceived and designed the study, conducted all the analyses, and wrote most of the manuscript. KLA helped with the study design and edited the manuscript.

Chapter 3 has been submitted for publication. **Wang SS**, Hammel A, Adams, KL Rapid evolution of sequence and expression of lincRNAs in flowering plants. I designed the study, conducted most of the analyses, and wrote the manuscript. AH helped with the study design and performed population genomics analyses. KLA conceived the study, helped with the study design, and edited the manuscript.

Chapter 4 is in preparation for publication. **Wang SS**, Adams KL Evolution of lincRNAs by duplication in plants. I designed and conducted all the analyses and wrote the manuscript. KLA conceived the study, helped with the study design, and edited the manuscript.

Chapter 5 is planned for publication. **Wang SS**, Desbiez-Piat A, Adams KL Expression divergence of lincRNAs after polyploidy and in response to abiotic stresses in *Brassica*. I designed and conducted most analyses, and wrote the manuscript. AD conducted some

analyses of gene expression. KLA conceived the study and edited the thesis chapter.

Similar information is listed in the footnotes on the first pages of these chapters

# Table of Contents

**List of Tables**

**List of Figures**

**Acknowledgements**

I would like to thank my supervisor, Keith Adams, for giving me the excellent opportunity to study several interesting questions in plant genome evolution as well as his great guidance throughout my Ph.D study. I would also like to thank my committee members Quentin Cronk, Carl Douglas, and Naomi Fast, and my examiners Jae-Hyeok Lee, Lewis Lukens , and Rosie Redfield for their helpful comments. I thank Aude Darracq, Grant De Jong, Arnaud Desbiez-Piat, Xu Qiu Guo, Alex Hammel, John Lee, Qianshi Lin, Yichun Qiu, David Tack, Yii Van Tay and other members in Keith Adams lab for having a nice time in the lab. I also want to thank Zhuoqing Fang, Weiyi Li and Yichun Qiu for their valuable discussions. I give my special thanks to my girlfriend Shuting Huo for her support and encouragement, without which I can hardly finish my Ph.D.

My PhD study was funded by a 4-Year Fellowship (4YF) from the University of British Columbia and grants from the Natural Science and Engineering Research Council (NSERC) to Keith Adams.

This dissertation is dedicated to my family for their everlasting support.

## 1. Introduction

The understanding of the genome is a central area of biological research. Genomics research has been a hotspot in the field of biology during the past two decades as more and more genomes have been sequenced. Comparative analysis of genomes across species is an important and powerful approach in genomics studies. By comparison of genomes of various species, we can track evolutionarily conserved sequences across lineages, investigate the genomic features underlying the differences between organisms, and provide insights into the evolutionary forces that drive the evolution of the genome (Eyre-Walker 1999; Hardison 2003).

One of the most surprising facts that genomics research has revealed is that only a small proportion of the entire genome is thought to be protein-coding genes despite the important functions of their products. Non-coding regions, which do not have the potential to encode proteins, were considered as "junk DNA" for a long period (Kung et al. 2013). However, advanced by the rapid development of high-throughput sequencing technology, a lot of studies have identified a large number of transcripts from non-coding regions of the genome (Liu et al. 2012; Necsulea et al. 2014; Hezroni et al. 2015; Iyer et al. 2015; Gallart et al. 2016). These regions are often transcribed at much lower levels than protein-coding genes and show high tissue- or condition- specificity in expression.. Researchers have characterized the functions of numerous long non-coding RNAs in various species, many of which likely have crucial roles in diverse pathways (Ponting et al. 2009; Chekanova 2015; Yamada 2017). These findings suggest that the non-coding

regions are not “silent”, but rather they are expressed and could be biologically important.

Another central topic in evolutionary genomics and comparative genomics is the origin of new genes. Several mechanisms have been shown for new genes to arise in the genome. Among them, gene duplication is believed to be the most common and important one, as evidenced by the large number of duplicated genes in most sequenced eukaryotic genomes (Zhang 2003; Flagel and Wendel 2009; Panchy et al. 2016). However, previous studies mainly focused on the duplication of protein-coding genes. The role of gene duplication in the evolution of non-coding RNAs, and the expression divergence between duplicated non-coding RNAs, is only starting to be explored.

My thesis investigated the evolution of non-coding RNAs (largely long intergenic non-coding RNAs) and duplicate gene/non-coding RNA evolution in plants. In the next section, I will provide a brief overview of how genes evolve after duplication, and the evolution of long non-coding RNAs (lincRNAs), with a focus on plants.

## 1.1 Evolution of genes after duplication

### 1.1.1 Classification of gene duplications

First promoted by Ohno (1970), gene duplication is now viewed as a major mechanism in the evolution of genes and genomes (Zhang 2003; Innan and Kondrashov 2010). The genomes of most sequenced eukaryotes to date, from protists to fungi, from plants to animals, have been shown to be abundant with duplicated genes (Taylor and Raes 2004;

Flagel and Wendel 2009). Based on the way in which duplicated genes originate, gene duplication can be classified into whole-genome duplication, tandem duplication, RNA-based duplication (retrotransposition) and other types of duplication (such as duplicative transposition).

Whole-genome duplication (WGD), also known as polyploidy, often arises from nondisjunction during meiosis (Soltis et al. 2009). It results in the duplication of genes as well as intergenic sequences from the whole genome. While many genes return to single copy after WGD over time, some genes may be kept with clear genomic co-linearity. Ancient WGD events (paleopolyploidy) have occurred in many eukaryotic lineages. Wolfe and Shields (1997) described the first case of ancient WGD, in budding yeast, based on detailed analysis of the sequence similarity and synteny of genes in the yeast genome. Two rounds of WGD (called 2R WGD) occurred early during vertebrate evolution, resulting in the expansion of many gene families (McLysaght et al. 2002). Furthermore, more recent WGD is observed in other species including *Xenopus*, teleost fishes and *Plasmodium* (reviewed in Van de Peer et al. 2009). Whole-genome duplication is most commonly seen in plants (Adams and Wendel 2005; Paterson et al. 2012; Schranz et al. 2012). All angiosperms are known to have undergone multiple rounds of whole-genome duplications (Jiao et al. 2011). Additionally WGD is of particular importance as it affects all genes in the genome and can make crucial contributions to the evolution of the species. For example, the whole-genome duplication event in budding yeast was shown to contribute greatly to the cellular robustness (Li et al. 2010) and

rewiring of the protein-protein interaction network (Presser et al. 2008). Whole-genome duplication is also known to facilitate rapid speciation by differential reciprocal loss of duplicated copies among different populations (Schnable et al. 2012).

Tandem duplicates are duplicated genes that are located next to each other in the chromosome (Hanada et al. 2008). Tandem duplicates most likely form from replication slippage and ectopic recombination. Tandem duplicates tend to have high sequence similarity due to their relatively recent origins (Liu et al. 2011). However, this does not necessarily mean that tandemly duplicated genes have the same functions. One reason is that, distinct from whole-genome duplication where the complete genome is duplicated, tandem duplication may not necessarily result in the duplication of the entire copy of the original gene. Regulatory sequences and even part of coding regions might be missing in tandem duplicates, which could lead to potential functional divergence (Ganko et al. 2007). For example, Fan et al. (2008) described two tandemly clustered duplicates that acquired new functions and underwent strong positive selection in *Drosophila melanogaster*. By systematically analyzing tandem duplicates genome-wide, Liu et al. (2011) showed that 38% of tandem duplicates show reciprocal expression in *Arabidopsis thaliana*, significantly higher than those derived from whole-genome duplication.

RNA-based duplication, or retroduplication, refers to the process where a duplicated gene is generated by retrotransposition of another gene (Marques et al. 2005; Kaessmann et al. 2009). In RNA-based duplication, a retrogene is inserted into the genome from reverse transcription of its parental gene (Zhang et al. 2011; Assis and Bachtrog 2013).

Typically, RNA-based duplicates comprise only a small proportion of all genes in the genome with no more than 200 copies in each species of mammals (Carelli et al. 2016). Genes generated by RNA-based duplication are thought to lack both introns and the original *cis*-regulatory elements (Sorourian et al. 2014). Therefore, it is expected that retrogenes may show more divergence in expression and function compared with their parental genes than other types of duplicates (Kaessmann et al. 2009).

In addition to the three categories of duplication described above, duplicated genes can originate in other ways, such as segmental duplication, small-scale duplication, and duplicative transposition (Casneuf et al. 2006; Guan et al. 2007). Duplicated genes arising from these mechanisms can be classified as neither whole-genome duplicates nor tandem duplicates because they neither show synteny in the flanking regions nor are located close to their paralogs. Hence, such duplicated genes are often regarded as interspersed duplicates (Arsovski et al. 2015).

### 1.1.2 Fates of duplicated genes

Genes can have various evolutionary fates after duplication (Zhang 2003). Many duplicated genes are pseudogenized and lost over time (Bowers et al. 2003; Byrne and Wolfe 2005; Thomas et al. 2006). Duplicated genes that are retained in the genome may show either redundancy or divergence. Previous studies have reported several functionally redundant duplicates with nearly identical nucleotide sequences that are likely mediated by gene conversion (Liao 1999; Gao and Innan 2004; Ezawa et al. 2006;

S.S. Wang et al. 2015). Gene conversion, by non-reciprocal exchange of DNA sequences, can result in the homogenization of the sequences of paralogous genes (Nei and Rooney 2005; Mano and Innan 2008). When transcribed simultaneously, redundant duplicated genes can lead to an increased amount of the same protein, which might be beneficial to the cell in some circumstances (Sugino and Innan 2006). In addition, functional redundancy is also considered as a way to ensure the uniformity of the protein in the whole complex, thereby avoiding potential interference caused by paralogs with divergent sequences (Baker et al. 2013). Moreover, redundant paralogs may benefit the cell because one of the duplicates can act as a back-up of the other in case the other one loses its function (Li et al. 2010).

Subfunctionalization is a mechanism by which duplicated genes can diverge. In this scenario, paralogs in a pair of duplicated genes may take on part of the original functions, or the original expression pattern, of the ancestral copy separately. One model to explain subfunctionalization is duplication-degeneration-complementation (DDC) (Force et al. 1999). In this model, degenerate mutations are accumulated by drift in both paralogs after duplication. After the fixation of degenerate mutations, paralogs may have non-overlapping functions. Therefore, the loss of either copy could be harmful to the cell. Another hypothesis is escape from adaptive conflict (EAC) (Hittinger and Carroll 2007). Based on the model of EAC, both paralogs may specialize (improve part of the ancestral functions) under positive selection if the ancestral copy carried out multiple functions that cannot be improved independently.

In addition to subfunctionalization, many duplicates diverge via neofunctionalization where one copy gains a new function(s), or expression pattern, whereas the other copy retains the original function and/or expression pattern (Tirosh and Barkai 2007; Assis and Bachtrog 2013). This process might be facilitated by positive selection where advantageous mutations are rapidly accumulated, or by relaxation of purifying selection. The concept of neofunctionalization can be well illustrated by a pair of duplicates, *BSK1* and *SSP* in *Arabidopsis thaliana*, derived from a WGD specific to Brassicaceae (Liu and Adams 2010). While *BSK1* kept the original function of brassinosteroid signal transduction, *SSP* evolved under relaxed purifying selection to acquire new functions in paternal control of zygote elongation and the first asymmetric cell division. *BSK1* and *SSP* show opposite expression patterns: *BSK1* is expressed in all tissue types expect for pollen, whereas *SSP* is only expressed in pollen. Liu and Adams (2010) further suggested that the neofunctionalization of *SSP* is likely due to the loss of the kinase domain and relaxed purifying selection.

## 1.2 Evolutionary genomics of lincRNAs

### 1.2.1 Identification of lincRNAs

LincRNAs, long intergenic non-coding RNAs, are a type of non-coding molecule that has attracted more and more interest in recent years. Typically lincRNAs should meet the following three requirements: i) their length must be > 200 nucleotides ii) they are located in intergenic regions iii) they should have no coding potential (Liu et al. 2012;

Ulitsky and Bartel 2013). The advances in high-throughput technology and bioinformatic tools allow for the large-scale identification of lincRNAs. To date, thousands of lincRNAs have been identified in a variety of organisms including human (Necsulea et al. 2014; Iyer et al. 2015; Kornienko et al. 2016), mouse (Guttman et al. 2010), zebrafish (Ulitsky et al. 2011), and Arabidopsis (Liu et al. 2012), among many others.

In the very early stage of the genome-wide identification of lincRNAs, lincRNAs were mainly identified using a combination of cDNAs, tilling arrays and RNA-seq (Ulitsky and Bartel 2013). With the raid development of high-throughput sequencing, now RNA-seq has become the dominant tool in the identification of lincRNAs. The number of lincRNAs varies greatly across species, from 50 in *Plasmodium falciparum* to 170 in *Caenorhabditis elegans*, and from ~6000 in *Arabidopsis thaliana* to more than 20000 in human (Ulitsky and Bartel 2013). However, this variation could be due, at least in part, to the tissue types and developmental stages that were surveyed, the quality and depth of sequencing, the tissue types and developmental stages that were surveyed, and the methods and criteria used for identification (Ulitsky and Bartel 2013).

Furthermore, it is worth noting that the concept of lincRNAs is still evolving. Some lincRNA loci identified in previous studies may not be regarded as lincRNAs based on a new definition of lincRNAs. For example, currently the lincRNA identification typically includes the removal of those with sequence similarity to housekeeping RNAs, transposable elements (TE), and protein-coding genes. However, such careful filtering of the lincRNA data sets was missing in some early studies (Nelson et al. 2016). Also, the

integration of ribo-seq and the use of more rigorous computational approaches have found that some previously identified lincRNAs possibly have the ability to code for short peptides (Ruiz-Orera et al. 2014).

### 1.2.2 Characteristics of lincRNAs

The large-scale identification of lincRNAs has revealed many unique features of lincRNAs. In general, lincRNAs have been shown to have shorter lengths and fewer introns than protein-coding genes. Another important characteristic of lincRNAs is the low sequence conservation across species (Washietl et al. 2014; Hezroni et al. 2015; Mohammadin et al. 2015). The expression of lincRNAs is also very distinct from protein-coding genes. Most lincRNAs are lowly expressed (Ulitsky and Bartel 2013; Kornienko et al. 2016). Moreover, a substantial proportion of lincRNAs are expressed in a highly tissue-, developmental stage- or condition-specific pattern (Ulitsky et al. 2011; Liu et al. 2012). A lot of lincRNAs were found to be actively expressed in response to specific stresses (Shuai et al. 2014; Wang et al. 2017). This implies that many lincRNAs likely function under certain conditions, and partly explains why many lincRNAs can only be detected under particular stress conditions. However, how the expression of stress-responsive lincRNAs is activated remains unclear.

Despite the pervasive expression, functions have been characterized for relatively few lincRNAs. Currently, lincRNAs are generally thought to function by serving as signals, decoys, molecular guides and scaffolds (Wang and Chang 2011; Rinn and Chang

2012). Most lincRNAs with known functions are reported in animals. One of the most well-known examples is *Xist*, a lincRNA that plays essential roles in X chromosome inactivation (XCI). During female development, acting as the signal of XCI by marking the time and space of XCI, *Xist* coats the inactive X chromosome, resulting in the repression of gene expression of the entire X chromosome (Lee 2009). Another example is HOTAIR. HOTAIR binds PRC2 and the LSD1-CoREST complex at the same time, thereby ensuring changes in histone modifications (Tsai et al. 2010). In plants, such functionally well characterized lincRNAs are very few (Kim and Sung 2012; Zhang and Chen 2013; Chekanova 2015; Yamada 2017). What roles lincRNAs play in plants and the differences in lincRNA functions between plants and animals still wait to be studied.

### 1.2.3 Evolutionary origin of lincRNAs

The origin of new genes is a crucial question in evolutionary genomics and molecular evolution. So far, most studies have focused on the origin of protein-coding genes, and the origin of lincRNAs remains poorly understood. It is proposed that lincRNAs can originate by the following three ways: *de novo* birth from intergenic regions, transformation from protein-coding genes or other genes, and duplication of other lincRNAs (Ulitsky and Bartel 2013).

Previous studies in animals found that lincRNAs rarely have sequence similarity to each other within the same species (Ulitsky et al. 2011; Derrien et al. 2012). Thus *de novo* birth is considered as the primary mechanism through which new lincRNAs are

generated in the genome (Ulitsky and Bartel 2013). A few studies have analyzed the origin of lincRNAs from previously existing protein-coding genes, and from duplication of other lincRNAs. Chen et al. (2015) analyzed the birth of new genes from intergenic sequences and proposed the evolutionary scenario of the conversion between lincRNAs and protein-coding genes in primates. Derrien et al. (2012) identified 194 lncRNA families with at least two members based on sequence similarity clustering, suggesting limited roles of gene duplication in the expansion of lincRNAs in human.

Though current studies tend to suggest *de novo* birth as the main mechanism for the origin of lincRNAs the field of lincRNA research is rapidly developing. Many of the above studies were performed with incomplete sets of lincRNAs, and a thorough analysis using the latest lincRNA data sets is still lacking. Additionally, most of them focus on animals and studies of the origin and evolution of lincRNAs in plants remain scarce. It is known that lincRNA repertoires vary greatly among species (Necsulea et al. 2014; Hezroni et al. 2015). So it would be tempting to speculate that the origin of lincRNAs in plants may show very distinct patterns from animals. Thus, the analysis of lincRNA origin in plants is of particular importance to better understand the evolutionary dynamics of lincRNAs.

## 1.3 Dissertation goals

### 1.3.1 To study the divergence of duplicated genes by changes in microRNA binding sites

One of the most important ways by which duplicated genes diverge is expression divergence. I asked the following questions about how duplicated genes diverge in their micoRNA binding sites: What is the proportion of gene pairs that show divergence in microRNA binding sites among different types of gene duplicates in plants? Is the divergence in microRNA binding sites associated with the divergence in expression? What are the contributions of evolutionarily young microRNAs to the divergence of microRNA binding sites between duplicated genes? To answer the above questions, I analyzed microRNA-target interactions in *Arabidopsis thaliana*.. I characterized duplicated gene pairs with divergent microRNA binding sites, analyzed the relationship between expression divergence and microRNA divergence, and identified interesting cases of gene families showing frequent changes in microRNA binding sites.

### 1.3.2 To investigate the evolution and duplication of lincRNAs

I conducted comprehensive analyses of the evolution and duplication of lincRNAs in flowering plants. I asked the following three major research questions. What is the sequence and expression conservation of lincRNAs across flowering plants? How many duplicated lincRNAs are there in flowering plant genomes and how do they diverge? How do lincRNAs evolve after polyploidization? To answer these questions, I performed

sequence alignments, analyzed population genetic data, compiled RNA-seq data-sets, and analyzed the expression level of lincRNAs. I also analyzed transcriptome data sets of polyploidy and diploid *Brassica* species to study lincRNA expression in an allotetraploid.

## 2 Duplicate Gene Divergence by Changes in MicroRNA Binding Sites in *Arabidopsis* and *Brassica*[1]

### 2.1 Introduction

Gene duplication is a major mechanism of new gene creation that has led to the evolution of new gene functions (Zhang 2003; Flagel and Wendel 2009). Duplicated genes can be generated by whole-genome duplication (WGD), tandem duplication (TD), retrotransposition, and other mechanisms. After gene duplication, paralogs may have multiple different fates (Semon and Wolfe 2008; Innan and Kondrashov 2010). Many paralogs show divergence in gene structure, expression pattern, and function. The functions of duplicated genes can diverge by the acquisition of new function, neofunctionalization, or partitioning of ancestral function, subfunctionalization (Hughes 1994; Force et al. 1999). Expression patterns of duplicated genes can diverge by changes in gene regulation, including gain of a new expression pattern relative to the ancestral state or partitioning of an ancestral expression pattern between the duplicates, also referred to as neofunctionalization and subfunctionalization, respectively (Force et al. 1999). Functional and expression divergence are widely regarded as important mechanisms for the retention of duplicated genes.

MicroRNAs (miRNAs), a kind of short noncoding RNA (Cuperus et al. 2011), play

[1] Chapter 2 has been published. **Wang SS**, Adams KL (2015) Duplicate Gene Divergence by Changes in MicroRNA Binding Sites in Arabidopsis and Brassica. *Genome Biology And Evolution* 7: 646-655.

important roles in the regulation of gene expression at the posttranscriptional level by transcript degradation or suppression of translation (Bonnet et al. 2006; Li and Mao 2007; Meng et al. 2011; Takuno and Innan 2011) and may provide a dynamic way to regulate gene expression in many eukaryotes (Berezikov 2011; Rogers and Chen 2013). In plants, gene silencing mediated by miRNAs is an important mechanism in regulating some developmental processes (Chen 2009; Rubio-Somoza and Weigel 2011) and the response to stress (Sunkar et al. 2012), among other functions. Some of the most common miRNA targets in plants include transcription factors and F-box domain-containing proteins (Rhoades et al. 2002; Jones-Rhoades et al. 2006).

Although several of the proteins in miRNA regulation systems are shared by a wide range of plants and animals, the molecular mechanism of the action of miRNAs has been shown to be different between animals and plants in many ways (Chen and Rajewsky 2007; Axtell and Bowman 2008; Voinnet 2009). One distinction is that miRNAs often tend to target protein-coding regions of mRNAs in plants but 3'-untranslated regions (UTRs) in animals (Filipowicz et al. 2008), implying that in plants the miRNA binding sites of protein-coding genes may be under stronger selective pressure and evolve more slowly (Chen and Rajewsky 2007; Guo et al. 2008). Another distinction lies in the mechanism of target recognition. In plants, the recognition of target sites often requires relatively extensive complementarity between miRNAs and target sites (Iwakawa and Tomari 2013; Rogers and Chen 2013). In animals, miRNA-target interactions are more tolerant to mismatches in pairing (Zeng and Cullen 2004; Bartel 2009). The high fidelity

of pairing between miRNAs and targets makes the prediction of target genes and their miRNA binding sites easier and more reliable in plants (Rhoades et al. 2002; Jones-Rhoades and Bartel 2004).

A few studies have examined miRNA-target interactions in duplicated genes. Li et al. (2008) found that miRNAs appear to preferentially regulate duplicated genes over singletons in mammals, based on miRNA binding site prediction results. This finding was further supported by another study where genes localized in CNV (copy number variation) regions were shown to have more miRNA-predicted targets in human (Felekkis et al. 2011). In *Arabidopsis*, Takuno and Innan (2008) showed a negative correlation between the copy numbers of miRNAs and the size of the gene families they regulate. Despite these studies, a genome-wide analysis characterizing the evolution of miRNA regulation in duplicated gene pairs has not been reported. Divergence in miRNA regulation between duplicated genes may be an important mechanism of divergence in expression and function.

We conducted a systematic analysis of the evolution of miRNA binding sites after gene duplication using duplicated genes in Brassicaceae, with a focus on *Arabidopsis* thaliana because of the large number of identified miRNAs and experimentally verified miRNA-target interactions in that species. We analyzed whole-genome duplicates from the alpha-WGD in the *Arabidopsis* lineage, tandem duplicates, and other types of duplicates. We also analyzed genes in *Brassica rapa* generated by the whole-genome triplication (WGT) in its lineage as another and more recent polyploidy event.

## 2.2 Materials and methods

### 2.2.1 Duplicate Gene Data Sets

Genes from *A. thaliana* used in this study were retrieved from TAIR (Lamesch et al. 2012). Sequences annotated as transposable elements were eliminated from the analyses based on TAIR annotation. An all-against-all BLASTP search was performed to identify duplicate and singleton genes in *A. thaliana*. Sequences with E values less than 1e-10 (as used for defining duplicates in (Casneuf et al. 2006; He and Zhang 2006; Su et al. 2006; Yang and Gaut 2011) and sequence coverage above 50% were defined as duplicates, and those having no nonself hits with E values less than 1e-3 were considered to be singletons (Amoutzias et al. 2010). Genes encoded by the mitochondrial genome or chloroplast genome were removed.

Duplicates derived from the alpha-WGD in *A. thaliana* were from the Blanc and Wolfe data set (Blanc et al. 2003) which contains 2,584 pairs of duplicates generated by the most recent WGD event (alpha-WGD) at the base of the Brassicaceae family. Also 1,096 pairs of tandem duplicate pairs were obtained from (Haberer et al. 2004). In addition, we identified 3,178 pairs of other types of duplicates, defined as those with best reciprocal hits and not overlapping WGD duplicates and tandem duplicates. In total, a set of 6,858 pairs of paralogous gene pairs from *A. thaliana* generated by different mechanisms was analyzed. Paralogous genes derived from the *Brassica* lineage-specific genome triplication and their syntenic information were obtained from Cheng et al.

(2012).

### 2.2.2 miRNA Data Sets

miRNA sequences from *A. thaliana* and *B. rapa* were downloaded from miRBase (Griffiths-Jones et al. 2006), a widely used database for miRNA resources which includes a large number of experimentally verified miRNAs in a wide range of species. The mature miRNA sequences were used to predict miRNA binding sites. To define young and ancient miRNAs, we performed a BLASTN search against the genomes of 23 plant species (see Table 2.1 for the full list). Young miRNAs were defined as those with no BLAST hits outside of the *Arabidopsis* genus at the E value cutoff of 1e-10, sequence coverage above 50%, and in addition without homologs outside of the *Arabidopsis* genus based on the annotation of miRBase. Other miRNAs were defined as ancient. Lists of young and ancient miRNAs are in Table 2.2.

### 2.2.3 Analysis of miRNA Target Genes

Computational methods have also been shown to be powerful tools in prediction of miRNA targets in plants (Jones-Rhoades and Bartel 2004; Wang et al. 2004). Many prediction tools have been developed for plant-specific miRNA target gene prediction in the past 5 years (Dai et al. 2011). In this study, we used the following three plant-specific miRNA binding sites prediction methods: psRNAtarget (Dai and Zhao 2011), Tapir (Bonnet et al. 2010), and the miRNA target prediction tool implemented in UEA sRNA

workbench (Stocks et al. 2012) to predict potential miRNA targets. All of the three prediction tools are considered as powerful tools in miRNA-target interaction predictions specific to plants and have been widely utilized (Jeong et al. 2011; Shivaprasad et al. 2012; McHale et al. 2013; Weiberg et al. 2013). The default cutoff value of the number of mismatched base pairs was used for each program: 3 for psRNAtarget, 3.5 for TAPIR, and 3 for sUEA. Each G:U and non-G:U mismatch is counted as 0.5 points and 1 point, respectively (Jones-Rhoades and Bartel 2004; Schwab et al. 2005; Lu et al. 2008). It is thought that the combination of the use of multiple methods would help to decrease the false positive rate of prediction methods and get more accurate results compared with using a single prediction method (Dai et al. 2011; Ding et al. 2012). Thus in this study we define a positive miRNA-target interaction when it is predicted by at least two of the three prediction programs in order to get predicted miRNA targets with higher confidence. The prediction data set is listed in https://figshare.com/articles/microRNA_targets_of_Arabidopsis_thaliana/4955447. When comparing the prediction data set with the experimental data set, we found that 112 of the 156 experimentally verified miRNA-target interactions were included in the prediction data set, which is 72% overlap between the two data sets.

Experimentally verified miRNA targets of *A. thaliana* were manually collected based on the combination of multiple publications and miRNA target databases (Sun et al. 2013; Hsu et al. 2014). The experimental data include miRNA-target interaction results from both degradome sequencing and low-throughput technologies. The final data set contains

156 experimentally verified miRNA-target interactions in 145 protein-coding genes (https://figshare.com/articles/microRNA_targets_of_Arabidopsis_thaliana/4955447).

### 2.2.4 Sequence and Expression Analyses

The alignment of paralogous genes was done using MUSCLE v3.8.31(Edgar 2004). The Yn00 program implemented in PAML v4.7 (Yang 2007) was used to calculate Ka/Ks values of duplicated genes. Normalized expression data from 63 different organs and developmental stages of *A. thaliana* were collected from AtGenExpress (http://arabidopsis.org/ servlets/TairObject?type=expression_set&id=1006710873, last accessed February 13, 2015) and were used to calculate the Pearson correlation coefficient of expression patterns between duplicates. Jacalin domain containing proteins were identified by using hmmscan (Eddy 1998) with a cutoff E value of 1e-10. The best-fit substitution model used in phylogenetic reconstruction was determined as WAG+G+F+I (Whelan and Goldman 2001) using Prottest (Darriba et al. 2011). Phylogenetic trees were constructed with RAxML v7.3.9 (Stamatakis 2006) and 1,000 bootstrap replicates were performed to obtain the support value for each node of the tree. The final tree was visualized using FigTree v1.3.1. The phylogenetic tree and the alignment of jacalin domain containing proteins in *A. thaliana* were deposited at TreeBase (Morell 1996) under the accession S16068. Sequence format processing was done with scripts written in Perl and Ruby (Goto et al. 2010) (available upon request).

## 2.3 Results

### 2.3.1 Duplicates Are More Often Targeted by miRNAs than Singletons

To determine whether duplicated genes or singletons in *A. thaliana* are more likely to be under miRNA regulation, we assembled defined sets of 22,054 duplicates and 3,520 singletons (see Materials and methods). We manually collected experimentally verified miRNA targets in *A. thaliana* from different publications and databases (see Materials and methods). The final data set of known miRNA targets contains 145 protein-coding genes with 156 miRNAtarget interactions. Surprisingly, only one of them was a singleton (Figure 2.1B). We found that 0.6% of duplicates and 0.03% of singletons are miRNA targets. Overall the analyses indicate that duplicated genes are indeed more likely to be targeted by miRNAs than singletons in *A. thaliana* based on the experimental data set (*P*–value < 1e-4, chi-square test).

It is possible that duplicated genes might be overrepresented in the experimentally verified data set for miRNAtarget interaction because they happened to be more highly studied than singletons. Also, all possible miRNA-target interactions in *A. thaliana* have not been experimentally identified. To further test whether miRNA targets are indeed more enriched in duplicates than in singletons, we analyzed all possible miRNA-target interactions genome-wide using prediction methods. Three plant-specific prediction methods: UEA sRNA (Stocks et al. 2012), psRNAtarget (Dai and Zhao 2011), and TAPIR (Bonnet et al. 2010) were used in this study. Given the inaccuracy caused by individual prediction programs, only those genes predicted to be the targets by at least two of three

programs are considered as potential targets. The combination of different computational tools is thought to be able to minimize the negative impact of using only one program to predict miRNA targets (Dai et al. 2011). Based on this criterion, 1,210 miRNA-target interactions including 1,125 target genes and 147 miRNAs were identified and considered as the miRNA binding site prediction data set. Most of the target genes have one predicted miRNA binding site (an average of 1.08 for duplicates and 1.02 for the singletons). We found that among all targets 92% are duplicates whereas 8% are singletons (Figure 2.1A). Consistent with the experimental data, this result shows that duplicates are more likely to be regulated by miRNAs than singletons in *A. thaliana* (*P*-value < 1e-6, chi-square test). To test whether the result might be affected by the stringent criterion used to predict miRNA targets, we did the same analysis using the three prediction methods separately. They gave similar results and reflected the same trends (*P*-value < 1e-7) (Figure 2.2). In addition, we repeated the same analyses using duplicated genes defined with the E-value cutoff as less than 1e-20 and 1e-30. In both analyses, duplicates are overrepresented in both the experimental data set and the binding site prediction data set (Table 2.3) the results from both prediction and experimental data indicate a preferential role of miRNA regulation in duplicated genes in *A. thaliana*.

### 2.3.2 miRNA Target Sites Have Diverged Extensively in Duplicated Genes

To assess the conservation of miRNA binding sites between duplicated genes, we analyzed all pairs of duplicates with at least one gene as an miRNA target to determine

whether they have the same or divergent miRNA binding sites. We used alpha whole-genome duplicates, tandem duplicates, and other types of duplicates in the analyses. Divergent miRNA binding site patterns were detected if only one of the two paralogous genes has an miRNA binding site, or if both of the genes have miRNA binding sites but the binding sites are different. In cases where at least one gene in a paralog pair is an miRNA target, 91% and 68% of the paralog pairs were observed to show divergent patterns of miRNA binding sites in the miRNA binding site prediction data set and experimental data set, respectively (Table 2.4). Among the paralog pairs with divergent patterns of miRNA binding sites, most of the pairs have only one gene as an miRNA target (95% and 93% for the miRNA binding site prediction data set and the experimental data set, respectively). Others show both duplicates with binding sites but these binding sites are by different miRNAs.

We also determined whether there is any difference in the proportion of divergent miRNA binding site patterns among all three classes of duplicated genes. Considering the small sample size of the experimental data set, the analysis was limited to the binding site prediction data set. We found that 91%, 89% and 90% of paralogous gene pairs were shown to have divergent miRNA binding sites for whole-genome duplicates, tandem duplicates and other types of duplicates, respectively (Table 2.4). No significant difference was detected among them ($p > 0.1$, chi-square test). Altogether, the above results indicate a large divergence of miRNA binding site patterns between duplicated genes, but different types of duplicated genes do not show differences in this regard.

### 2.3.3 Divergence in miRNA Binding Sites in Genes Derived from whole genome triplication in *Brassica rapa*

To extend the study to another species and to analyze miRNA binding sites in duplicated genes derived from a more evolutionarily recent WGD event than the alpha-WGD in the Brassicaceae, we used the WGT event that occurred in the ancestor of extant *Brassica* species after the split with the *Arabidopsis* lineage at about 17–20 Mya (Yang et al. 1999; Lysak et al. 2005; Parkin et al. 2005). Duplicated genes derived from the WGT have been identified (Wang et al. 2011). We used *B. rapa* for analysis because it has the largest number of currently identified miRNA genes among *Brassica* species in miRBase. Considering the limited number experimentally verified miRNA targets in *Brassica*, only the three miRNA binding site prediction methods were used. Similar to the analyses in *A. thaliana*, protein-coding genes predicted to be miRNA targets by at least two of three prediction programs were included in the prediction data set for *B. rapa*. After genome triplication, some triplicated genes retained three copies whereas others retained only one or two copies. In total, there are 70 pairs and triplets of genes derived from the WGT with at least one member predicted to be an miRNA target. Among them, 52 paralog pairs/triplets show divergence of miRNA binding sites (Table 2.5; https://figshare.com/articles/List_of_pairs_triplets_of_genes_with_the_same_and_divergent_microRNA_binding_sites_patterns_in_Brassica_rapa/4955450). Among the retained triplicates, there were more cases of two genes having an miRNA binding site than all

three or just one. Thus, consistent with *A. thaliana*, the majority of duplicated genes analyzed in *B. rapa* have extensively diverged in their miRNA binding sites patterns. Moreover, the proportion of paralogous gene pairs with divergent miRNA binding sites patterns derived from the Brassica-specific WGT is significantly lower than that of *A. thaliana* for the prediction data set ($p < 0.05$, chi-square test). This could be due to the lower divergence time of paralogous genes formed by the Brassica-specific genome triplication than the alpha-WGD specific to Brassicaceae.

### 2.3.4 Duplicated Genes with Divergent miRNA Regulation Patterns Show More Divergence in Expression Patterns in *A. thaliana*

To determine whether there is a relationship between miRNA binding site divergence and expression divergence in duplicated genes, we analyzed the expression correlation between paralogous genes in *Arabidopsis* using both the binding site prediction data set and the experimental data set. (We used *Arabidopsis* and not *Brassica* for the expression analysis because much more expression data are available for *Arabidopsis*.) We used microarray data from 63 different organs and developmental stages of *A. thaliana* (see Materials and methods). Paralog pairs with divergent miRNA binding sites show more divergence in expression patterns than those with the same miRNA target sites, indicated by their significantly lower Pearson correlation coefficient for both the target site prediction data set and experimental data set (Figure 2.3). Although the expression correlation coefficients vary between the two data sets, similar patterns are apparent.

Thus, the divergence of miRNA binding site patterns s associated with the divergence in gene expression in *A. thaliana*.

It is possible that the group of paralog pairs with the same miRNA binding sites could show more similar expression patterns if they were formed more recently. To determine whether paralog pairs with the same binding sites are on average younger than those with divergent binding sites, we calculated Ks values for the two sets of paralog pairs. Paralog pairs with the same binding sites were detected to be younger, as a whole, than those with divergent miRNA binding sites patterns as inferred by Ks values of 1.65 for pairs with divergent binding sites and 1.16 for pairs with the same binding sites (*P*-value < 0.01). This suggests that younger duplicates, in general, have less divergent miRNA binding sites that could contribute to less divergence in expression patterns.

### 2.3.5 Evolutionarily Recent miRNAs Make Major Contributions to the Divergence of miRNA Binding Patterns between Duplicates

To investigate to what extent evolutionarily recent miRNA genes contribute to the divergence of miRNA regulation of paralogous genes, we analyzed duplicated gene pairs in *A. thaliana* for targets of miRNAs that are restricted to the *Arabidopsis* genus (young miRNAs) versus those that are present in other species outside of the *Arabidopsis* genus (ancient miRNAs). We used *Arabidopsis* because of the large number of miRNAs identified in *A. thaliana* and *Arabidopsis lyrata*; in contrast, fewer miRNAs have been identified in *Brassica* species. We classified miRNAs in *Arabidopsis* as young miRNA

genes or ancient miRNA genes according to whether they have homologs outside of the *Arabidopsis* genus at E value of 1e-10 and also based on the annotation of miRBase (see Materials and methods). Young miRNAs in *A. thaliana* were defined as those with homologs only present in *A. thaliana* and/or A. lyrata. Those with homologs found outside the *Arabidopsis* genus were defined as ancient miRNAs. We analyzed the alpha whole-genome duplicates because it is known that they formed at the base of the Brassicaceae family, using miRNA targets from the binding site prediction data set.

Out of 201 duplicated gene pairs that have divergent miRNA binding sites, 104 pairs (51%) are targets of young miRNAs. In contrast, 28% (6 of 21) of paralog pairs with the same miRNA binding sites are targets of the evolutionarily young miRNAs. To see whether the results could be due to the criteria used in the identification of young miRNAs, another list of young miRNAs was generated with a BLASTN E value of 1e-3. No new young miRNAs were discovered and thus the results were the same. As alpha whole-genome duplicates formed at the base of the Brassicaceae family, the regulation by these young miRNAs is clearly indicative of gain of binding by miRNAs after gene duplication. This analysis demonstrates that the birth of new miRNA genes can give rise to the diversification of miRNA regulation and create differences in regulation between duplicated genes.

### 2.3.6 Phylogenetic Analysis of Jacalin Domain Containing Proteins in *Arabidopsis* Reveals Dynamic Evolution of miRNA Targets

Based on our miRNA target predictions, we found that a family of proteins called jacalins is enriched in miRNA binding sites. Jacalins are a large family containing 56 members in *A. thaliana*. Jacalins are thought to be involved in the response to biotic or abiotic stimuli but their detailed functions are poorly understood (Yamaji et al. 2012). AT5G28520, a protein-containing jacalin domain, was found to be regulated by miR842 and miR846 (Jia and Rock 2013). In our prediction results, 18 of 49 jacalin protein sequences are predicted to be targets of at least one miRNA, with four sequences having two different miRNA binding sites. Two miRNAs, miR842 and miR846, were predicted to be miRNAs that target jacalins. Both miR842 and miR846 are only found in *A. thaliana* and *A. lyrata* indicating their recent origin after the divergence of the *Arabidopsis* genus and other species in Brassicaceae.

To explore how miRNA binding sites have changed after gene duplications within the jacalin family, we reconstructed the phylogenetic history of jacalins in *Arabidopsis* and then mapped the miRNA binding sites predicted to be present in each gene on the phylogenetic tree. It appears that multiple gains and losses of miRNA binding sites events have happened during the evolution of jacalin domain containing proteins in *Arabidopsis*, although the exact number is difficult to assess. In one branch of the tree (the lower left side of Figure 2.4), many closely related genes potentially generated by recent duplication events show very different patterns of miRNA regulation. Some very closely

related genes are targeted by different miRNAs, whereas distantly related paralogs can be regulated by the same miRNA. For example, AT5G49850, AT5G49860, and AT5G49870 were generated through TD and form one clade in the phylogenetic tree. AT5G49850 and AT5G49870 are predicted to be targeted by miR846, whereas AT5G49860 is not shown to have any miRNA binding sites possibly due to the absence of the first jacalin domain present in AT5G49850 and AT5G49870. The phylogenetic analysis of the jacalin family provides a nice example of the dynamic evolution, including multiple gains and losses, of miRNA binding sites after duplications within a gene family.

## 2.4 Discussion

### 2.4.1 Duplicates Are More Likely to be Targeted by miRNAs than Singletons

Our analyses revealed a higher fraction of duplicates as potential targets for miRNA regulation in *Arabidopsis*, indicated by both experimentally verified and predicted miRNA targets. These observations suggest an important role of miRNAs in regulating the expression of duplicated genes in *Arabidopsis*. Our study provides the first reported evidence for the preferential regulation of duplicated genes over singletons by miRNAs in plants. Our findings are consistent with a computational study in mammals (Li et al. 2008). Thus, the miRNA regulation of duplicated genes in plants and animals shows similar trends in this regard.

It has been shown that the reduction of expression levels can facilitate the retention of duplicated genes by buffering the toxic effect caused by imbalanced gene dosage (Qian

et al. 2010). Hence, the enrichment of miRNA regulation in duplicated genes in *A. thaliana* suggests their contributions to maintaining gene expression balance by silencing and downregulating paralogous genes. The downregulation of expression of duplicated genes may play an important role in retention of some of them. It is possible that some genes with miRNA binding sites may avoid the negative effect caused by imbalanced dosage and be more likely to be retained after duplication. In addition, the preferential regulation of duplicates by miRNAs might be attributed to the ability of miRNA regulation to lead to tissue-specific expression divergence between paralogs. Neofunctionalization and subfunctionalization of expression patterns of duplicated genes, facilitated by miRNA regulation, could lead to retention of some duplicated genes.

### 2.4.2 Divergence of miRNA Binding Site Patterns after Gene Duplication

After duplication genes can show divergence in expression patterns and functions. In this study, we show that a large majority of duplicated genes in *Arabidopsis* show divergent patterns of miRNA binding sites. For the data set of duplicates with experimental evidence for miRNA targeting, 68% of duplicate pairs with at least one miRNA target show clear divergence of miRNA binding sites. For the data set based on prediction results, the number increased to 87%. These results demonstrate that a large majority of duplicates show different miRNA regulation patterns no matter which data set was utilized in the analyses. We did not find a significant difference among the different types of duplicates (WGDs, tandems, other duplicates) in regards to their miRNA binding site

divergence levels. Thus, the mechanism of gene duplication probably does not have an effect on the evolution of miRNA binding sites.

To extend the study to another species and examine a more recent case of polyploidy, we studied genes duplicated by the WGT in *Brassica*. Similar to duplicates in *A. thaliana*, triplicated genes in *B. rapa* have diverged extensively with respect of their miRNA binding sites. As there can be up to three paralogs derived from the Brassica-specific WGT event retained in the genome of *B. rapa*, one could hypothesize that the genes might have more divergent miRNA regulation. However, our analysis shows that the extent to which miRNA binding sites have diverged in *B. rapa* is less than in whole-genome duplicate pairs in *A. thaliana*. We think that this is possibly because the Brassica-specific genome triplication occurred more recently than the alpha-WGD specific to the Brassicaceae family. The shorter divergence time for triplicated genes in *B. rapa* may lead to less divergence in their miRNA regulation compared with *A. thaliana*. However, it should be noted that miRNA genes identified in *B. rapa* are likely incomplete. A more comprehensive analysis of miRNA binding site divergence after genome triplication might be performed when a more complete set of miRNA genes is available in *B. rapa* as well as other species within the *Brassica* genus.

Divergence in miRNA binding sites between duplicated genes may have an impact on their expression patterns and functions. Our observation that paralogs with divergent miRNA binding sites tend to show a greater divergence in expression profiles supports that possibility. In some cases, the divergent patterns of miRNA regulation may lead to

the differential expression between paralogs. For example, in *Arabidopsis* allopolyploids, nonadditive expression of duplicated miRNAs led to expression level differences between their duplicated target genes in some cases (Ha et al. 2009).

### 2.4.3 Evolutionarily Recent Gain of miRNA Regulation

We identified miRNAs that are specific to the *Arabidopsis* genus after the divergence of its lineage from the *Brassica* lineage within the Brassicaceae family that we refer to as young miRNAs. We present evidence that 51% of divergent miRNA regulation patterns between paralogs derived from WGD, analyzed in *A. thaliana*, can be attributed to young miRNAs that were born after the paralogs originated by duplication. Thus, it could be inferred that the divergence in miRNA binding sites between paralogs can occur by gain of miRNA regulation by the binding of a newly born miRNA. Thus, sequence changes in the coding region or UTR would not necessarily be needed for miRNA regulation to be gained. Because miRNA binding sites are often localized in coding regions in plants instead of in 30 -UTRs as in animals (Millar and Waterhouse 2005; Chen and Rajewsky 2007), it is thought that it is more difficult for genes in plants to gain regulation by an miRNA by the accumulation of point mutations (Chen and Rajewsky 2007). However, if divergent miRNA binding site patterns are caused by miRNAs born after the gene duplication occurred, point mutations would not be needed. There are several ways in which new miRNAs can arise in plants (Nozawa et al. 2012). miRNAs could be generated through the duplication of preexisting miRNAs (Maher et al. 2006), transition

of miniature inverted-repeat transposable elements (Piriyapongsa and Jordan 2008), inverted duplication of protein-coding genes (Allen et al. 2004), and spontaneous mutations in intergenic regions (De Felippes et al. 2008). The inverted duplication of protein-coding genes is of particular interest in terms of duplicated genes gaining miRNA regulation. This is because a newly born miRNA through this mechanism will have the same sequence as the protein-coding gene from which it originates (Allen et al. 2004). Therefore, the protein-coding gene from which the miRNA originates may become an miRNA target without changes in the coding sequences. Additionally, it is plausible that a new miRNA happens to have nearly perfect complementary to a sequence of a protein-coding gene through random mutations allowing for miRNA targeting. Thus, there are several ways in which new miRNAs can be created. Our results emphasize the important role of young miRNAs in regulation of duplicated genes.

**Table 2.1** List of organisms whose genomes were used to identify evolutionarily young microRNAs.

| **Organisms** |
|---|
| *Aquilegia coerulea* |
| *Arabidopsis lyrata* |
| *Brachypodium distachyon* |
| *Brassica rapa* |
| *Citrus clementina* |
| *Chlamydomonas reinhardtii* |
| *Capsella rubella* |
| *Cochliobolus sativus* |
| *Eucalyptus grandis* |
| *Eriocaulon parvulum* |
| *Glycine max* |
| *Gossypium raimondii* |
| *Malus domestica* |
| *Mimulus guttatus* |
| *Oryza sativa* |
| *Physcomitrella patens* |
| *populus trichocarpa* |
| *Prunella vulgaris* |
| *Ricinus communis* |
| *Solanum lycopersicum* |
| *Selaginella moellendorffii* |
| *Vitis vinifera* |
| *Zea mays* |

**Table 2.2** Lists of young and ancient microRNAs in *Arabidopsis thaliana* and the plant species whose nuclear genomes were searched to identify young and ancient microRNAs. Young microRNAs are defined as those with homologs only found in the *Arabidopsis* genus (*Arabidopsis thaliana* and *Arabidopsis lyrata*). Ancient microRNAs are those with homologs found outside of the *Arabidopsis* genus (see Materials and methods).

| **Young microRNAs** | **Ancient microRNAs** |
|---|---|
| miR163 | miR156 |
| miR1887 | miR390 |
| miR1888 | miR157 |
| miR2933 | miR158 |
| miR2934 | miR159 |
| miR2936 | miR160 |
| miR2937 | miR161 |
| miR2938 | miR162 |
| miR2939 | miR164 |
| miR3434 | miR165 |
| miR3932 | miR166 |
| miR3933 | miR167 |
| miR401 | miR168 |
| miR402 | miR169 |
| miR404 | miR170 |
| miR405 | miR171 |
| miR406 | miR172 |
| miR407 | miR173 |
| miR420 | miR1886 |
| miR4221 | miR2111 |
| miR4227 | miR2112 |
| miR4228 | miR319 |
| miR4239 | miR3440 |
| miR4243 | miR391 |
| miR4245 | miR393 |
| miR447 | miR394 |
| miR5012 | miR395 |
| miR5013 | miR396 |
| miR5014 | miR397 |

| Young microRNAs | Ancient microRNAs |
|---|---|
| miR5015 | miR398 |
| miR5016 | miR399 |
| miR5017 | miR400 |
| miR5018 | miR403 |
| miR5019 | miR408 |
| miR5020 | miR413 |
| miR5021 | miR414 |
| miR5022 | miR415 |
| miR5023 | miR416 |
| miR5024 | miR417 |
| miR5025 | miR418 |
| miR5026 | miR419 |
| miR5027 | miR4240 |
| miR5028 | miR426 |
| miR5029 | miR472 |
| miR5595 | miR5640 |
| miR5628 | miR5651 |
| miR5629 | miR5654 |
| miR5630 | miR5658 |
| miR5631 | miR824 |
| miR5632 | miR825 |
| miR5633 | miR827 |
| miR5634 | miR828 |
| miR5635 | miR840 |
| miR5636 | miR845 |
| miR5637 | miR854 |
| miR5638 | miR857 |
| miR5639 | miR858 |
| miR5641 | miR860 |
| miR5642 | miR862 |
| miR5643 | |
| miR5644 | |
| miR5645 | |
| miR5646 | |
| miR5647 | |
| miR5648 | |
| miR5649 | |
| miR5650 | |
| miR5652 | |

| Young microRNAs | Ancient microRNAs |
|---|---|
| miR5653 | |
| miR5655 | |
| miR5656 | |
| miR5657 | |
| miR5659 | |
| miR5660 | |
| miR5661 | |
| miR5662 | |
| miR5663 | |
| miR5664 | |
| miR5665 | |
| miR5666 | |
| miR5995 | |
| miR5996 | |
| miR5997 | |
| miR5998 | |
| miR5999 | |
| miR771 | |
| miR773 | |
| miR774 | |
| miR775 | |
| miR776 | |
| miR777 | |
| miR778 | |
| miR779 | |
| miR780 | |
| miR781 | |
| miR782 | |
| miR822 | |
| miR823 | |
| miR826 | |
| miR829 | |
| miR830 | |
| miR831 | |
| miR832 | |
| miR833 | |
| miR834 | |
| miR835 | |
| miR836 | |

| Young microRNAs | Ancient microRNAs |
|---|---|
| miR837 | |
| miR838 | |
| miR839 | |
| miR841 | |
| miR842 | |
| miR843 | |
| miR844 | |
| miR846 | |
| miR847 | |
| miR848 | |
| miR849 | |
| miR850 | |
| miR851 | |
| miR852 | |
| miR853 | |
| miR855 | |
| miR856 | |
| miR859 | |
| miR861 | |
| miR863 | |
| miR864 | |
| miR865 | |
| miR866 | |
| miR867 | |
| miR868 | |
| miR869 | |
| miR870 | |

**Table 2.3** The results of the proportion of duplicated genes vs. singletons as microRNA targets using E-value cutoffs of 1e-20 and 1e-30 along with at least 50% sequence coverage.

| | **e-20** | | | **e-30** | | |
|---|---|---|---|---|---|---|
| | duplicates | singletons | p-values | duplicates | singletons | p-values |
| **total** | 20775 | 3520 | | 19684 | 3520 | |
| **exp** | 132 | 1 | 1.10E-05 | 127 | 1 | 9.60E-06 |
| **2_outof_3** | 687 | 54 | 2.10E-08 | 663 | 54 | 9.50E-09 |
| **psRNAtarget** | 1017 | 67 | 2.65E-15 | 982 | 67 | 7.00E-16 |
| **tapir** | 963 | 93 | 1.05E-07 | 927 | 93 | 4.60E-08 |
| **UEA** | 992 | 83 | 1.51E-10 | 957 | 83 | 5.10E-11 |

**Table 2.4** Conservation and divergence of microRNA binding site patterns in duplicated genes in *Arabidopsis thaliana*.

| | WGD | TD | Others | total |
|---|---|---|---|---|
| **microRNA binding site prediction dataset** | | | | |
| Same | 21 | 8 | 22 | 51 |
| Divergent | 211 | 65 | 231 | 507 |
| Total | 232 | 73 | 253 | 558 |
| **experimental dataset** | | | | |
| Same | 12 | 1 | 7 | 20 |
| Divergent | 14 | 9 | 20 | 43 |
| Total | 26 | 10 | 27 | 63 |

The numbers of paralog pairs showing the same or divergent microRNA binding site patterns based on the microRNA binding site prediction dataset and the experimental dataset are indicated. Each category (same, divergent and total) of microRNA binding site pattern is divided into three classes corresponding to the three types of duplicated genes, from left to right, whole genome duplicates (WGD), tandem duplicates (TD) and other types of duplicates (others).

**Table 2.5** Conservation and divergence of microRNA binding site patterns in whole genome duplicates and triplicates in *Brassica rapa*.

| | Duplicates | | Triplicates | | | Total |
|---|---|---|---|---|---|---|
| No. of miRNA targets | 1 | 2 | 1 | 2 | 3 | |
| Same | - | 17 | - | - | 1 | 18 |
| Divergent | 34 | 1 | 0 | 14 | 3 | 52 |
| Total | 34 | 18 | 0 | 14 | 4 | 70 |

The numbers of paralog pairs and triplicates showing the same or divergent microRNA binding site patterns based on the microRNA binding site prediction dataset for *Brassica rapa* are indicated. Genes generated via WGT are divided into duplicates and triplicates based on how many genes are retained. 'No. of targets' indicates how many genes are microRNA targets (1or 2 for duplicates and 1, 2 or 3 for triplicates).

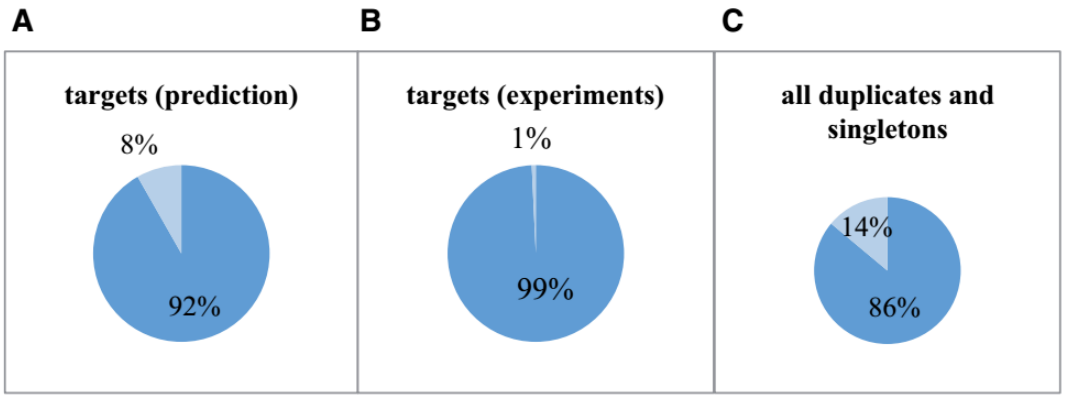


**Figure 2.1** Duplicated genes are more likely to be targeted by miRNAs than singletons. The proportions of duplicates and singletons among all miRNA targets based on binding site prediction data set (A) and experimental data set (B) are indicated. The proportions of all duplicates and singletons in the genome are shown in (C). Lighter and darker portions of the pie charts represent singletons and duplicates, respectively.

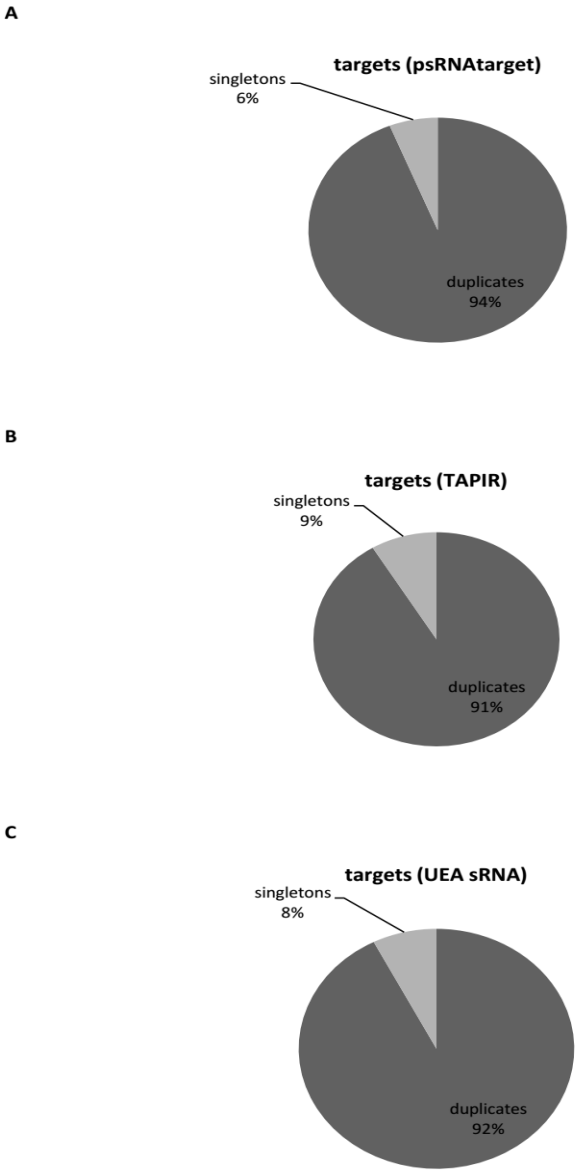


**Figure 2.2** Proportion of duplicates and singletons in the targets identified by individual prediction programs. A. Results based on psRNAtarget. B. Results based on TAPIR. C. Results based on UEA sRNA. The parts of the pie chart representing duplicates and singletons are colored in deep grey and light grey, respectively.

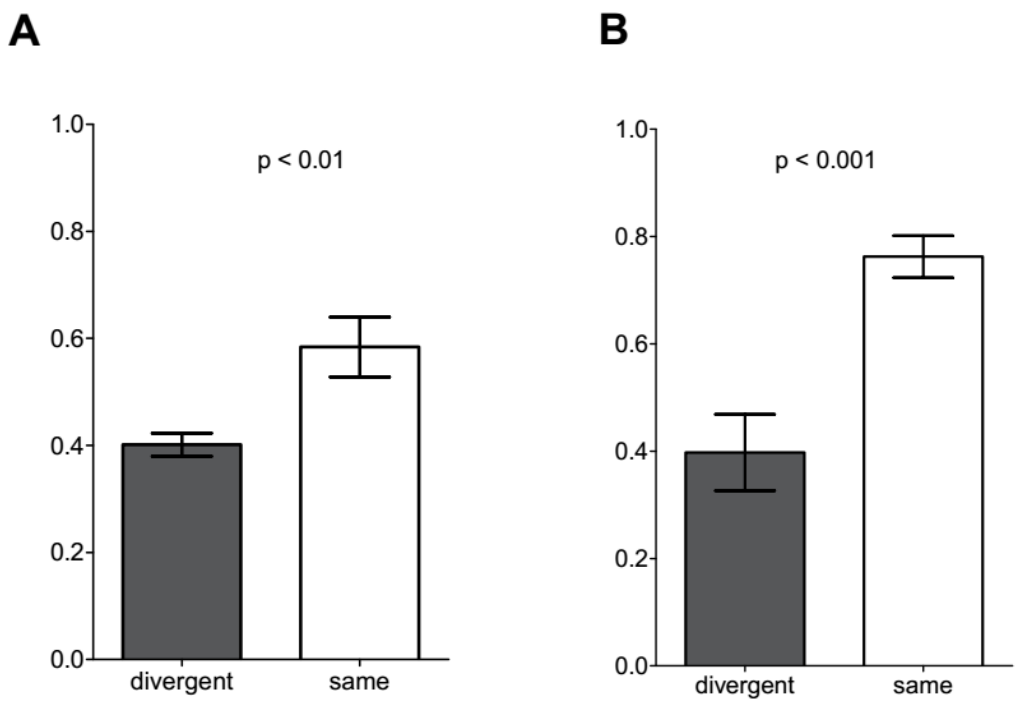


**Figure 2.3** Expression correlation analysis between paralog pairs with the same and divergent miRNA regulation patterns. All paralog pairs with at least one gene targeted by an miRNA are classified into two categories based on whether they show divergent miRNA regulation patterns for both miRNA binding site prediction data set (A) and experimental data set (B). The Pearson correlation coefficient between two paralogous genes is calculated based on the microarray data with 63 different organ types and developmental stages (see Materials and methods).

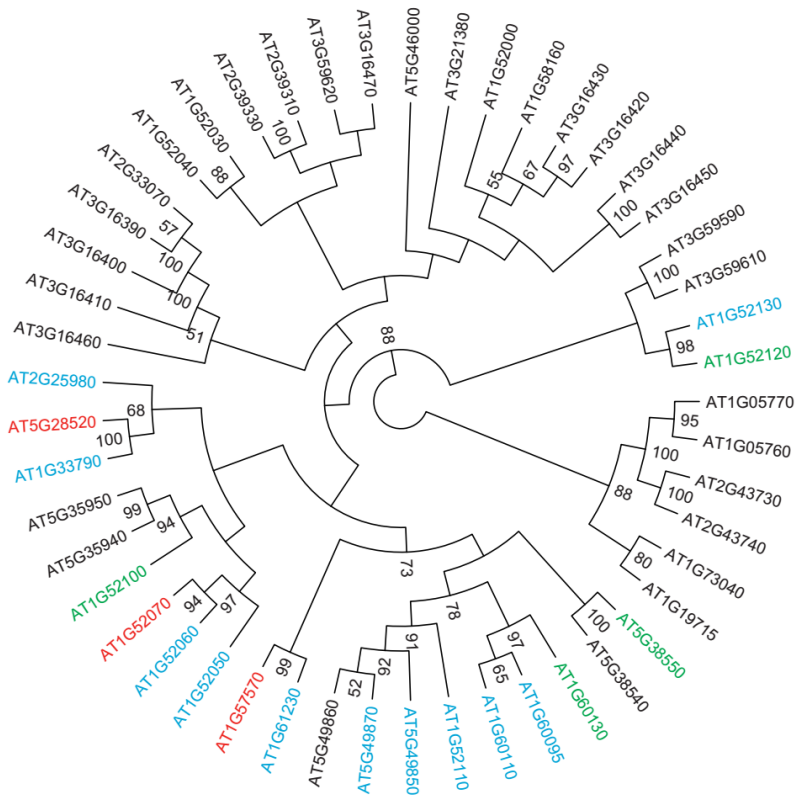


**Figure 2.4** Phylogenetic analysis reveals dynamic evolution of miRNA regulation in jacalin family genes in *Arabidopsis thaliana*. Maximum-likelihood analysis was performed using RAxML. WAG+G+F+I was chosen as the most suitable substitution model based on the result of ProtTest before the phylogenetic reconstruction. Gene symbols with the color of green, blue, and red indicate targeting by miR842, miR846, and both miRNAs, respectively. Numbers next to the nodes correspond to bootstrap values obtained from 1,000 bootstrap replicates. Only the nodes with bootstrap values greater than or equal to 50 are shown in the tree.

## 3. Rapid Evolution of Sequence and Expression in Flowering Plant LincRNAs[1]

### 3.1 Introduction

With more and more transcriptome data available, it has been shown that areas of many intergenic regions in the genome of plants and other eukaryotes can be transcribed into non-coding RNAs (ncRNAs) (Axtell and Bowman 2008; Necsulea et al. 2014; Wang et al. 2014; Gallart et al. 2016). Long intergenic non-coding RNAs (lincRNAs) are defined as non-coding RNAs with sequence length of more than 200 nucleotides located in intergenic regions. As a group, lincRNAs are often lowly expressed in a tissue-specific pattern (Liu et al. 2012; Ulitsky and Bartel 2013). LincRNAs participate in a range of biological processes and act by various molecular mechanisms (Wang and Chang 2011). LincRNAs have been implicated in such functions as molecular signals of activation or silencing of transcription with temporal and spatial specificity, epigenetics, and the mediation of the interaction between nucleotides and proteins by acting as modular scaffolds or guides to recruit chromatin-modifying machinery to target sites (reviewed in (Wang and Chang 2011)).

Although the functions of most lincRNAs have yet to be characterized in plants, some lincRNAs have been shown to play important roles in a wide variety of processes (Ariel et al. 2015; Chekanova 2015; Yamada 2017). For example, *IPS1* inhibits the

[1] A version of Chapter 3 has been submitted for publication. **Wang SS**, Hammel A, Adams, KL Rapid evolution of sequence and expression of lincRNAs in flowering plants.

activity of ath-miR399 by mimicking its target during phosphorus starvation, and regulates Pi homeostasis in *Arabidopsis thaliana* (Franco-Zorrilla et al. 2007). Many other microRNA-mimicking lincRNAs have been discovered in Arabidopsis and rice, hinting their widespread roles in the regulation of microRNA activity (Wu et al. 2013). *ENOD40*, a lincRNA gene well conserved in legumes, is able to form a highly structured RNA and mediate the relocalization of MtRBP1 by direct physical interaction, thereby regulating root symbiotic nodule organogenesis (Campalans et al. 2004). Another example is *LDMAR*, a lincRNA with sequence length over 1,000 nucleotides. It is believed to participate in the regulation of photoperiod-sensitive male sterility by mediating changes in DNA methylation and transcription specifically under long-day conditions in rice (Ding et al. 2012).

Since the first genome-wide identification of 6,480 lincRNA transcripts in *Arabidopsis thaliana* (Liu et al. 2012), a large number of lincRNAs have been identified in many plants, unveiling much more complex plant lincRNA repertoires than were previously appreciated. The evolution of lincRNAs in plants is only beginning to receive attention. Mohammadin et al. 2015 studied lincRNAs in Brassicaceae including basal *Aethionema*, and the Cleomaceae, but only found a small number of lincRNAs with detectable sequence similarity shared between species. Instead, they found many positionally conserved lincRNAs without sequence similarity indicating the importance of the conservation of not only the sequence but also the genomic context in plant lincRNA evolution. Nelson et al. (2016) analyzed the sequence conservation and

evolution of Arabidopsis lincRNAs in Brassicaceae, showing a relatively low level of lincRNA conservation across the family. They also examined the origin of lincRNAs in Brassicaceae by duplication. Despite these two recent studies, the sequence conservation, evolution, and expression patterns of lincRNAs in plants, especially in lineages outside Brassicaceae, are still poorly understood.

In this study, we performed a large-scale analysis to study the dynamics of sequence and expression evolution of lincRNAs across 55 plants. We used lincRNA sequence data sets from five species: *Arabidopsis thaliana*, *Oryza sativa japonica, Zea mays*, *Medicago truncatula* and *Solanum lycopersicum*. These species were chosen to give a broad taxonomic range across the eudicots and monocots among those plants where lincRNAs have been identified on a large scale. We also integrated more than 70 RNA-seq data sets from many different tissue types to explore the evolution of lincRNA expression among both closely and distantly related species.

## 3.2 Materials and methods

### 3.2.1 Sources of sequences

We utilized seven plant lincRNA data sets: *Arabidopsis thaliana* (Liu et al. 2012; Jin et al. 2013), *Oryza sativa japonica* (Zhang et al. 2014), *Zea mays* (Li et al. 2014), *Medicago truncatula* (T.Z. Wang et al. 2015) and *Solanum lycopersicum* (Zhu et al. 2015). For maize lincRNA data set, only sequences annotated as high-confidence lincRNAs from (Li et al. 2014) were used. Sequences in the lincRNA data sets with similarity to

housekeeping RNAs were removed using infernal v1.1 (Nawrocki and Eddy 2013) with the *P*-value cutoff of 1e-5. We further discarded lincRNAs with protein-coding potential using BLASTX (Altschul et al. 1997) with the e-value cutoff of 1e-5 and using RNAcode (Washietl et al. 2011) with the *P*-value cutoff of 0.05. We also eliminated potential transposon- or small RNA-derived lincRNAs using RepeatMasker (http://www.repeatmasker.org, last accessed September, 2014). Sequences and genomic coordinates were retrieved for each lincRNA data set. For lincRNA locus with more than one size of transcript (a few lincRNAs had evidence of alternative spicing), the longest one was used in subsequent analyses. In total, 5176, 1020, 1091, 3449 and 1855 lincRNA sequences were used in subsequent analyses for Arabidopsis, rice, maize, Medicago and tomato respectively.

MicroRNA sequences were downloaded from miRBase (Kozomara and Griffiths-Jones 2014). Genome sequences, genome annotation and protein sequences for each plant were downloaded from public databases (Table 3.1; Figure 3.1). It is worth noting that the genome of *Carica papaya* might be of relatively low quality due to low sequencing coverage (Lyons and Tang 2014). However, given its important phylogenetic position as an outgroup species to Brassicales, we included it in this study.

### 3.2.2 Identification of homologous loci and conserved synteny of lincRNAs and estimation of evolutionary gain and loss across plant families

Homology-based genome-wide searching has been shown to be powerful to identify

homologous loci of non-coding RNAs in other species (Nozawa et al. 2012; Taylor et al. 2014). BLASTN (Altschul et al. 1997) searches were performed first to identify homologous lincRNA loci across 55 plant genomes. Given the relatively low sequence conservation of lincRNAs (Ulitsky et al. 2011; Necsulea et al. 2014), the argument of 'wordsize' in BLASTN was adjusted from its default 11 to 7 to help improve the sensitivity of searches (Ulitsky et al. 2011). Homologous loci of a lincRNA in a certain plant were defined as those meeting one of the following criteria. i) Hits have BLAST e-values lower than or equal to 1e-10 ii) Hits have BLAST e-values between 1e-5 and 1e-10 and sequence similarity of at least 80% and aligned sequence of at least 50 nucleotides (Gaiti et al. 2015; Hezroni et al. 2015; Ye et al. 2015). Otherwise, the lincRNA was considered not to have homologous loci in that plant. Additionally, we used a more stringent e-value cutoff of 1e-10 and 1e-20 to identify homologous loci of lincRNAs to see the consistency of results.

We identified syntenic lincRNAs by looking for their neighboring syntenic protein-coding genes as (Zhang et al. 2009; Abrouk et al. 2012; Nelson et al. 2016). The synteny information of protein-coding genes from Plant Genome Duplication Database (PGDD) (Lee et al. 2013) was retrieved. For species not included in PGDD, syntenic blocks were generated following the same procedure as done for PGDD using MCScanX (Wang et al. 2012) as used in (Xu et al. 2014; Huang et al. 2016). Syntenic blocks with fewer than twenty genes were discarded from subsequent analyses. We also identified syntenic lincRNAs without removal of syntenic blocks with less than twenty genes and

obtained similar results. For a lincRNA and its conserved locus in a related species, ten upstream and ten downstream orthologous protein–coding genes were retrieved for both of them, respectively. Conserved lincRNA loci with at least ten out of the twenty surrounding genes located in a syntenic block were defined as syntenic conserved lincRNA loci. If fewer than ten surrounding genes from the upstream or downstream region were extracted for any of the two genes, a conserved lincRNA locus was detected if at least half of the surrounding genes were found in a syntenic block. Syntenic loci of tomato lincRNAs were not identified in other species due to the lack of sufficient genomes with annotations of high quality in Solanaceae.

Evolutionary gains and losses of genes were estimated using the parsimony method (Nozawa et al. 2012; Washietl et al. 2014; Nitsche et al. 2015). Briefly, a gain of a gene is placed adjacent to the node leading to the last common ancestor of all taxa with the gene. A gene is considered to be lost along the branch if it is not present in that branch and all leaves (tips) derived from that branch. The divergence time of species was retrieved from TimeTree (Hedges et al. 2006).

### 3.2.3 Identification of homologous protein-coding genes and microRNA genes

Protein sequences were obtained from the genomic resources listed in Table 3.1. Homologous protein-coding genes were identified using BLASTP with the e-value cutoff of 1e-10 (Casneuf et al. 2006; He and Zhang 2006). Syntenic protein-coding genes were collected from PGDD (Lee et al. 2013) or identified following the same procedure as

used by PGDD using MCScanX (Wang et al. 2012) as used in (Xu et al. 2014; Huang et al. 2016). Homologous microRNA loci were determined using the same criteria as lincRNAs (Will et al. 2007; Wang and Adams 2015). Only representative gene models (without alternative splicing) were used in analyses.

### 3.2.4 Identification of gene families of lincRNAs, miRNAs and proteins

LincRNAs were clustered using OrthoMCL (Li et al. 2003) as used in (Mohammadin et al. 2015) and additionally SILIX (Miele et al. 2011). Protein families were identified using OrthoMCL (Li et al. 2003). MicroRNA families were determined based on miRBase (Kozomara and Griffiths-Jones 2014).

### 3.2.5 Identification of conserved regions across species in lincRNAs

Conserved regions of lincRNAs met the following criteria: i) the lincRNAs in which the conserved region is found must have alignments in at least three plants from different genera to the target species ii) the region has overlap in all of these species. Coordinates extraction and format conversion were assisted by BEDTools v2.25.0 (Quinlan and Hall 2010), BEDOPS v2.4.2 (Neph et al. 2012) and custom scripts written in Ruby (Goto et al. 2010).

### 3.2.6 Population genetics analysis

Single nucleotide variant (SNV) data including 80 ecotypes of *Arabidopsis thaliana* (Cao

et al. 2011) were collected from the Arabidopsis 1001 Genomes Project (http://1001genomes.org, last accessed July, 2013). Nucleotide diversity, π, was calculated based on (Nei and Li 1979). We inferred the derived alleles (new mutations that have arisen among populations) following the procedure of (Keinan et al. 2007). *A. lyrata* and *C. rubella* were used as outgroup species to infer ancestral and derived alleles of *A. thaliana*. Their whole genome alignments with *A. thaliana* were performed by LASTZ (Harris 2007) (non-default parameters: --notransition --step=20). The ancestral alleles were determined as alleles which are shared by both *A. lyrata* and *C. rubella* identified using MafFilter v1.0.0 (Dutheil et al. 2014) and also coincide with one of the alleles in *A. thaliana* at the corresponding site on the chromosome. Alleles distinct from the ancestral allele were defined as derived alleles. Alleles present only in *A. lyrata* or only in *C. rubella* were discarded.

### 3.2.7 Expression analyses of plant lincRNAs

Transcriptome data for from various species were retrieved from NCBI's SRA (Sequence Read Archive) database (Table 3.2). RNA-seq reads were trimmed by Cutadapt v1.3 (Martin 2011) to filter out those with sequence length less than 20 nucleotides and sequencing quality less than 20 before they were mapped to the reference genome. Processed reads were mapped to the reference genome using STAR v2.4.2 (Dobin et al. 2013). Read counts were calculated using HTSeq v0.6.1 (Anders et al. 2015) and FPKM for each gene locus were calculated with Cufflinks v2.1.1 (Trapnell et al. 2012). FPKM

values from multiple biological replicates were averaged for each gene. LincRNAs and proteins with FPKM higher than 0.01 in all biological replicates were considered as expressed genes.

An expression tissue specificity index and expression breadth index were used to measure expression specificity of lincRNAs and protein-coding genes. Expression tissue specificity τ (Yanai et al. 2005) is defined by

$$\tau = \frac{\sum_{j=1}^{n}(1-[log_2 S(i,j)/log_2 S(i,max)])}{n-1},$$

where n denotes the number of tissue types and S(i, max) denotes the highest expression of gene i across the n tissues. Expression tissue specificity index ranges from 0 to 1, with a higher value indicating higher tissue specificity. Expression breadth is defined as the number of tissue types in which a gene is expressed. Co-expressed gene pairs were defined as those with co-expression Pearson correlation coefficient higher than 99% of randomly selected gene pairs (Zhan et al. 2006). Gene co-expression network was visualized with Cytoscape v3.4 (Shannon et al. 2003).

## 3.3 Results

### 3.3.1 LincRNAs exhibit considerably low conservation

To study the evolutionary conservation patterns of lincRNAs, we used lincRNA data sets from five plant species (*Arabidopsis thaliana*, *Oryza sativa japonica*, *Zea mays*, *Medicago truncatula*, and *Solanum lycopersicum*) from published studies. LincRNAs that are likely derived from protein-coding genes, housekeeping RNAs, small RNAs and

transposons were removed (see Materials and methods). We conducted a large-scale BLAST search using lincRNA data sets from the five species against 55 plant genomes including 50 flowering plants and 5 non-flowering plants (Table 3.2; Figure 3.1). Homologous lincRNA loci were defined as regions with e-value no higher than 1e-10 or lower than 1e-5 but with 80% identity with at least 50 bp in the BLASTN alignment. Syntenic lincRNAs were defined as homologous loci with at least ten syntenic genes among 20 flanking genes between related species (see Materials and methods).

We chose to do a detailed study of lincRNA evolution in the Brassicaceae family because of the availability of mostly complete genome sequences from several diverse genera across the family. LincRNA loci from *Arabidopsis thaliana* were well conserved in *Arabidopsis lyrata*, *Camelina sativa* and *Capsella rubella*, with more than 60% homologous loci and over 45% syntenic loci present (Figure 3.2A). *Leavenworthia alabamica* displayed a decrease in lincRNA conservation with only 42% homologous loci and 18% syntenic loci (Figure 3.2A). The conservation level of *Arabidopsis thaliana* lincRNA loci displayed another drastic decrease in *Aethionema abamicum* with 20% homologous loci and 11% syntenic loci detected (Figure 3.2A).

We also examined lincRNA evolution within the Poaceae family (grasses). LincRNAs from *Oryza sativa ssp. japonica* show a moderate level of conservation (Figure 3.2B). More than 98% of all lincRNA loci in the Japonica cultivar had at least one homologous locus detected in both *Oryza sativa ssp. indica* and *Oryza glaberrima* (two domesticated rice species), suggesting very high lincRNA conservation level among

domesticated rice. *Oryza brachyantha*, a wild rice that is the basal species in the *Oryza* genus and has diverged from *Oryza sativa* for about 15 million years, displays a major decrease in lincRNA conservation with 55% of homologous loci detected. 41% and 22% maize lincRNAs have homologous loci in *Sorghum bicolor* and *Setaria italica*, respectively. No more than 2% of homologous loci were detected in species outside of the Poaceae (Table 3.3).

Though many homologous lincRNA loci are located in syntenic regions, some are found in non-syntenic regions with no gene collinearity detected (Figure 3.2). Homologous lincRNA loci found in non-syntenic regions might result from translocation or duplication followed by deletion of the syntenic loci. We found that lincRNAs with syntenic loci in other species are more conserved than non-syntenic lincRNAs as indicated by higher aligned length (Figure 3.3). This suggests that syntenic lincRNAs are more conserved than those located at non-syntenic regions.

### 3.3.2 Comparison of lincRNA conservation with microRNAs and protein-coding genes

We compared the conservation of lincRNAs with protein-coding genes and microRNAs in Brassicaceae, Poaceae, Fabaceae and Solanaceae. Protein-coding genes showed significantly higher proportions of homologous loci (Figure 3.4A-D) than lincRNAs (Figure 3.2) in all species examined. In Brassicaceae, even the most basal lineage *Aethionema abamicum* contains homologous loci of 88% of the protein-coding genes in

*Arabidopsis thaliana* whereas no more than 80% of lincRNA loci could be found outside the genus *Arabidopsis*. Similarly, maize also shows a striking contrast between proteins and lincRNAs where protein-coding genes showed four times more homologous loci in most grasses species.

The proportion of homologous loci of microRNAs is roughly similar within a family and higher outside of a family in comparison with lincRNAs in all five families analyzed. For example, within Brassicaceae the percentages of homologous loci for microRNAs and lincRNAs were about the same (around 53%) (Table 3.3; Figure 3.4E-H). However in *Tarenaya hassleriana*, the closest sister lineage to Brassicaceae, 26% of Arabidopsis microRNAs were shown to have homologous loci whereas only 9% of Arabidopsis lincRNAs showed homologous sequences in this lineage (Figure 3.2A; Figure 3.5A; Figure 3.4E). The low conservation of lincRNAs was more obvious as divergence time increases: no more than 2% of Arabidopsis lincRNA homologous loci could be detected in more distantly related species (Table 3.3). In contrast, about 10% of the homologous loci of microRNAs from *Arabidopsis thaliana* could be found in species from both rosids and asterids (Table 3.3). We repeated all the analyses using more stringent criteria to identify homologous loci for both lincRNAs and microRNAs across species and the patterns remained unchanged (Table 3.4-3.5).

### 3.3.3 Recent origins and extensive lineage-specific loss or decay of plant lincRNAs

The low sequence conservation level of lincRNAs suggests that many lincRNAs were

likely gained recently or lost in a lineage-specific manner during evolution. To evaluate these possibilities, we mapped the evolutionary history of the homologous loci of lincRNAs from five plants with lincRNA data sets available (*Arabidopsis*, *Oryza*, *Zea*, *Medicago*, and *Solanum*) onto a species phylogeny in each of the families and traced their potential gain and loss, and then compared the patterns with protein-coding genes and microRNAs.

The phylogenetic patterns of gain and loss clearly indicate that the vast majority of plant lincRNAs potentially originated within the family in both the Poaceae and Brassicaceae (Figure 3.5). The homologous loci of nearly half of maize lincRNAs were not found in any other species analyzed and only 12% date back to the origin of Poaceae (Figure 3.5C). 34% of rice lincRNAs likely originated before the divergence of all Poaceae species analyzed in this study (Figure 3.5B). In the Brassicaceae where lincRNAs generally show the highest conservation level within a family, 22% of Arabidopsis lincRNAs date back to the origin of the Brassicaceae. The last common ancestor of the analyzed Solanaceae contained more than 53% of the tomato lincRNA loci examined. Similar to lincRNAs, a large number of microRNAs were also found to originate within a family, although more microRNAs are shared by more distantly related species in general (Table 3.3; Table 3.4-3.5). Relatively few protein-coding genes originated within a family. Even in Fabaceae where the evolution of protein-coding gene pools displays the largest fluidity, the last common ancestor of the family was estimated to have 76% of the protein-coding genes found in Medicago (Figure 3.6), which is much

higher than for lincRNAs (Figure 3.5) and microRNAs (Figure 3.7).

Homologous loci of lincRNAs also show frequent loss or decay specific to particular lineages (Figure 3.5). For instance, *Leavenworthia alabamica* from Brassicaceae has undergone a rapid lineage-specific loss of lincRNA loci (Figure 3.5A). This is consistent with previous observations of the reduction of non-coding regions in this lineage that has undergone a recent whole-genome triplication event (Haudry et al. 2013). The most extensive loss or decay of homologous lincRNA loci in Fabaceae occurred in the branch leading to the common ancestor of *Phaseolus vulgaris* and *Vigna radiata* (Figure 3.5D). These results further confirm the recent origins and extensive loss or decay of lincRNA loci across plants.

A more detailed view of how lincRNA loci are potentially lost in a lineage-specific manner can be obtained from Figure 3.8. We counted the taxa in which a certain lincRNA gene is lost and the frequency of this event for *Arabidopsis thaliana*, rice, maize, Medicago and tomato. In all cases examined, lincRNAs appear to show a higher frequency of gene loss compared to protein-coding genes and a roughly similar frequency compared to microRNAs. It should be also noted that homologs of some genes, particularly lincRNAs, evolve rapidly and can be difficult to detect. Also, some lincRNAs might have more ancient origins but have diverged too much in sequence to be detected in early-splitting lineages. These factors could affect our estimates of lincRNA locus gains and losses, and potentially lead to over-estimates of lincRNA loss. Thus losses should be regarded as loss or decay to the point that the homologous locus of a lincRNA

is undetectable with our methods.

In addition, we compared the conservation of lincRNA homologous loci between plants and animals using human lincRNA sequences. We analyzed the sequence evolution of 1383 human lincRNAs shorter than 1000 bp based on the annotation of GENCODE v19 (Harrow et al. 2012) using the same criteria as plant lincRNAs in five mammals. More homologous loci of human lincRNAs appeared to have originated in more ancient time (Figure 3.9). Human lincRNAs were also much less likely to undergo lineage-specific loss/decay than plants between species with similar divergence time (Figure 3.9). These results indicate much higher conservation of lincRNA loci at sequence level in human.

### 3.3.4 LincRNAs with ancient origins

Although the majority of lincRNAs originated within a family, some lincRNA loci were conserved in species outside of a family. In contrast to the conservation level within a family, the conservation level of lincRNA sequences outside a family is among the lowest in Brassicaceae. In *Tarenaya hassleriana*, the closest sister group of Brassicaceae with a genome available, 9% (451) of Arabidopsis lincRNAs have homologous loci detected (Figure 3.2). No more than 2% of Arabidopsis lincRNAs were shown to have homologous loci in any other rosids (Table 3.3). Similarly, lincRNAs of rice showed drastic decrease of conservation in species outside Poaceae (Table 3.3). The presence of homologous lincRNA loci in distantly related species points to likely ancient origins of

many lincRNA loci.

Considerably fewer lincRNAs were conserved in syntenic regions outside of a family. In *Tarenaya hassleriana*, 200 out of 451 Arabidopsis lincRNA homologous loci were found to be located in syntenic blocks and only one and four syntenic loci could be detected in *Carica papaya* and *Theobroma cacao*, respectively. For grasses, only one lincRNA of rice has homologous loci found in *Musa acuminata* in synteny and no syntenic loci of maize lincRNAs were detected in species outside of Poaceae.

### 3.3.5 LincRNA loci share short highly conserved motifs across species that are selectively constrained

MicroRNAs show more sequence identity within their functional sites and exhibit higher evolutionary flexibility at surrounding sites (Fahlgren et al. 2010). We investigated whether lincRNAs also share highly conserved regions (see Materials and methods for definition) across species. These short highly conserved regions have to be shared by at least three species outside the same genus of the species whose lincRNA data set was analyzed. Conserved regions were detected in 25% of lincRNAs on average for each species (from 9% in maize to 46% in Arabidopsis). The lengths of conserved motifs varied from a few nucleotides to more than 200 nucleotides (Figure 3.10A). The average lengths of highly conserved motifs among these five species range from 57 nt in tomato and 85 nt in rice63 nt in maize and 121 nt in tomato. These results indicate that the lincRNAs have small highly conserved regions with the majority displaying extensive

flexibility in sequence.

While most of these highly conserved regions were shared by species within a family, some could be detected in species having diverged for more than 100 million years. In particular, a few such regions are more than 100 bp long and are shared by many species outside the family (Figure 3.10B,C). These remarkable deeply conserved regions indicate their ancient origins and suggest potential functional significance across plants.

To further explore if highly conserved motifs are under stronger selective constraints than surrounding regions, we compiled a SNV (single nucleotide variation) data set from (Cao et al. 2011) including 80 different ecotypes of *Arabidopsis thaliana*. We found that highly conserved motifs exhibit significantly reduced nucleotide diversity than surrounding regions and intron sequences, indicating stronger constraints at the population level (Figure 3.10D). We also focused on the distribution of derived allele frequency (DAF) and minor allele frequency (MAF) among 80 Arabidopsis ecotypes. We identified ancestral and derived alleles in *Arabidopsis thaliana* using genomes of *Arabidopsis lyrata* and *Capsella rubella* as an outgroup (see Materials and methods). We observed a significant excess of rare (<5%) derived alleles and minor alleles for conserved motifs (*P*-value < 1e-5, Fisher's exact test; *P*-value < 1e-5, Kolmogorov-Smirnov test) (Figure 3.10E,F). Moreover, the shifts in both DAF and MAF towards rarer alleles in highly conserved regions of lincRNAs were pronounced when compared to introns, indicative of purifying selection on these short conserved sequences in lincRNAs (Figure 3.10E,F). These results suggest that plant lincRNAs share many

short conserved motifs across species which are more strongly selectively constrained than surrounding regions and potentially under negative selection, which might implicate functional domains of lincRNAs.

### 3.3.6 Evolutionarily ancient lincRNAs are overrepresented in expressed lincRNAs across tissues and show lower tissue specificity than lineage-specific lincRNAs

To gain insights into the evolution of expression of plant lincRNAs, we analyzed expression data with multiple biological replicates from different tissue types for lincRNAs from *Arabidopsis thaliana*, *Oryza sativa japonica* and *Zea mays.* First we investigated the relationship between expression and evolutionary age of lincRNAs. We compiled multiple RNA-seq data sets in various organ types (see Materials and methods) and investigated the relationship between the evolutionary age and conservation level and expression pattern of lincRNAs. We classified lincRNAs into ancient and lineage-specific lincRNAs based upon whether they are specific to the genus (*Arabidopsis, Oryza or Zea*) or found outside of the genus (Wang and Adams 2015). Expressed lincRNAs were defined as those with FPKM higher than 0.01 in all biological replicates. We observed that ancient lincRNAs were significantly overrepresented among the expressed lincRNAs in all tissue types analyzed in the three species (*P*-value < 1e-5, Chi-squared test) except for anther in maize (Figure 3.11A-C). Also, ancient lincRNAs were more highly expressed than lineage-specific lincRNAs in general (Figure 3.12A). In addition, ancient lincRNAs exhibit lower tissue specificity (*P*-value < 0.05, Wilcoxon signed-rank test)

(Figure 3.11D) and higher expression breadth (*P*-value < 0.01, Wilcoxon signed-rank test) (Figure 3.12B) in all organ types examined in *Arabidopsis*, rice and maize, suggesting ancient lincRNAs are more broadly expressed. To see if the results could be affected by the definition of lineage-specific and ancient lincRNAs, we redefined lineage-specific and ancient lincRNAs as those specific and non-specific to the family (Brassicaceae or Poaceae) and repeated all analyses above. We focused on Arabidopsis and rice as there are few maize lincRNAs non-specific to the family. We found the same patterns with strong statistical significance (Figure 3.13).

We also analyzed the differences between lineage-specific lincRNAs and ancient lincRNAs from the perspective of gene expression network. We constructed co-expression network consisting of lincRNAs and protein-coding genes for Arabidopsis, rice and maize. Ancient lincRNAs were more overrepresented in the network (*P*-value < 0.05 in all cases, Chi-squared test) (Figure 3.14A), suggesting that lineage-specific lincRNAs are be less integrated in the gene co-expression network than ancient ones. No significant difference between lineage-specific lincRNAs and ancient lincRNAs was observed in terms of their co-expression correlation coefficient with flanking genes (*P*-value > 0.1, t test) (Figure 3.14B). We also showed several examples of lincRNAs co-expressed with genes involved in epigenetic and transcription regulation, pointing to potential roles of these lincRNAs in the pathways of genes they are co-expressed with (Figure 3.14C).

### 3.3.7 Rapid divergence of lincRNA expression across species

To explore the evolutionary dynamics of lincRNA expression across species, we estimated the proportions of expressed lincRNA and protein-coding gene loci that are shared between species. Collectively, we obtained RNA-seq data sets for *Arabidopsis thaliana*, *Oryza sativa japonica*, *Zea mays* and *Medicago truncatula* and their related species covering multiple tissue types. For all expressed lincRNAs and protein-coding genes with syntenic loci in related species, we calculated the proportion of their syntenic regions that are also expressed, as performed in (Kutter et al. 2012). On average, 50% of syntenic lincRNAs expressed in *A. thaliana* were also expressed in *A. lyrata* among the tissue types analyzed, whereas 95% of syntenic protein-coding genes from *A. thaliana* were also expressed in *A. lyrata* (*P*-value < 1e-10, Chi-squared test) (Figure 3.15A). 70% and 65% of syntenic lincRNAs that are expressed in *Oryza sativa japonica* are expressed in *Oryza sativa indica* and *Oryza glaberrima* respectively, significantly lower than protein-coding genes (*P*-value < 1e-10, Chi-squared test) (Figure 3.15B). Similar patterns were observed for lincRNAs from maize and Medicago (Figure 3.15C, D). In addition, lincRNAs show more divergence in the proportion of shared expressed loci among different tissue types than protein-coding genes (Figure 3.15), possibly resulting from the tissue-specific expression pattern of lincRNAs. Also, to account for different sizes of RNA-seq libraries, for each comparison we normalized sizes of different libraries by resampling identical number of RNA-seq reads per tissue type and species (Kutter et al. 2012; Necsulea et al. 2014) and similar results were obtained (Figure 3.16A-D).

In addition, we calculated the Spearman correlation coefficients of expression level between lincRNAs and protein-coding genes and their syntenic genes in closely related species, as used in (Hezroni et al. 2015). We found that lincRNAs show lower correlation in expression than protein-coding genes (Figure 3.16E). Taken together, these results indicate that lincRNAs show more divergent expression patterns across species than proteins.

## 3.4 Discussion

### 3.4.1 Plant lincRNAs generally exhibit low conservation levels with potential considerable variation in different lineages

Through systematic and thorough analyses in 55 plant species, we successfully identified a large number of lincRNA sequences that are conserved across closely related species. This study revealed that lincRNA loci from several plant families likely originated within a family, and they appear to evolve rapidly in sequence. LincRNAs also showed extensive loss or decay during evolution. The extensive fluidity of lincRNA loci in plants leads to patchy distributions of lincRNAs where some of them are conserved in more distantly related species but are lost in closely related lineages. All of these evolutionary patterns are in stark contrast with genes coding for proteins that are generally deeply conserved.

Nelson et al. 2016 compared the sequence evolution of syntenic loci between lincRNAs and protein-coding genes in Brassicaceae, revealing the rapid sequence

evolution of lincRNAs in Brassicaceae. Our study investigated the evolution of both syntenic and non-syntenic loci in different plant families, expanding previous findings in plant lincRNA evolution into a broader context. Our analyses reveal that plant lincRNAs show potential large variation in their conservation levels within a family. The Brassicaceae tend to have the highest level of lincRNA sequence conservation among those families studied. LincRNA loci from maize were the least conserved across species with extensive loss and only 12% loci date back to the origin of the grass family. The low conservation level of maize lincRNAs and their synteny might result from the large genomic rearrangements in maize (Schnable et al. 2009). Also, transposable elements are known to be associated with the generation and expansion of lincRNAs in animals (Kelley and Rinn 2012; Kapusta et al. 2013). The highly transposon-enriched genome of maize (Schnable et al. 2009) might also contribute to the lineage-specific birth and rearrangements of lincRNAs, thereby leading to the low conservation of maize lincRNAs among other grasses.

Syntenic lincRNAs show stronger conservation than non-syntenic lincRNAs, consistent with the observation for protein-coding genes (Glover et al. 2015). Non-syntenic lincRNAs could be generated in the following two ways. One is that they are translocated from the original locus. Alternatively, the orthologous locus might be deleted and it is the paralog of the lincRNA that was detected. It should be noted that the number of syntenic lincRNAs might be underestimated in papaya whose genome assembly is not as high quality as some genomes (Lyons and Tang 2014).

Our study also compared sequence conservation between lincRNAs and microRNAs. LincRNAs appeared to be less conserved than microRNAs in sequence in plants (Table 3.3). It should be noted that non-coding RNAs could evolve very fast in sequence. The criteria used to identify homologous loci for protein-coding genes may not be suitable for non-coding RNAs. Thus we used relatively permissive criteria (see Materials and methods) in the identification of homologous loci of lincRNAs and microRNAs to increase the sensitivity of homology search (Barquist et al. 2016). Some of the homologous loci in lincRNAs and microRNAs might not represent "true" homologs. We additionally used more stringent criteria to repeat the analyses. The basic conclusion remains unchanged (Table 3.4-3.5).

LincRNAs of plants are less conserved in sequence than human lincRNAs between species with similar divergence time (Figure 3.9). We think that the difference is likely due to the following reasons. First, plant genomes have undergone many rounds of polyploidy events and extensive genomic rearrangement (Adams and Wendel 2005; Soltis et al. 2009). These factors might facilitate the fast evolution of lincRNAs in plants after large structural change at genomic level. Second, the removal of functionless DNA is more efficiently in plants than in animals, which could result in the rapid sequence evolution of lincRNAs in plants (Freeling and Subramaniam 2009; Burgess and Freeling 2014).

### 3.4.2 Conserved lincRNAs with ancient origins

Our analyses revealed the highly dynamic evolutionary history and small number of lincRNA loci conserved between distantly related species. Indeed, homologous loci of lincRNAs conserved in species outside of a family were much fewer than those detected within a family. Due to frequent changes in genome structure over long periods and the rapid sequence evolution, homologous non-coding sequences might be expected to not be detectable in syntenic regions. However, we still detected some conserved lincRNAs with ancient origins and a few of them were located in syntenic blocks though the number of homologous loci decreases drastically in distantly related species. Plant lincRNAs with ancient origins, in particular those expressed and located in syntenic regions, might have broad regulatory functions which are rather deeply conserved during evolutionary history (Necsulea et al. 2014).

### 3.4.3 Highly conserved regions across plants implicate sequence-specific functional motifs for lincRNAs

By searching for conserved sites in a wide variety of species, we found a large number of conserved short motifs surrounded by relatively unconstrained sequences in plant lincRNAs, consistent with recent findings in lincRNAs of vertebrates (Hezroni et al. 2015). The highly conserved regions in plant lincRNAs average about 70 nucleotides and their coverage relative to the full length of lincRNA sequence varies from 11% to 26%. While most of these highly conserved motifs are short, some are more than 200 nucleotides, implying potential diverse roles of these highly conserved motifs and

lincRNAs in plants. A few short conserved regions of lincRNAs are shared by several distantly related plants, pointing to their potential ancient origins. Conserved regions in distantly related species are generally shorter and have more substitutions, indicating gradual changes of sequences of lincRNA homologous loci.

We analyzed single nucleotide variant data set across 80 ecotypes of *Arabidopsis thaliana* to test selection constraints acting on these short motifs. The lower nucleotide diversity, larger number of derived alleles and higher minor allele frequency in the highly conserved regions shared across species all support the hypothesis that highly conserved motifs are under stronger selection constraints than surrounding regions and introns, indicating that they are potentially under purifying selection. This finding also suggests that the selection on lincRNAs might be mainly limited to only a short region, although the results need to be regarded cautiously. It is possible that these highly conserved regions are necessary for lincRNA functions while the function of the rest of the sequence depends more on secondary structure thereby allowing more divergence in sequence (Pang et al. 2006). This pattern might be analogous to that of microRNAs where target binding sites and the nucleotides to which they are paired often evolve slowly while the surrounding sequences play more structural roles and are under more relaxed selection (Axtell and Bowman 2008; Ehrenreich and Purugganan 2008; Fahlgren et al. 2010). Several lincRNAs have been shown to carry short potential functional units with high conservation (Chureau et al. 2002; Marques and Ponting 2009; Ulitsky et al. 2011). It was reported that only one tenth of the full sequence is needed for functions of some

lincRNAs (Quinn et al. 2014). Our identification of short motifs and the evidence for their selective constraints implicate that short highly conserved motifs are embedded in a highly flexible architecture of the full locus and implicate functional regions in plant lincRNAs (Xu et al. 2017).

### 3.4.4 Large divergence of lincRNA expression across species

Our study reveals that non-lineage-specific lincRNAs, those that are older and have more ancient origins, are overrepresented in expressed lincRNAs and have lower expression tissue specificity in Arabidopsis, rice and maize. This pattern is basically consistent with protein-coding genes (Zhang and Yang 2015). Lineage-specific lincRNAs likely have recent origins. They might have gained expression recently and are expressed in a limited number of organs in a way similar to proteins.

LincRNAs show rapid evolution in expression patterns between mammals (Kutter et al. 2012; Washietl et al. 2014). We observed that the expression of lincRNAs also shows very rapid change between closely related plant species. For example, 70% of syntenic lincRNAs expressed in *O. sativa japonica* are expressed in *O. sativa indica*, and 50% of lincRNAs expressed in *A. thaliana* are expressed in *A. lyrata*. That is considerably lower than in protein-coding genes, which suggests that lincRNAs show rapid change in expression across different plants. It should be noted that we only analyzed the expression conservation of syntenic loci for both lincRNAs and protein-coding genes as done in (Kutter et al. 2012). Examination of more plants with more tissue types and

deeper sequencing might help to gain more insights into the turnover of lincRNA expression across lineages. This result also demonstrates the large divergence of expression level of lincRNAs in plants. Some plant lincRNAs have been shown to be involved in the regulation of gene expression, either in *cis* (on neighboring genes) or *trans* (on distal genes) (Marques and Ponting 2014). The rapid divergence of lincRNA expression might contribute to lineage-specific changes in the regulation of gene expression and possibly underlie some phenotypic variation among lineages (Wang et al. 2016; Wang et al. 2017).

**Table 3.1.** Genomic resources used in the study.

| Species | Genome sequence and genome annotation |
|---|---|
| *Arabidopsis thaliana* | TAIR10 |
| *Arabidopsis lyrata* | JGI v1.0 |
| *Brassica oleracea* | JCVI v1.0 |
| *Capsella rubella* | JGI v1.0 |
| *Camelina sativa* | v2.0 |
| *Leavenworthia alabamica* | v1.0 |
| *Sisymbrium irio* | v1.0 |
| *Aethionema arabicum* | v1.0 |
| *Eutrema salsugineum* | v1.0 |
| *Tarenaya hassleriana* | v1.0 |
| *Carica papaya* | Hawaii Agriculture Research Center v1.0 |
| *Populus trichocarpa* | *P. trichocarpa* v2.2 |
| *Gossypium raimondii* | JGI v2.1 on assembly v2.0 |
| *Lotus japonicas* | Kazusa 2.5 |
| *Medicago truncatula* | JCVI v4.0 |
| *Glycine max* | JGI v1.1 |
| *Vitis vinifera* | genescope v1 |
| *Beta vulgaris* | RefBeet v1.1 |
| *Ricinus communis* | JCVI v1.0 |
| *Manihot esculenta* | Cassava v4 |
| *Theobroma cacao* | v1.1 |
| *Citrus sinensis* | JGI v1.0 |
| *Eucalyptus grandis* | JGI assembly v1.1 |
| *Cucumis melo* | Melonomics v3.5 |
| *Cucumis sativus* | Chinese long v2 |
| *Prunus persica* | JGI v1.0 |
| *Malus domestica* | IASMA |
| *Fragaria vesca* | Strawberry Genome 1.0 |
| *Amborella trichopoda* | Amborella v1.0 |
| *Citrullus lanatus* | Cucurbit Genomics Database v1 |
| *Cicer arietinum* | v1.0 |
| *Cajanus cajan* | v5.0 |
| *Phaseolus vulgaris* | v1.0 |
| *Vigna radiata* | v6 |
| *Solanum lycopersicum* | SL2.50 |
| *Slanum tuberosum* | PGSC v3.4 |
| *Solanum pennellii* | v2.0 |

| Species | Genome sequence and genome annotation |
|---|---|
| *Capsicum annuum CM334* | v1.55 |
| *Nicotiana tabacum TN90 Burley* | Sol genomics network |
| *Solanum melongena L.* | v2.5.1 |
| *Mimulus guttatus* | v2.0 |
| *Oryza sativa ssp. japonica* | RGAP v7 |
| *Oryza sativa ssp. indica* | 9311_BGF_2005 |
| *Oryza brachyantha* | v1.4 |
| *Oryza glaberrima* | AGI v.1 |
| *Sorghum bicolor* | JGI v1.4 |
| *Setaria italica* | JGI v2.1 |
| *Brachypodium distachyon* | v2.1 |
| *Zea mays* | v2 |
| *Musa acuminata* | Genescope/Cirad |
| *Picea abies* | v1.0 |
| *Selaginella moellendorffii* | JGI v1.0 |
| *Physcomitrella patens* | JGI v1.6 |
| *Chlamydomonas reinhardtii* | JGI v5.3.1 |
| *Ostreococcus lucimarinus* | JGI v2.0 |

**Table 3.2** RNA-seq data sets used in this study.

| **Organism** | **Tissue type** | **RNA-seq data set** |
|---|---|---|
| *Arabidopsis thaliana* | rosette | SRR2039793-SRR2039794 |
| *Arabidopsis thaliana* | inflorescence | SRR1657473-SRR1657475 |
| *Arabidopsis thaliana* | floral bud | SRR800754-SRR800755 |
| *Arabidopsis thaliana* | anther | SRR1559345-SRR1559346 |
| *Arabidopsis thaliana* | leaf | SRR1283943-SRR1283945 |
| *Arabidopsis thaliana* | seedling | SRR1119205-SRR1119206 |
| *Arabidopsis lyrata* | rosette | SRR2039795-SRR2039796 |
| *Arabidopsis lyrata* | inflorescence | SRR1657476-SRR1657478 |
| *Arabidopsis lyrata* | floral bud | SRR800644-SRR800645 |
| *Oryza sativa ssp. japonica* | seedling shoot | ERR008651, ERR008652, ERR008657, ERR008658, ERR008663, ERR008664 |
| *Oryza sativa ssp. japonica* | leaf | SRR305891-SRR305893 |
| *Oryza sativa ssp. Japonica* | endosperm | SRR2338866-SRR2338867 |
| *Oryza sativa ssp. japonica* | panicle | SRR1633182-SRR1633187 |

| **Organism** | **Tissue type** | **RNA-seq data set** |
|---|---|---|
| *Oryza sativa ssp. japonica* | root | SRR1537554-SRR1537556 |
| *Oryza sativa ssp. indica* | seedling shoot | ERR008647, ERR008648, ERR008653, ERR008654, ERR008659, ERR008660 |
| *Oryza glaberrima* | leaf | SRR1174376 |
| *Oryza glaberrima* | panicle | SRR1174378 |
| *Oryza brachyantha* | root | SRR351196 |
| *Zea mays* | anther | SRR2078791-SRR2078793 |
| *Zea mays* | endosperm | SRR1169634-SRR1169635 |
| *Zea mays* | seed | SRR1170939-SRR1170940 |
| *Zea mays* | embryo | SRR531916, SRR531917, SRR531919 |
| *Zea mays* | ear | SRR2078794-SRR2078796 |
| *Sorghum bicolor* | anther | SRR349769 |
| *Sorghum bicolor* | embryo | SRR959765 |
| *Sorghum bicolor* | endosperm | SRR349768 |
| *Sorghum bicolor* | seed | DRR030758- DRR030760 |

| **Organism** | **Tissue type** | **RNA-seq data set** |
|---|---|---|
| *Medicago truncatula* | root | SRR1726578, SRR1726542, SRR1726506, SRR1726470 |
| *Cicer arietinum* | root | SRR1066056 |

**Table 3.3** Percentages of homologous lincRNA and microRNA loci in different plant lineages. Percentages shown are the percentages of lincRNAs or microRNAs found in at least one species in each category (within the family, in rosids, in asterids, and in monocots).

| | lincRNA | | | | miRNA | | | |
|---|---|---|---|---|---|---|---|---|
| **species** | **within family** | **rosids** | **asterids** | **monocots** | **within family** | **rosids** | **asterids** | **monocots** |
| Arabidopsis | 51.1% | 1.2% | 0.6% | 0.2% | 53.7% | 11.3% | 9.6% | 1.8% |
| rice | 48.9% | 3.2% | 3.9% | 5.0% | 51.0% | 3.0% | 2.8% | 6.4% |
| maize | 14.3% | 0.7% | 0.5% | 0.8% | 58.8% | 3.2% | 4.1% | 19.8% |
| Medicago | 18.0% | 3.6% | 4.2% | 1.7% | 22.3% | 3.6% | 4.5% | 0.7% |
| tomato | 69.3% | 5.1% | 6.1% | 3.2% | 77.9% | 20.2% | 17.0% | 6.5% |

**Table 3.4** Proportion of homologous lincRNA and microRNA loci in different plant lineages using the e-value cutoff of 1e-10 solely.

| | lincRNA | | | | miRNA | | | |
|---|---|---|---|---|---|---|---|---|
| **species** | **within family** | **rosids** | **asterids** | **monocots** | **within family** | **rosids** | **asterids** | **monocots** |
| Arabidopsis | 45.6% | 0.5% | 0.5% | 0.1% | 48.1% | 7.4% | 4.0% | 0.9% |
| rice | 43.0% | 1.7% | 1.9% | 2.8% | 38.3% | 1.8% | 1.6% | 4.6% |
| maize | 9.7% | 0.3% | 0.2% | 0.5% | 56.0% | 1.5% | 2.0% | 16.4% |
| Medicago | 12.0% | 2.0% | 2.5% | 0.9% | 15.7% | 3.0% | 4.5% | 0.7% |
| tomato | 63.3% | 2.9% | 3.6% | 1.9% | 73.2% | 12.3% | 13.0% | 3.2% |

**Table 3.5** Proportion of homologous lincRNA and microRNA loci in different plant lineages using the e-value cutoff of 1e-20.

| | lincRNA | | | | miRNA | | | |
|---|---|---|---|---|---|---|---|---|
| **species** | **within family** | **rosids** | **asterids** | **monocots** | **within family** | **rosids** | **asterids** | **monocots** |
| Arabidopsis | 33.0% | 0.2% | 0.1% | 0.1% | 33.9% | 1.3% | 0.3% | 0.1% |
| rice | 36.6% | 0.5% | 0.6% | 1.2% | 27.5% | 0.7% | 0.4% | 1.2% |
| maize | 5.1% | 0.0% | 0.0% | 0.0% | 30.1% | 0.0% | 0.0% | 2.9% |
| Medicago | 6.2% | 0.9% | 1.3% | 0.3% | 3.7% | 0.8% | 2.7% | 0.2% |
| tomato | 53.0% | 1.3% | 1.4% | 0.9% | 57.1% | 2.7% | 7.8% | 0.2% |

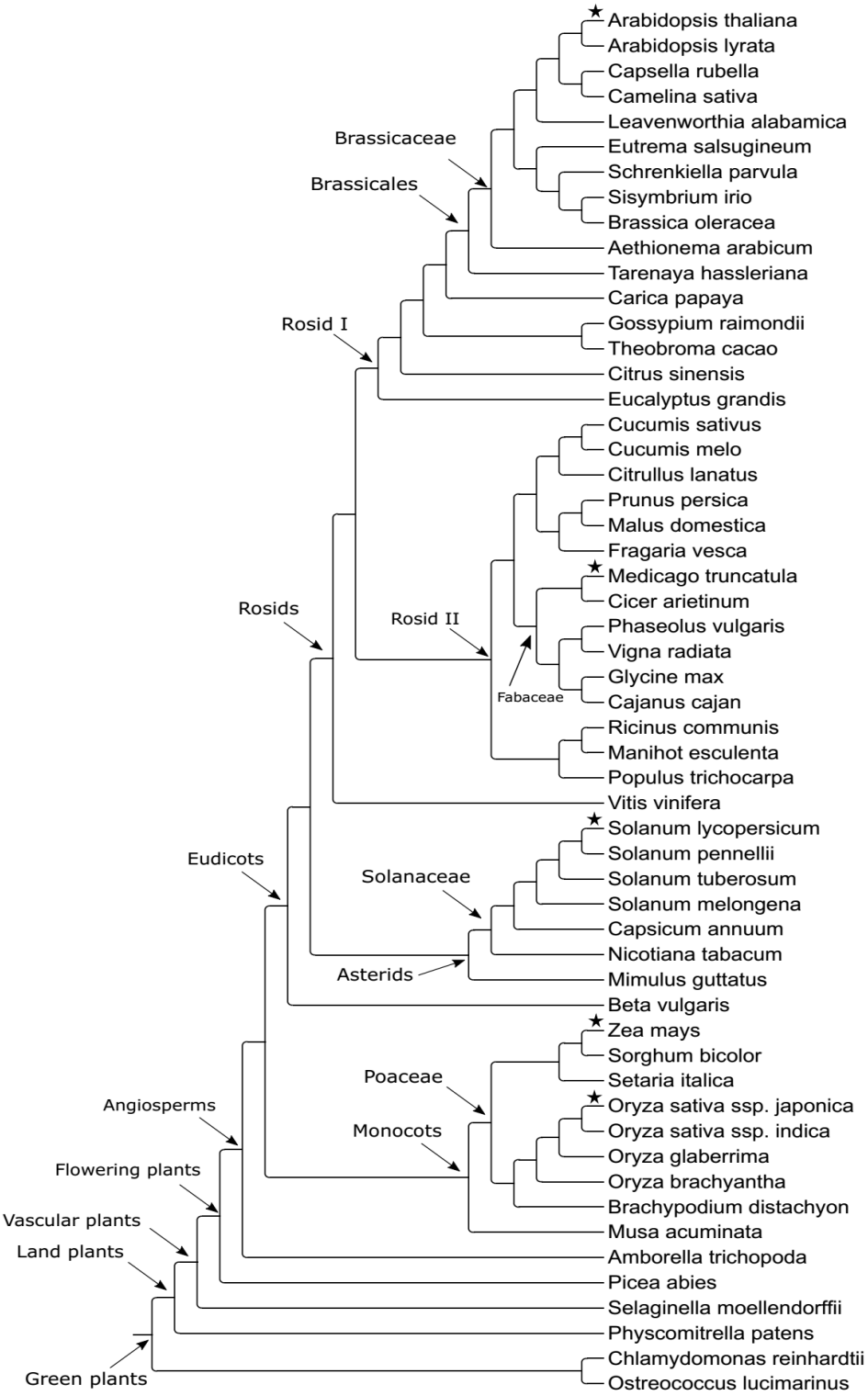


**Figure 3.1** The phylogeny of all 55 plants whose nuclear genomes were used in the study.

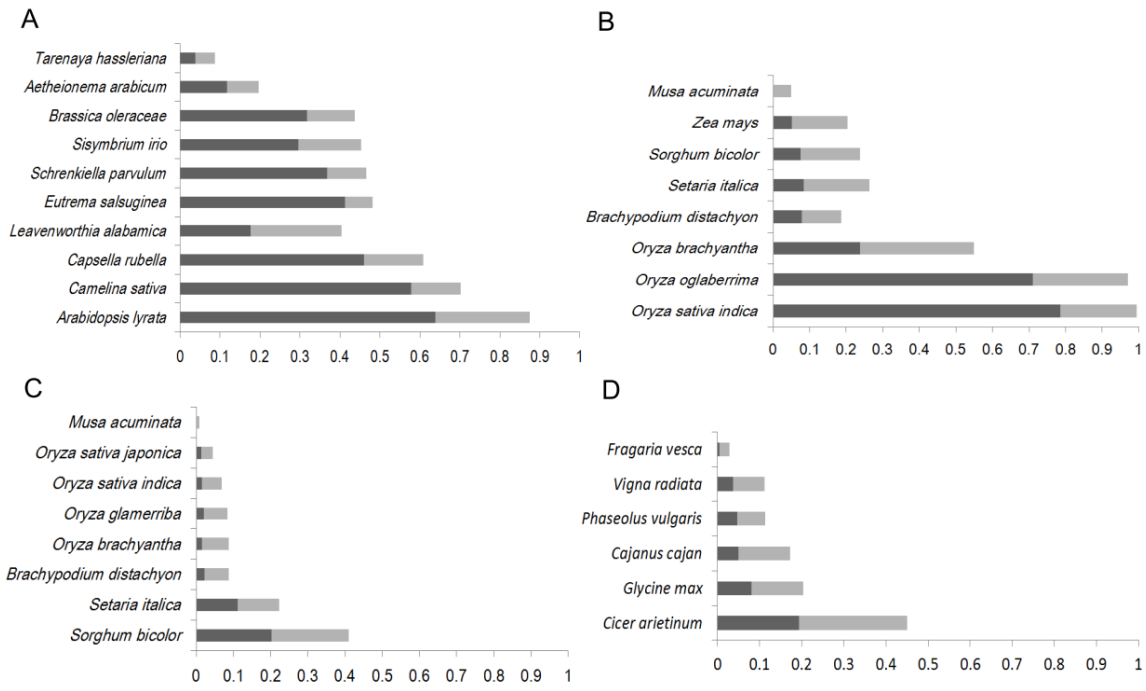


**Figure 3.2** Proportion of syntenic and non-syntenic homologous lincRNA loci from different species within a family. Syntenic and non-syntenic homologous lincRNA loci are in dark grey and light grey respectively. (A) *Arabidopsis thaliana* (B) *Oryza sativa ssp. japonica* (C) *Zea mays* (D) *Medicago truncatul*a.

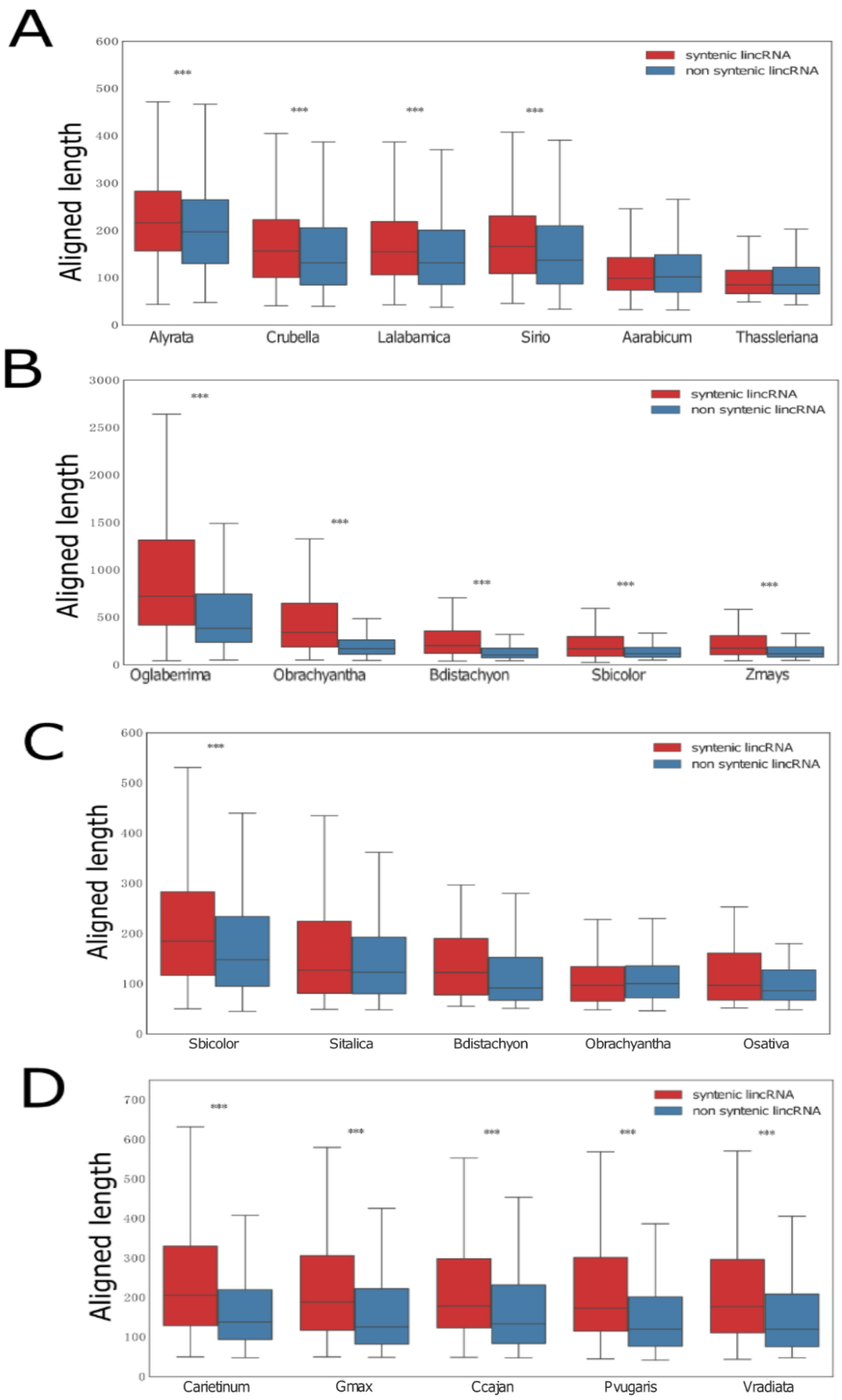
A
B
C
D
syntenic lincRNA
non syntenic lincRNA
Aligned length
***
Alyrata
Crubella
Lalabamica
Sirio
Aarabicum
Thassleriana
Oglaberrima
Obrachyantha
Bdistachyon
Sbicolor
Zmays
Sitalica
Osativa
Carietinum
Gmax
Ccajan
Pvugaris
Vradiata

**Figure 3.3** Comparison of aligned length of lincRNAs with and without syntenic loci in representative species. (A) Arabidopsis (B) rice (C) maize (D) Medicago. Syntenic and non-syntenic lincRNA loci are in red and blue, respectively. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to $\pm 1.5$ interquartile range (IQR). Asterisks: * *P*-value <0.05, ** *P*-value <0.01 and *** *P*-value <0.005.

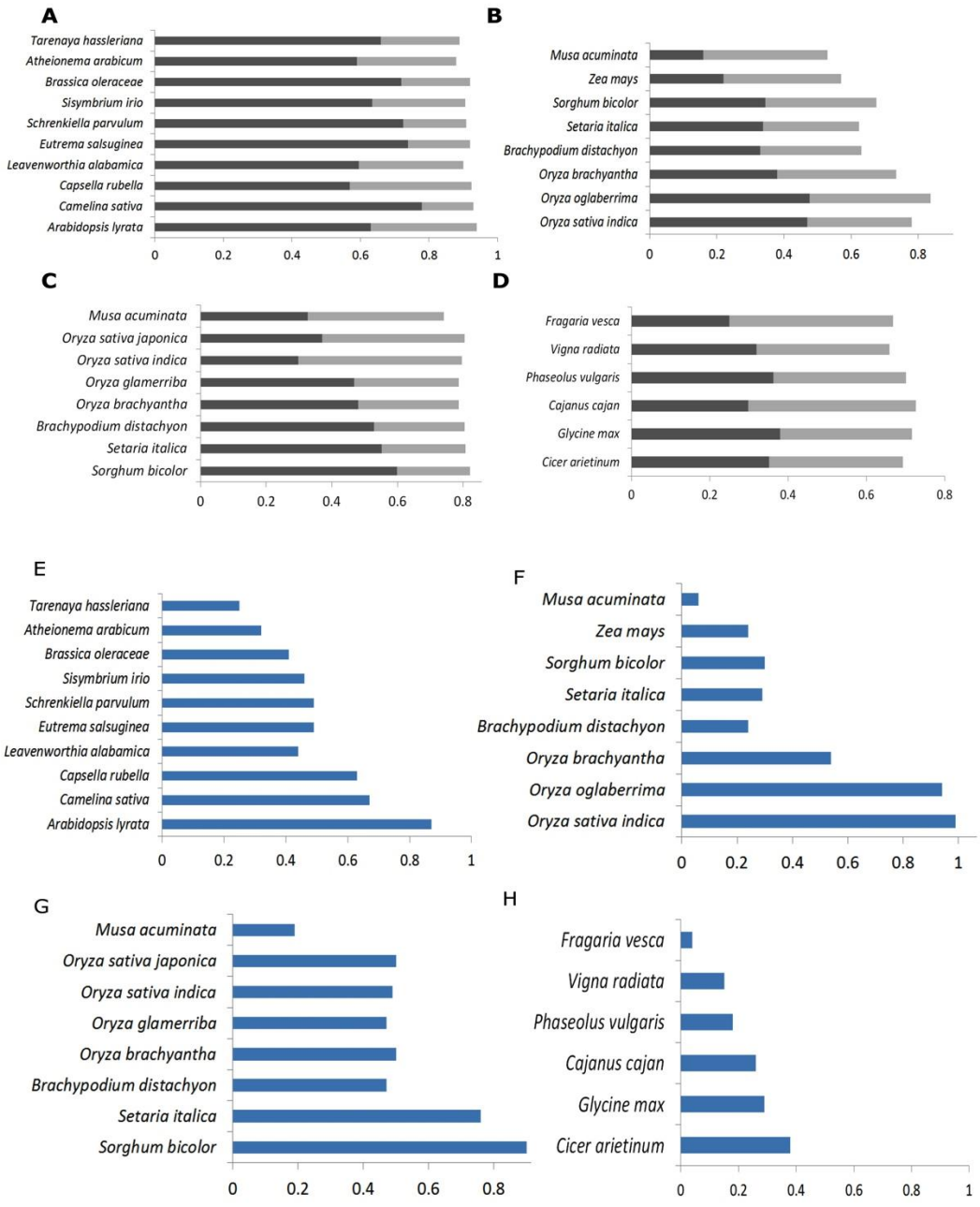


**Figure 3.4** Proportion of homologous (A-D) protein-coding gene loci and (E-H) microRNA gene loci in the indicated species. The graphs show comparisons to: (A) Arabidopsis, (B) rice, (C) maize, (D) Medicago, (E) Arabidopsis, (F) rice, (G) maize, (H) Medicago.

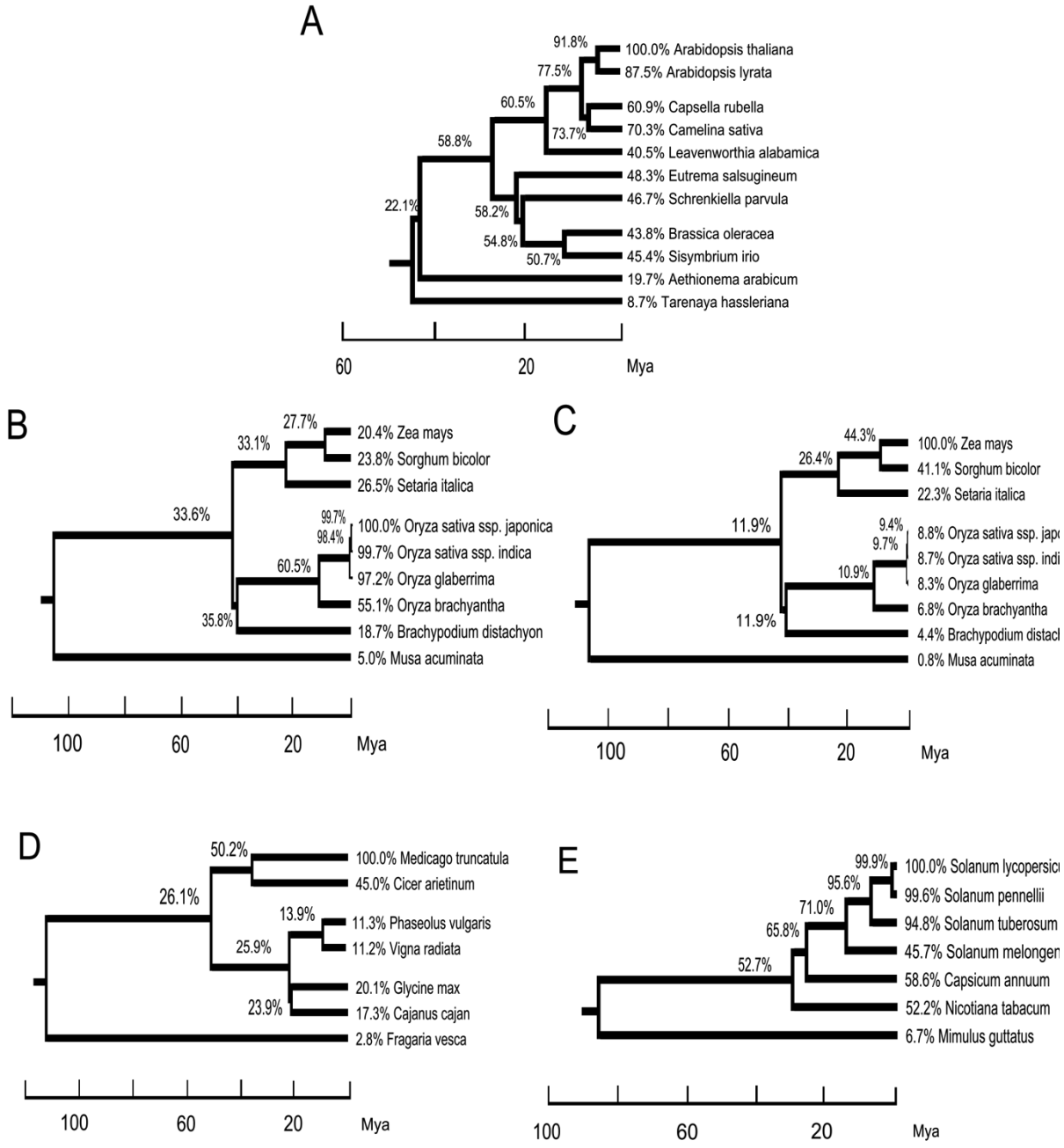


**Figure 3.5** Evolutionary gain and loss of lincRNAs from different species within a family. The value adjacent to each node refers to the inferred percentage of homologous lincRNA loci found at the corresponding ancestral plant. (A) *Arabidopsis thaliana*, (B) *Oryza sativa ssp. Japonica*, (C) *Zea mays*, (D) *Medicago truncatul*a, (E) *Solanum lycopersicum*.

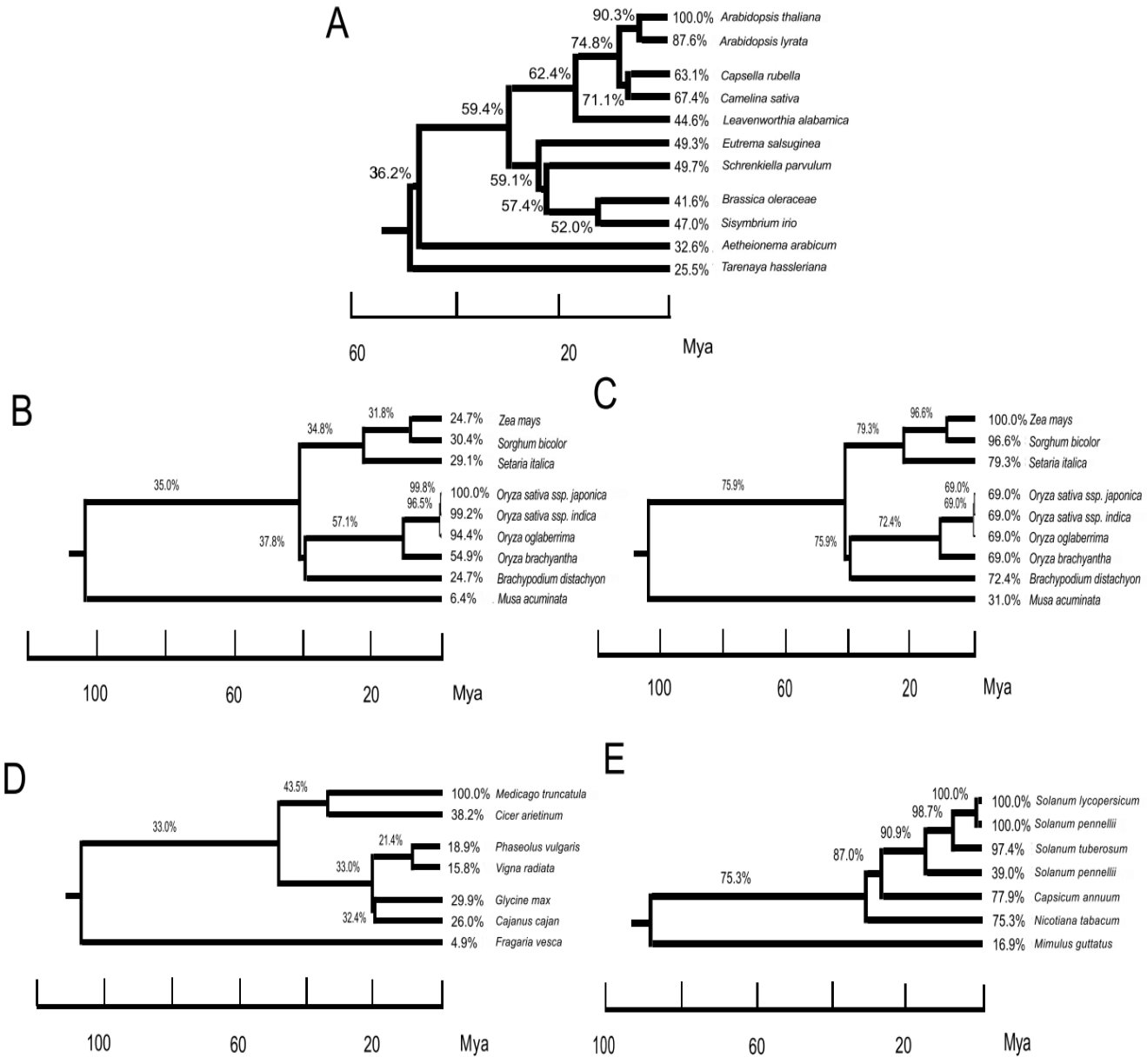


**Figure 3.6** Evolutionary gain and loss of microRNA genes from different species within a family. The value adjacent to each node refers to the inferred percentage of homologous microRNA loci found at the corresponding ancestral plant. (A) *Arabidopsis thaliana,* (B) *Oryza sativa ssp. Japonica*, (C) *Zea mays*, (D) *Medicago truncatul*a, (E) *Solanum lycopersicum.*

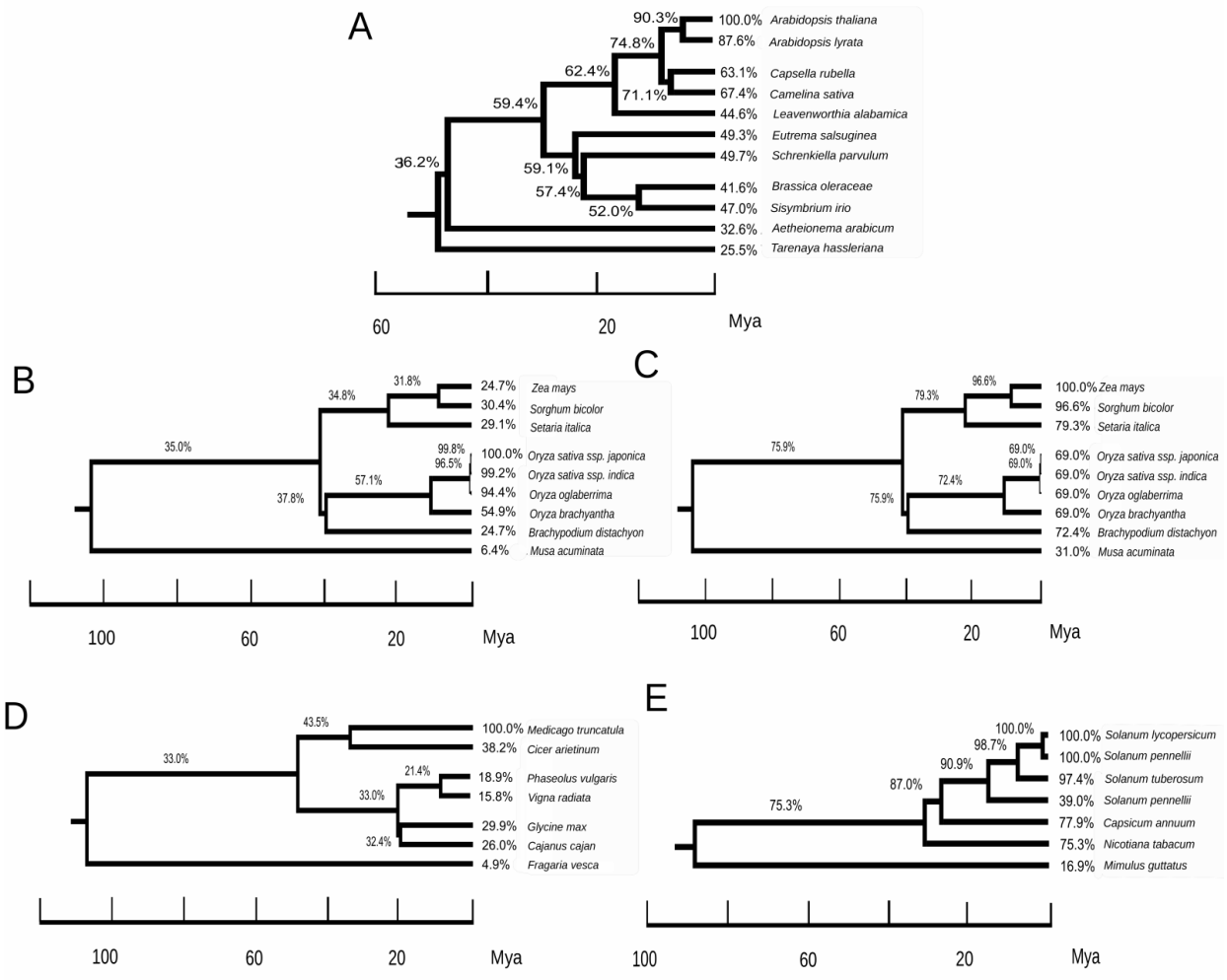


**Figure 3.7** Evolutionary gain and loss of homologous loci of lincRNAs from different species within a family. The value adjacent to each node refers to the inferred percentage of homologous lincRNA loci found at the corresponding ancestral plant. (A) *Arabidopsis thaliana*, (B) *Oryza sativa ssp. Japonica*, (C) *Zea mays*, (D) *Medicago truncatul*a, (E) *Solanum lycopersicum*.

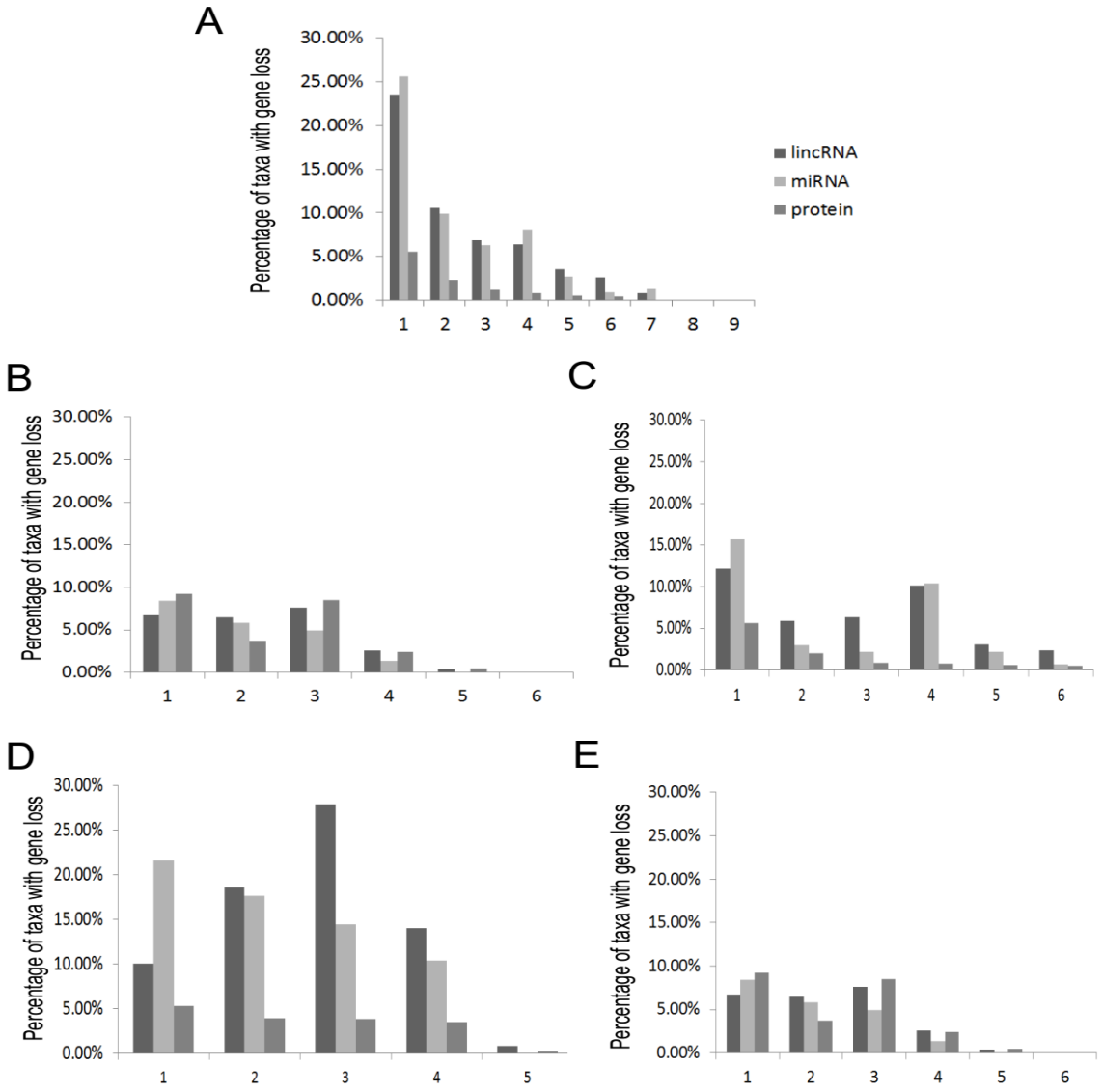


**Figure 3.8** The percentage of lincRNAs, microRNAs and protein-coding genes with potential lineage-specific loss from different species within a family. The X-axis represents the number of taxa while the Y-axis indicates the percentage of genes with the corresponding number of taxa with potential gene loss. The bars in dark, light and medium grey denote lincRNAs, microRNAs and protein-coding genes, respectively. (A) *Arabidopsis thaliana*, (B) *Oryza sativa ssp. Japonica*, (C) *Zea mays*, (D) *Medicago truncatul*a (E) *Solanum lycopersicum*.

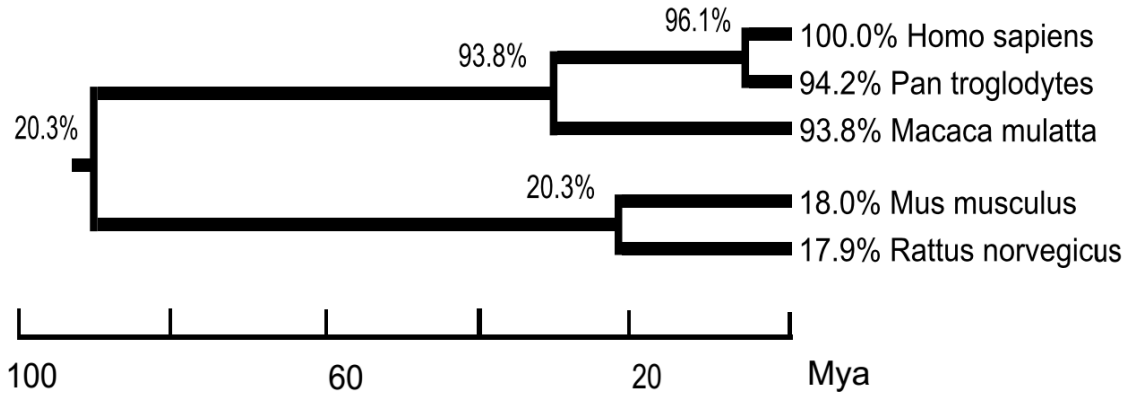


**Figure 3.9** Evolutionary gain and loss of human lincRNA homologous loci in mammals.

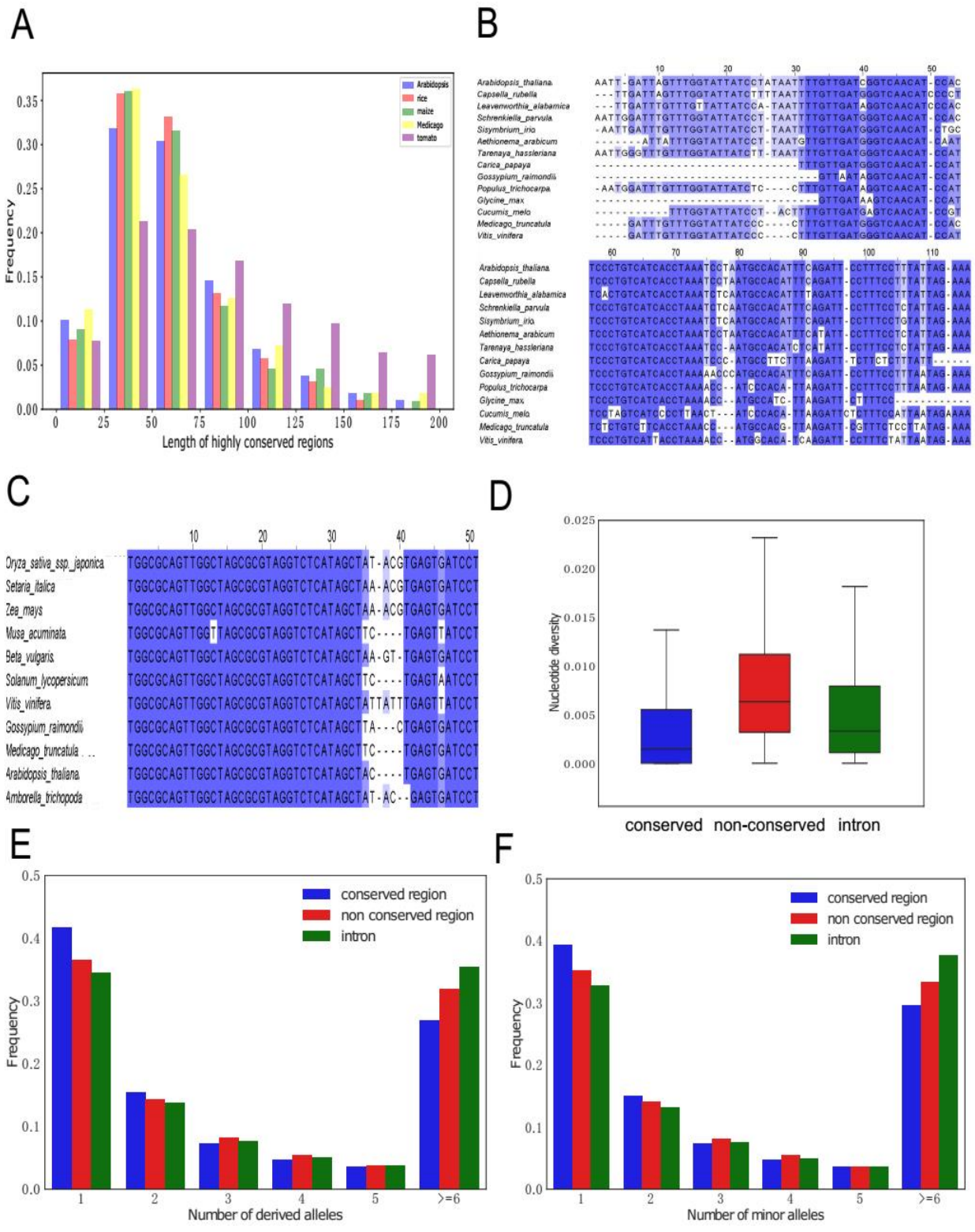


**Figure 3.10** Highly conserved regions in lincRNA sequences. (A) Distribution of the lengths of highly conserved regions in lincRNAs for *Arabidopsis thaliana*, *Oryza sativa*

*ssp. japonica*, *Zea mays*, *Medicago truncatul*a and *Solanum lycopersicum*. The X-axis represents the sequence length of the highly conserved regions identified in lincRNAs from the five species in the previous sentence. The Y-axis represents the frequency of the conserved regions with the corresponding sequence length. Only highly conserved regions with no more than 200 nt in length are shown. (B) The alignment of the conserved region in lincRNA XLOC_026561 from *Arabidopsis thaliana* across lineages. **(C)** The alignment of the conserved region in lincRNA XLOC_052616 from *Oryza sativa ssp. japonica* across lineages. (D) Nucleotide diversity of highly conserved regions and surrounding regions (non-conserved regions) of lincRNAs and introns of protein-coding genes in *Arabidopsis thaliana*. (E) The distribution of derived allele frequency (DAF) of highly conserved regions and surrounding regions (non-conserved regions) of lincRNAs and introns of protein-coding genes in *Arabidopsis thaliana*. (F) The distribution of minor allele frequency (MAF) of highly conserved regions and surrounding regions (non-conserved regions) of lincRNAs and introns of protein-coding genes in *Arabidopsis thaliana*.

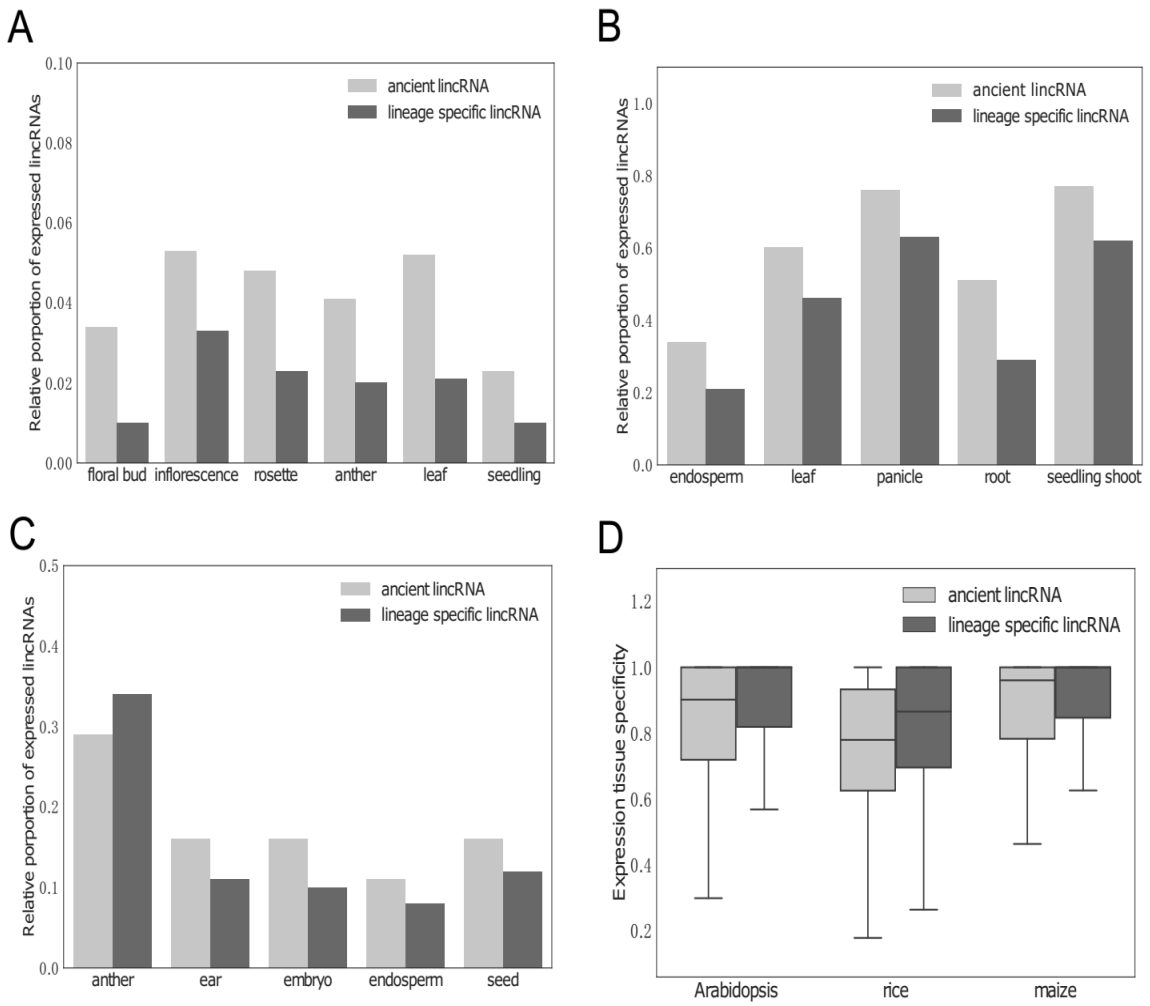


**Figure 3.11** Expression differences between ancient and lineage-specific lincRNAs. (A-C) Relative proportion of expressed lincRNAs in ancient lincRNA (non-genus-specific lincRNA) and lineage-specific lincRNA (genus-specific lincRNA) for (A) *Arabidopsis thaliana*, (B) *Oryza sativa japonica* and (C) maize. Relative proportion of expressed lincRNAs is defined as the number of expressed lincRNAs divided by the number of the total number of lincRNAs in each category (ancient lincRNA or lineage-specific lincRNA). (D) Box plots of the expression tissue specificity of ancient and lineage-specific lincRNAs in *Arabidopsis thaliana*, *Oryza sativa japonica* and maize.

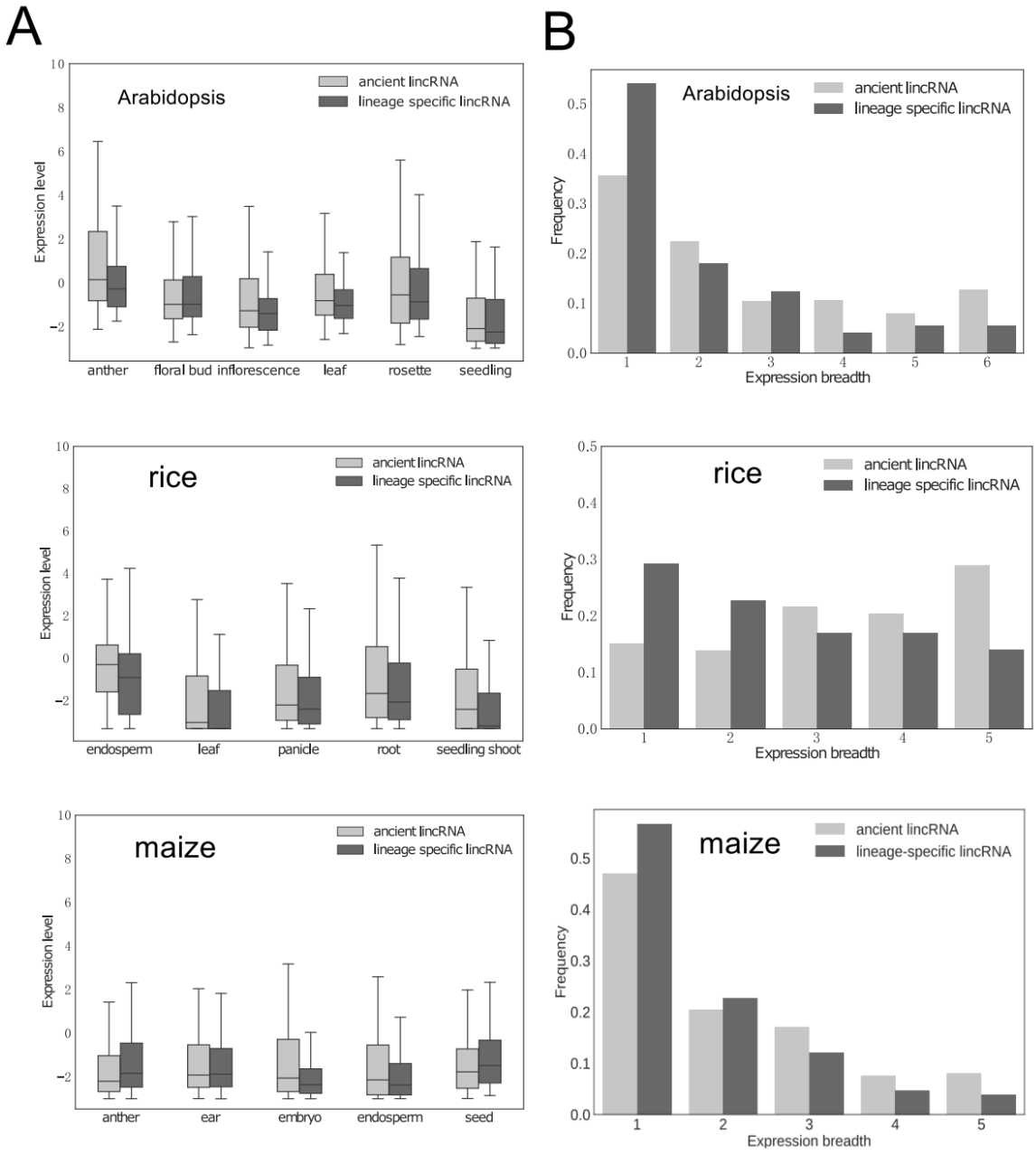


**Figure 3.12** Differences between ancient (non-genus-specific) and lineage-specific (genus-specific) lincRNAs in expression. (A) expression level; (B) expression breadth.

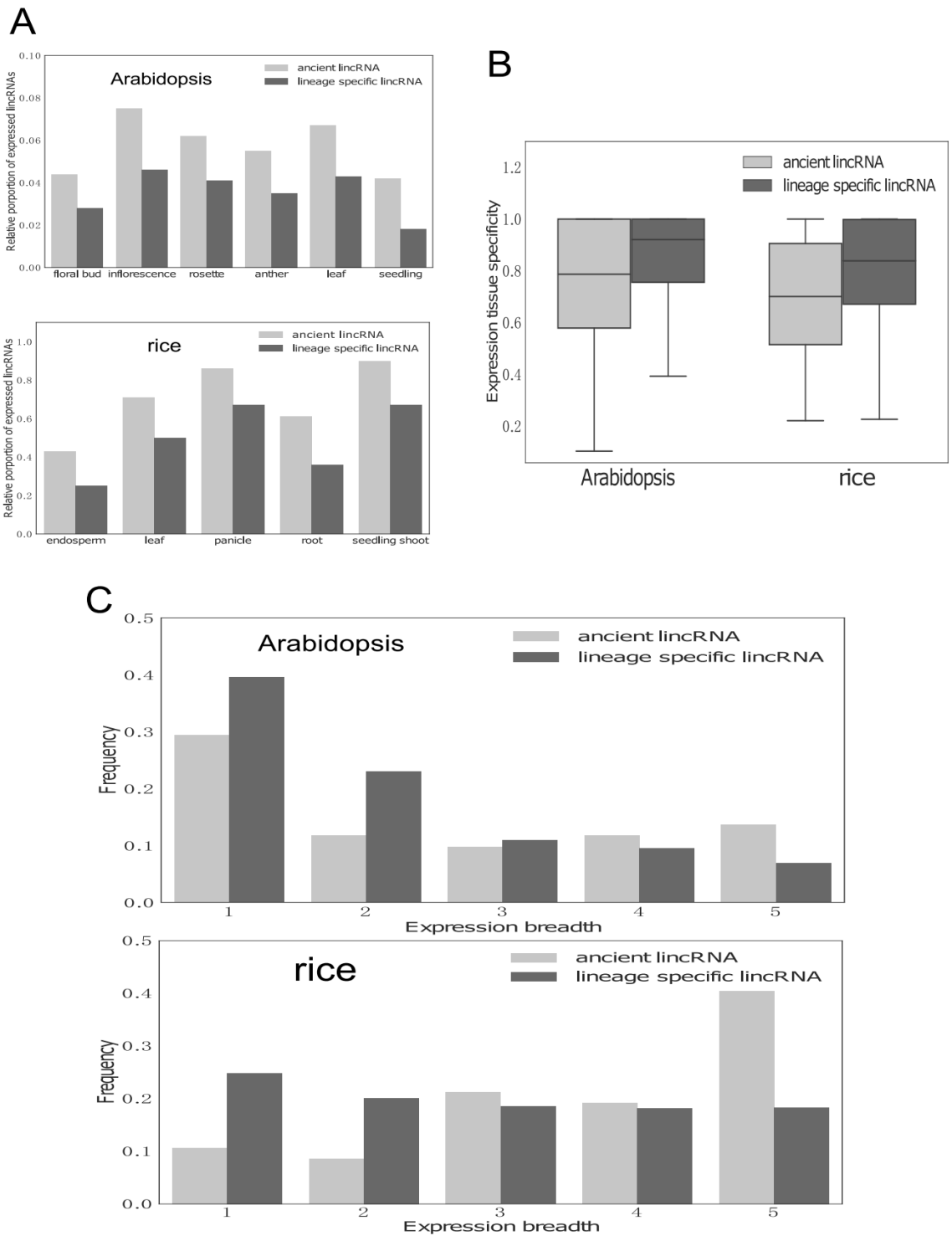


**Figure 3.13** Differences between ancient (non-family-specific) and lineage-specific (family-specific) lincRNAs in expression. (A) expression level; (B) expression tissue specificity; (C) expression breadth.

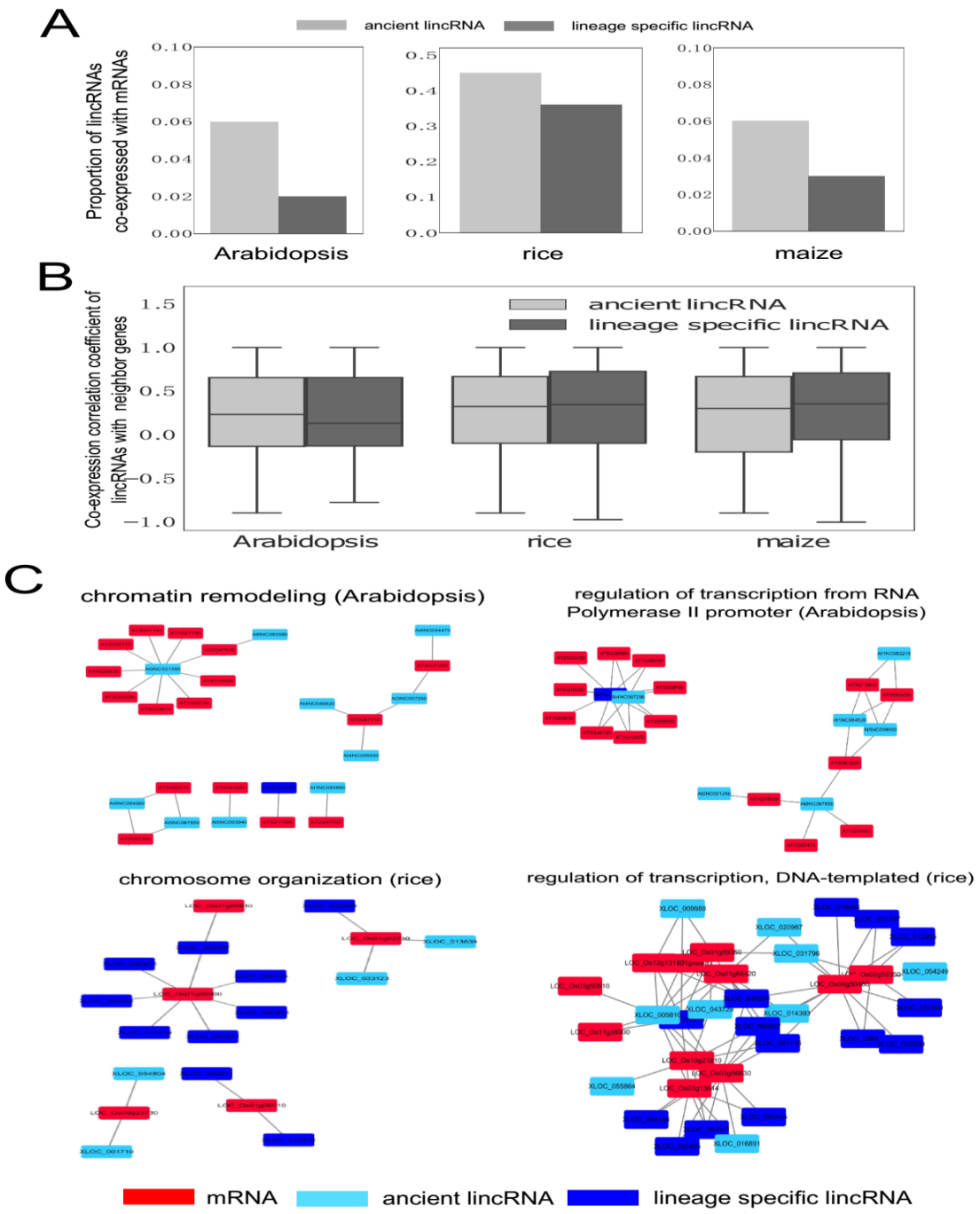


**Figure 3.14** Gene co-expression network analysis of ancient and lineage-specific lincRNAs. (A) Proportion of lincRNAs co-expressed with protein-coding genes. (B) Difference in co-expression co-efficient between lineage-specific and ancient lincRNAs. (C) Examples of co-expression network involving lincRNAs.

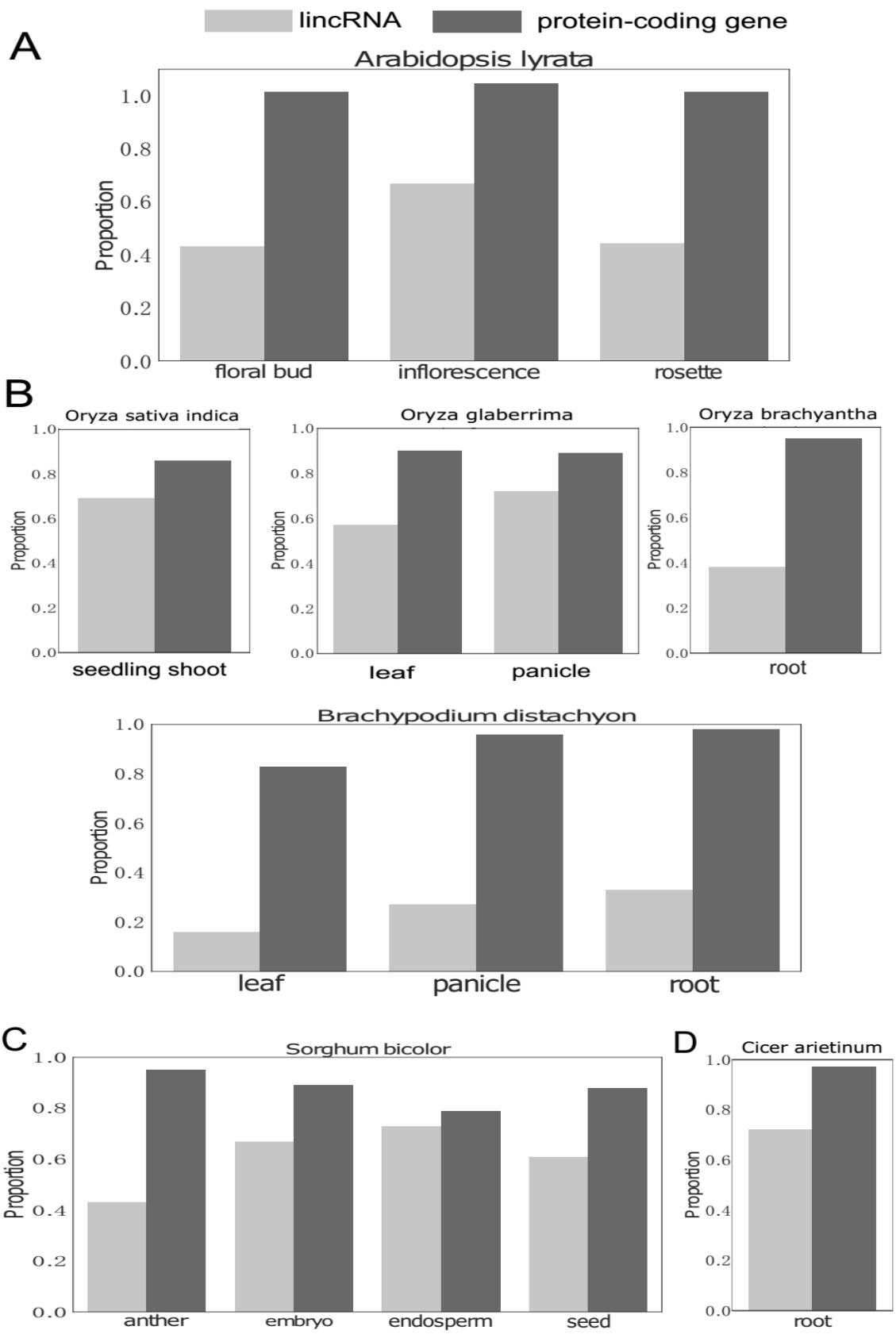
lincRNA
protein-coding gene
A
Arabidopsis lyrata
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
floral bud
inflorescence
rosette
B
Oryza sativa indica
seedling shoot
Oryza glaberrima
leaf
panicle
Oryza brachyantha
root
Brachypodium distachyon
leaf
panicle
root
C
Sorghum bicolor
anther
embryo
endosperm
seed
D
Cicer arietinum
root

**Figure 3.15** Proportions of lincRNAs and protein-coding genes expressed across species. Proportions of orthologous loci of lincRNAs (light grey) and protein-coding genes (dark grey) in (A) *Arabidopsis thaliana*, (B) *Oryza sativa japonica*, (C) *Zea mays* and (D) *Medicago truncatula* with evidence of expression in other indicated species are shown.

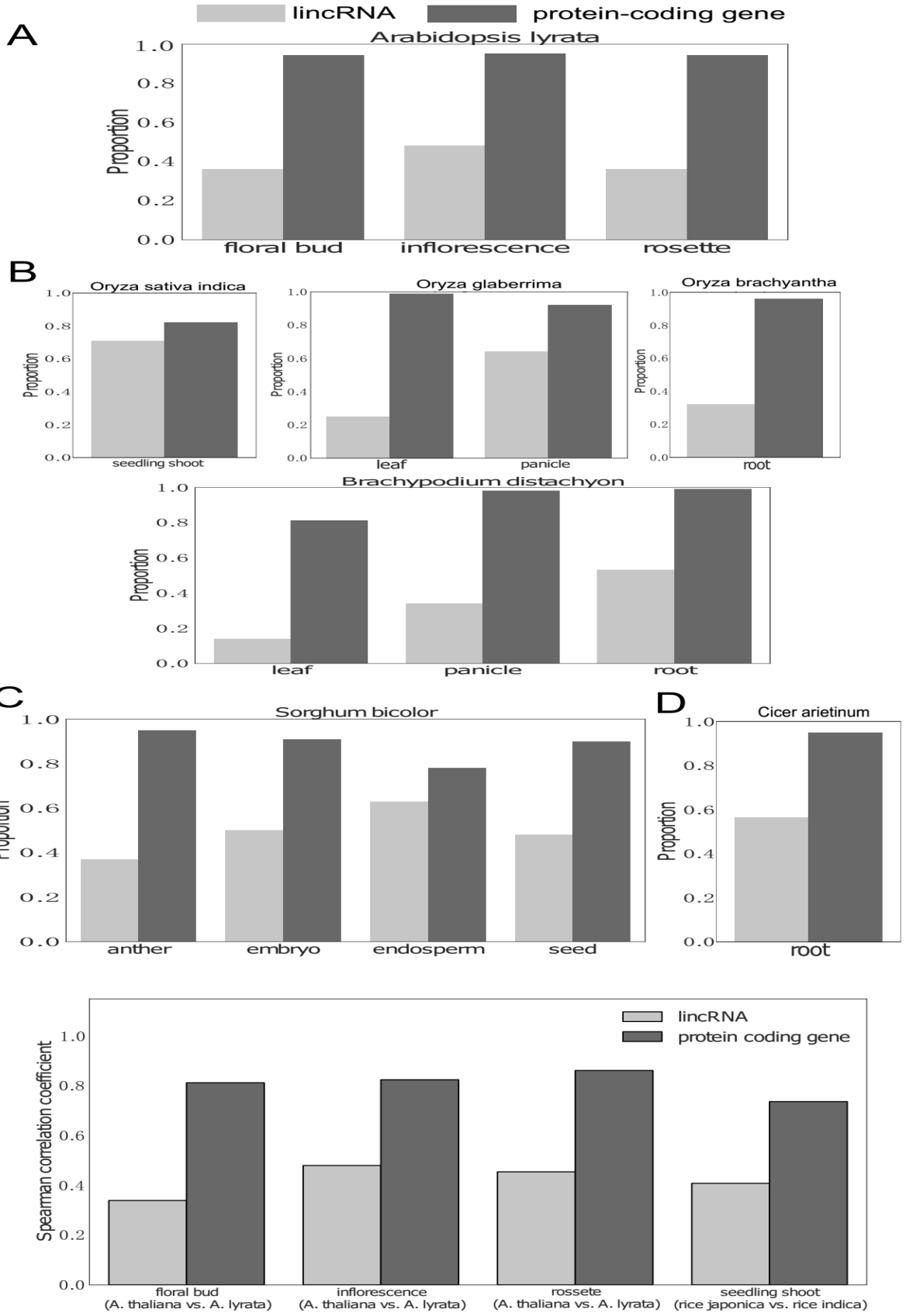
A
lincRNA
protein-coding gene
Arabidopsis lyrata
Proportion
floral bud
inflorescence
rosette
B
Oryza sativa indica
seedling shoot
Oryza glaberrima
leaf
panicle
Oryza brachyantha
root
Brachypodium distachyon
leaf
panicle
root
C
Sorghum bicolor
anther
embryo
endosperm
seed
D
Cicer arietinum
root
Spearman correlation coefficient
lincRNA
protein coding gene
floral bud
(A. thaliana vs. A. lyrata)
inflorescence
(A. thaliana vs. A. lyrata)
rossete
(A. thaliana vs. A. lyrata)
seedling shoot
(rice japonica vs. rice indica)

**Figure 3.16** Differences between lincRNAs and protein-coding genes in expression conservation across species.

# 4 Evolution of LincRNAs by Duplication in Plants[1]

## 4.1 Introduction

Gene duplication is a prominent mechanism for the evolution of protein-coding genes (Ohno 1970). Genomic data have revealed the abundance of a large number of duplicated genes in most sequenced eukaryotic genomes (Friedman and Hughes 2001; Zhang 2003; Flagel and Wendel 2009). Duplicated genes can be generated by whole-genome duplication (WGD), tandem duplication (TD) and interspersed duplication (Doyle et al. 2008). Angiosperms have undergone several rounds of polyploidy events, leading to a large number of whole-genome duplicates (Adams and Wendel 2005; Van de Peer et al. 2009; Schranz et al. 2012). Genes may experience different fates in expression after duplication. Some duplicates tend to be co-expressed and show redundancy while others might become divergent in expression giving rise to opportunities for evolutionary novelties (Zhang 2003). Of particular interest is reciprocal expression and silencing of paralogs among different tissue types where one copy is expressed in some organ types while the other copy is expressed in other organ types, resulting in differential contributions to the transcriptome between the two paralogs (Adams et al. 2003; Liu and Adams 2007).

LincRNAs (long intergenic non-coding RNAs) are derived from intergenic loci and they are common in many eukaryotic genomes (Ulitsky and Bartel 2013; Johnsson et al.

[1] Chapter 4 is in preparation for publication. **Wang SS**, Adams KL Evolution of lincRNAs by duplication in plants.

2014; Marques and Ponting 2014; Yang et al. 2014). Recently a growing body of evidence has revealed that lincRNAs play important roles in the cell (Ietswaart et al. 2012; Kim and Sung 2012; Quinn and Chang 2016). LincRNAs can act as molecular signals, decoys and scaffolds and participate in a variety of biological processes including epigenetics, microRNA target mimics and gene expression regulation (Wang and Chang 2011; Ulitsky and Bartel 2013). Although lincRNAs have been discovered for a long time, it was not until recent years that researchers could perform genome-wide identification, as well as demonstrate molecular functions for them (Necsulea et al. 2014; Chekanova 2015). The pace of lincRNA research has been greatly accelerated by the rapid development of high-throughput genomic technology and molecular tools. Using next-generation sequencing and genomic arrays, many studies have provided detailed transcriptional landscapes and revealed a large number of lincRNAs in various species (Liu et al. 2012; Li et al. 2014; Necsulea et al. 2014; Hezroni et al. 2015; Chen et al. 2016).

Previous studies, mostly in animals, have shown that lincRNAs evolve quickly and show rapid changes in expression (Kutter et al. 2012; Necsulea et al. 2014; Nitsche et al. 2015; Kornienko et al. 2016). In addition, some lincRNAs show positional conservation without sequence similarity detectable (Ulitsky et al. 2011; Mohammadin et al. 2015; H. Wang et al. 2015). Moreover, lincRNAs might serve as a repository for the generation of protein-coding genes during evolution (Ruiz-Orera et al. 2014).

However, the role of gene duplication in lincRNA evolution is poorly understood. It

was found that small-scale duplication appears to be the dominant type of duplication for lincRNAs in Brassicaceae (Nelson et al. 2016). Other attempts to explore this topic are mostly from animals based upon a limited number of species. It was found that only a very small proportion of lincRNAs were generated by duplication in human, and de novo birth was conceived to be the main driving force in the expansion of lincRNA gene pools (Derrien et al. 2012). The growing number of lincRNAs identified on a large scale in plants, whose genomes are extensively shaped by rounds of WGD, provides the opportunity to investigate how gene duplication influences the evolution of lincRNAs and compare with protein-coding genes in different plant lineages.

In this study, we explored the evolutionary role of duplication in plant lincRNAs. We collected lincRNA data sets from multiple representative plant lineages including *Arabidopsis thaliana*, *Populus trichocarpa* and *Cucumis sativus* from rosids, *Solanum lycopersicum* from asterids and *Oryza sativa ssp. japonica* and *Zea mays* from grasses. We characterized duplicated lincRNAs and compared between different plants. We classified duplicated lincRNAs into different types based on how they were derived. We also analyzed the expression patterns of duplicated lincRNAs and compared with protein-coding genes. Moreover, we investigated factors that were likely associated with the expression divergence between duplicated lincRNAs.

## 4.2 Materials and methods

### 4.2.1 Sources of lincRNAs and genomic sequences

Sequences of lincRNAs were retrieved from *Arabidopsis thaliana, Populus trichocarpa, Cucumis sativus, Solanum lycopersicum, Oryza sativa,* and *Zea mays* (Table 4.1). For loci of lincRNAs and protein-coding genes with more than one transcript or isoform, only the longest one was used for analyses.

### 4.2.2 Identification of duplicated lincRNAs

All-against-all BLASTN was performed to identify duplicated lincRNAs. Considering the relatively fast evolutionary rate of lincRNAs (Ulitsky et al. 2011; Necsulea et al. 2014), the word size of 7 was used in the BLASTN search to improve the sensitivity of the identification of duplicated lincRNAs (Ulitsky et al. 2011). LincRNAs were defined as duplicates if they had at least one non-self BLAST hit with e-value lower than or equal to 1e-10. Additionally, an e-value cutoff of 1e-5 together with the minimum aligned length of 50 bp and sequence identity above 80% were used to identify more divergent lincRNA duplicates and to see if the use of less stringent criteria would affect analyses.

Duplicated lincRNAs were paired with their best non-self BLAST hit defined by those with the lowest e-value to generate duplicated lincRNA pairs.

### 4.2.3 Identification of duplicated protein-coding genes

All-against-all BLASTP searches were run to search for duplicates of protein-coding

genes. Duplicated genes were identified if they have non-self BLAST hits with an e-value less than or equal to 1e-10 (Casneuf et al. 2006; Yang and Gaut 2011). Pairs of protein-coding genes were generated based on the same principle as for lincRNAs mentioned above.

### 4.2.4 Classification of duplicated lincRNAs

Duplicated lincRNAs were classified into different categories according to the mechanisms via which they originated. Whole-genome duplicates were defined according to the following procedure. We collected WGD blocks and protein-coding genes found in WGD blocks for Arabidopsis, poplar, tomato, rice and maize from corresponding publications (Table 4.2). No WGD duplicates from cucumber were analyzed as no recent WGD event has been reported in this lineage (Huang et al. 2009). First, all duplicated lincRNAs which are found in WGD blocks were extracted. Second, duplicated lincRNAs found in any WGD block were searched for homologous lincRNAs found in the sister WGD block. Last, for those whose homologs were found in the sister WGD block, 10 upstream and 10 downstream protein-coding genes from each lincRNA derived from WGD were extracted to get the set of surrounding genes. If at least 10 out of the 20 surrounding protein-coding genes were found to be WGD-derived duplicated pairs, the two duplicated lincRNAs were considered to be a pair of WGD-derived lincRNAs. Triplicates derived from whole-genome triplication (WTD) in tomato were identified similarly. The genomic contexts of all lincRNAs derived from WGD or WTD were

examined with GenomeDiagram (Pritchard et al. 2006) implemented in Biopython and further displayed using Circos (Krzywinski et al. 2009).

Tandemly duplicated lincRNAs were defined following the procedure of (Hanada et al. 2008; Zou et al. 2009) using the following criteria: (i) they meet the requirement of duplicated lincRNAs as defined above, (ii) they are located within 100 kb on the same chromosome. Duplicated lincRNA pairs not included in whole-genome duplicates and tandem duplicates are defined as interspersed duplicates.

### 4.2.5 Evolutionary rates calculation

The aligned regions of each pair of duplicated lincRNAs were retrieved based on the BLAST result and aligned using MUSCLE v3.8 (Edgar 2004) with default parameters. Sequence divergence for all alignments was calculated using the ape package in R (Paradis et al. 2004) based on the evolutionary model of K80 (Kimura 1980). Sequence format conversion and processing were performed using seqmagick (http://fhcrc.github.com/seqmagick) and custom scripts written in Perl and Ruby (Goto et al. 2010).

### 4.2.6 Expression divergence comparison

Transcriptomic RNA-seq data from public datasets were retrieved from NCBI SRA (Sequence Read Archive) database (Kodama et al. 2012) (Table 4.3). RNA-seq reads were trimmed by Cutadapt v1.3 (Martin 2011) to filter out those with sequencing quality

less than 20 and sequence length less than 20 bp before being mapped to the genome. Processed reads were mapped to the reference genome using STAR v2.4.2 (Dobin et al. 2013) (non-default parameters: --alignIntronMax 25,000). The FPKM value for each gene was calculated using Cufflinks v2.1.1 (Trapnell et al. 2012) and was further normalized via log-transformation. For organ types and developmental stages with more than one replicate, the average of all replicates was used as the expression level of a gene. The Pearson correlation coefficient was calculated for each gene pair based on log2-transferred values of FPKM obtained from different organ types or developmental stages as used in (Blanc and Wolfe 2004). Expression Euclidean distance was calculated based on the following formula (Pereira et al. 2009)

$$EucD = \sqrt{\sum_{j=1}^{k} \left(x_{1j} - x_{2j}\right)^2}$$

where $x_{ij}$ denotes the expression level of the gene under consideration in species *i* and in tissue *j* and *k* denotes the number of tissues or developmental stages.

Duplicated genes with reciprocal expression were identified based on the procedure described in (Liu et al. 2011). Briefly, a pair of reciprocally expressed genes should meet the following criteria: First, both genes should be expressed in at least one tissue/organ type. Second, in a certain condition gene 1 is specifically expressed (i.e. gene 2 is not expressed) and in another condition gene 2 is specifically expressed (i.e. gene 1 is not expressed). For organ types or developmental stages with at least two replicates, a gene was considered to be expressed only if the expression was detected in all replicates.

Reciprocally expressed duplicates are considered to show completely reciprocal expression if they show no overlap in expressed organ types and have fully partitioned spatial expression (Figure 4.1).

Tissue expression complementarity (TEC) was used as an index to measure the level of complementary expression (Huerta-Cepas et al. 2011). TEC was calculated by calculating the relative number of tissues or developmental stages where only one gene is expressed over the total number of tissues or developmental stages in which each gene is expressed:

$$TEC_{ij} = \frac{d_i/t_i + d_j/t_j}{2}$$

where $d_i$ denotes the number of tissues in which gene *i* is expressed whereas gene *j* is not expressed and $t_i$ denotes the total number of tissues where gene *i* is expressed. The greater the value of TEC, the higher the level of tissue complementarity of the gene pair (Huerta-Cepas et al. 2011).

## 4.3 Results

### 4.3.1 Highly variable components of duplicated lincRNAs in different species

We first performed genome-wide identification of duplicated lincRNAs in different plant genomes. We chose six plant species from three representative clades of plant phylogeny, including *Arabidopsis thaliana, Populus trichocarpa, Cucumis sativus, Solanum lycopersicum, Oryza sativa*, and *Zea mays*, to see whether different species displayed the same or different evolutionary patterns in lincRNA duplication (see Methods).

All-against-all BLAST searches were performed to identify duplicates of lincRNAs. Duplicated lincRNAs were defined as those with at least one non-self hit with e-value less than or equal to 1e-10 (See Methods). Distinct from previous studies in animals where a very small proportion (4.3%) of duplicated lincRNAs was identified (Derrien et al. 2012), we found that some plant species, particularly rice and maize, contain a large proportion of lincRNAs that are duplicated (Figure 4.2). In contrast, all three species from rosids displayed a very low proportion of lincRNA duplicates with *Arabidopsis thaliana* having only 10% duplicated lincRNAs. Tomato showed a medium level of lincRNA duplication with a proportion of lincRNA duplicates of 51%. The huge variation of the amount of duplication of lincRNAs among different lineages is in sharp contrast to protein-coding genes where different plant genomes encode relatively similar proportions of duplicates (Figure 4.2).

We compared the proportions of duplicated genes between lincRNAs and protein-coding genes. We found that duplicated genes for lincRNAs showed significantly lower proportions than genes for proteins ($p$ < 1e-20 in all species, chi-squared test) (Figure 4.2). This trend was most prominent in *Arabidopsis* where protein-coding genes were about seven times more enriched for duplicates than lincRNAs. Even in rice and maize where about 65% of lincRNAs are duplicated genes, protein-coding genes still showed higher proportions of duplicated genes than lincRNAs (Figure 4.2). Since lincRNAs are often rapidly evolving, some duplicated lincRNAs may not be detected using the e-value cutoff of 1e-10. To see whether the results were affected by the e-value

cutoff, we extended the e-value cutoff from 1e-10 to 1e-5 (see Methods). We repeated the same analyses and got similar results (Figure 4.3). These results suggest that fewer lincRNAs are duplicated compared with protein coding genes in plants.

### 4.3.2 Duplicated lincRNAs are mainly generated by interspersed duplication

Duplicated genes can be born in a variety of ways. The way duplicates are created might have an important impact on their evolutionary fates. To investigate the influences of different types of duplication on lincRNAs, we classified duplicated lincRNAs into whole genome duplicates, tandem duplicates and interspersed duplicates (see Methods). For simplicity, we only focused on the most recent polyploidy event in each lineage. Contrary to protein-coding genes, only a small fraction of duplicate lincRNA pairs were found to be derived from whole-genome duplication (Table 4.4; Figure 4.4; https://figshare.com/articles/Duplicated_lincRNAs/4959176). *Populus* contains the most abundant WGD-derived duplicated lincRNA pairs (8%), possibly because the Salicoid-specific WGD occurred recently in this lineage (Tuskan et al. 2006). We detected 11% of the duplicated lincRNAs are in tandem gene clusters in *Populus*. On average 15% of tandem duplicate clusters contained more than two members (https://figshare.com/articles/Duplicated_lincRNAs/4959176). The rest of the duplicated lincRNA pairs (80%) are interspersed duplicates. Rice has only 0.7% WGD-derived lincRNAs. 95% of duplicated lincRNAs were generated by interspersed duplication in rice. The high proportions of interspersed duplicated lincRNA pairs suggest that

duplicated lincRNAs are mainly generated via interspersed duplication.

### 4.3.3 Extensive expression divergence between paralogous lincRNAs

It has been shown that many duplicated protein-coding genes show expression divergence. To explore how paralogous lincRNA genes diverge in expression after duplication and compare with protein-coding genes, we analyzed 58 RNA-seq data sets from Arabidopsis, rice and maize including multiple tissue types and developmental stages (Table 4.3). We analyzed the co-expression pattern for duplicates of lincRNAs and protein-coding genes. A higher value of expression coefficient indicates a higher tendency of co-expression between two duplicated genes. Co-expression correlation coefficients of duplicated lincRNAs are on average 0.35, 0.34 and 0.32 for Arabidopsis, rice and maize, respectively, suggesting extensive expression divergence between lincRNA paralogs (Figure 4.5a). The co-expression correlation coefficients of paralogous lincRNAs are significantly higher than randomly chosen lincRNAs pairs in all the three species (Figure 4.6), indicating that the expression of paralogous lincRNAs was significantly correlated in general. As the expression of lincRNAs was reported to be correlated with their neighboring protein-coding genes in plants (Liu et al. 2012), we analyzed the co-expression pattern of lincRNAs and the protein-coding genes most physically adjacent to them on the chromosome. Duplicates of lincRNAs (compared with each other) were still found to show significantly higher co-expression correlation coefficients than duplicates compared with their neighboring protein coding genes

(Figure 4.7).

We further calculated co-expression correlation coefficients between paralogs of protein-coding genes using the same RNA-seq data sets. The average correlation coefficients of protein duplicates were 0.68, 0.51 and 0.61 for Arabidopsis, rice and maize, respectively, all of which were significantly higher than those of lincRNAs ($p <$ 1e-5, Wilcoxon test) (Figure 4.5a).

To see whether the results could be affected by the index used to measure expression divergence, we calculated the expression Euclidean distance, another commonly used index to measure gene expression divergence, for duplicates of both lincRNAs and protein-coding genes. Again, paralogous lincRNAs showed more expression divergence than protein-coding genes indicated by higher values of expression Euclidean distance ($p$ < 1e-7, Wilcoxon test) (Figure 4.5b). To see if the divergent expression of lincRNA duplicates could be affected by the e-value cutoff used, we repeated the analyses with a new set of lincRNA duplicates identified with the e-value cutoff of 1e-20. The results were similar and the same conclusion held true (Figure 4.8-4.9). Therefore the above results demonstrated that duplicated genes of lincRNAs showed more extensive expression divergence than protein-coding genes.

Duplicated genes derived from different mechanisms may show differences in expression. To explore the differences among different modes of lincRNA duplicates, we calculated co-expression correlation coefficients and expression Euclidean distances for whole-genome duplicates, tandem duplicates and interspersed duplicates of lincRNAs.

We focused analyses on rice and maize as the numbers of lincRNA pairs derived from WGD and TD in other species are too small to make comparisons. For simplicity, tandem gene clusters with more than two members were not analyzed. We found that paralogs derived from tandem duplication showed higher co-expression correlation coefficients and lower expression Euclidean distances (Figure 4.10), indicating their higher expression similarity in comparison to the other two types of duplicated lincRNAs.

### 4.3.4 Widespread tissue-specific complementary expression of duplicated lincRNAs

Additionally, we explored the expression divergence of duplicated lincRNAs in terms of tissue types and developmental stages. Complementary expression is an important indicator of expression divergence in different tissues or developmental stages between duplicated genes. The most extreme case of complementary expression is reciprocal expression where one paralog is only expressed in some organ types or developmental stages whereas the other copy is only expressed in others (Liu et al. 2011). LincRNAs tend to be expressed in specific tissue types, developmental stages or in response to certain stimuli (Ulitsky and Bartel 2013; Chekanova 2015). It could be hypothesized that the highly tissue-specific expression pattern might lead to extensive complementary expression between duplicated lincRNAs. To test this hypothesis, we calculated the proportion of duplicated gene pairs showing reciprocal expression patterns in Arabidopsis, rice and maize for both lincRNAs and protein-coding genes. The proportions of reciprocally expressed lincRNA gene pairs were 0.28, 0.30 and 0.41 in Arabidopsis, rice

and maize respectively, all of which were significantly higher than protein-coding genes ($p$ < 1e-8, chi-squared test) (Figure 4.11a). We also calculated the proportions of gene pairs with completely reciprocal expression where the expression of both paralogs is completely separated (see Methods). Again, lincRNA duplicates exhibited significantly higher proportions of completely reciprocally expressed gene pairs than protein-coding genes ($p$ < 1e-15, chi-squared test) (Figure 4.11b).

To further quantify the extent of complementary expression, we calculated tissue expression complementarity (TEC). TEC is an index to quantitatively measure the level of complementary expression and is briefly defined as the number of mutually exclusive tissues where two paralogs are expressed divided by the combined expression breadth of the pair of duplicates (Huerta-Cepas et al. 2011) (see Methods). A higher value of TEC is indicative of higher level of complementary expression (Huerta-Cepas et al. 2011). We found that the values of TEC for lincRNAs were significantly lower than protein-coding genes in all species examined ($p$ < 0.0001, Wilcoxon test) (Figure 4.11c), suggesting more extensive complementary expression for lincRNAs. Furthermore, we calculated the proportion of tissue or developmental stages where both copies of genes are expressed (see Methods). We observed that the proportions for lincRNAs were 0.23, 0.43 and 0.34 for Arabidopsis, rice and maize respectively, significantly lower than protein-coding genes ($p$ < 1e-9, Wilcoxon test) (Figure 4.11d). Therefore, all these results demonstrated that lincRNAs showed more divergent expression pattern in terms of tissues or developmental stages and exhibited more extensive complementary expression compared

to proteins.

We also explored the differences among different modes of duplicates of lincRNAs. Tandem duplicates showed a significantly higher frequency of complementary expression than whole-genome duplicates and interspersed duplicates (Figure 4.12-4.15). The differences between whole-genome duplicates and tandem duplicates were less prominent in rice, possibly due to the relatively small number of duplicated lincRNA pairs derived from WGD in rice.

### 4.3.5 Expression divergence of duplicated lincRNAs is not correlated with sequence divergence

The expression divergence of duplicates of proteins has been shown to be correlated with sequence divergence (Gu et al. 2002; Ganko et al. 2007). To investigate whether the same conclusion holds true for lincRNAs, we calculated sequence divergence for duplicated lincRNAs. To ensure a higher reliability of alignment, which is important for calculating sequence evolutionary rates, only duplicated lincRNAs with the alignment coverage higher than or equal to 50% were used in the analysis. Contrary to proteins, we did not detect any significant correlation between sequence divergence and expression divergence for lincRNAs in any of the analyzed species (Table 4.5). To examine whether the results were affected by the cutoff of alignment coverage, we repeated the same analysis for all duplicated lincRNAs with no cutoff of alignment coverage required. We found similar results (Table 4.6). Therefore, these results showed that expression

divergence is not correlated with sequence divergence for duplicated lincRNAs in plants, which is different from proteins.

## 4.4 Discussion

### 4.4.1 Duplicated lincRNAs display extensive variation among plant lineages and distinct patterns from protein-coding genes

Our study highlights several striking patterns in lincRNA duplication. First, we found that lincRNAs have a lower proportion of duplicated genes than protein-coding genes. Second, lincRNAs show highly distinct patterns across species. Different plants showed highly variable fractions of duplicated lincRNAs in contrast to smaller differences in the fractions of duplicated proteins across species (Figure 4.2). Gene duplicability of protein-coding genes is considered to be highly consistent across angiosperms (Li et al. 2016). The highly variable repertoires of duplicated genes of lincRNAs are in sharp contrast to protein-coding genes.

Whole-genome duplication contributes greatly to the expansion of protein-coding genes. For instance, approximately 25% of duplicated genes derived from the most recent whole-genome duplication event in Brassicaceae are estimated to be retained in duplicate pairs in the genome of *Arabidopsis thaliana* (Bowers et al. 2003). My results showed that distinct from protein-coding genes, only a very small proportion of lincRNA duplicates are likely derived from whole-genome duplication in all species examined, consistent with a previous study (Nelson et al. 2016). Either there has been extensive loss of

duplicated lincRNAs after the whole-genome duplication, or the lincRNAs derived by whole-genome duplication have diverged considerably such that they are no longer recognizable. The majority of duplicated lincRNAs belong to interspersed duplicates. This indicates that the generation of duplicates of lincRNAs is likely different from protein-coding genes. Also, lincRNAs tend to show rapid turnover rates in evolution (Kutter et al. 2012; Mohammadin et al. 2015). Many duplicated lincRNAs might have relatively recent origins prior to the most recent polyploidy event and were generated in other ways instead of WGD. Besides, many lincRNAs are known to be rapidly evolving, so the rapid divergence of duplicated lincRNAs might result in accelerated evolutionary rates for either one or both copies. Thus the number of WGD-derived lincRNA pairs detected might be underestimated in this study.

### 4.4.2 Extensive expression divergence of duplicated lincRNAs

By analyzing expression data sets from different tissue types and developmental stages, we found that duplicated lincRNAs showed lower co-expression correlation coefficient than protein-coding genes in Arabidopsis, rice and maize (Figure 4.5a). Because the expression profiles of paralogous genes are supposed to be virtually identical just after duplication, the initial co-expression correlation coefficient of the duplicates should be equal to 1 (Gu et al. 2002). Therefore our observation of the low co-expression correlation coefficient of lincRNA duplicates indicates that duplicated lincRNAs have substantially diverged in expression. The extensive divergence of duplicated lincRNAs

was further confirmed when the expression divergence was measured using expression Euclidean distance (Figure 4.5b). Interestingly, although the proportions of duplicated lincRNAs vary greatly across plants, their co-expression correlation coefficients are highly consistent among Arabidopsis, rice and maize. It is possible that duplicated lincRNAs exhibit more extensive expression divergence than protein-coding genes for the following reasons. First, one way in which rapid expression divergence between paralogs could happen is by incomplete duplication which might cause the *cis*-regulatory elements of two duplicated genes to be initially dissimilar (Ganko et al. 2007). Incomplete duplication applies to tandem duplicates and interspersed duplicates but not whole-genome duplicates. As most duplicated lincRNA pairs were likely derived from interspersed duplication, the potential for incomplete duplication might lead to more expression divergence between duplicated lincRNAs. Second, the change in expression for lowly expressed genes is thought to cause less harm than those expressed highly (Zhang and Yang 2015). The transcription level of lincRNAs is much lower than protein-coding genes. So the transcriptional change of newly duplicated lincRNAs might be more tolerated than newly duplicated protein-coding genes, thereby causing more expression divergence for duplicated lincRNAs.

Expression divergence could be asymmetric where one paralog is always more highly expressed than the other one (Casneuf et al. 2006). Otherwise, the expression pattern of paralogous genes can be complementary, where the duplicate which is more highly expressed varies by tissue type or developmental stage (Adams et al. 2003). Here

we showed that lincRNA duplicates exhibited more extensive complementary expression than protein-coding genes with several different expression indexes (Figure 4.11). Many lincRNAs are known to show high tissue specificity in expression and only expressed under certain conditions (Chekanova 2015). Thus the high expression tissue specificity of lincRNAs may contribute to the extensive complementary expression between lincRNA paralog pairs. Complementary expression can be indicative of regulatory subfunctionalization where each copy takes parts of the ancestral expression pattern (Adams et al. 2003). The higher level of complementary expression of lincRNAs suggests that subfunctionalization plays a more important role for the divergence of duplicated genes for lincRNAs than proteins. Although the relationship between expression and function of lincRNAs is yet to be clearly characterized, our results from expression analyses provide important clues to extensive functional diversification of duplicated lincRNAs.

It should be noted the expression data sets used in this study only represent a sampling of all possible organ types or developmental stages of a plant. Thus our power to compare the expression level between paralogous lincRNAs may be somewhat limited by the RNA-seq data sets currently available for plants. Analyses with more abundant expression data from more species might help gain deeper insights into the divergence of expression profile for duplicated lincRNAs in the future.

### 4.4.3 Differences in expression among different types of duplicated lincRNAs

Duplicated genes with different origins may have different evolutionary fates. To study the differences in expression among different modes of lincRNA duplicates, we compared the expression divergence for whole-genome duplicates (WGD), tandem duplicates (TD) and interspersed duplicates of lincRNAs in rice and maize. We found that tandemly duplicated lincRNAs showed less expression divergence than the other two types of duplicates (Figure 4.10, 4.12-4.15). This pattern is distinct from protein-coding genes where tandem duplicates often show larger expression divergence (Ganko et al. 2007; Liu et al. 2011). It is possible that the different patterns between lincRNAs and proteins are because many lincRNAs are involved in epigenetic regulation and interact with chromatin proteins to regulate gene expression in *cis* (Guil and Esteller 2012; Marques and Ponting 2014). A lot of lincRNAs are reported to be co-regulated with their neighboring genes (Guttman et al. 2009; Liu et al. 2012). So tandemly duplicated lincRNAs, which are physically located adjacent to each other and share a similar chromatin environment might be more likely to be under co-regulation than those generated via other mechanisms.

### 4.4.4 Expression divergence of lincRNAs is not correlated with sequence divergence

For protein-coding genes, sequence divergence is positively correlated with expression divergence in many eukaryotic species (Gu et al. 2002; Makova and Li 2003; Ganko et al. 2007). However, we found that this pattern does not appear to be true for lincRNAs, at

least in the three species we examined, as no significant correlation was detected between sequence divergence and expression divergence for lincRNA duplicates (Table 4.5; Table 4.6). This finding implies that we cannot simply infer the expression divergence based on sequence similarity of duplicated genes for lincRNAs. Although the relationship between sequence similarity and evolutionary age is not well illustrated for lincRNAs, the lack of correlation between sequence divergence and expression divergence suggests that paralogous lincRNAs might diverge rapidly in expression soon after duplication. The distinct patterns of the relationship of sequence divergence and expression divergence between lincRNAs and protein-coding genes also suggest that some evolutionary principles for proteins may not be applicable to lincRNAs. More work is necessary to dissect the detailed mechanisms of the expression divergence for plant lincRNAs.

### 4.4.5 Conclusion

In this study, we analyzed the sequence and expression evolution of duplicated lincRNAs in plants. We found that the proportion of duplicated lincRNAs in some plant species was much higher than previously reported in animals, suggesting a more important role of duplication in the evolution of plant lincRNAs. Our analysis also revealed several differences between lincRNAs and protein-coding genes in terms of gene duplication. i) LincRNAs diverge more in expression after gene duplication. ii) Tandem lincRNA duplicates showed the highest expression similarity. iii) Sequence divergence and expression divergence are not correlated between paralogous lincRNAs. These distinct

features of duplicated lincRNAs suggest different factors that might affect the evolution of lincRNAs. Also, the rapid divergence between duplicated lincRNAs might underlie some lineage-specific features in plants.

**Table 4.1** Sources of gene sequences and genomic information.

| Species | LincRNA sequence | No. of lincRNAs | Genome sequence and genome annotation |
|---|---|---|---|
| *Arabidopsis thaliana* | (Liu et al. 2012) | 5799 | TAIR10 (Lamesch et al. 2012) |
| *Populus trichocarpa* | (Shuai et al. 2014) | 3153 | *P. trichocarpa* v2.2 (Tuskan et al. 2006) |
| *Cucumis sativus* | (Hao et al. 2015) | 3298 | Chinese long v2 (Huang et al. 2009) |
| *Solanum lycopersicum* | (Zhu et al. 2015) | 3133 | SL2.50 (Fernandez-Pozo et al. 2015) |
| *Oryza sativa ssp. japonica* | (Zhang et al. 2014) | 11364 | RGAP v7 (Kawahara et al. 2013) |
| *Zea maize* | (Li et al. 2014) | 12476 | Maize genome v2 (Schnable et al. 2009) |

**Table 4.2** Sources of WGD and WTD blocks

| Species | Sources |
|---|---|
| Arabidopsis | (Bowers et al. 2003) |
| poplar | (Lee et al. 2013) |
| tomato | (Sato et al. 2012) |
| rice | (Throude et al. 2009) |
| maize | (Schnable et al. 2011) |

**Table 4.3** RNA-seq data sets

| | | |
|---|---|---|
| *Arabidopsis thaliana* | Rosette | SRR2039793-SRR2039794 |
| *Arabidopsis thaliana* | inflorescence | SRR1657473-SRR1657475 |
| *Arabidopsis thaliana* | floral bud | SRR800754-SRR800755 |
| *Arabidopsis thaliana* | anther | SRR1559345-SRR1559346 |
| *Arabidopsis thaliana* | leaf | SRR1283943-SRR1283945 |
| *Arabidopsis thaliana* | seedling | SRR1119205-SRR1119206 |
| *Arabidopsis thaliana* | endosperm | SRR1039915 |
| *Arabidopsis thaliana* | embryo | SRR1039914 |
| *Oryza sativa ssp. japonica* | seedling shoot | ERR008651, ERR008652, ERR008657, ERR008658, ERR008663, ERR008664 |
| *Oryza sativa ssp. japonica* | leaf | SRR305891-SRR305893 |
| *Oryza sativa ssp. japonica* | endosperm | SRR2338866-SRR2338867 |
| *Oryza sativa ssp. japonica* | panicle | SRR1633182-SRR1633187 |
| *Oryza sativa ssp. japonica* | root | SRR1537554-SRR1537556 |
| *Oryza sativa ssp. japonica* | callus | SRR358795- SRR358798 |
| *Oryza sativa ssp. japonica* | shoot apical meristem | DRR021355 |
| *Oryza sativa ssp. japonica* | anther | SRR1618546 |
| *Oryza sativa ssp. japonica* | pistil | SRR1618547 |
| *Oryza sativa ssp. japonica* | seed | SRR1618548 |
| *Zea mays* | anther | SRR2078791-SRR2078793 |
| *Zea mays* | pollen | SRR1028862 |
| *Zea mays* | shoot apical meristem | SRR2078797-SRR2078798 |
| *Zea mays* | endosperm_18DAP | SRR1169634-SRR1169635 |
| *Zea mays* | endosperm_24DAP | SRR1169639-SRR1169640 |
| *Zea mays* | seed_18DAP | SRR1170939-SRR1170940 |
| *Zea mays* | seed_24DAP | SRR1170944-SRR1170945 |
| *Zea mays* | embryo_18DAP | SRR531916, SRR531917, SRR531919 |
| *Zea mays* | embryo_24DAP | SRR533845, SRR533848, SRR533835 |
| *Zea mays* | ear | SRR2078794-SRR2078796 |

**Table 4.4** Numbers of different types of duplicated lincRNA pairs. Numbers shown in parentheses represent the proportion of each type of duplicates in all duplicated pairs of lincRNAs.

| Organisms | WGD | TD | Interspersed |
|---|---|---|---|
| Arabidopsis | 10 (3%) | 37 (12%) | 284 (85%) |
| poplar | 37 (9%) | 49 (11%) | 341 (80%) |
| cucumber | - | 11 (2%) | 512 (98%) |
| tomato | 6 (0.5%) | 106 (9%) | 1120 (90.5%) |
| rice | 50 (0.7%) | 306 (4%) | 6922 (95.3%) |
| maize | 113 (1%) | 789 (10%) | 6712 (89%) |

**Table 4.5** Spearman correlation coefficients (*r*) and associated *P*-values between sequence divergence and expression divergence for duplicated lincRNAs in different organisms.

| Organism | Correlation coefficient (*r*) | *P*-value |
|---|---|---|
| Arabidopsis | 0.02 | 0.81 |
| Rice | -0.01 | 0.67 |
| Maize | 0.02 | 0.23 |

**Table 4.6** Spearman correlation coefficients (*r*) and associated *P*-values between sequence divergence and expression divergence for duplicated lincRNAs without alignment coverage cutoff in different organisms.

| Organism | Correlation coefficient (*r*) | *P*-value |
|---|---|---|
| Arabidopsis | -0.02 | 0.71 |
| Rice | -0.01 | 0.50 |
| Maize | 0.01 | 0.40 |

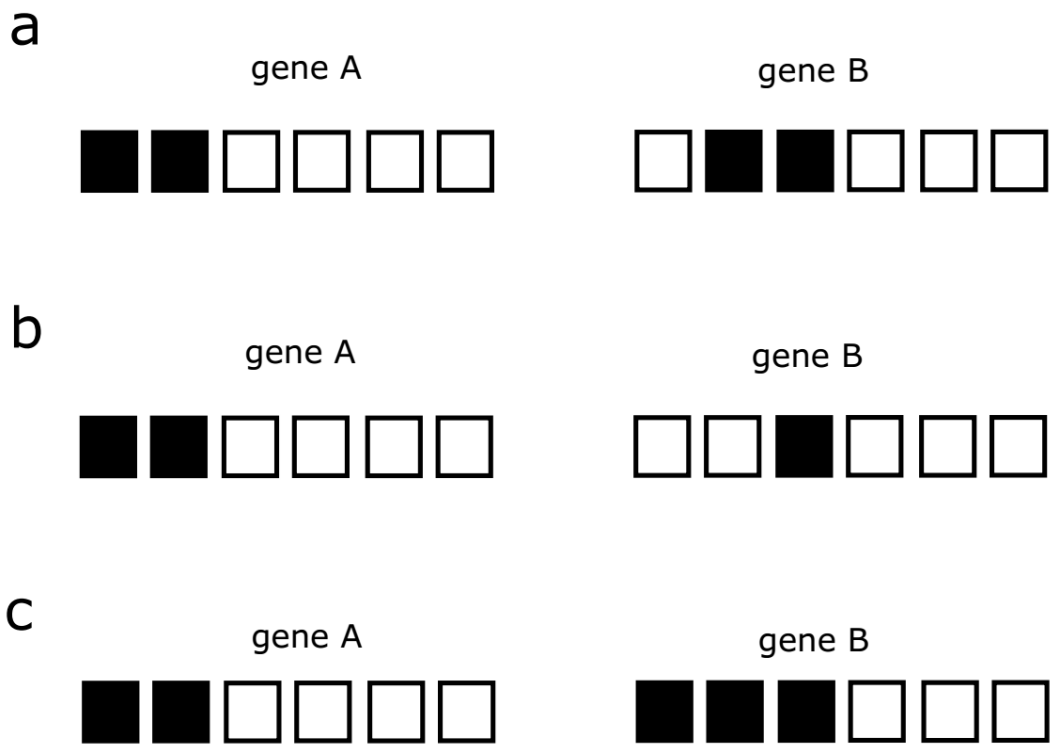


**Figure 4.1** Schematic figure of (a) reciprocal expression, (b) completely reciprocal expression and (c) non-reciprocal expression. Gene A and gene B are a pair of duplicates. Black and white squares indicate tissue types or developmental stages where a gene is and is not expressed, respectively.

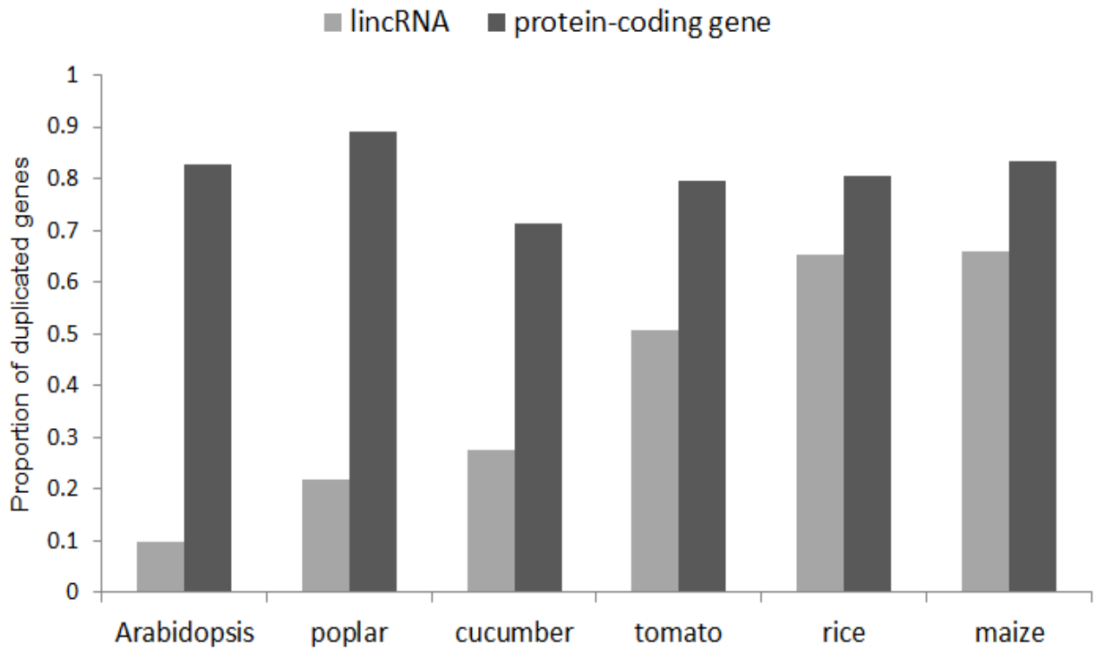


**Figure 4.2** Proportions of duplicated lincRNAs and protein-coding genes in different plant genomes.

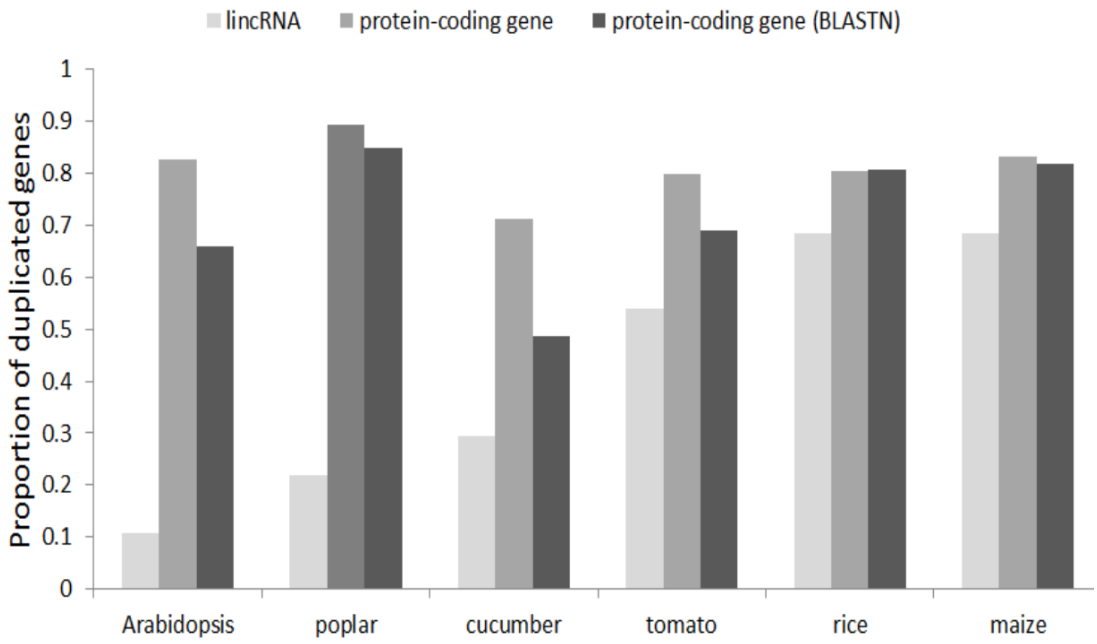


**Figure 4.3** Proportions of duplicated genes for lincRNAs and protein-coding genes. Duplicated lincRNAs identified with the BLASTN e-value cutoff of 1e-5, protein-coding genes identified using BLASTP and protein-coding genes identified with the same procedure as lincRNAs using BLASTN are shown.

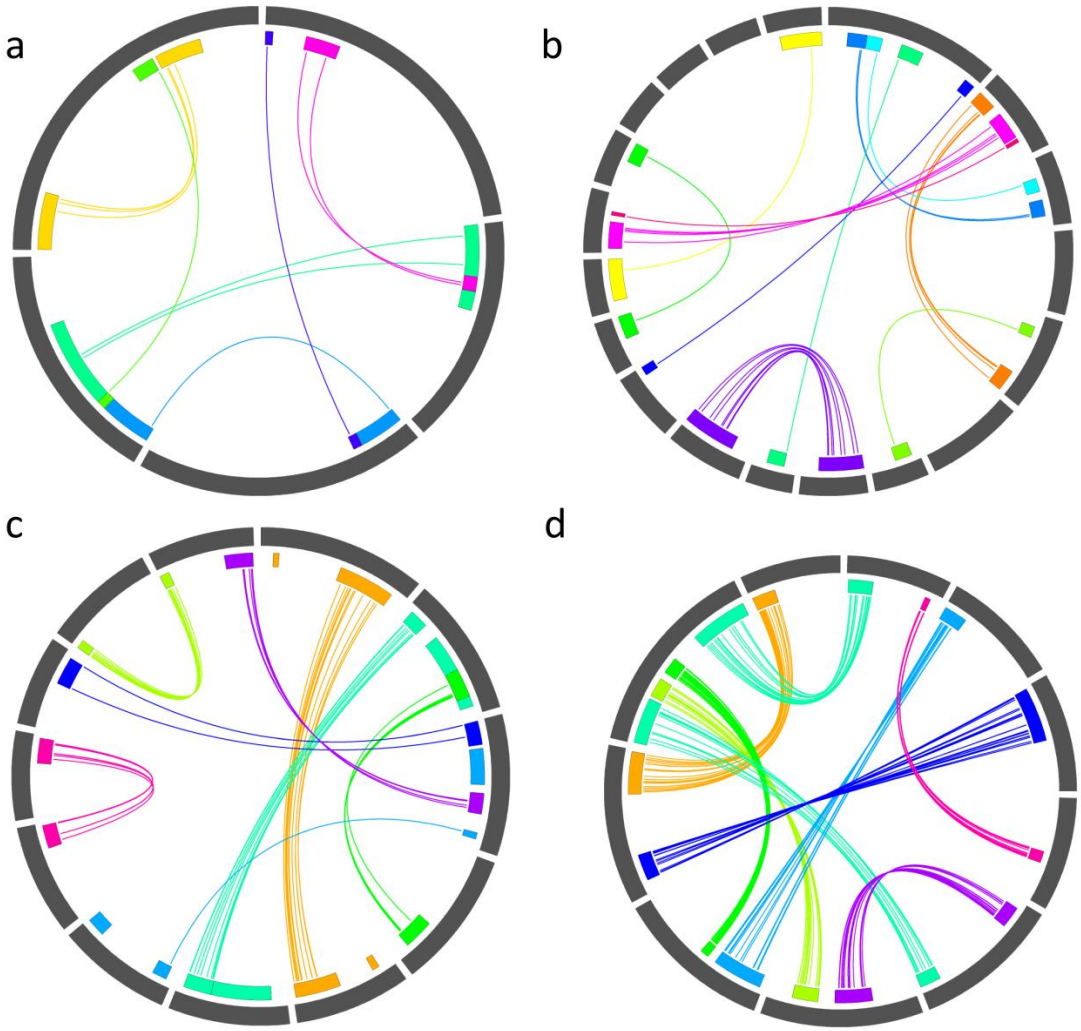


**Figure 4.4** Visualization of WGD-derived duplicated lincRNAs in (a) Arabidopsis, (b) poplar, (c) rice and (d) maize. The outer ring displays chromosomes arranged end to end. Bars in the inner ring depict WGD blocks. Links between different blocks represent WGD-derived gene pairs.

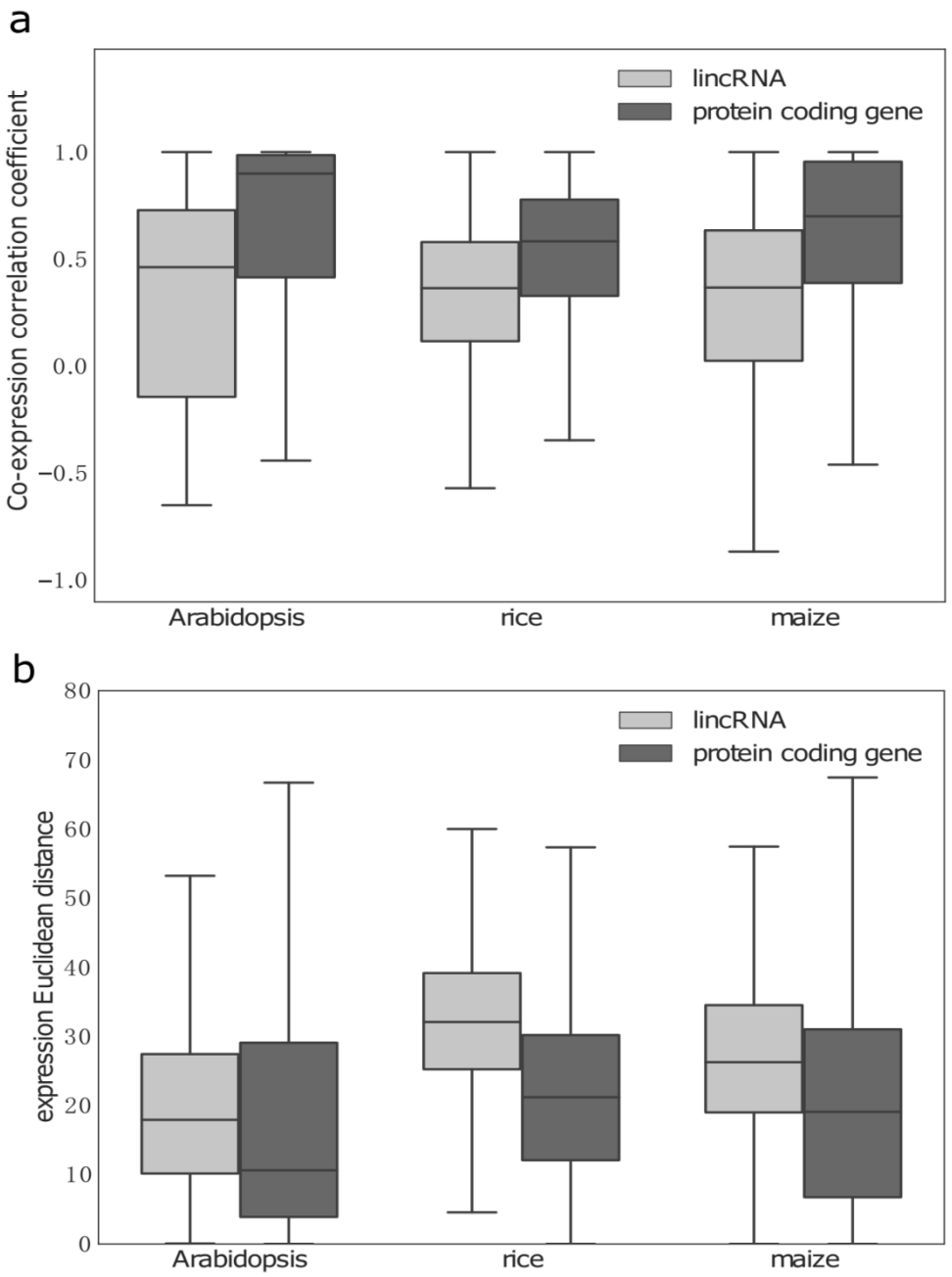


**Figure 4.5** (a) Co-expression correlation coefficient and (b) expression Euclidean distance of duplicated gene pairs for lincRNAs and protein-coding genes in Arabidopsis, rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is shown by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

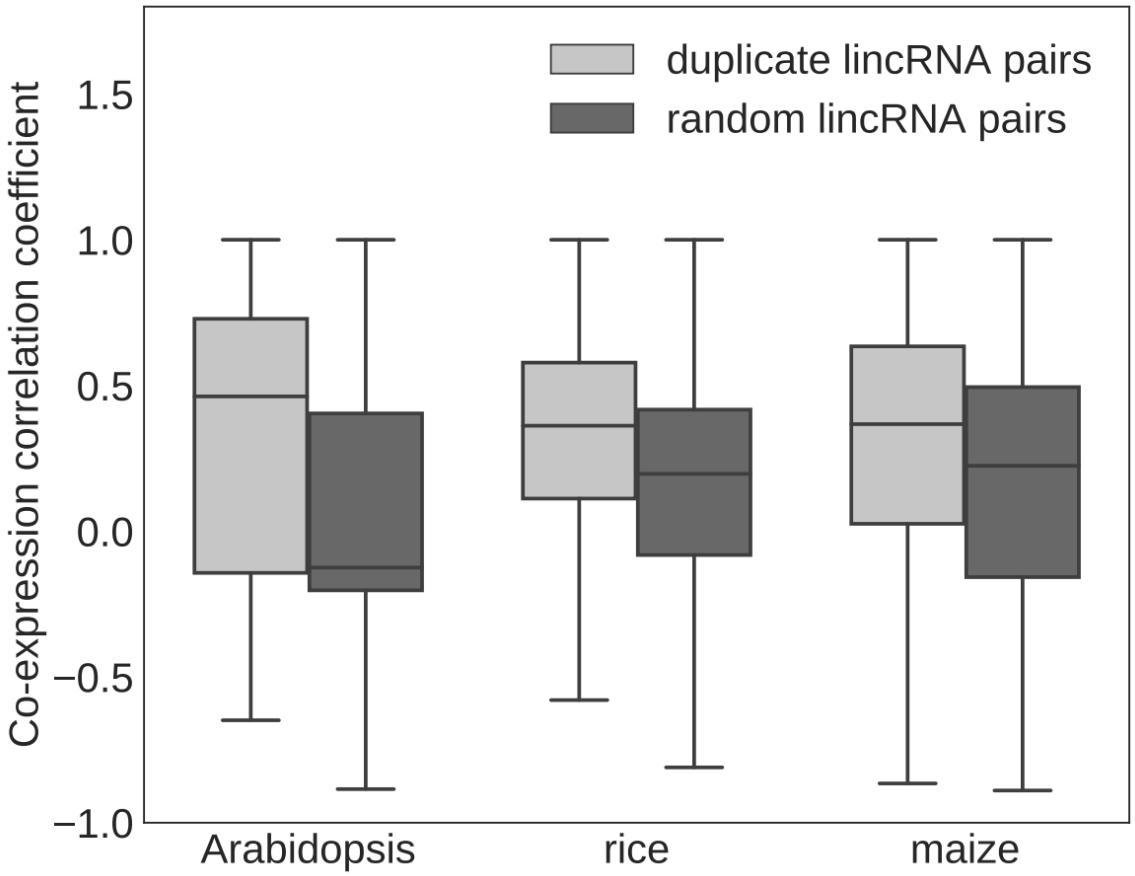


**Figure 4.6** Co-expression correlation coefficient of duplicated gene pairs for lincRNAs and 3000 randomly chosen pairs in Arabidopsis, rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is shown by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

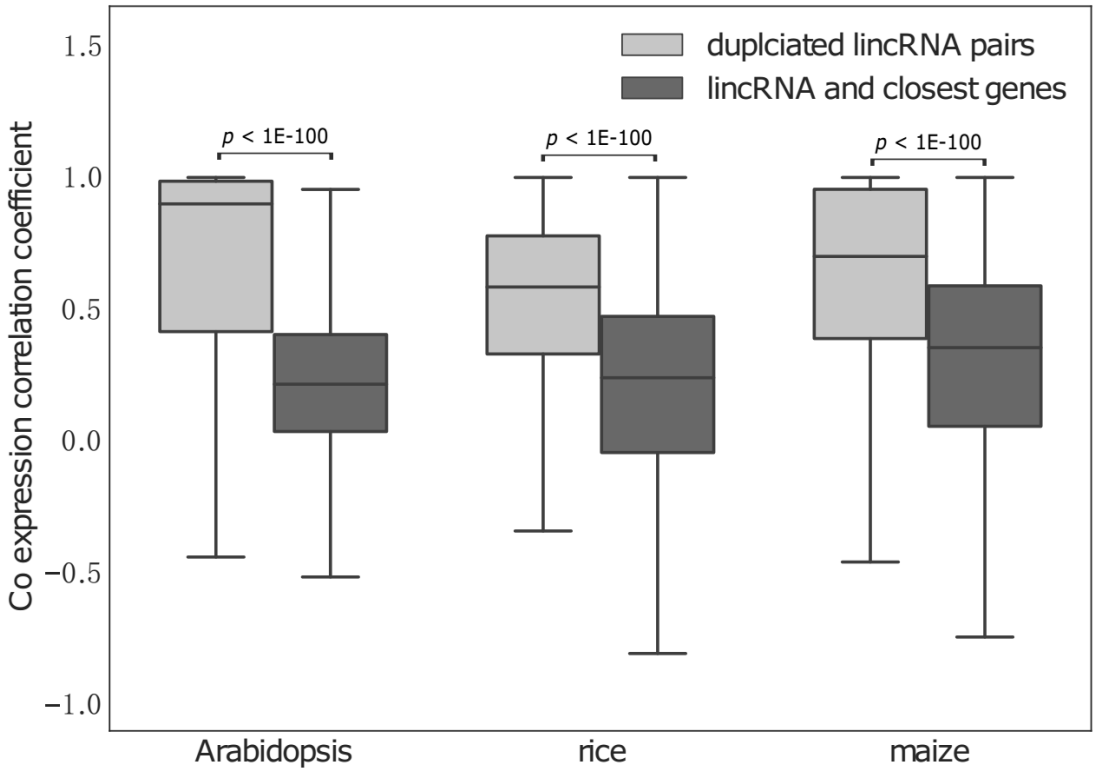


**Figure 4.7** Co-expression correlation coefficient of duplicated gene pairs for lincRNAs and lincRNA-neighboring gene pairs in Arabidopsis, rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

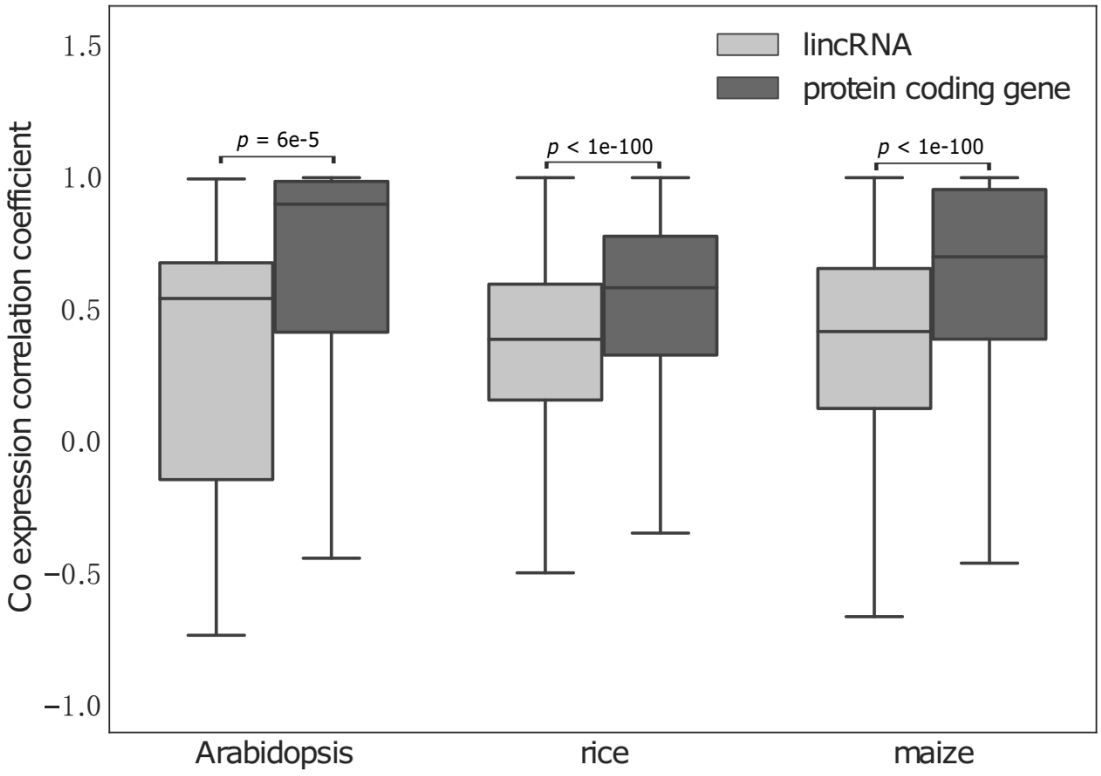


**Figure 4.8** Co-expression correlation coefficient of duplicated gene pairs for lincRNAs identified using the e-value cutoff of 1e-20 and protein-coding genes in Arabidopsis, rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

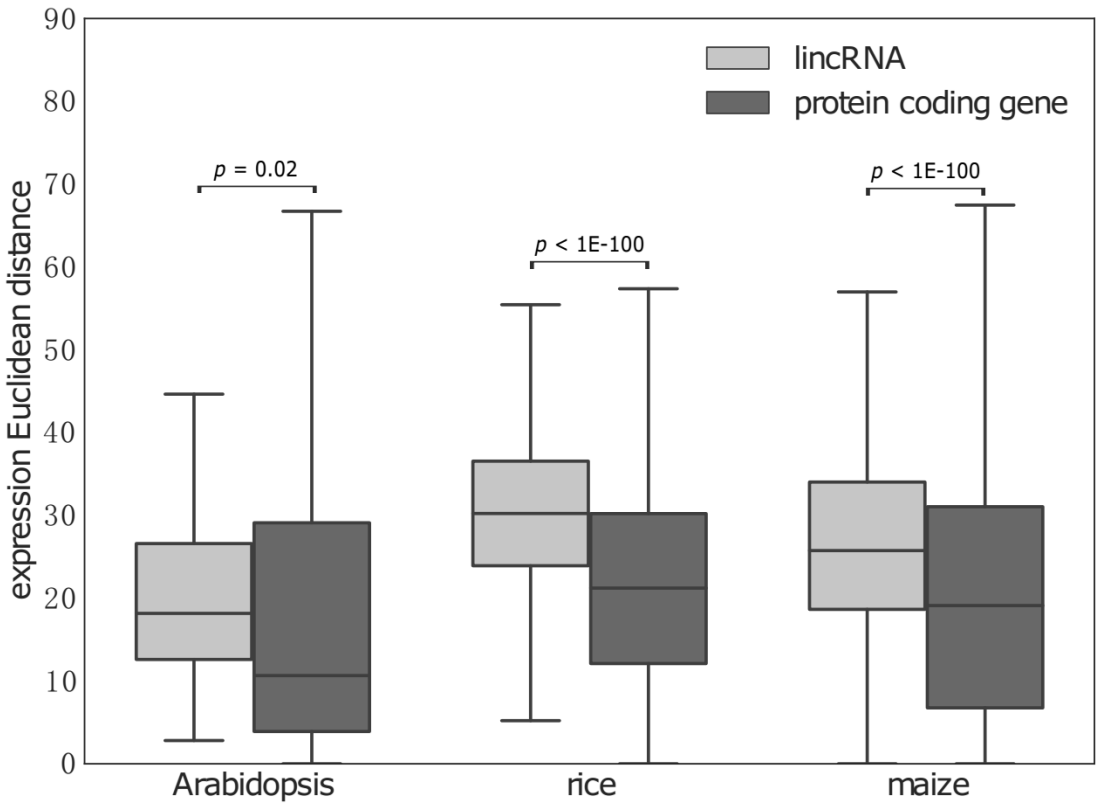


**Figure 4.9** Expression Euclidean distance of duplicated gene pairs for lincRNAs identified using the e-value cutoff of 1e-20 and protein-coding genes in Arabidopsis, rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

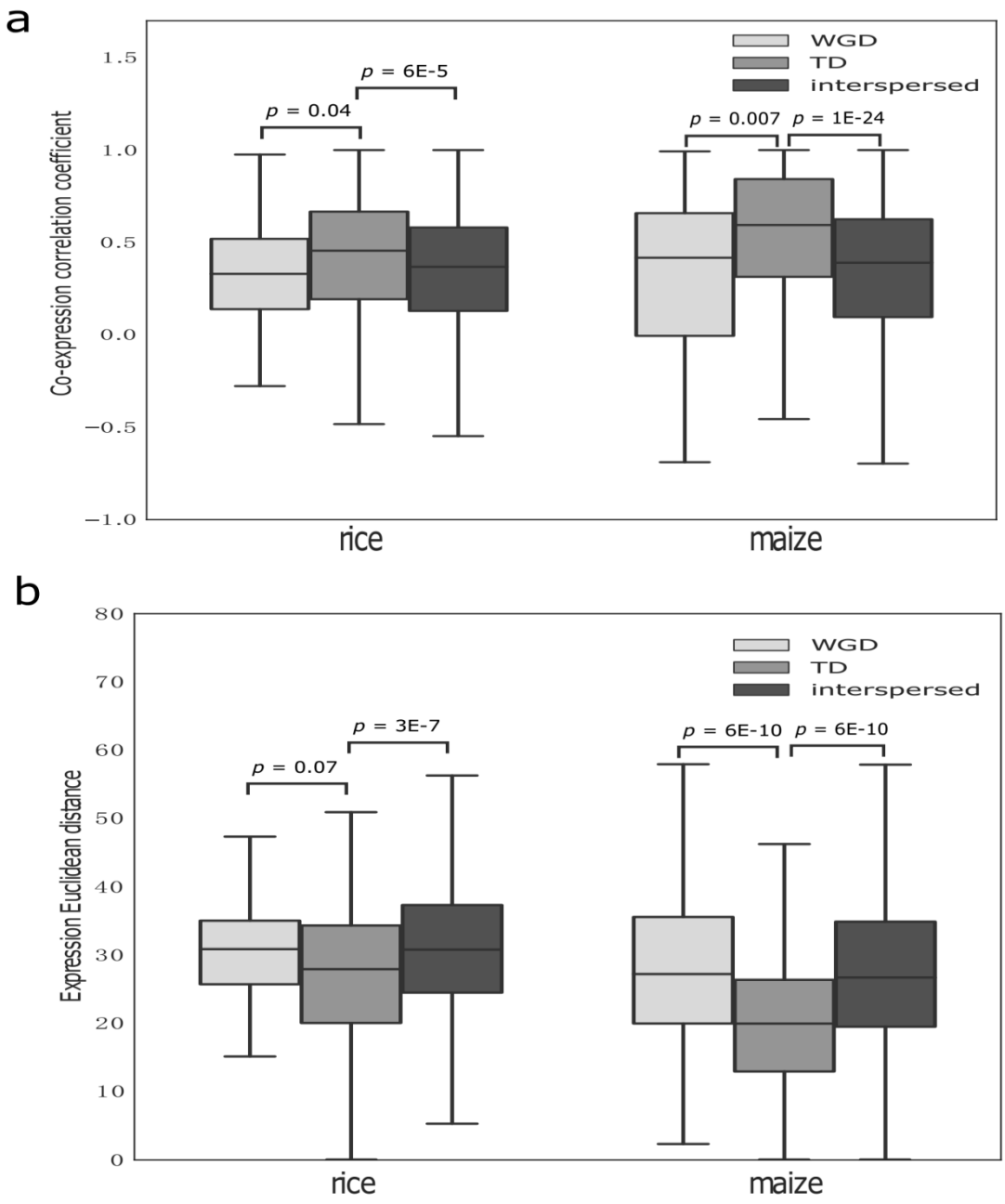


**Figure 4.10** Co-expression pattern among different types of duplicated lincRNAs. (a) Co-expression correlation coefficient and (b) Euclidean distance of duplicated genes for lincRNAs in rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

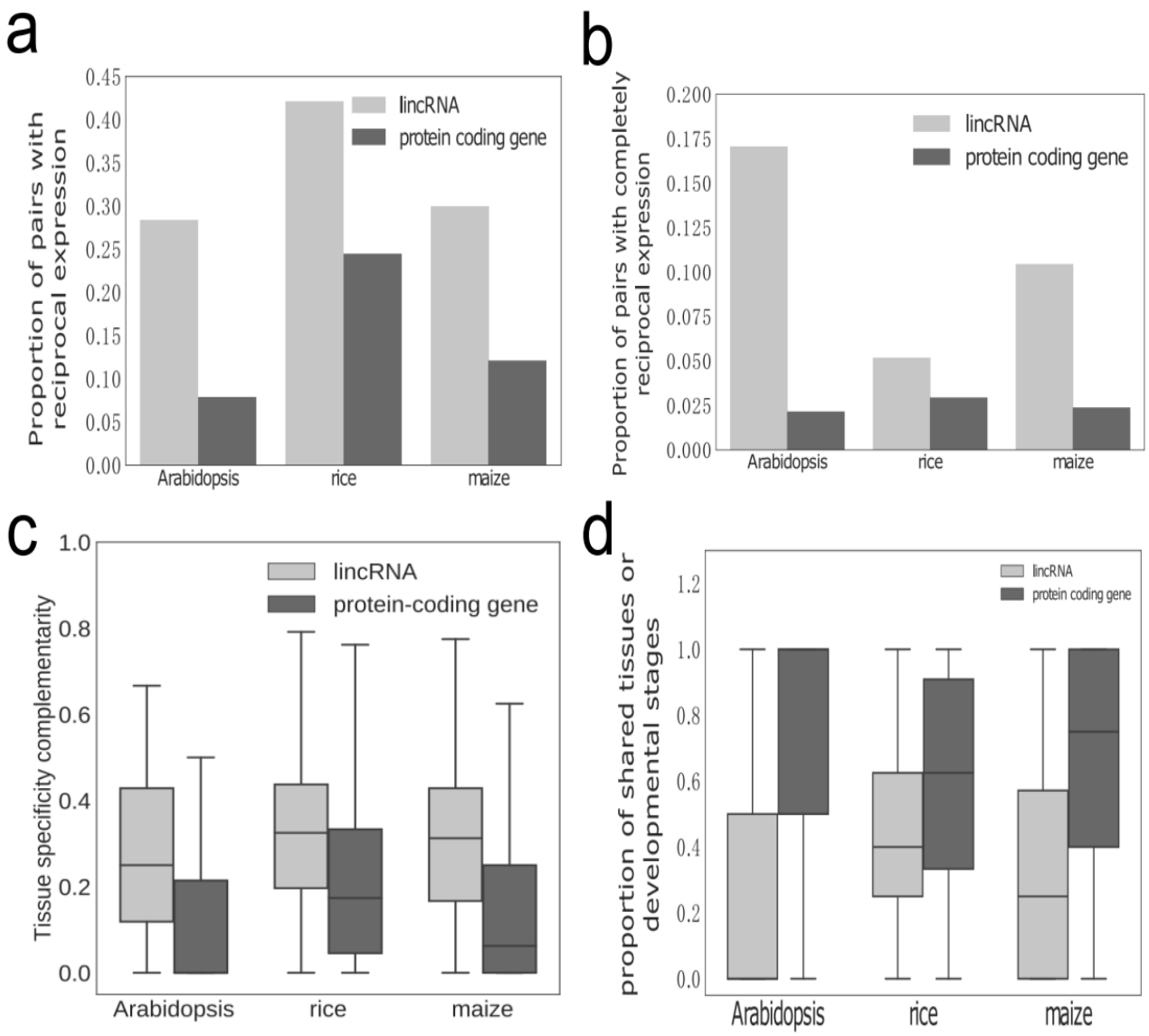


**Figure 4.11** Complementary expression of duplicated lincRNAs and protein-coding genes in Arabidopsis, rice and maize. (a) Proportion of duplicate pairs with reciprocal expression for lincRNAs and protein-coding genes. (b) Proportion of duplicate pairs with completely reciprocal expression for lincRNAs and protein-coding genes. (c) Tissue expression complementarity (TEC) for lincRNAs and protein-coding genes in different species. A higher value of TEC indicates that duplicates are more likely to show a complementary expression pattern. Boxes extend from the first quartile (Q1) to the third

quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR). (d) Proportion of tissue types or developmental stages shared by both copies in a duplicate pair for lincRNAs and protein-coding genes in different species. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

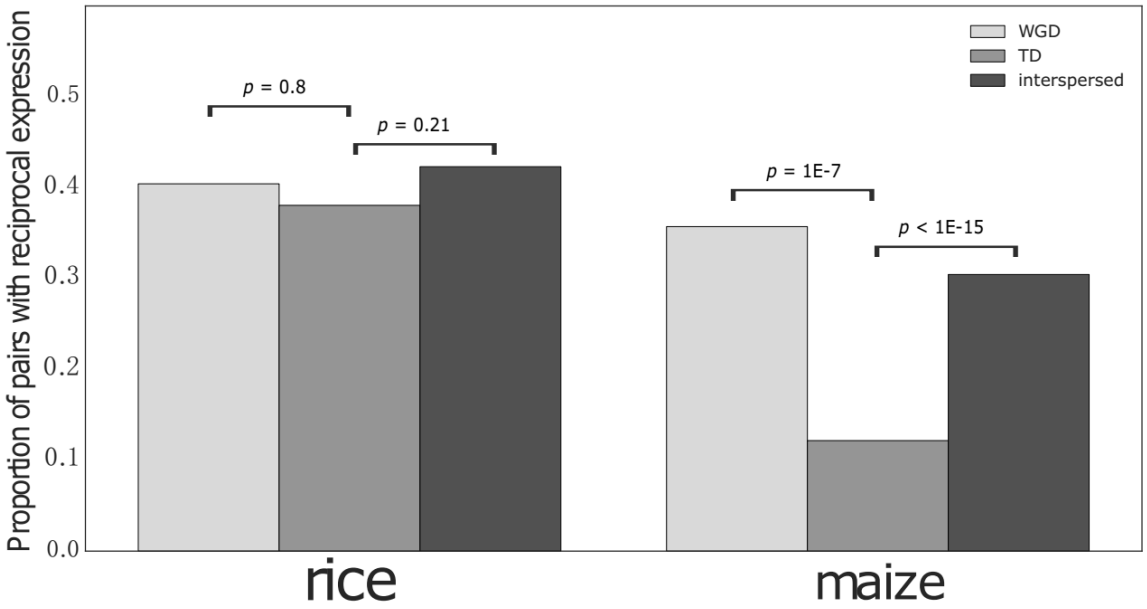


**Figure 4.12** Proportions of duplicate pairs with reciprocal expression for different types of duplicated lincRNA pairs in rice and maize.

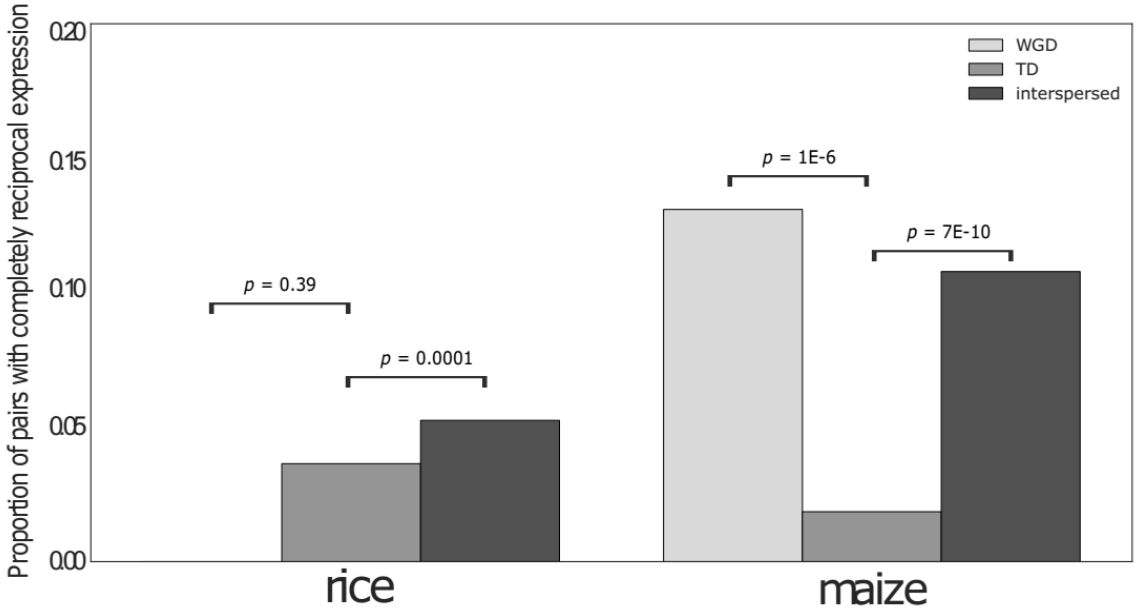


**Figure 4.13** Proportions of duplicate pairs with completely reciprocal expression for different types of duplicated lincRNA pairs in rice and maize.

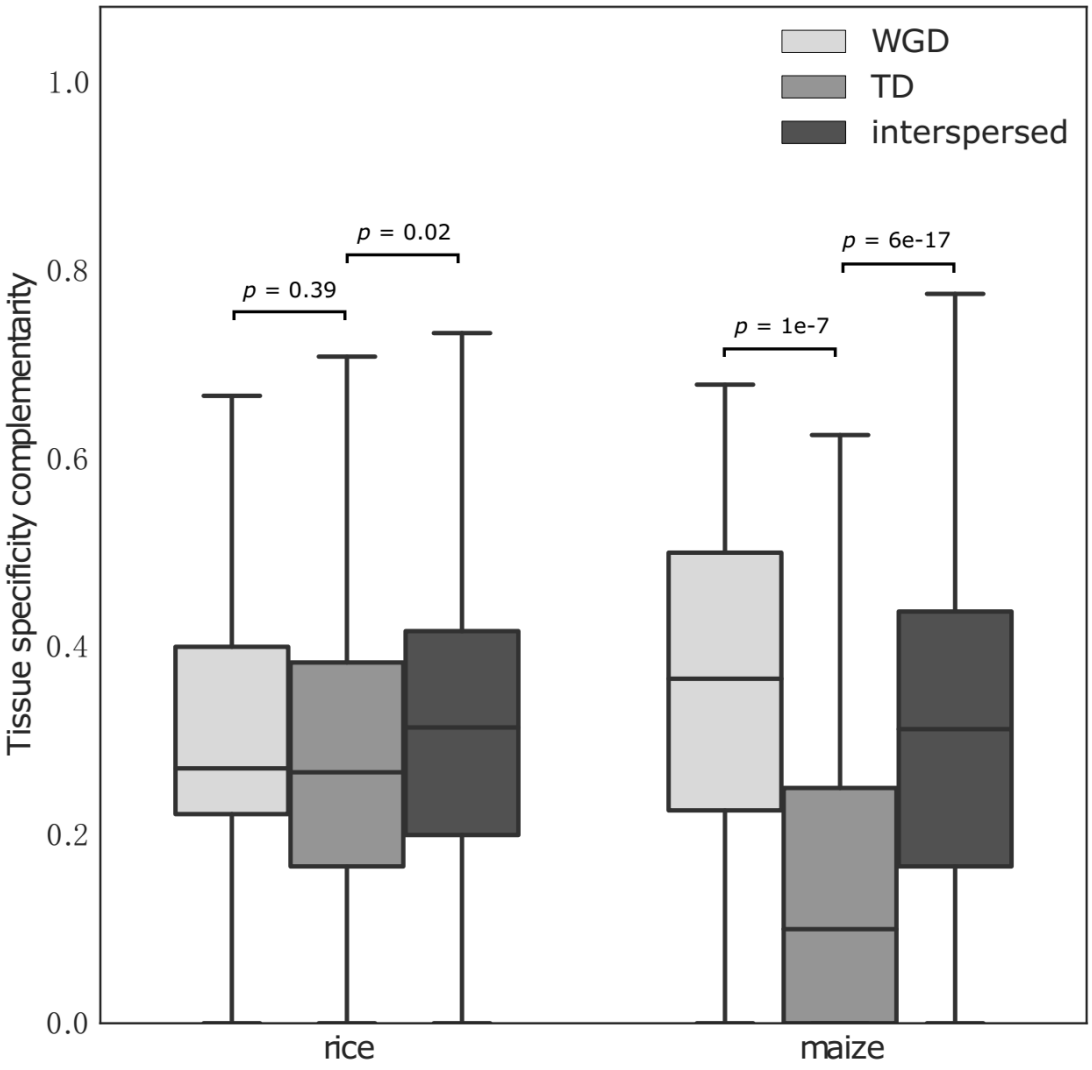


**Figure 4.14** Tissue expression complementarity (TEC) for different types of duplicated lincRNA pairs in rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

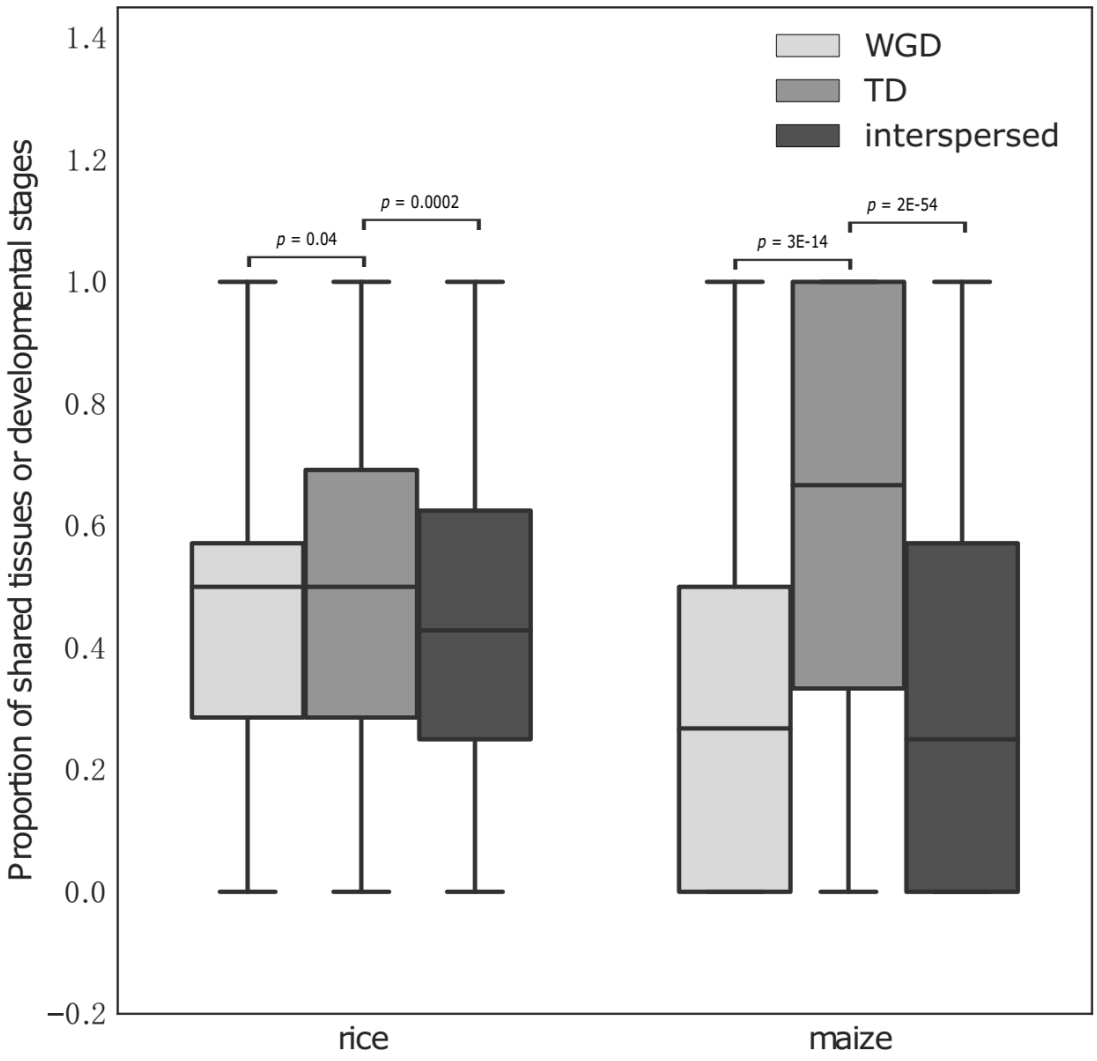


**Figure 4.15** Proportions of tissue types or developmental stages with shared expression by both copies in a duplicate pair, for different types of duplicated lincRNA pairs in rice and maize. Boxes extend from the first quartile (Q1) to the third quartile (Q3). The median is identified by a line inside the box. Whiskers extend to ± 1.5 interquartile range (IQR).

# 5 Expression Analysis of Long Intergenic Non-Coding RNAs in Polyploid and Diploid *Brassica* Species

## 5.1 Introduction

Polyploidy is common in angiosperms and an important force in driving the evolution of plant genomes (reviewed in Schranz et al. 2012; Van de Peer et al. 2009). Genomes of all angiosperms have undergone multiple rounds of polyploidy events (Jiao et al. 2011). Some polyploidy events took place relatively recently (within the last ~2 million years), while others occurred much earlier which could date back to more than 100 Mya (Bowers et al. 2003; Jiao et al. 2011). After polyploidy, genes can diverge in expression between the polyploid and parental diploid species. Many studies have shown major changes in expression patterns and biased or lack of expression of one duplicated gene following allopolyploidy, often in an organ-specific manner (reviewed in Doyle et al. 2008; Madlung and Wendel 2013; Yoo et al. 2014). However nearly all studies published to date have examined expression patterns of protein coding genes and small RNAs, but not long non-coding RNAs.

LincRNAs (long intergenic non-coding RNAs) are a type of molecule that has attracted growing interest in recent years (Ariel et al. 2015; Yamada, 2017). LincRNAs are defined as non-coding RNAs transcribed from intergenic regions with length of longer than 200 nucleotides (Chekanova, 2015; Marques and Ponting, 2014; Yang et al. 2014). In comparison to protein-coding genes, lincRNAs are often lowly expressed in a

tissue-specific and condition-specific manner. LincRNAs participate in diverse biological pathways and play many important cellular roles. For example, *IPS1* (*Induced by Phosphate Starvation 1*), plays important roles in regulating Pi homeostasis by mimicking the target of ath-miR399 under phosphorus starvation in *Arabidopsis thaliana* (Franco-Zorrilla et al. 2007). Another example is *LDMAR* (*long-day–specific male-fertility–associated RNA*) that participates in the regulation of photoperiod-sensitive male sterility in rice (Ding et al. 2012).

Since the first large-scale identification of lincRNAs in *Arabidopsis thaliana* (Liu et al. 2012), genome-wide identification of lincRNAs have been performed using transcriptome data in various plants, including rice (Zhang et al. 2014), maize (Li et al. 2014), poplar (Shuai et al. 2014), tomato (Wang et al. 2016; Zhu et al. 2015) and cotton (Wang et al. 2015). In addition to the annotation of plant lincRNAs, these studies also reveal the basic features, including the short transcript length, highly tissue-specific expression pattern and low expression level in a variety of plant species.

In this study, we performed genome-wide identification of lincRNAs in polyploid and diploid *Brassica* species, and then analyzed the expression of the identified lincRNAs after polyploidization and in response to different abiotic stresses. We analyzed the expression of lincRNAs in three resynthesized lines of polyploid *Brassica napus* and its diploid parents to identify differentially expressed genes. We also analyzed the effects of three abiotic stress conditions (heat, cold, and drought) on expression of lincRNAs. The results from the lincRNA analyses were compared with protein coding genes to

identify similarities and differences.

## 5.2 Materials and methods

### 5.2.1 RNA-seq data

The RNA-seq data of three lines of the recently resynthesized polyploid *Brassica napus* lines (P47, P48, and P52; from Gaeta et al. 2007), along with the diploid parental species *Brassica rapa* and *Brassica oleracea*, were generated by Tack and Adams (unpublished data). Briefly, plants were grown together in growth chambers and three biological replicates of the first true leaves were harvested. RNA was extracted and then sequenced by Illumina HiSeq, with 100 bp paired end reads, at Genome Quebec. The RNA-seq data of the natural *Brassica napus* (cultivar Sentry summer rape) under abiotic stress treatments were generated by Lee and Adams (unpublished data). Three week old seedlings were subjected to the following treatments: Heat treatment was 35°C for 24 hours, cold treatment was 4°C for 24 hours, and drought treatment was 22°C with 25% PEG-6000 for 24 hours. Three biological replicates were collected for each treatment and for untreated plants. RNA was extracted and then sequenced by Illumina HiSeq, with 100 bp paired end reads, at Genome Quebec.

### 5.2.2 Genome-wide identification of lincRNAs

Reads were mapped using GSNAP (Wu and Nacu, 2010) after the removal of adaptor sequences. Cufflinks v2.1.1 (Trapnell et al. 2014) was used to assemble the transcripts.

Transcripts from different samples were combined using StringTie v1.2.2 (Pertea et al. 2015). Transcripts derived from protein-coding genes or regions within the 500 bp upstream of protein-coding genes were removed using BEDTools v2.25.0 (Quinlan and Hall, 2010). Based on the definition of lincRNAs, transcripts with length shorter than or equal to 200 bp were discarded. BLASTX (Altschul et al. 1997) was run to eliminate transcripts with sequence similarity to protein-coding genes. Transcripts with e-value higher than or equal to 1e-3 were removed. Sequence analyses and format conversion were performed using BEDOPS v2.4.14 (Neph et al. 2012), Sambamba v0.5.9 (Tarasov et al. 2015), BamTools v2.3.0 (Barnett et al. 2011) and in-house scripts written in Python (Cock et al. 2009), Ruby (Goto et al. 2010) and R.

### 5.2.3 Identification of lincRNA transcripts with sequence similarity to housekeeping RNAs, small RNAs and transposons

LincRNAs that are potentially derived from housekeeping RNAs were identified using infernal v1.1 (Nawrocki and Eddy, 2013) with the cutoff value of 1e-5. Small RNAs and transposons in lincRNAs were detected using RepeatMasker (http://www.repeatmasker.org, last accessed September, 2014) with default settings.

### 5.2.4 Identification of differentially expressed genes

The numbers of reads mapped to each position were counted using featureCounts implemented in Subread v1.4.6 (Liao et al. 2013). Differentially expressed genes were

identified with DESeq (Anders and Huber, 2010) and edgeR (Robinson et al. 2010). LincRNAs or protein-coding genes with *P*-value < 0.05 and fold change > 2.0 were defined as differentially expressed genes. Only genes with at least 10 reads mapped in each replicate were used in the analysis of differential expression.

## 5.3 Results

### 5.3.1 Genome-wide identification of lincRNAs in multiple *Brassica* species

To facilitate the identification of lincRNAs in *Brassica*, we developed a user-friendly pipeline linc_finder. Linc_finder performs automatic identification of lincRNAs form a large number of transcripts with a series of filtering including protein-coding sequences and transcripts derived from UTRs (Figure 5.1). We applied linc_finder to identify lincRNAs transcribed in the leaf tissue of several *Brassica* species (Materials and methods).

The RNA-seq data were generated prior to this study. The data sets include the transcriptome data of three lines of the recently resynthesized (synthetic) *Brassica napus* and their diploid parental species (*Brassica rapa* and *Brassica oleracea*), referred to as the synthetic *Brassica* data set, and also the RNA-seq data of four different conditions of the natural *Brassica napus*, referred to as the natural *Brassica napus* data set. There were three biological replicates for the RNA-seq data of both of the synthetic *Brassica* data set and the natural *Brassica napus* data set. The transcriptome was assembled for each biological replicate. Transcripts from all transcriptomes of synthetic *Brassica* species and

the natural *Brassica napus* were merged respectively, which were then used in the identification of lincRNAs (see Materials and methods).

By using linc_finder, we identified a total of 1,184 lincRNA loci in the synthetic *Brassica* species and 2,534 lincRNA loci in the natural *Brassica napus* (in four different conditions). In the lincRNAs of the synthetic *Brassica* species, 492 and 692 transcripts were derived from the *Brassica rapa* and *Brassica oleracea* subgenome, respectively. In the natural *Brassica napus* lincRNA sequence set, 962 and 1572 lincRNA loci were transcribed in the *Brassica rapa* subgenome and *Brassica oleracea* subgenome, respectively. For both the synthetic Brassica and the natural *Brassica napus* lincRNA sets, the subgenome of *Brassica oleracea* appeared to have more lincRNAs (*P*-value < 1e-8, binomial test), suggesting that lincRNAs were more likely to be derived from the *B. oleracea* subgenome than the *B. rapa* subgenome.

### 5.3.2 Basic characteristics of *Brassica* lincRNAs

To better understand the features of lincRNAs in *Brassica*, we analyzed the sequence length, the GC content, the intron number, the expression level and the expression specificity of lincRNAs and compared them with protein-coding genes. We found that lincRNAs appeared to have shorter sequence length and lower GC content (Figure 5.2; Figure 5.3). LincRNAs also contain fewer introns with 57% having a single exon (Figure 5.2b; Figure 5.3b). Furthermore, we found that lincRNAs were expressed at lower level than mRNAs (Figure 5.42; Figure 5.3d). These results were consistent with previous

findings regarding the basic characteristics of lincRNAs in plants, suggesting a high-quality set of lincRNAs that we generated.

### 5.3.3 Divergence of lincRNA expression after polyploidization

To assess the expression divergence of lincRNAs between the polyploid *Brassica napus* and its parental species from a quantitative perspective, we analyzed the differential expression pattern of lincRNAs and compared the pattern with protein-coding genes. Only genes with at least 10 reads mapped in all three replicates of a genotype were included in the differential expression analysis. Differentially expressed genes were identified using DESeq by comparing the expression level between the parental species (*Brassica rapa* and *Brassica oleracea*) and the synthetic *Brassica napus* (see Materials and methods). Differentially expressed genes were defined as genes with at least a two-fold change in the expression level and a *P*-value (FDR adjusted) lower than 0.05 (see Materials and methods).

The proportion of differentially expressed lincRNAs between the synthetic *Brassica napus* and the diploid parental species *Brassica rapa* ranged from 19.1% in *Brassica napus* P47 to 27.4% in *Brassica napus* P52, with an average of 23% of all of the three lines (Table 5.1). For protein-coding genes, on average 17.8% of genes were differentially expressed between *Brassica rapa* and the synthetic *Brassica napus* (Table 5.1). *Brassica napus* P52 showed the highest proportion of differentially expressed protein-coding genes, and those from *Brassica napus* P48 were the least likely to show

differential expression with 15.4% protein-coding genes differentially expressed (Table 5.1). These results indicated that lincRNAs were more likely to be differentially expressed than protein-coding genes between *Brassica rapa* and the synthetic polyploids. Also, there was more variation in regards to the proportion of differentially expressed lincRNAs than protein-coding genes derived from the *Brassica rapa* subgenome after polyploidization (Table 5.1).

Among those genes that were differentially expressed, 14% of the lincRNAs were more highly expressed in the parental species *Brassica rapa*, compared with 6% of lincRNAs that were more highly expressed in the synthetic *Brassica napus* (Table 5.1). This result suggested that the expression of lincRNAs derived from the *B. rapa* subgenome was more likely to be down-regulated after polyploidization (*P*-value < 1e-5, binomial test). 13% of lincRNAs were differentially expressed between *Brassica oleracea* and the synthetic *Brassica napus* (Table 5.1), much lower than that of lincRNAs from the *B. rapa* subgenome (*P*-value < 1e-10, chi-squared test), indicating a lower expression divergence of lincRNAs derived from the *B. oleracea* subgenome between the synthetic polyploid and the parental species.

We repeated all the above analyses using edgeR. The same criteria (see Materials and methods) were used to identify differentially expressed genes. The general trends remained the same (Table 5.2). The consistent patterns obtained by using edgeR further confirmed the above results.

### 5.3.4 Divergence of lincRNA expression in the natural *Brassica napus* after different stress treatments

We investigated the expression divergence of lincRNAs in the natural *Brassica napus* among different abiotic stress conditions: drought, cold and heat. The fractions of differentially expressed lincRNAs varied greatly among different stress conditions. 11%, 35% and 22% of lincRNAs were differentially expressed in the drought, the cold, and the heat treatments, respectively. More differentially expressed lincRNAs were down-regulated in the cold and drought treatments whereas the opposite pattern was observed in the heat treatment (Table 5.3). In addition, we compared differentially expressed genes between lincRNAs and protein-coding genes. We found that lincRNAs were more likely to be differentially expressed in all of the three stress conditions than were protein coding genes (Table 5.3). We repeated all the above analyses by using edgeR in the identification of differentially expressed genes and similar results were observed (Table 5.4).

We also analyzed the contribution of different subgenomes to the differential expression of lincRNAs in the natural *Brassica napus*. The subgenome of *Brassica oleracea* was found to have more differentially expressed lincRNAs than the subgenome of *Brassica rapa* (Table 5.5). We also calculated the relative proportion of differentially expressed lincRNAs for each subgenome to examine whether lincRNAs from one subgenome were more likely to be differentially expressed (Table 5.5). No significant difference was found in regards to the relative proportion of differentially expressed

lincRNAs between the two subgenomes. We repeated the analysis using edgeR which rendered similar results (Table 5.6).

## 5.4 Discussion

### 5.4.1 Identification of lincRNA transcripts from transcriptome data

There are several bioinformatic tools that have been designed for the identification of lincRNAs. However, specific caution needs to be paid when existing software are directly used to identify lincRNAs for the following reasons. First, lincRNA research is a novel area and the definition of lincRNAs is still evolving. Some 'lincRNAs' identified in the very early studies may not meet the current requirements for lincRNAs. For example, a careful analysis of the lincRNA data sets identified by (Liu et al. 2012) revealed that 10% of the identified lincRNAs are likely derived from protein-coding genes. Second, different categories of long non-coding RNAs (e.g. lincRNAs, intronic lncRNAs, and anti-sense lncRNAs) may be mixed all together in some software (Li et al. 2014). A careful examination of the results may be necessary to distinguish between different categories of long non-coding RNAs. Last, software originally designed for animal species might not be suitable for plants.

To generate a set of lincRNA sequences of high quality for *Brassica*, I developed a computational pipeline linc_finder. linc_finder integrates the results of multiple software, and generates a set of high-quality lincRNAs after a series of rigorous filtering steps. Users can specify different arguments in the identification of lincRNAs including the

minimum distance between a lincRNA and a protein-coding gene, the e-value cut off in BLASTX searches, and the stringency level of lincRNAs.

### 5.4.2 Expression divergence of lincRNAs between the synthetic polyploid *Brassica napus* lines and their diploid parents

By analyzing the lincRNA data sets in the three resynthesized *Brassica napus* lines and their parental species *Brassica rapa* and *Brassica oleracea*, we show considerable expression divergence of lincRNAs after polyploidization. A considerable proportion of protein-coding genes have been shown previously to diverge in expression after polyploidization, compared with the parental diploid species, in various resynthesized and natural polyploid species. By comparing between lincRNAs and protein-coding genes, we found that lincRNAs showed greater divergence in expression than protein-coding genes.

One of the most important functions of lincRNAs is in the regulation of the expression of genes in either a *cis*- or *trans*- manner (e.g., Ariel et al. 2015). Hence, the rapid change in the expression pattern of lincRNAs after polyploidization might affect the expression of many other genes. This highlights potential roles of lincRNAs in rewiring gene expression networks in the allopolyploid. What factors might cause expression changes in the polyploids compared with their parents? Divergent interactions between regulatory factors derived from one parental subgenome in the allopolyploids with regulatory elements from the other parental subgenome may alter expression levels of

lincRNAs. Other factors involved could be epigenetic. DNA methylation and histone modifications can change after polyploidy (e.g., Chen, 2007; Wang et al. 2013) and those epigenetic marks are known to have effects on gene expression.

In addition, we found that lincRNAs transcribed from the subgenome of *B. rapa* were more likely to show differential expression than those transcribed from the *B. oleracea* subgenome. The same trend was also found for protein coding genes. This suggests different impacts of polyploidization on the genes transcribed from different subgenomes and that the evolutionary fates of both lincRNAs and protein-coding genes can be affected by the genomic environment, e.g. the subgenome from which they are derived. This might reflect the fractionation after allopolyploidization for lincRNAs as proposed in a recent study in cotton (Wang et al. 2015).

### 5.4.3 LincRNA expression in response to stress in the natural *Brassica napus*

Previous studies of diploid plants have shown differential expression of lincRNAs in response to stress conditions, suggesting that some lncRNAs play important roles in stress-response pathways (reviewed in Zhang and Chen, 2013). For example, Di et al. (2014) characterized 303 lncRNAs differentially expressed under stress conditions. in *Arabidopsis thaliana*. In an earlier study, Xin et al. (2011) identified 125 lncRNAs that were expressed under heat stress and powdery mildew infection, among which four lncRNAs were potential precursors of miRNA.

By analyzing deep RNA-seq transcriptome data from polyploid *Brassica napus*

subjected to three different abiotic stress treatments, we characterized lincRNAs specifically expressed in response to stresses. We found that lincRNAs exhibited greater divergence in expression across different stresses than protein-coding genes. There were higher differences in drought and heat stressed plants than in cold stressed plants. Some of the stress-responsive lincRNAs might be involved in the regulation of responses to different abiotic stresses. These genes provide good candidates for future functional studies to further elucidate the gene expression landscape in response to abiotic stress.

**Table 5.1** Proportions and numbers of differentially expressed lincRNAs and protein-coding genes in the three lines of the synthetic *Brassica napus* vs. their parents identified by DESeq.

| | **LincRNA** | | | | **protein-coding gene** | | | |
|---|---|---|---|---|---|---|---|---|
| | No. of genes | Up | Down | DEGs | Up | Down | DEGs | Significance (lincRNA DEG vs. protein DEG) |
| BR vs. P47 | 435 | 0.039 | 0.152 | 0.191 | 0.085 | 0.088 | 0.173 | NS |
| BR vs. P48 | 466 | 0.077 | 0.15 | 0.227 | 0.080 | 0.074 | 0.154 | *** |
| BR vs. P52 | 445 | 0.058 | 0.216 | 0.274 | 0.11 | 0.097 | 0.206 | *** |
| BO vs. P47 | 426 | 0.068 | 0.054 | 0.122 | 0.040 | 0.040 | 0.081 | ** |
| BO vs. P48 | 449 | 0.094 | 0.038 | 0.132 | 0.045 | 0.028 | 0.073 | *** |
| BO vs. P52 | 444 | 0.04 | 0.04 | 0.08 | 0.063 | 0.041 | 0.104 | NS |

BR: *B. rapa*, BO: *B. oleracea*, P51: *B. napus P51*, P48: *B. napus P48*, P52: *B. napus P52*.

DEG: differentially expressed gene

Up: up-regulated gene

Down: down-regulated gene

**Table 5.2** Proportions and numbers of differentially expressed lincRNAs and protein-coding genes in the three lines of the synthetic *Brassica napus* vs. their parents identified by Edger.

| | **LincRNA** | | | | **protein-coding gene** | | | |
|---|---|---|---|---|---|---|---|---|
| | No. of genes | Up | Down | DEGs | Up | Down | DEGs | Significance (lincRNA DEG vs. protein DEG) |
| BR vs. P47 | 435 | 0.06 | 0.198 | 0.258 | 0.10 | 0.097 | 0.201 | ** |
| BR vs. P48 | 466 | 0.086 | 0.202 | 0.288 | 0.094 | 0.088 | 0.182 | *** |
| BR vs. P52 | 445 | 0.07 | 0.24 | 0.31 | 0.12 | 0.11 | 0.227 | *** |
| BO vs. P47 | 426 | 0.092 | 0.077 | 0.169 | 0.059 | 0.055 | 0.114 | *** |
| BO vs. P48 | 449 | 0.116 | 0.058 | 0.174 | 0.056 | 0.042 | 0.098 | *** |
| BO vs. P52 | 444 | 0.095 | 0.081 | 0.176 | 0.074 | 0.059 | 0.132 | ** |

BR: *B. rapa*, BO: *B. oleracea*, P51: *B. napus P51*, P48: *B. napus P48*, P52: *B. napus P52*.

DEG: differentially expressed gene

Up: up-regulated gene

Down: down-regulated gene

**Table 5.3** Proportions and numbers of differentially expressed lincRNAs and protein-coding genes in the natural *Brassica napus* across stress conditions identified by DESeq.

| | **LincRNA** | | | | **protein-coding gene** | | | |
|---|---|---|---|---|---|---|---|---|
| | No. of genes | Up | Down | DEGs | Up | Down | DEGs | Significance (lincRNA DEG vs. protein DEG) |
| drought | 1752 | 0.042 | 0.066 | 0.108 | 0.020 | 0.045 | 0.065 | *** |
| cold | 1514 | 0.131 | 0.221 | 0.352 | 0.16 | 0.16 | 0.32 | ** |
| heat | 1788 | 0.124 | 0.091 | 0.215 | 0.059 | 0.081 | 0.14 | *** |

DEG: differentially expressed gene

Up: up-regulated gene

Down: down-regulated gene

**Table 5.4** Proportions and numbers of differentially expressed lincRNAs and protein-coding genes in the natural *Brassica napus* across stress conditions identified by Edger.

| | **LincRNA** | | | | **protein-coding gene** | | | |
|---|---|---|---|---|---|---|---|---|
| | No. of genes | Up | Down | DEGs | Up | Down | DEGs | Significance (lincRNA DEG vs. protein DEG) |
| drought | 1752 | 0.046 | 0.067 | 0.113 | 0.019 | 0.045 | 0.065 | *** |
| cold | 1514 | 0.128 | 0.227 | 0.355 | 0.15 | 0.17 | 0.32 | ** |
| heat | 1788 | 0.125 | 0.092 | 0.217 | 0.060 | 0.080 | 0.14 | *** |

DEG: differentially expressed gene

Up: up-regulated gene

Down: down-regulated gene

**Table 5.5** Distribution of differentially expressed lincRNAs identified by DESeq between different stresses in the *B. rapa* subgenome (BR) and *B. oleracea* subgenome (BO) of polyploid *B. napus*.

| | **Total** | **BR** | **BO** | **BR_DEG** | **BO_DEG** | **BR DEG relative proportion** | **BO DEG relative proportion** |
|---|---|---|---|---|---|---|---|
| **drought** | 1752 | 640 | 1112 | 69 | 119 | 0.107 | 0.107 |
| **cold** | 1514 | 562 | 952 | 177 | 356 | 0.314 | 0.37 |
| **heat** | 1788 | 674 | 1114 | 149 | 235 | 0.221 | 0.211 |

DEG: differentially expressed gene

**Table 5.6** Distribution of differentially expressed lincRNAs identified by Edger between different stresses in the *B. rapa* subgenome (BR) and *B. oleracea* subgenome (BO)

| | Total | BR | BO | BR_DEG | BO_DEG | BR DEG relative proportion | BO DEG relative proportion |
|---|---|---|---|---|---|---|---|
| **drought** | 1752 | 640 | 1112 | 73 | 125 | 0.114 | 0.112 |
| **cold** | 1514 | 562 | 952 | 181 | 357 | 0.322 | 0.375 |
| **heat** | 1788 | 674 | 1114 | 152 | 235 | 0.226 | 0.211 |

DEG: differentially expressed gene

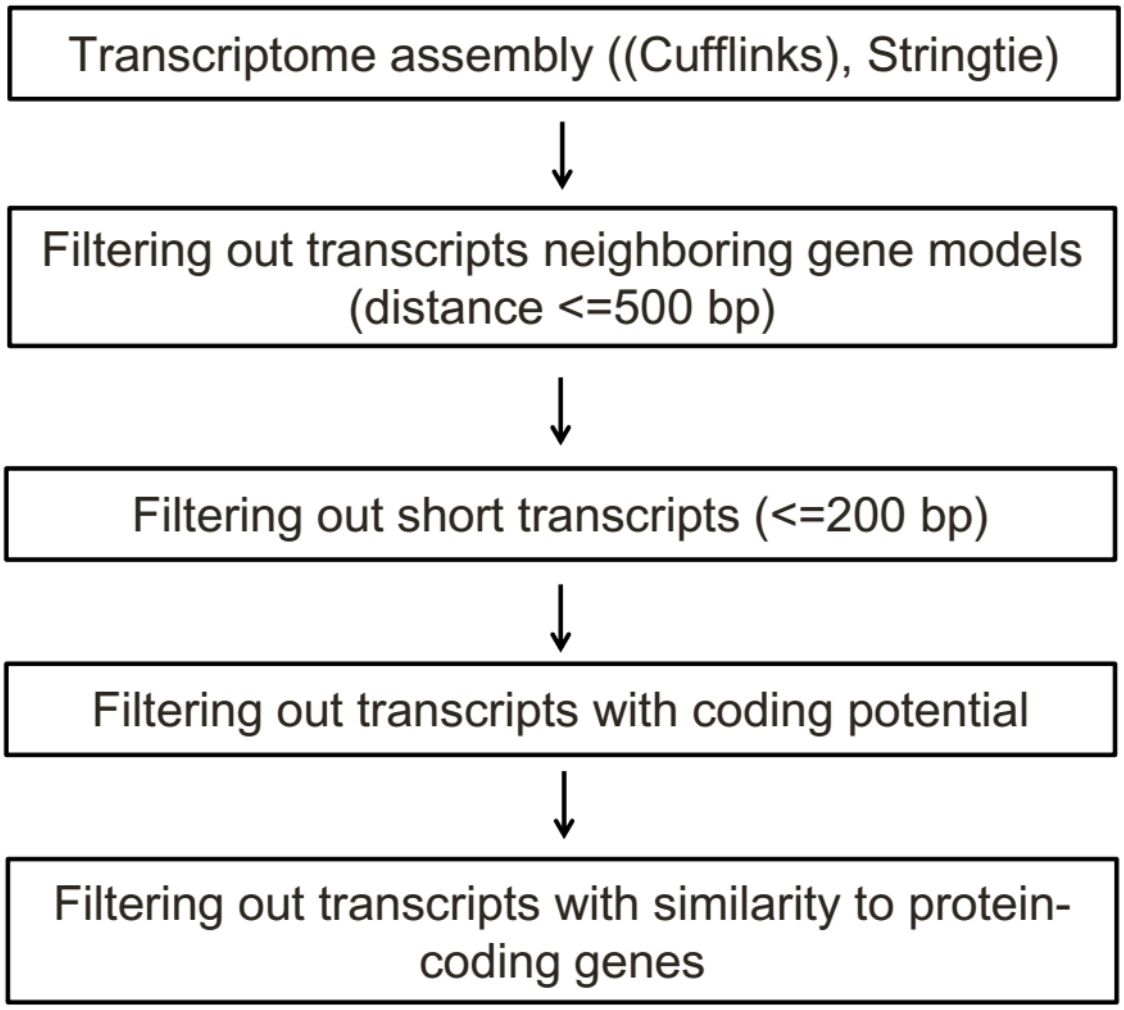


**Figure 5.1** Flow chart of linc_finder.

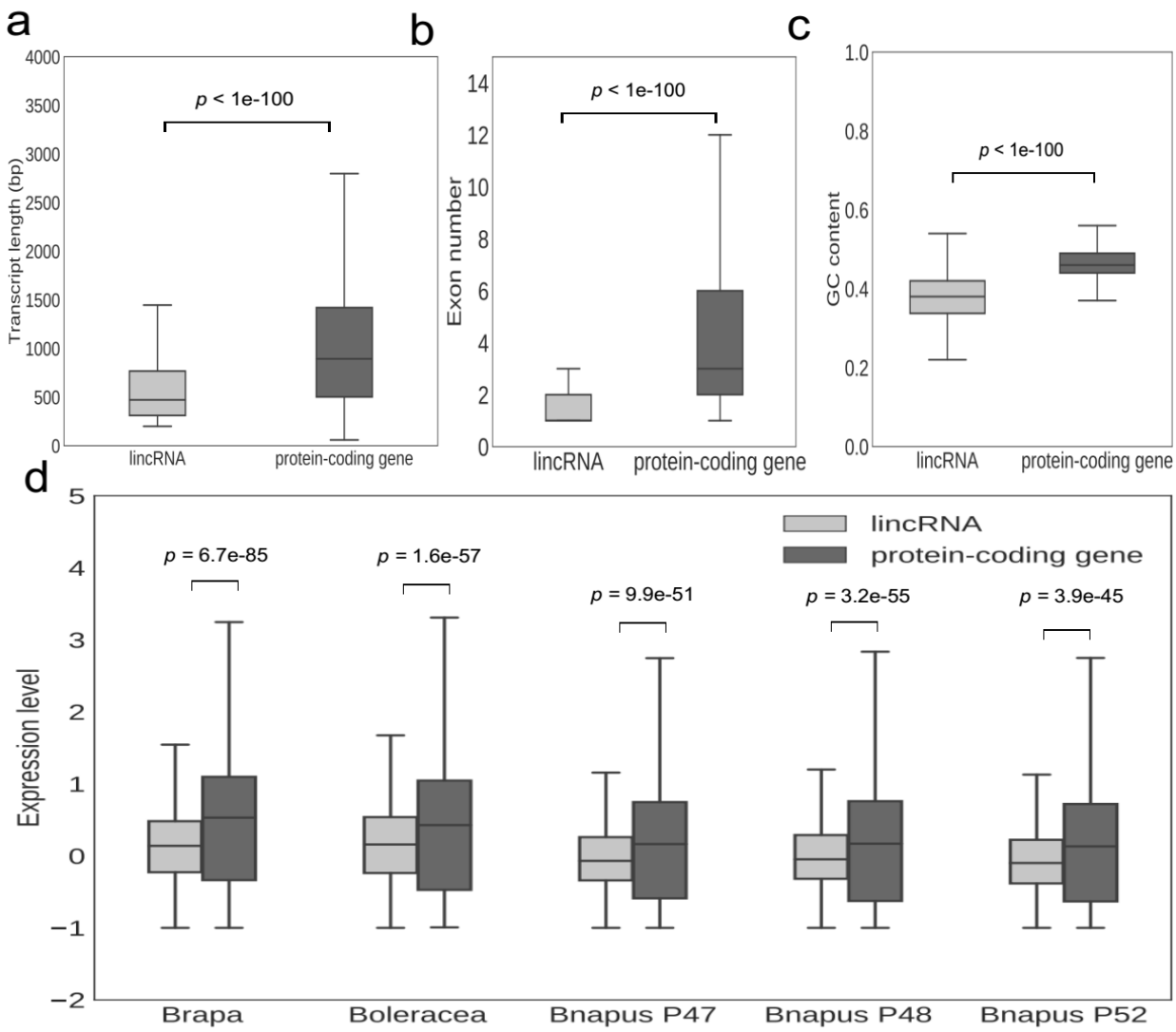


**Figure 5.2** Transcript length (A) , exon number (B), GC content (C), and expression level (D) of lincRNAs and protein-coding genes in the three lines of the polyploid *Brassica napus* and their parents (*Brassica rapa* and *Brassica oleracea*).

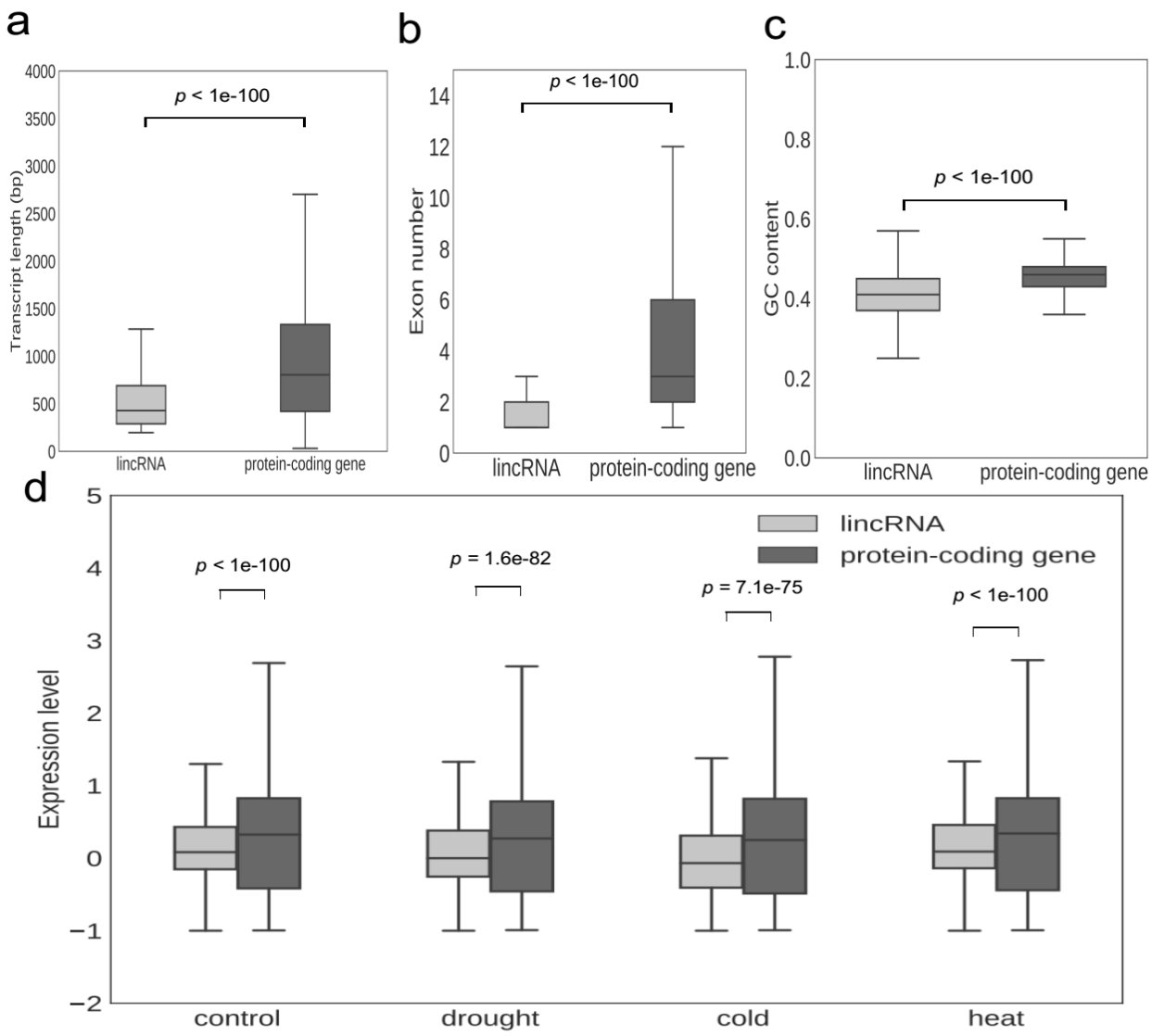


**Figure 5.3** Transcript length (A) , exon number (B), GC content (C), and expression level (D) of lincRNAs and protein-coding genes in the natural *Brassica napus*.

## 6. Concluding chapter

My Ph.D thesis has two themes: gene evolution after duplication in plants (Chapter 2, 4 and 5) and the evolution of plant lincRNAs (Chapter 3, 4 and 5). Here I briefly summarize the contributions of my studies to the fields for each chapter. Also, the limitations of each study and possible future directions are discussed below.

### 6.1 Divergence of duplicated genes through microRNA binding site divergence

In the second chapter, I investigated the impact of microRNA binding site divergence on the evolution of duplicated genes in *Arabidopsis thaliana* and Brassica. Through the analysis of microRNA binding site divergence in different types of duplicated genes, I quantified the proportion of duplicate gene pairs with divergent microRNA binding sites. Most duplicated genes were found to show divergent microRNA binding sites. My analyses reveal the contribution of microRNA binding sites to the divergence of genes after duplication. Duplicated genes with divergent microRNA binding sites show more divergence in expression, suggesting the impact of divergence of microRNA regulation on the expression divergence of duplicated genes.

The following questions are yet to be solved in regards to miRNA binding sites in duplicated genes. First, there are two scenarios by which duplicated genes can show microRNA binding site divergence. One is that the ancestral state of the duplicated genes had a microRNA site and the binding site became divergent later in evolution. The other possibility is that neither of the two duplicated genes was regulated by microRNA(s) and

it is the gain of a microRNA binding site in one of them that leads to the binding site divergence. Although I assessed the contribution of evolutionarily young microRNAs to the gain of binding sites, those gains are more difficult to assess for more ancient microRNAs and it will require a more detailed phylogenetic approach to determine the ancestral state of the microRNA binding sites. Second, although the vast majority of duplicated genes show divergent microRNA binding sites, the genes with microRNA binding sites are not always the copy that is more highly expressed in a pair of duplicates. I think that this observation could be accounted for by the following two reasons. First, the expression of many miRNAs is highly tissue-specific. Thus the tissue types or developmental stages we sampled to analyze gene expression levels might not reflect the expression of genes in the tissue types where the microRNA happens to be expressed. Second, microRNA targets might still show a higher expression level compared with their non-target paralogs even though they are under the regulation of microRNAs. For example, it has been found that genes with microRNA binding sites have more optimal codons which might be associated with an increased efficiency of translation in *Arabidopsis thaliana* (Takuno and Innan 2011).

### 6.2 LincRNA evolution in flowering plants

In Chapter 3 and 4 I characterized evolutionary features of lincRNAs in flowering plants. I analyzed the sequence conservation of lincRNAs in Chapter 3. Sequences of lincRNAs diverge very rapidly compared with protein-coding genes and microRNAs. The analysis

of selection constraints on lincRNAs might be important to understanding the functions of lincRNAs. Unlike protein-coding genes where methods of the analysis of selection on the sequence have been well developed, fewer methods are available to quantify the level of selection on lincRNA sequences. To the best of my knowledge, the analyses on the identification of conserved motifs in Chapter 3 provide the first insights into the highly conserved regions in plant lincRNAs that are shared by multiple species. Based on the identification of these conserved regions, I analyzed SNP data set to assess the selection constraints on the conserved regions of lincRNAs. Stronger selection constraints were detected in the conserved regions than surrounding regions and introns. These results suggest that the conserved regions are likely under purifying selection and provide new clues to understanding lincRNA functions. It could be hypothesized that these regions might play sequence-specific roles whereas flanking sequences are more important at the structure level and are thus more flexible in sequence.

In Chapter 3 I analyzed lincRNA data sets from five representative species of four plant families. However, as the original lincRNA transcripts were not always identified using the same criteria, and were not based on the same tissue types in all studies, caution should be taken when making comparison between different families. Thus the vast majority of the chapter aims to compare among lincRNAs, protein-coding genes and microRNAs from the same species rather than making conclusions on the comparison between different plant species and families.

Another issue that needs to be taken into consideration is that a large portion of the

analyses in Chapter 3 is focused on lincRNA evolution at the sequence level. The expression conservation of lincRNAs was only examined between closely related species using the RNA-seq data from the same organ type. This is mainly due to the limited number of high-quality RNA-seq data sets available in plants. It is plausible that lincRNA homologous loci without evidence for transcription in my analyses are transcribed in other organ types or conditions that were not analyzed. Analysis with deeper sequencing data from more tissue types may help to reveal a more detailed expression landscape of lincRNAs. Also, although we have identified many short conserved regions in lincRNAs, the functions of them still remain unknown and wait to be validated by functional studies in the future.

### 6.3 The evolution of lincRNAs by gene duplication

Very few studies have investigated the origin of lincRNAs. In general, lincRNAs may originate by the following three ways: conversion from previous protein-coding genes, de novo birth, and duplication of other lincRNAs (Ulitsky and Bartel 2013). I characterized duplicated lincRNAs in several plant species in Chapter 4. The proportion of duplicated lincRNAs is generally low and varies greatly among lineages. Also, only a small proportion of duplicated lincRNAs are likely derived from whole-genome duplication. All the above characteristics are in sharp contrast to protein-coding genes.

Based upon the detailed identification of duplicated lincRNA data sets, I analyzed the expression divergence of lincRNAs and compared this pattern with protein-coding genes,

revealing many unique features of duplicated lincRNAs. First, the expression of duplicated lincRNAs showed much less correlation than for mRNAs. Second, lincRNAs are much more likely to show reciprocal expression patterns. Third, tandem duplicates of lincRNAs show the highest similarity of expression. Last, the expression divergence is not correlated with sequence divergence.

The above unique features of lincRNAs are important to understand the duplication of lincRNAs. The large expression divergence of lincRNAs indicates that lincRNAs likely diverge rapidly in expression after duplication. Previous studies have pointed out that the expression of lincRNAs is more affected by the local chromatin environment (Chekanova 2015). As most duplicated lincRNAs are likely generated by interspersed duplication, the chromatin environment is likely to be different between the two duplicates, thereby resulting in the large divergence of lincRNAs at the expression level. The observations in regards to the high expression similarity between tandem duplicates of lincRNAs are very different from protein-coding genes. The finding that tandem duplicates of lincRNAs show the highest expression similarity might also be due to their similar chromatin environment.

### 6.4 LincRNA expression evolution after polyploidization

To better understand how the expression of lincRNAs diverges after polyploidization, as well as in response to abiotic stresses in a polyploid, I conducted analyses of lincRNA expression divergence after polyploidization in the *Brassica* polyploid and diploid system

(Chapter 5). My analyses indicate large expression divergence of lincRNAs not only after polyploidization but also across different stress conditions. In addition, lincRNAs show more divergence in expression compared with protein-coding genes. The expression divergence of lincRNAs between the resynthesized *Brassica napus* and parental species suggest the rapid divergence of lincRNAs soon after polyploidization. Based on the expression change of lincRNAs, it could be hypothesized that some lincRNAs might be involved in lineage-specific features, potentially underlying the phenotypic variation between the diploid and polyploid species, and between responses to different stress conditions.

Chapter 5 raises further questions on gene expression after polyploidization and under abiotic stress. Due to the availability of data, I only analyzed the gene expression in leaves in the *Brassica* polyploid system. Considering the lineage-specific and tissue-specific manner of lincRNA expression, it would be desirable to characterize the expression of lincRNAs in more tissue types and in other species. Another possible future direction is to investigate the mechanisms underlying the rapid expression divergence of lincRNAs after polyploidization and in response to stresses. It could be hypothesized that lincRNAs might be more affected by the changes of epigenetic status after polyploidization. To test this hypothesis, one may need to perform genome-wide identification of changes in DNA methylation and histone modification that are potentially associated with expression divergence of lincRNAs.

## 6.5 Concluding remarks on lincRNA evolution

One of the important results in my dissertation is the low conservation level sequence and expression conservation of lincRNAs in flowering plants. Homology searches and phylogenetic analysis pointed out that the homologous loci of the vast majority of plant lincRNAs examined can only date back to species within the family. While a small proportion of lincRNAs likely have more ancient origins, they do not show the same conservation level as protein-coding genes, as indicated by their lower sequence similarity and shorter aligned sequences. The lack of conservation of the sequences of lincRNAs can be due to their rapid evolution or their recent origins. Some studies have suggested that some lincRNAs may be conserved in expression while diverging in sequence, pointing out the positional conservation of lincRNAs (Ulitsky et al. 2011; Mohammadin et al. 2015). Others suggested that the recent activity of transposons may drive the origins of lincRNAs, supporting the second scenario (Kapusta et al. 2013; Wang et al. 2016). Further analyses are needed to answer this question. In addition, orthologs of plant lincRNAs displayed very rapid expression turnover, both qualitatively and quantitatively,, even between very closely related species. The expression conservation level is, in general, higher between those species that are closely related species than those that are distantly related. Additionally, lineage-specific lincRNAs are expressed at lower levels than those conserved in distantly related species, suggesting gains of lincRNA expression gradually.

My dissertation also investigated the evolution of plant lincRNAs after duplication

and polyploidization, providing some of the first insights into the evolution of lincRNAs by gene duplication in plants. The proportion of lincRNAs that are likely derived from gene duplication is much higher in plants than previously reported in animals. This result suggests different mechanisms of lincRNA evolution between plants and animals. Note that the proportion of duplicated lincRNAs might be larger than estimated in my dissertation. This is because most lincRNAs are subject to rapid sequence evolution, and the currently annotated lincRNA repertoires might represent a proportion of lincRNAs that exist due to their low expression level and high expression tissue specificity. In general, lincRNAs exhibit higher expression divergence between duplicates than protein-coding genes in the plant species surveyed in my dissertation. These results indicate the rapid divergence of expression after the duplication of lincRNAs. Also, lincRNAs showed more expression divergence than protein-coding genes between the diploid and polyploid *Brassica* species. Expression divergence between duplicated lincRNAs may contribute to functional divergence between them in some cases.

Evolutionary conservation in sequences can sometimes give clues as to which regions are functional. I characterized a large number of short motifs that are evolutionarily conserved across various flowering plants, consistent with studies in animals (Hezroni et al. 2015). I also uncovered evidence for purifying selection on these sequences. These findings suggest that only a small part of the entire transcript can be involved in the function of the lincRNA. One possible scenario is that these conserved motifs may be responsible for sequence recognition whereas their fast-evolving flanking

sequences might provide structural support. Future functional studies could test these hypotheses.

**References**

Abrouk M, Zhang RZ, Murat F, Li AL, Pont C, Mao L, Salse J. 2012. Grass MicroRNA Gene Paleohistory Unveils New Insights into Gene Dosage Balance in Subgenome Partitioning after Whole-Genome Duplication. Plant Cell 24:1776–1792.

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. U. S. A. 100:4649–4654.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8:135–141.

Allen E, Xie ZX, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat. Genet. 36:1282–1290.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 12:A1326–A1326.

Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y. 2010. Posttranslational regulation impacts the fate of duplicated genes. Proc. Natl. Acad. Sci. U. S. A. 107:2967–2971.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Anders S, Pyl PT, Huber W. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

Ariel F, Romero-Barrios N, Jegu T, Benhamed M, Crespi M. 2015. Battles and hijacks: noncoding transcription in plants. Trends Plant Sci. 20:362–371.

Arsovski AA, Pradinuk J, Guo XQ, Wang S, Adams KL. 2015. Evolution of Cis-Regulatory Elements and Regulatory Networks in Duplicated Genes of Arabidopsis. Plant Physiol. 169:2982–2991.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 110:17409–17414.

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

Axtell MJ, Westholm JO, Lai EC. 2011. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biol. 12:221.

Baker CR, Hanson-Smith V, Johnson AD. 2013. Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. Science (80-. ). 342:104–108.

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27:1691–1692.

Barquist L, Burge SW, Gardner PP. 2016. Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families. Curr Protoc Bioinforma. 54:12.13.1–12.13.25.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. Cell 136:215–233.

Berezikov E. 2011. Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet. 12:846–860.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res 13:137–144.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691.

Bonnet E, He Y, Billiau K, Van de Peer Y. 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. Bioinformatics 26:1566–1568.

Bonnet E, Van de Peer Y, Rouze P. 2006. The small RNA world of plants. New Phytol. 171:451–468.

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

Burgess D, Freeling M. 2014. The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. Plant Cell 26:946–961.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.

Campalans A, Kondorosi A, Crespi M. 2004. Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. Plant Cell 16:1047–1059.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. 43:956-U60.

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 26:301–314.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. 7:R13.

Chekanova JA. 2015. Long non-coding RNAs and their functions in plants. Curr Opin Plant Biol 27:207–216.

Chen J, Shishkin AA, Zhu XP, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M. 2016. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 17.

Chen JY, Shen QS, Zhou WZ, Peng J, He BZ, Li Y, Liu CJ, Luan X, Ding W, Li S, et al.

2015. Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. PLoS Genet. 11:1–24.

Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. 8:93–103.

Chen XM. 2009. Small RNAs and their roles in plant development. Annu. Rev. Cell Dev. Biol. 25:21–44.

Chen ZJ. 2007. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu. Rev. Plant Biol. 58:377–406.

Cheng F, Wu J, Fang L, Sun SL, Liu B, Lin K, Bonnema G, Wang XW. 2012. Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of Brassica rapa. PLoS One 7:e36442.

Chureau C, Prissette M, Bourdet A, Barbe V, Cattolico L, Jones L, Eggen A, Avner P, Duret L. 2002. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. Genome Res. 12:894–908.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Dai X, Zhao PX. 2011. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39:W155-9.

Dai XB, Zhuang ZH, Zhao PXC. 2011. Computational analysis of miRNA targets in plants: current status and challenges. Brief. Bioinform. 12:115–121.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 22:1775–1789.

Ding JH, Lu Q, Ouyang YD, Mao HL, Zhang PB, Yao JL, Xu CG, Li XH, Xiao JH, Zhang QF. 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc. Natl. Acad. Sci. U. S. A. 109:2654–2659.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary Genetics of Genome Merger and Doubling in Plants. Annu. Rev. Genet. 42:443–461.

Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics 15:53.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Ehrenreich IM, Purugganan MD. 2008. Sequence variation of microRNAs and their binding sites in Arabidopsis. Plant Physiol. 146:1974–1982.

Eyre-Walker A. 1999. Evolutionary genomics. Trends Ecol. Evol. 14:176.

Ezawa K, OOta S, Saitou N. 2006. Genome-wide search of gene conversions in duplicated genes of mouse and rat. Mol. Biol. Evol. 23:927–940.

Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, Smith LM, Dasenko M, Givan SA, Weigel D, et al. 2010. MicroRNA Gene Evolution in Arabidopsis lyrata and Arabidopsis thaliana. Plant Cell 22:1074–1089.

Fan C, Chen Y, Long M. 2008. Recurrent tandem gene duplication gave rise to functionally divergent genes in Drosophila. Mol. Biol. Evol. 25:1451–1458.

Felekkis K, Voskarides K, Dweep H, Sticht C, Gretz N, Deltas C. 2011. Increased number of microRNA target sites in genes encoded in CNV regions. Evidence for an

evolutionary genomic interaction. Mol Biol Evol 28:2421–2424.

De Felippes FF, Schneeberger K, Dezulian T, Huson DH, Weigel D. 2008. Evolution of Arabidopsis thaliana microRNAs from random sequences. Rna-a Publ. Rna Soc. 14:2455–2459.

Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, Bombarely A, Fisher-York T, Pujar A, Foerster H, et al. 2015. The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. Nucleic Acids Res. 43:D1036–D1041.

Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9:102–114.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183:557–564.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

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

Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. Curr. Opin. Plant Biol. 12:126–132.

Friedman R, Hughes AL. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. 11:373–381.

Gaiti F, Fernandez-Valverde SL, Nakanishi N, Calcino AD, Yanai I, Tanurdzic M, Degnan BM. 2015. Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity. Mol. Biol. Evol. 32:2367–2382.

Gallart AP, Pulido AH, de Lagran IAM, Sanseverino W, Cigliano RA. 2016. GREENC: a

Wiki-based database of plant lncRNAs. Nucleic Acids Res. 44:D1161–D1166.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in Arabidopsis. Mol. Biol. Evol. 24:2298–2309.

Gao LZ, Innan H. 2004. Very low gene duplication rate in the yeast genome. Science (80-. ). 306:1367–1370.

Glover NM, Daron J, Pingault L, Vandepoele K, Paux E, Feuillet C, Choulet F. 2015. Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biol 16:188.

Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. 2010. BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics 26:2617–2619.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34:D140–D144.

Gu ZL, Nicolae D, Lu HHS, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. 18:609–613.

Guan YF, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics 175:933–943.

Guil S, Esteller M. 2012. Cis-acting noncoding RNAs: friends and foes. Nat. Struct. Mol. Biol. 19:1068–1075.

Guo XY, Gui YJ, Wang Y, Zhu QH, Helliwell C, Fan LJ. 2008. Selection and mutation on microRNA target sequences during rice evolution. BMC Genomics 9:454.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific

transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. 28:756.

Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, Carrington JC, Chen XM, Wang XJ, Chen ZJ. 2009. Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc. Natl. Acad. Sci. U. S. A. 106:17835–17840.

Haberer G, Hindemitt T, Meyers BC, Mayer KFX. 2004. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of arabidopsis. Plant Physiol. 136:3009–3022.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148:993–1003.

Hao ZQ, Fan CY, Cheng T, Su Y, Wei Q, Li GL. 2015. Genome-Wide Identification, Characterization and Evolutionary Analysis of Long Intergenic Noncoding RNAs in Cucumber. PLoS One 10:e0121800.

Hardison RC. 2003. Comparative Genomics. PLOS Biol. 1:e58.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. . Ph.D. Thesis, Pennsylvania State Univ.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 22:1760–1774.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45:891–898.

He XL, Zhang JZ. 2006. Higher duplicability of less important genes in yeast genomes. Mol. Biol. Evol. 23:144–151.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Rep. 11:1110–1122.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681.

Huang SW, Li RQ, Zhang ZH, Li L, Gu XF, Fan W, Lucas WJ, Wang XW, Xie BY, Ni PX, et al. 2009. The genome of the cucumber, Cucumis sativus L. Nat. Genet. 41:1275–1281.

Huang ZN, Duan WK, Song XM, Tang J, Wu P, Zhang B, Hou XL. 2016. Retention, Molecular Evolution, and Expression Divergence of the Auxin/Indole Acetic Acid and Auxin Response Factor Gene Families in Brassica Rapa Shed Light on Their Evolution Patterns in Plants. Genome Biol. Evol. 8:302–316.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief. Bioinform. 12:442–448.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256:119–124.

Ietswaart R, Wu Z, Dean C. 2012. Flowering time control: another window to the connection between antisense RNA and chromatin. Trends Genet. 28:445–453.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108.

Iwakawa H, Tomari Y. 2013. Molecular insights into microRNA-mediated translational repression in plants. Mol. Cell 52:591–601.

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. 2015. The landscape of long noncoding RNAs in the human

transcriptome. Nat. Genet. 47:199–208.

Jeong DH, Park S, Zhai JX, Gurazada SGR, De Paoli E, Meyers BC, Green PJ. 2011. Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. Plant Cell 23:4185–4207.

Jia F, Rock CD. 2013. MIR846 and MIR842 comprise a cistronic MIRNA pair that is regulated by abscisic acid by alternative splicing in roots of Arabidopsis. Plant Mol. Biol. 81:447–460.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97-U113.

Jin JJ, Liu J, Wang H, Wong L, Chua NH. 2013. PLncDB: plant long non-coding RNA database. Bioinformatics 29:1068–1071.

Johnsson P, Lipovich L, Grander D, Morris K V. 2014. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim. Biophys. Acta-General Subj. 1840:1063–1071.

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

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

Kaessmann H, Vinckenbosch N, Long MY. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. 10:19–31.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo XY, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. Plos Genet. 9:e1003470.

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S,

Schwartz DC, Tanaka T, Wu JZ, Zhou SG, et al. 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6:1–10.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39:1251–1255.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 13:R107.

Kim ED, Sung S. 2012. Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci. 17:16–21.

Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates Of Base Substitutions Through Comparative Studies Of Nucleotide-Sequences. J. Mol. Evol. 16:111–120.

Kodama Y, Shumway M, Leinonen R, C INSD. 2012. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 40:D54–D56.

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol. 17.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42:D68–D73.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Kung JTY, Colognori D, Lee JT. 2013. Long Noncoding RNAs: Past, Present, and Future. Genetics 193:651–669.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid Turnover of Long Noncoding RNAs and the Evolution of

Gene Expression. Plos Genet. 8:e1002841.

Lamesch P, Berardini TZ, Li DH, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40:D1202–D1210.

Lee JT. 2009. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev. 23:1831–1842.

Lee TH, Tang HB, Wang XY, Paterson AH. 2013. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res. 41:D1152–D1158.

Li AL, Mao L. 2007. Evolution of plant microRNA gene families. Cell Res. 17:212–218.

Li J, Musso G, Zhang Z. 2008. Preferential regulation of duplicated genes by microRNAs in mammals. Genome Biol. 9:R132.

Li J, Yuan Z, Zhang Z. 2010. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet 6:e1001187.

Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, et al. 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 15:R40.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. Plant Cell 28:326–344.

Liao DQ. 1999. Concerted evolution: Molecular mechanism and biological implications. Am. J. Hum. Genet. 64:24–30.

Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41.

Liu J, Jung C, Xu J, Wang H, Deng SL, Bernad L, Arenas-Huertero C, Chua NH. 2012.

Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis. Plant Cell 24:4333–4345.

Liu SL, Adams KL. 2010. Dramatic Change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the brassicaceae. Mol. Biol. Evol. 27:2817–2828.

Liu SL, Baute GJ, Adams KL. 2011. Organ and Cell Type-Specific Complementary Expression Patterns and Regulatory Neofunctionalization between Duplicated Genes in Arabidopsis thaliana. Genome Biol. Evol. 3:1419–1436.

Liu ZL, Adams KL. 2007. Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. Curr Biol 17:1669–1674.

Lyons E, Tang H. 2014. Syntenic Sequence Conservation Between and Within Papaya Genes. In: Ming R, Moore HP, editors. Genetics and Genomics of Papaya. Vol. 10. New York: Springer New York.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Maher C, Stein L, Ware D. 2006. Evolution of Arabidopsis microRNA families through duplication events. Genome Res. 16:510–519.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. 13:1638–1645.

Mano S, Innan H. 2008. The evolutionary rate of duplicated genes under concerted evolution. Genetics 180:493–505.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. Plos Biol. 3:1970–1979.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10:R124.

Marques AC, Ponting CP. 2014. Intergenic lncRNAs and the evolution of gene expression. Curr Opin Genet Dev27:48–53.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:10–12.

McHale M, Eamens AL, Finnegan EJ, Waterhouse PM. 2013. A 22-nt artificial microRNA mediates widespread RNA silencing in Arabidopsis. Plant J 76:519–529.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31:200–204.

Meng YJ, Shao CG, Chen M. 2011. Toward microRNA-mediated gene regulatory networks in plants. Brief. Bioinform. 12:645–659.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Millar AA, Waterhouse PM. 2005. Plant and animal microRNAs: similarities and differences. Funct Integr Genomics 5:129–135.

Mohammadin S, Edger PP, Pires JC, Schranz ME. 2015. Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Plant Biol. 15.

Morell V. 1996. TreeBASE: The roots of phylogeny. Science (80-. ). 273:569.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635–640.

Nei M, Li WH. 1979. Mathematical-Model for Studying Genetic-Variation In Terms Of Restriction Endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76:5269–5273.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39:121–152.

Nelson AD, Forsythe ES, Devisetty UK, Clausen DS, Haug-Batzell AK, Meldrum AM, Frank MR, Lyons E, Beilstein MA. 2016. A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants. G3 6:2881–2891.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919–1920.

Nitsche A, Rose D, Fasold M, Reiche K, Stadler PF. 2015. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. Rna 21:801–812.

Nozawa M, Miura S, Nei M. 2012. Origins and Evolution of MicroRNA Genes in Plant Species. Genome Biol. Evol. 4:230–239.

Ohno S. 1970. Evolution by gene duplication. New York, Springer-Verlag

Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171:2294–2316.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22:1–5.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics 171:765–781.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin DC, Llewellyn D, Showmaker KC, Shu SQ, Udall J, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427.

Van de Peer Y, Maere S, Meyer A. 2009. OPINION The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. 10:725–732.

Pereira V, Waxman D, Eyre-Walker A. 2009. A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence. Genetics 183:1597–1600.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33:290–+.

Piriyapongsa J, Jordan IK. 2008. Dual coding of siRNAs and miRNAs by plant transposable elements. Rna-a Publ. Rna Soc. 14:814–821.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and Functions of Long Noncoding RNAs. Cell 136:629–641.

Presser A, Elowitz MB, Kellis M, Kishony R. 2008. The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. Proc. Natl. Acad. Sci. U. S. A. 105:950–954.

Pritchard L, White JA, Birch PRJ, Toth IK. 2006. GenomeDiagram: a python package for the visualization of large-scale genomic data. Bioinformatics 22:616–617.

Qian WF, Liao BY, Chang AYF, Zhang JZ. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. 26:425–430.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. 17:47–62.

Quinn JJ, Ilik IA, Qu K, Georgiev P, Chu C, Alchtar A, Chang HY. 2014. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. Nat. Biotechnol. 32:933–940.

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. 2002. Prediction of plant microRNA targets. Cell 110:513–520.

Rinn JL, Chang HY. 2012. Genome Regulation by Long Noncoding RNAs. Annu. Rev. Biochem. Vol 81 81:145–166.

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

Rogers K, Chen XM. 2013. Biogenesis, Turnover, and Mode of Action of Plant MicroRNAs. Plant Cell 25:2383–2399.

Rubio-Somoza I, Weigel D. 2011. MicroRNA networks and developmental plasticity in plants. Trends Plant Sci. 16:258–264.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. Elife 3:e03523.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.

Schnable JC, Freeling M, Lyons E. 2012. Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses. Genome Biol. Evol. 4:265–277.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U. S. A. 108:4069–4074.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science (80-. ). 326:1112–1115.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. 15:147–153.

Semon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. Proc. Natl. Acad. Sci. U. S. A. 105:8333–8338.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski

B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–2504.

Shivaprasad P V, Chen HM, Patel K, Bond DM, Santos BA, Baulcombe DC. 2012. A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. Plant Cell 24:859–874. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22408077

Shuai P, Liang D, Tang S, Zhang ZJ, Ye CY, Su YY, Xia XL, Yin WL. 2014. Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in Populus trichocarpa. J. Exp. Bot. 65:4975–4983.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy And Angiosperm Diversification. Am. J. Bot. 96:336–348.

Sorourian M, Kunte MM, Domingues S, Gallach M, Ozdil F, Rio J, Betran E. 2014. Relocation Facilitates the Acquisition of Short Cis-Regulatory Regions that Drive the Expression of Retrogenes during Spermatogenesis in Drosophila. Mol. Biol. Evol. 31:2170–2180.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V. 2012. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics 28:2059–2061.

Su ZX, Wang JM, Yu J, Huang XQ, Gu X. 2006. Evolution of alternative splicing after gene duplication (vol 16, pg 182, 2006). Genome Res. 16:557.

Sugino RP, Innan H. 2006. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. Trends Genet. 22:642–644.

Sunkar R, Li YF, Jagadeeswaran G. 2012. Functions of microRNAs in plant stress responses. Trends Plant Sci. 17:196–203.

Takuno S, Innan H. 2008. Evolution of complexity in miRNA-mediated gene regulation systems. Trends Genet. 24:56–59.

Takuno S, Innan H. 2011. Selection fine-tunes the expression of microRNA target genes in Arabidopsis thaliana. Mol. Biol. Evol. 28:2429–2434.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034.

Taylor JS, Raes J. 2004. Duplication and divergence: The evolution of new genes and old ideas. Annu. Rev. Genet. 38:615–643.

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ. 2014. Evolutionary history of plant microRNAs. Trends Plant Sci. 19:175–182.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.

Throude M, Bolot S, Bosio M, Pont C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, et al. 2009. Structure and expression analysis of rice paleo duplications. Nucleic Acids Res. 37:1248–1259.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol 8:R50.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7:562–578.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science (80-. ). 329:689–693.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N,

Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science (80-. ). 313:1596–1604.

Ulitsky I, Bartel DP. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. Cell 154:26–46.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell 147:1537–1550.

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

Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK. 2017. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants. Plant J 90: 133-146.

Wang H, Niu QW, Wu HW, Liu J, Ye J, Yu N, Chua NH. 2015. Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. Plant J 84:404–416.

Wang KC, Chang HY. 2011. Molecular Mechanisms of Long Noncoding RNAs. Mol. Cell 43:904–914.

Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, Wang P, Liu N, Lindsey K, Zhang X. 2015. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol 207:1181-1197.

Wang SS, Adams KL. 2015. Duplicate Gene Divergence by Changes in MicroRNA Binding Sites in Arabidopsis and Brassica. Genome Biol. Evol. 7:646–655.

Wang SS, Chen YH, Cao QH, Lou HQ. 2015. Long-Lasting Gene Conversion Shapes the Convergent Evolution of the Critical Methanogenesis Genes. G3-Genes Genomes Genet. 5:2475–2486.

Wang TZ, Liu M, Zhao MG, Chen RJ, Zhang WH. 2015. Identification and characterization of long non-coding RNAs involved in osmotic and salt stress in

Medicago truncatula using genome-wide high-throughput sequencing. BMC Plant Biol. 15:131.

Wang X, Ai G, Zhang CL, Cui L, Wang JF, Li HX, Zhang JH, Ye ZBA. 2016. Expression and diversification analysis reveals transposable elements play important roles in the origin of Lycopersicon-specific lncRNAs in tomato. New Phytol. 209:1442–1455.

Wang X, Wu R, Lin X, Bai Y, Song C, Yu X, Xu C, Zhao N, Dong Y, Liu B. 2013. Tissue culture-induced genetic and epigenetic alterations in rice pure-lines, F1 hybrids and polyploids. BMC Plant Biol. 13:77.

Wang XJ, Reyes JL, Chua NH, Gaasterland T. 2004. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biol. 5:R65.

Wang XW, Wang HZ, Wang J, Sun RF, Wu J, Liu SY, Bai YQ, Mun JH, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1139.

Wang YP, Tang HB, DeBarry JD, Tan X, Li JP, Wang XY, Lee TH, Jin HZ, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49.

Wang YQ, Wang XC, Deng W, Fan XD, Liu TT, He GM, Chen RS, Terzaghi W, Zhu DM, Deng XW. 2014. Genomic Features and Regulatory Roles of Intermediate-Sized Non-Coding RNAs in Arabidopsis. Mol. Plant 7:514–527.

Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. Rna 17:578–594.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res. 24:616–628.

Weiberg A, Wang M, Lin FM, Zhao HW, Zhang ZH, Kaloshian I, Huang HD, Jin HL. 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA

interference pathways. Science (80-. ). 342:118–123.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. Plos Comput. Biol. 3:680–691.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.

Wu HJ, Wang ZM, Wang M, Wang XJ. 2013. Widespread Long Noncoding RNAs as Endogenous Target Mimics for MicroRNAs in Plants. Plant Physiol. 161:1875–1884.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881.

Xu P, Zhang XF, Wang XM, Li JT, Liu GM, Kuang YY, Xu J, Zheng XH, Ren LF, Wang GL, et al. 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nat. Genet. 46:1212–1219.

Xu Q, Song Z, Zhu C, Tao C, Kang L, Liu W, He F, Yan J, Sang T. 2017. Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change. BMC Plant Biol 17:42.

Yamada M. 2017. Functions of long intergenic non-coding (linc) RNAs in plants. J. Plant Res. 130:67–73.

Yamaji Y, Maejima K, Ozeki J, Komatsu K, Shiraishi T, Okano Y, Himeno M, Sugawara K, Neriya Y, Minato N, et al. 2012. Lectin-mediated resistance impairs plant virus infection at the cellular level. Plant Cell 24:3482.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange

transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Yang L, Froberg JE, Lee JT. 2014. Long noncoding RNAs: fresh perspectives into the RNA world. Trends Biochem Sci 39:35–43.

Yang L, Gaut BS. 2011. Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes. Mol. Biol. Evol. 28:2359–2369.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J. Mol. Evol. 48:597–604.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591.

Ye CY, Chen L, Liu C, Zhu QH, Fan LJ. 2015. Widespread noncoding circular RNAs in plants. New Phytol. 208:88–95.

Zeng Y, Cullen BR. 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Res. 32:4776–4785.

Zhan S, Horrocks J, Lukens LN. 2006. Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. Plant J. 45:347–357.

Zhang JZ. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhang JZ, Yang JR. 2015. Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16:409–420.

Zhang LF, Chia JM, Kumari S, Stein JC, Liu ZJ, Narechania A, Maher CA, Guill K, McMullen MD, Ware D. 2009. A Genome-Wide Characterization of MicroRNA Genes in Maize. Plos Genet. 5:e1000716.

Zhang YC, Chen YQ. 2013. Long noncoding RNAs: New regulators in plant development. Biochem. Biophys. Res. Commun. 436:111–114.

Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, Qu LH, Shu WS, Chen YQ. 2014. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol. 15:512.

Zhang YE, Vibranovski MD, Krinsky BH, Long MY. 2011. A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. Bioinformatics 27:1749–1753.

Zhu B, Yang Y, Li R, Fu D, Wen L, Luo Y, Zhu H. 2015. RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. J Exp Bot 66:4483–4495.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009. Evolution of Stress-Regulated Gene Expression in Duplicate Genes of Arabidopsis thaliana. Plos Genet. 5:e1000581.