GENE CONTENT EVOLUTION IN PLANT GENOMES: STUDIES OF WHOLE GENOME DUPLICATION, INTERGENIC TRANSCRIPTION AND EXPRESSION EVOLUTION IN BRASSICACEAE AND POACEAE SPECIES

By

Gaurav Dilip Moghe

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Genetics – Doctor of Philosophy
Quantitative Biology – Dual Degree

2013

ABSTRACT

GENE CONTENT EVOLUTION IN PLANT GENOMES: STUDIES OF WHOLE GENOME DUPLICATION, INTERGENIC TRANSCRIPTION AND EXPRESSION EVOLUTION IN BRASSICACEAE AND POACEAE SPECIES

By

Gaurav Dilip Moghe

Phenomena that create new genes and influence their diversification are important contributors to evolutionary novelty in living organisms. My research has focused on addressing the following questions regarding such phenomena in plants. First, what are the patterns of evolution of duplicate genes derived via whole genome duplication (WGD)? Second, do transcripts originating from intergenic regions constitute novel genes? Third, how do expression patterns of orthologous genes evolve in plants? I have addressed these questions using comparative genomic and transcriptomic analyses of species in the Brassicaceae and Poaceae families.

To understand the evolution of WGD derived duplicate genes, we sequenced and annotated the genome of wild radish (*Raphanus raphanistrum*), a Brassicaceae species which experienced a whole genome triplication (WGT) event ~24-29 million years ago. Through comparative genomic analyses of sequenced Brassicaceae species, I found that most WGT duplicate genes were lost over time. Duplicates that are still retained were found to undergo sequence and expression level divergence. Interestingly, while duplicate copies tend to diverge in expression level, one of the copies tends to maintain its original expression state in the tissue studied. Furthermore, duplicates that are retained in extant species tend to have higher expression levels, greater expression breadth, higher network connectivity and tend to be involved in

functions such as transcription factor activity, stress response and development. Functional diversification of such duplicates can assist in evolution of novel characters in plants post WGD.

To understand the nature of intergenic transcription, I analyzed multiple transcriptome datasets in *Arabidopsis thaliana* as well as in species of the Poaceae family. My results suggest that plant genomes do not show any evidence of pervasive intergenic transcription. Although thousands of intergenic transcripts can be found in each species, most of these transcripts have low breadths of expression, tend not to be conserved within or between species and show a significant bias in being located very close to genes or in open chromatin regions. My results suggest that most intergenic transcripts may be associated with transcription of the neighboring genes or may be produced as a result of noisy transcription. Properties of intergenic transcripts identified in my research will be useful in distinguishing functionally relevant transcripts from noise.

To understand expression evolution, I analyzed patterns of evolution of orthologous genes between Poaceae species and found that sequence divergence is strongly associated with level and breadth of expression, and very weakly with expression divergence. Both sequence and expression evolution were found to be constrained for genes involved in core biological processes such as metabolism, transcription, photosynthesis and transport.

Overall, the results of this research are broadly applicable to the field of gene annotation and increase our understanding of evolution of gene content in plant genomes.

This dissertation is dedicated…
…to my parents, for opening my mind and letting me pursue my interests
…to my dear sister, for the joy she has brought to my life
…to my wife, for her unconditional love and support

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Shin-Han Shiu for being an exemplary mentor. Shin-Han has immensely helped me grow as a scientist and as an individual, and has taught me many a life lessons, which will be useful to me throughout my career.

I would also like to acknowledge my committee members – Drs. David Arnosti, Titus Brown, Robin Buell and Barry Williams – for their valuable feedback on my projects, especially Dr. Barry Williams and Dr. Robin Buell for giving me the resources and the opportunity to work on very interesting questions. I would also like to thank Jeannine Lee and Dr. Barb Sears for their continuous support and advice.

Many thanks to the Shiu lab members, especially Melissa, David and Alex, for being excellent collaborators. I also value the help I have received over the years from everyone else in the lab – Andy, Cheng, Guangxi, Kelian, Ming-Jung, Johnny, Nick and Sahra. The Shiu lab has been a great place to work in because of you folks!

I owe a great deal of my contentment in my PhD years to my friends in East Lansing and elsewhere in the US with whom I could hang out and have fun. The list is too long to note here, but I will dearly cherish the memories of all the activities we did together over the past six years.

Finally, it is impossible for me to overstate the contribution of my family - my wife Ashwini, who has been a constant source of love and enthusiasm and a pillar of emotional support for me, and my parents and sister 9000 miles away in Mumbai. Without their love, support and advice, I couldn't have made it this far.

TABLE OF CONTENTS

CHAPTER THREE

CHAPTER FOUR

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
WGD	Whole Genome Duplication
WGT	Whole Genome Triplication
AT	*Arabidopsis thaliana*
AL	*Arabidopsis lyrata*
BR	*Brassica rapa*
RR	*Raphanus raphanistrum*
OS	*Oryza sativa*
BD	*Brachypodium distachyon*
SB	*Sorghum bicolor*
ZM	*Zea mays*
TxFrag	Transcript Fragment
ITF	Intergenic Transcribed Fragment
ITFR	Intergenic Transcribed Fragment Region
EST	Expressed Sequence Tag
Ka or dN	Non-synonymous substitution rate
Ks or dS	Synonymous substitution rate
KS	Kolmogrov-Smirnov
GO	Gene Ontology

CHAPTER ONE

Introduction

What is a gene? Traditionally, a gene has been defined as the unit of heredity. However, the definition of a gene has changed several times over the past century, and the answer to this question is still under active debate (Pearson, 2006; Gerstein et al., 2007; Pesole, 2008; Djebali et al., 2012). Understanding which genomic elements constitute genes and understanding the evolutionary characteristics of such elements is an active area of research. Functional elements in the genome include those producing a functional product such as protein-coding and RNA genes, as well as regulatory features such as transcription factor binding sites, nucleosome binding sites, DNA methylation sites and insulator regions. In addition to these elements, the genome also consists of other features with ambiguous functionality such as repeats, pseudogenes and unannotated transcribed regions. The advent of high throughput DNA and RNA sequencing methods as well as advances in bioinformatics in the last decade have provided us with an unprecedented ability to ask questions regarding the characteristics and evolution of these elements using comparative genomic and transcriptomic approaches. In my thesis, I have made use of such approaches to gain insight into two phenomena that potentially influence gene content in the plant kingdom, namely whole genome duplication and intergenic transcription.

ORIGINATION OF NOVEL GENES

Acquisition of new genes is an important driver of evolutionary novelty. The mechanisms by which new genes arise in a genome have been the focus of research since the early 20th century. Pioneering studies in Drosophila (Muller, 1935; Bridges et al., 1936) and maize (McClintock, 1950) suggested that novel phenotypes can arise through insertion, deletion and translocation of genomic elements (e.g.: the mosaic color patterns in maize seeds due to transposon hopping), possibly creating new genes in the process. However, further research over the next thirty years suggested gene duplication as the primary driver of novel gene creation, an argument summarized in Susumo Ohno's seminal book *Evolution by Gene Duplication* (Ohno, 1970). Today, we know that novel genes in eukaryotic genomes tend to arise via six principal processes – Gene duplication, retroposition (insertion of processed, reverse transcribed RNA back into the genome), exon shuffling, *trans*-splicing (fusion of partial preRNAs of two distant genes during RNA processing), horizontal gene transfer and *de novo* generation (Ding et al., 2012). Further discussion, as well as my research, focuses on two of these processes – gene duplication and *de novo* generation.

Origination via gene duplication

Gene duplication is a potent mechanism for creation of novel genes. Duplicate genes can arise via four processes – tandem duplication, segmental duplication, transposition and whole genome duplication. In the model plant *Arabidopsis thaliana,* there are several examples of lineage-specific gene family expansions caused via tandem duplications such as expansions in the Receptor-Like Kinase/Pelle, F-Box and Ubiquitin ligase gene families (Hanada et al., 2008). Segmental duplications also occur in plants; however, their frequency of occurrence is not clear, and genes arising via segmental duplication are significantly less likely to be retained than expected (Cannon et al. 2004). The third mode of gene duplication involves duplicating the entire genome (also termed polyploidization or whole genome duplication (WGD)). Polyploidization is known to occur very frequently in the plant kingdom. Recent research using Expressed Sequence Tag (EST) sequences from multiple plants discovered evidence for two ancient duplications – one in the ancestor of all seed plants ~350 million years ago (mya) and the other ~230 mya after angiosperms diverged from gymnosperms (Jiao et al., 2011), suggesting that all seed and flowering plants have evolved from a polyploid ancestor. In addition, 30-50% of the extant flowering plant species are believed to be polyploids, and ~15% of angiosperm speciation events are estimated to be due to polyploidization (Ramsey and Schemske, 1998; Wood et al., 2009). Polyploidization has also been postulated as an answer to "Darwin's abominable mystery" – what caused the rapid acceleration in diversification of angiosperms ~100 mya (Friedman, 2009; Jiao et al., 2011)? Genes duplicated via such polyploidization events may contribute significantly to adaptation and evolution of novel characters in the plant kingdom.

Duplicate genes are a source of evolutionary novelty because over time, given the presence of an extra gene copy, their functions may diverge from each other. The evolution of duplicate genes may occur via three routes: 1) they may gain new functions (neo-functionalization)(Ohno, 1970), 2) they may partition the functions of the ancestral gene between themselves (sub-functionalization)(Force et al., 1999) or 3) one copy may lose all functions and may become a pseudogene (pseudogenization)(Li, 1984). Despite occurrence of multiple WGD events in the evolutionary history of plants, most plant genomes contain only ~25,000-45,000 genes. This indicates that a large proportion of duplicated genes are lost through time, and only a fraction of the genes are retained. Distinguishing the characteristics of genes that are retained from those that are lost, as , understanding whether these characteristics are consistent between different polyploidization events and between different lineages derived from the same WGD event would contribute significantly to our understanding of plant genome evolution.

De novo origination of genes

Although there has been much skepticism regarding novel genes arising *de novo* (Ohno, 1970; Jacob, 1977), comparative genomics approaches in organisms such as yeast (Dujon, 1996), Drosophila (Domazet-Loso and Tautz, 2003) and *A. thaliana* (Lin et al., 2010) have revealed that ~10% of annotated protein-coding genes in each species tend to be lineage-specific, with no known homologs outside their lineage. Based on recent studies (Birney et al., 2007; Djebali et al., 2012), it seems intergenic or intronic transcripts produced by RNA polymerase transcription may likely be a source of such novel genes.

The Human Genome Project released the first version of the human genome sequence in 2001 and annotated ~32,000 protein-coding genes (Lander et al., 2001). This number was subsequently revised to 24,500 genes (Pennisi, 2003). Compared to the human genome, the genomes of *Caenorhabditis elegans* and *Arabidopsis thaliana*, released around the same time, were found to have 19,000 and 25,498 genes (C. elegans Sequencing Consortium, 1998; Arabidopsis Genome Initiative, 2000). The similarity in the numbers of annotated protein coding genes despite an observably large difference in the organismal complexity led to a debate on whether novel, yet unannotated genes with features different from traditional protein-coding genes lay in the genomes of more complex organisms, and whether such features could be identified based on the transcriptome. Around the same time, global transcriptome profiles of mammals (Bertone et al., 2004a; Carninci et al., 2005a), *Saccharomyces cerevisiae* (David et al., 2006) and *A. thaliana* (Yamada et al., 2003a) revealed a complex transcriptional landscape, with several thousand transcripts lying in the introns or in intergenic regions, or possessing alternatively spliced isoforms of the primary transcript. Over the past decade, several studies have suggested presence of novel types of functional non-coding RNA such as microRNA, long non-coding RNA, piwiRNA, vaultRNA etc. (Esteller, 2011) as well as novel small open reading frames (Basrai et al., 1997; Hanada et al., 2007; Pruitt et al., 2007) in the genomes of multiple species.

Despite evidence for the existence of several novel genomic features, some of which are transcribed, it is not clear whether they contribute to organismal complexity. A large proportion of these novel features is not conserved across species and is lineage-specific. Novel intergenic or intronic transcripts tend to be expressed at very low levels (<0.01 transcripts/cell) and have a very narrow breadth of expression (Djebali et al., 2012). Despite several hundred publications on non-coding RNA in the past decade, only a handful of non-coding RNAs, such as *Xist*, *RepA*, *Air*, *Hotair*, *Coldair*, have been rigorously shown to be functional (Kim and Sung, 2012). Very few studies (Guttman et al., 2009; van Bakel et al., 2010) have systematically investigated the relative abundance of novel functional transcripts and their characteristics *vis-a-vis* known transcripts. Most of these studies have been conducted in mammalian model systems, with the plant kingdom being significantly under-sampled. Hence, the percentage of the transcriptome that contributes to evolutionary novelty, especially in plants, is still not clearly understood.

The questions raised above regarding the patterns of evolution of WGD derived duplicate genes and functionality of novel unannotated transcripts can be addressed using a comparative genomics/transcriptomics approach.

COMPARATIVE GENOMICS AND TRANSCRIPTOMICS AS TOOLS TO STUDY GENE CONTENT EVOLUTION

Comparative analysis of species has been used since historical times to understand biological principles. For example, Charles Darwin famously used this strategy to illustrate the concept of evolution in his book *On The Origin Of Species* (Darwin, 1859). After the development of the first DNA and protein sequencing methods, molecular information in the form of nucleotide and amino acid sequences began to be used for understanding relationships between species and even between genes. Such data allowed researchers to develop models of sequence evolution, techniques for assessing selection acting on sequences (Kimura, 1983; Ohta, 1992; Li, 1997) and methods of phylogenetic reconstruction (Felsenstein, 1989). Such methods can now be used to assess whether a sequence shows a signature of functionality. In recent times, comparative genomics and transcriptomics have also received a boost with the advent of technologies such as high throughput sequencing and rapid advances in computing power.

Illumina and 454 sequencing – together referred to as next-generation or second-generation sequencing approaches – have, over the past six years, enabled high-throughput and economical collection of genome and transcriptome data from organisms with or without reference genomes. By 2012, the cost for sequencing using Illumina had dropped to $0.5/Mb (Loman et al., 2012), compared to ~$2000/Mb using traditional Sanger sequencing (Liu et al., 2012). The ability to sequence cheaply has resulted in an explosion of genomic data from a variety of model organisms, permitting comparative genomics between different populations of the same species (Weigel and

Mott, 2009) and between multiple closely related species. Similarly, the availability of transcriptomic data has permitted us to explore gene expression under multiple conditions. In my research, I have used both genomic and transcriptomic data combined with comparative approaches in the plant Brassicaceae and Poaceae families. to address the questions of origination of novel genes and their evolution.

PLANT FAMILIES OF INTEREST

In my research, I have conducted comparative genomic and transcriptomic analyses on two plant families – Brassicaceae and Poaceae.

The Brassicaceae family is a large family of flowering plants consisting of ~330 genera and ~3700 species. The hallmark characteristic of this family is the production of glucosinolate compounds, a group of secondary metabolites useful for herbivore defense. This family contains the model organism *A. thaliana* as well as economically important crops such as *Brassica rapa* (canola)*, B. oleraceae* (cabbage)*, B. napus* (oilseed rape) and *B. nigra* (black mustard). The *A. thaliana* genome, released in 2000, was the first plant genome sequenced and was found to contain 25,498 genes. Over the past decade, ten updates of the genome assembly have been released taking the gene count to 27,416. A wealth of information is available for *A. thaliana* including several RNA expression datasets, DNA methylation and histone maps under specific conditions, gene network data, population structure and genomic data for >500 accessions as well as several experimental tools for genetic manipulation. Hence, *A. thaliana* is an attractive system for comparative genomic and transcriptomic studies.

In addition to *A. thaliana*, the genome of *B. rapa* has also been recently sequenced (Wang et al., 2011). The genome sequence, coupled with information from previous studies, suggests that a genome triplication event (termed as the α' WGD event) occurred ~25 million years ago (mya), after the lineage separated from the Arabidopsis lineage (Chapter 2). To understand the evolutionary patterns of duplicate genes and pseudogenes created as a result of the α' WGD event, we sequenced and annotated the genome of wild radish (*Raphanus raphanistrum*), a Brassicaceae species closely related to *B. rapa* and the cultivated radish *Raphanus sativus*. Given that the α' WGD event was common to both *B. rapa* and *R. raphanistrum*, the genome sequence of these two species allows us to compare and contrast the patterns of evolution of duplicate genes in different lineages.

The second plant family I have studied is the Poaceae family. This monocot family, popularly known as the grasses, is the fifth largest plant family, consisting of >600 genera and >10,000 species spread over most of earth's landmass. It is also the most economically important family with species such as cultivated rice (*Oryza sativa*), wheat (*Triticum aestivum*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), oats and bamboo as members. The genomes of eight Poaceae species have been sequenced so far (*Sorghum bicolor, Zea mays, Setaria italica, Panicum virgatum, Oryza sativa, Brachypodium distachyon, Hordeum vulgare, Triticum aestivum*). Although several studies have looked into the evolution of genes between Poaceae species (Paterson et al., 2009; Schnable et al., 2009; International Brachypodium Initiative, 2010), the evolution of gene expression patterns as well as intergenic transcripts in these species which have large genome sizes, is poorly understood. In my research (Chapter 4), I have made use of high-throughput RNA sequencing data from multiple tissues of four

Poaceae species to understand the characteristics and evolution of genic and intergenic transcription In Poaceae.

SIGNIFICANCE

Over the past ten years, several plant genomes have been sequenced using traditional Sanger-based as well as high-throughput sequencing technologies, revealing the extent of the influence of WGD in shaping plant genomes. Such genome sequence information can now be used to fill in important gaps in our understanding of duplicate gene evolution as a result of large evolutionary distances between sequenced genomes or unavailability of data. In my research (Chapter 2), I make use of four closely related Brassicaceae species to address fundamental questions regarding duplicate gene loss and retention in independent lineages derived from the same WGD event, rates of pseudogenization as well as characteristics of retained duplicates, and ask whether duplicate gene retention can be predicted using the characteristics of genes. Answers to these questions will increase our understanding of duplicate gene evolution in plants. In addition, the radish genome sequence and the transcriptomic resources made available as part of this study will be useful for geneticists and breeders studying radish as well as for comparative genomics in Brassicaceae.

In addition to genome sequencing, high-throughput RNA sequencing allows us to detect several thousand unannotated intergenic transcripts, but whether these constitute novel, functional genes is a matter of intense debate (van Bakel et al., 2010, 2011; Clark et al., 2011; Dinger et al., 2009). Addressing this question is not only important for updating existing gene annotations but also for fine-tuning gene prediction programs, which currently have a high success rate in finding canonical protein-coding genes with hallmark features such as translation start and stop codons, easily distinguishable intron-exon boundaries, protein domains and dinucleotide bias, but perform poorly when identifying small peptides, RNA genes, pseudogenes and other non-canonical genomic features. In addition, identifying non-canonical genes is important for understanding the nature of the genetic variation that, for example, causes a human to appear and behave different from a worm even though both possess a comparable number of canonical protein-coding genes.

Taken together, this research furthers our understanding of gene content evolution in the plant kingdom. My studies of Brassicaceae and Poaceae families, reveal principles that are broadly applicable to all flowering plants.

CHAPTER TWO

Genome sequencing of wild radish and the evolution of polyploidy-derived duplicate genes and pseudogenes in Brassicaceae[1]

[1] The work described in this chapter has been submitted for publication:

Gaurav Moghe, David Hufnagel, Haibao Tang, Yongli Xiao, Ian Dworkin, Christopher Town, Jeffrey K. Conner, and Shin-Han Shiu *(submitted)* The genome sequence of wild radish reveals the patterns of evolution of whole genome duplicate genes and pseudogenes in Brassicaceae.

ABSTRACT

Polyploidization events are frequent among flowering plants; however, whether duplicates of the same event follow similar evolutionary trajectories in independent lineages is not clear. To address this question, we sequenced the genome of wild radish (*Raphanus raphanistrum*), a Brassicaceae species that experienced a whole genome triplication event (referred to as α' WGT), 24-29 million years ago prior to diverging from *Brassica rapa*. We found that ~66% of the orthologous groups experienced gene loss since α' WGT in both species, either via gene deletion or pseudogenization. Although gene deletion may occur immediately after polyploidization, we did not find evidence for a immediate pseudogenization after α' WGT. Among retained duplicates, we found evidence for both sequence and expression level divergence, with sequence evolution being largely consistent between *B. rapa* and *R. raphanistrum*. Analysis of expression levels between *Arabidopsis thaliana* and radish flowers suggested that divergence among duplicates occurs primarily via decrease in flower expression, however, one of the copies still tends to maintain the original expression state. We also asked whether the genes whose WGD paralogs were lost had different characteristics than retained duplicates and found biases in function, sequence composition, expression patterns, network connectivity and rates of evolution. Using a machine learning approach, we then created a framework for predicting whether a duplicate would be retained after WGD. Overall, our study suggests a convergent pattern of duplicate gene loss and retention between *B. rapa* and *R. raphanistrum* and provides new insights into mode of evolution of duplicate genes post polyploidization.

INTRODUCTION

Polyploidization is a frequent occurrence in the plant world. Over 70% of angiosperms are polyploids or have experienced a polyploidization event in their evolutionary history (Ramsey and Schemske, 1998). In addition, ancient polyploidy is common to all flowering plant families and is correlated with dramatic increases in plant species richness in several angiosperm lineages (Soltis et al., 2009). A polyploidization event at least doubles the entire repertoire of a plant's gene content. The duplicated genes may remain functionally redundant briefly, but eventually may gain new functions (neo-functionalization,(Ohno, 1970)), be retained due to partition of ancestral functions (sub-functionalization, (Force et al., 1999)) or be lost via deletion of the gene segment or pseudogenization (Li et al., 1981). The mode of evolution of a duplicated gene pair, especially those derived from polyploidization, has been shown to be dependent on several features, including gene function (Blanc and Wolfe, 2004; Hanada et al., 2008), gene complexity (Chapman et al., 2006; Jiang et al., 2013), levels of gene expression (Pál et al., 2001), dominance of one of the parental genomes (Schnable et al., 2011) and network connectivity (Thomas et al., 2006). Despite correlations of these features with duplicate retention, it remains unclear to what extent these features may allow prediction of duplicate retention. This issue can be addressed in greater detail in Brassicaceae, given the close evolutionary relationship between the Brassiceae tribe species including the wild radish *Raphanus raphanistrum* (RR) and *Brassica rapa* (BR) and the *Arabidopsis* genus (43 mya, (Beilstein et al., 2010)), the recent hexaploidization in the Brassiceae lineage and the availability of a broad range of molecular data in

Arabidopsis thaliana (referred to as AT) that can be used to inferred the potential roles of Brassiceae duplicates.

In Brassicaceae, studies of duplicate genes in AT suggest three rounds of whole genome duplication (WGD) after its lineage diverged from the monocot lineage. The most recent WGD event (α) occurred 50-65 million years ago (mya) (Bowers et al., 2003; Beilstein et al., 2010), prior to the divergence of species in the Brassicaceae family. Notably, a further hexaploidization event (referred to as the α' whole genome triplication (WGT) event) occurred recently in the common ancestor of BR and RR (Lagercrantz and Lydiate, 1996; Lysak et al., 2005; Yang et al., 2006; Town et al., 2006; Wang et al., 2011). Among Brassiceae species, much of the knowledge about the evolution of α' duplicates is derived from species in the *Brassica* genus, particularly the recently sequenced BR genome (Wang et al., 2011). Since the occurrence of a hexaploidization event in BR's evolutionary history, >50% of the duplicated genes may have been lost via the processes of deletion and pseudogenization (Wang et al., 2011). Among retained duplicates, those involved in transcriptional regulation, hormonal signaling and response to environmental stresses were found to be over-represented (Wang et al., 2011). These findings provide a baseline understanding of WGD duplicate evolution. They also led to the question if the pattern of duplicate gene evolution will be similar in another Brassiceae tribe species that also experienced the α' event. To address this question, we chose to sequence the genome of wild radish, a species closely related to the *Brassica* species (Arias and Pires, 2012).

RR is a relative of the cultivated radish (*R. sativus*), a commercially important crop consumed primarily in Asia. RR, which is native to the Mediterranean region and likely the ancestor of *R. sativus*, has evolved a weedy form that has become a serious global agricultural pest. RR is difficult to control owing to prolific seed production and high levels of resistance to drought and herbicides (Warwick and Francis, 2005). Wild radish is also a model system in ecology and evolution (Conner, 2002; Conner et al., 2009). Availability of genomic and transcriptomic resources for *Raphanus* will contribute to a better understanding of the molecular basis and evolutionary characteristics of weediness as well as aid in improvement of cultivated radish. In addition, these resources enables comparisons of two post α' WGT species allowing us to pinpoint common and divergent trends in duplicate evolution.

In this study, we report a draft assembly and annotation of the RR genome. After establishing the orthologous relationships between AT, *A. lyrata* (AL), BR and RR genes, we asked three major questions: (1) Did evolution of duplicate genes post α' WGT follow similar trajectories in BR and RR? (2) What are the patterns of pseudogenization and duplicate gene divergence since the α' WGT event? and (3) Comparing properties of duplicates derived from α WGD and α' WGT, can we predict which genes would be retained or lost? Our results suggest that the evolution of WGD duplicate genes follows similar trends of losses and retention in independent descendant species and that the retention process may possess certain biases which can be uncovered through computational modeling.

RESULTS AND DISCUSSION

Sequencing and assembly of the wild radish genome

As the first step in creating a draft assembly for the RR genome, we estimated the genome size of RR using flow cytometry *(see Methods)*. The estimated size of 515 Mb is comparable to genome size estimates of related species including BR (529 Mb), *Brassica oleraceae* (696 Mb) and *Raphanus sativus* (573 Mb) (Johnston et al., 2005). Because RR is an obligate out-crosser with high heterozygosity, we sequenced the genome of a 5th generation inbred plant using paired end Illumina and mate-paired 454 sequencing strategies at 47X and 2.5X coverage of the estimated genome size, respectively. Reads were assembled with a hybrid approach using multiple assembly programs (see Methods, Figure 2.1). The final assembly size of 254 Mb represented 49.3% of the estimated genome size, with a N50 contig size of 10.1 kb (Table 2.1). This is comparable to the draft BR genome where the assembly is 283.8 Mb or 53.7% of the estimated genome size despite its significantly better sequencing coverage at 72X (Wang et al., 2011).

The reason that our genome assembly size is only around half of the RR genome size is likely because a large proportion of the missed sequence was highly repetitive and/or heterochromatic. The size of the euchromatic space in BR is estimated to be ~220 Mb (Mun et al., 2009). In addition, ~30% of all BR chromosomes are comprised of centromeric repeats that occupy ~50% of all heterochromatic domains (Lim et al., 2007). Assuming that most of this heterochromatin consists of repetitive, non-genic regions and RR is similar to BR in its heterochromatin content, it is likely that we captured most of the genic space in our RR assembly. The coverage of the gene space in our RR and the published BR assemblies was further assessed using Expressed Sequence Tags (ESTs) and using the Core Eukaryotic Gene Mapping Approach (CEGMA) (Parra et al., 2007). We found that 93.3% and 78.4% of the BR and RR ESTs could be mapped on to their cognate assemblies (see Methods, Table 2.1). In addition, the BR and RR assemblies contained complete matches for 248 (100%) and 241 (97.2%) CEGMA proteins, respectively (Table 2.1). These observations suggest that the RR assembly is less complete than BR. However, a significant proportion of the gene space in RR is covered in the draft assembly. Using the MAKER annotation pipeline (Cantarel et al., 2008), we predicted 38,174 proteins in the RR assembly (see Methods, Figure 2.2 AB). Finally, for comparing the gene space across species, we employed a combination of similarity-based as well as synteny-based approaches to define orthologous groups (OGs) between AT, AL, BR and RR protein-coding genes (see Methods).

To understand whether duplicate genes in BR and RR have evolved independently, it would be important to know the BR-RR speciation time in relation to the timing of the α' WGT event. Using our definitions of orthologous and paralogous relationships between genes the four Brassicaceae species, we first estimated the divergence time between BR and RR.

Timing the speciation and polyploidization events in Brassicaceae

In order to determine the amount of time for which the evolution has been occurring independently in BR and RR, we first sought to estimate the timing of the BR-RR speciation event in the context of the timings of the α' WGT and AT-BR speciation events. Previous studies have suggested a broad range of timings for speciation and

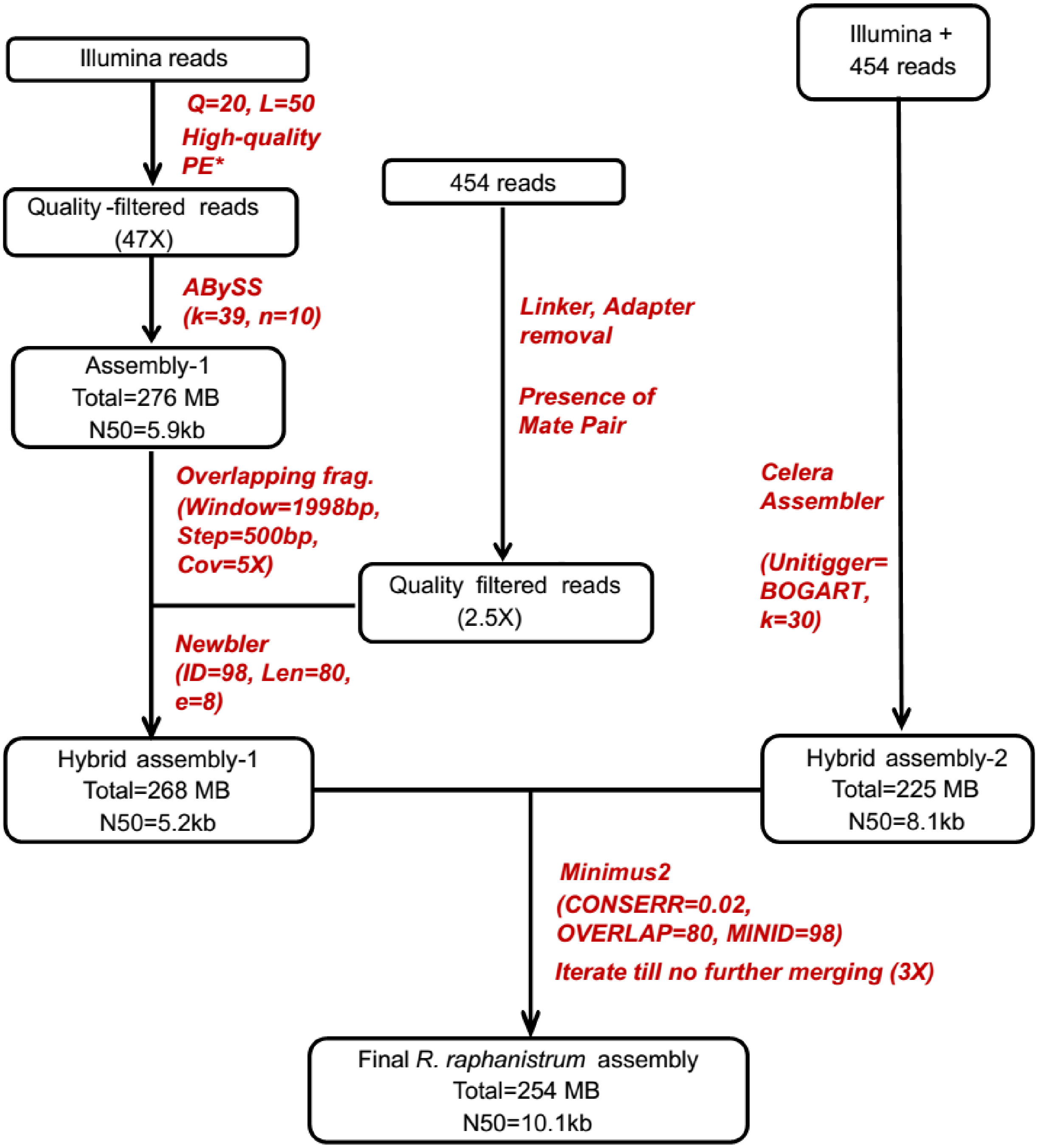


Figure 2.1: Pipeline implemented for assembling the RR genome. Software and parameters used for each step are noted in red. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Table 2.1: Comparison between RR and BR assemblies

	R. raphanistrum contigs	*B. rapa* contigs	*B. rapa* scaffolds
Sequencing technology and coverage	47X Illumina, 500bp PE + 2.5X 454, 3kb mate pair	72X Illumina, multiple insert sizes + Sanger BAC-end sequences	
Assembly size	254.6 Mb	264.1 Mb	283.8 Mb
Number of contigs	68,331	60,521	40,549
N50	10.1 kb	27.2 kb	1.9 Mb
Median contig size	1166 bp	173 bp	140 bp
Completeness of highly conserved eukaryotic genes[1]	97.2%	NA	100.0%
% consensus transcripts mapping to assembly[2]	78.4%	NA	93.3%

[1] Conservation of 248 Core Eukaryotic Genes (CEGs) in the R. raphanistrum assembly.

[2] 150,524 Raphanus and 85,508 Brassica ESTs were downloaded from NCBI dbEST and merged into 83,214 and 79,830 unique consensus transcripts using a custom merging pipeline (see Methods).

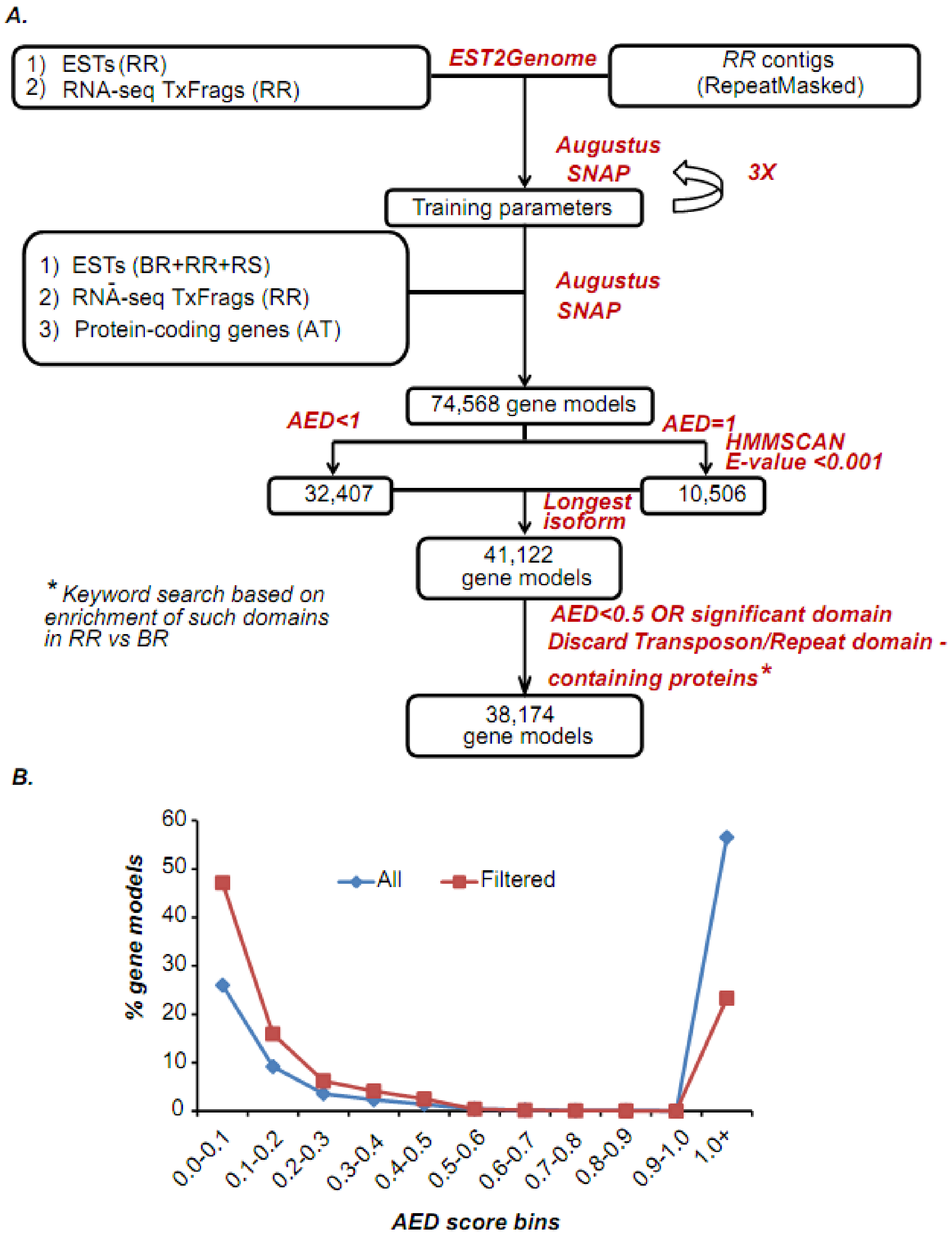
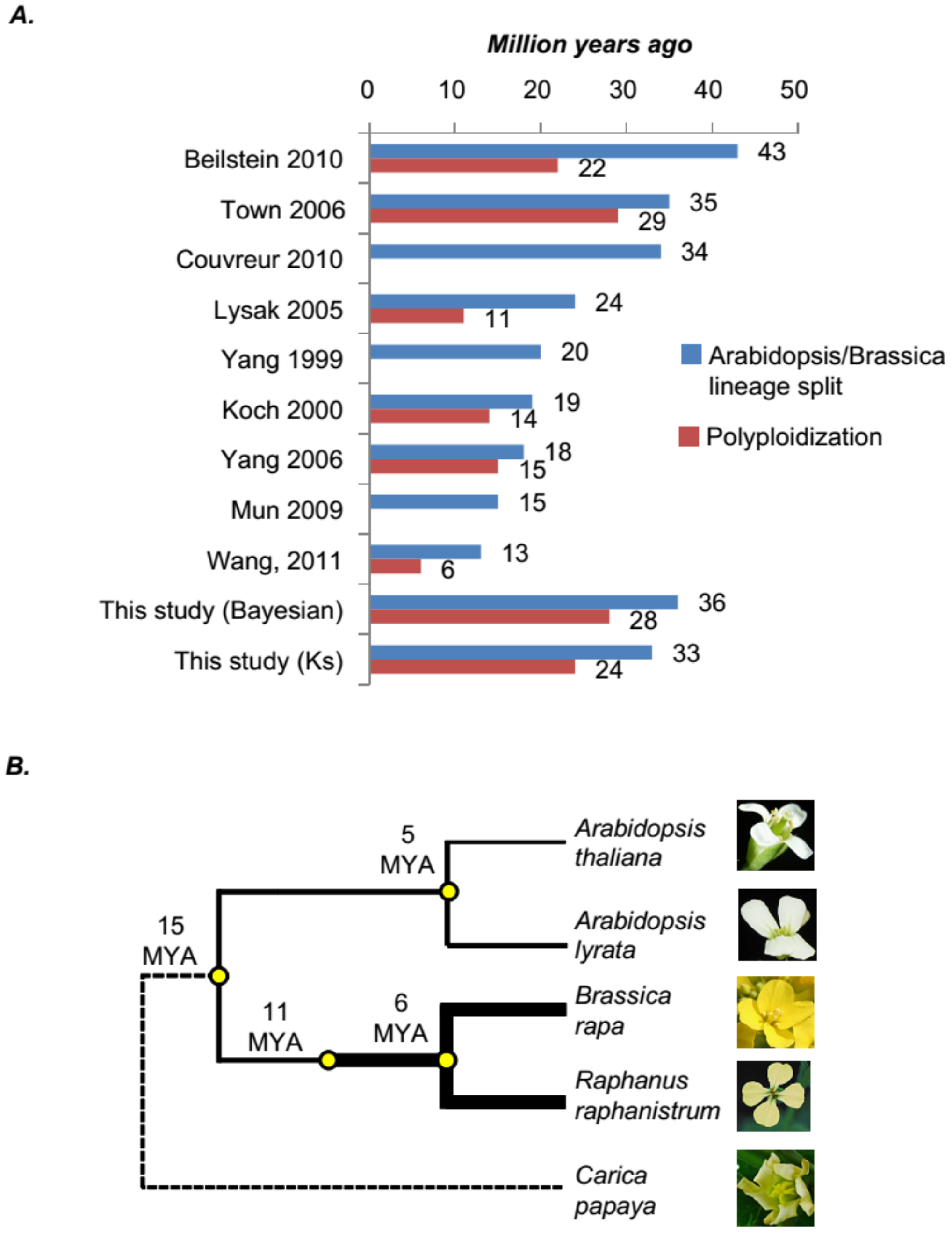


Figure 2.2: Gene prediction pipeline. (A): All protein domains related to repetitive elements were discarded in the last step. (B): Distributions of the Annotated Edit Distance (AED) values before and after the penultimate filtering step.

the WGT events in the Brassicaceae family (Figure 2.3A, (Yang et al., 1999; Koch et al., 2000; Lysak et al., 2005; Town et al., 2006; Mun et al., 2009; Couvreur et al., 2010; Beilstein et al., 2010)) and some of these estimates have been revised based on availability of new data (Beilstein et al., 2010). In the absence of consistent times and due to the methodological differences between these studies, we re-estimated the timing of α' WGT event using most updated data in addition to estimating the timing of the BR-RR speciation event.

In this study, using a lower limit of AT-BR divergence time of 30 mya (Beilstein et al., 2010) as well as a neutral substitution rate of $7*10^{-3}$ substitutions/site/million years (Ossowski et al., 2010), we performed Bayesian dating with a prior of 36 mya for the AT-BR divergence time (Town et al., 2006). We also obtained divergence times based on the synonymous substitution rate (*dS*) (Figure 2.4A). Using these two methods, we estimated the median divergence time between BR and RR to be 13-19 mya, prior to the divergence of AT and AL (10-11 mya) and much later than the divergence time between AT-BR lineages (32-36 mya) (Table 2.2). These estimates are significantly older than some of the previous estimates (Figure 2.3A), partly due to the prior and the lower limit for the divergence between AT and BR/RR lineages set at 36 and 30 million years respectively based on most recent fossil data, and partly due to our use of a lower neutral substitution rate than that used by Koch et al. (Koch et al., 2000) (Figure 2.3B). Our estimates are most similar to three other studies (Town et al., 2006; Couvreur et al., 2010; Beilstein et al., 2010). Using α' WGT-derived BR and RR duplicates, we estimated that the WGT event took place 24-29 mya (Figure 2.4B). Taken together, our results suggest that the polyploidization event likely occurred 3 to 12 million years after



Timing of events based on Ks values, assuming a substitution rate of 15*10e-3 substitutions/site/million years (Koch et al, 2000)

Figure 2.3: Divergence time estimates (A): Timing of the AT-BR split and the triplication event, as per previous studies. (B): Timing of various events based on the formula T= *dS* /(2 x Rate) using a rate of 15x10^{-3} substitutions/site/million years (Koch et al., 2000).

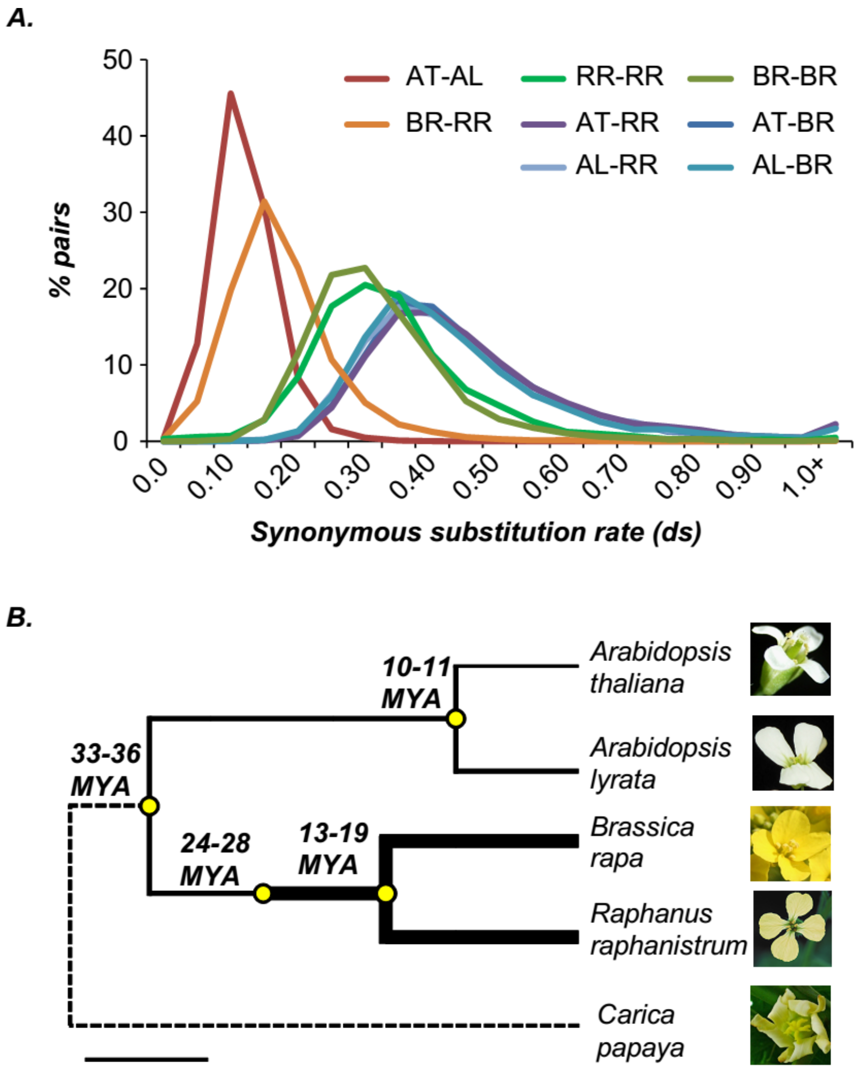


Figure 2.4: Relationships between Brassicaceae species. (A): Synonymous substitution rate (*dS*) distributions between pairs of orthologs and paralogs in Brassicaceae species. (B): Timing of various events, indicated by yellow dots, in the Brassicaceae family. The lower number for each time corresponds to the median time obtained using the formula T=(*dS*/2*rate) while the upper number corresponds to the median time obtained using a Bayesian dating approach (multidivtime). Thickness of the lines corresponds to the estimated genome sizes, assuming an ancestral genome size of 200 Mb. The image for *A. lyrata* is copyrighted (© Ya-Long Guo, Max Planck Institute for Developmental Biology) and used with permission.

Table 2.2: Descriptive statistics of the speciation and WGD times

	Multidivtime					***Synonymous rate (dS)***			
	Mean	***Median***	***95% CI of mean***	***SD***		***Mean***	***Median***	***95% CI of mean***	***SD***
AT-AL	11.1	11.3	10.5-11.2	1.3		11.3	10.1	10.2-12.5	46.1[1]
AT-BR	36.7	36.5	36.5-36.8	1.6		34.3	31.5	34.0-34.6	13.0
AT-RR	Same as AT-BR					35.2	32.1	34.9-35.6	15.0
BR-BR	27.4	28.2	27.1-27.7	4.5		24.5	23.1	24.2-24.6	7.6
RR-RR	27.8	29.0	27.4-28.2	4.9		26.4	24.9	26.1-26.9	11.1
BR-RR	19.0	18.8	18.5-19.7	5.4		14.4	13.5	14.3-14.6	7.5

[1] *SD is high due to the large deviation of the AT-AL dS distribution from normality*

the separation of the AT-BR lineages, and that the *Raphanus* genus may have been diverging from BR for a longer time than previously estimated (Yang et al., 2002; Lysak et al., 2005). In addition, we can also surmise that the α' duplicates may have shared 5-16 mya of common descent, followed by 13-19 mya of independent evolution in BR and RR. Thus, the BR and RR duplicates may be used to assess whether the patterns of retention and loss of genes post WGD occur in a similar fashion in related post WGD species.

Patterns of loss and retention of duplicate genes post α' WGT

AT and AL have 27,416 and 32,670 annotated genes, respectively. Assuming that the common ancestor of AT/AL/BR/RR had ~30,000 genes, the α' event should have created ~90,000 genes. Considering that there are 41,174 BR and 38,174 RR genes annotated, only ~42-45% genes in the ancestral hexaploid are retained. The extent of gene loss is evident at the protein domain level because there are on average 1.4 times more domain family members in both BR and RR vs AT instead of three times more (Figure 2.5A). After the α' event, the BR and RR lineages have evolved independently for 13-19 million years. Thus the patterns of gene retention and loss may have followed different trajectories in these species. There can be two extreme scenarios. The first is a completely random pattern of duplicate retention and loss in these two lineages. The second is that a gene retained in BR is always retained in RR. To see which scenario better describes lineage-specific evolution of α' duplicates, we examined the patterns of duplicate gene retention at the orthologous group (OG) level.

Each OG specifies one ancestral gene common between AT, AL, BR, and RR. In addition, the phylogenetic relationships of genes in an OG provide information of

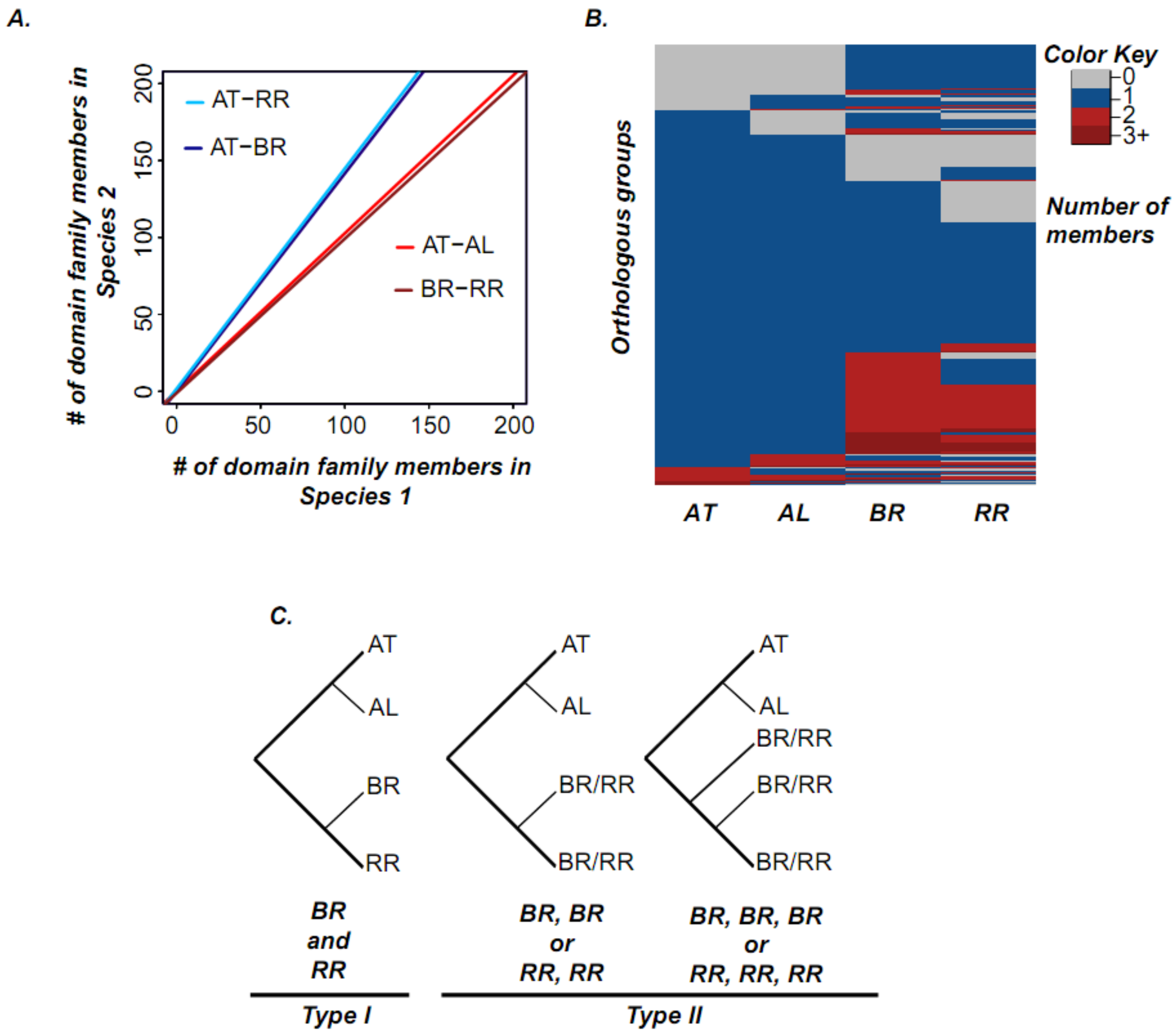


Figure 2.5: Patterns of evolution of α' duplicates. (A): Comparison of PFAM domain family memberships between pairs of species. (B): Comparison of orthologous groups between the four species, indicating a preponderance of unexpanded or lost OGs among all orthologous groups. (C) Schematic representations of Type I and Type II. Tree structures which could not be classified into Type I and Type II were classified as Type III.

speciation and α' duplication nodes that allow inference of whether α' duplicates are retained or lost. We identified 16,567 OGs containing high-confidence BR and RR genes derived from the α' WGT event (see Methods, Figure 2.5B). Using these OG definitions, we first asked whether α' duplicates tend to be lost in parallel between BR and RR i.e. in cases where BR duplicates are lost, are the RR duplicate lost too (and vice versa)? We found that in 10,521 and 8871 OGs where BR and RR duplicates, respectively, returned to a singleton state, 6235 (70.3%) cases were common, significantly higher than random expectation (Fisher Exact Test p<1e-16), suggesting occurrence of parallel losses in BR and RR.

For further analyses, we classified the OGs into three types (Figure 2.5C). The first type consists of 4702 OGs where α' duplicates are mostly retained in BR and/or RR. Specifically, type I OGs are defined as those with 1 member each from AT and AL and 2 or 3 members from BR or RR. The type II OGs (2534) are those with no α' duplicate retained: only 1 member from AT and AL and 1 member from BR and RR. The number of type II OGs is significantly lower than the common singleton cases reported above (6235) because we also excluded OGs where the BR or RR genes had putative tandem or segmental duplicates (see Methods). Such OGs were included in type III, which consists of the rest of OGs (9331). BR and RR genes in type I and type II OGs are referred to as retained duplicates and singletons, respectively. We found that retained duplicates tend be involved in biotic and abiotic stress response, hormonal signaling, development as well as regulation of transcription compared to singletons (Table 2.3). In contrast, singletons were enriched in processes such as DNA repair, cell division, metabolic processes as well as RNA modification and processing (Table 2.3).

Table 2.3: Gene Ontology categories over-represented as per Fisher Exact Test

GO categories over-represented in retained duplicates compared to singletons	***GO categories over-represented in singletons compared to retained duplicates***
Response to salt stress	Metabolic process
Response to chitin	Oxidation-reduction process
Response to cadmium ion	DNA repair
Response to cold	RNA processing
Response to wounding	RNA methylation
Defense response to fungus	RNA modification
Response to nematode	RNA processing
Response to water deprivation	Transcription initiation, DNA-dependent
Regulation of transcription	Regulation of transcription by RNA Pol II
Positive regulation of transcription	Embryo development ending in seed dormancy
Negative regulation of transcription	Cell division
Response to auxin stimulus	Vegetative to reproductive phase transition in meristem
Response to jasmonic acid stimulus	Proteolysis
Response to salicylic acid stimulus	Protein peptidyl-prolyl isomerization
Response to ethylene stimulus	Thylakoid membrane organization
Response to gibberelin stimulus	GPI anchor biosynthetic process
Jasmonic acid signaling pathway	
Auxin efflux	
Cytokinin mediation signaling pathway	
Unidimensional cell growth	
Stomatal movement	
Plant cell wall loosening	
Regulation of timing of reproductive trans.	
Photomorphogenesis	
Multicellular organismal development	
Seed development	
Floral organ abscission	
Seed germination	
Root hair cell development	
Secondary cell wall biogenesis	
Protein phosphorylation	
Small GTPase mediation signal transduc.	
Activation of MAPKK signaling	
MAPK cascade	
Circadian rhythm	
Transmembrane protein transport	
Glycolysis	
Fatty acid biosynthetic process	

Table 2.3 (cont'd)

GO categories over-represented in retained duplicates compared to singletons	***GO categories over-represented in singletons compared to retained duplicates***
Chlorophyll biosynthesis	
Cellulose biosynthesis	
Carbohydrate biosynthesis	

Overall, we find that a large percentage of OGs (~70%) experienced losses in BR and RR, returning them to a singleton state or deleting the entire gene lineage altogether. Such a behavior may be expected as the polyploid returned to a diploid state over the past 25 million years. The process of gene loss may occur via complete deletion of the gene segment or via accumulation of mutations leading to pseudogenes that can be identified. To better understand the process of gene loss, we identified pseudogenes in the BR and RR genomes to address questions regarding the properties of pseudogenes and the timing of their pseudogenization.

Pseudogenization of duplicate genes and timing of pseudogenization

Comparison of PFAM domain family sizes and sizes of OGs between BR/RR and AT/AL, suggests that extensive gene losses and pseudogenization have occurred after the α' WGT in the BR/RR lineage. To estimate the extent of pseudogenization and assess the properties of pseudogenes, we identified 39,659 BR and 21,226 RR pseudogenes that are fragments of their paralogs and/or contain premature stops/frameshifts (Figure 2.6A,B). Overall, the putative pseudogenes in BR and RR had significantly higher *dN/dS* values compared to functional ortholog and paralog pairs (KS test p<1e-15, Figure 2.6C), consistent with their assignment as neutrally evolving pseudogenes. However, some pseudogenes also had *dN/dS* values comparable to functional duplicate genes. These pseudogenes contain in-frame stops and/or frameshifts or are short fragments (Figure 2.6B), suggesting that they are not simply false positives but may have been created recently.

In addition, when compared to their functional paralogs, we find that 10,778 BR (13.4%) and 4898 RR (21.6%) had their *dS* values between 0.2 and 0.6, the 25th and 75th percentiles of the duplicate gene *dS* distributions. These pseudogenes could potentially have arisen via the α' WGT event.

Studies on synthetic polyploids suggest that newly formed polyploid undergoes rapid genomic arrangements in the first few generations, which can result in instantaneous loss of several thousand genes from the genome via deletion (Tian et al., 2010; Matsushita et al., 2012). In addition, nonsense/frameshift/indel mutations may also accumulate in the gene body leading to pseudogenization of gene copies. It is not known whether pseudogenization of whole genome duplicates, like deletion of genes, occurs rapidly after the polyploidization event or whether the fraction of duplicates that escapes deletion can be tolerated for some time. To address this question, we estimated the timing of pseudogenization for the pseudogenes derived from the α' WGT event.

Firstly, we stringently defined pseudogenes derived from α' WGT as those lying in homeologous regions with their functional paralogs (see Methods). Thus, 2268 BR and 1261 RR pseudogenes were identified as α' derived pseudogenes. To estimate timing, we used a previously published method (Chou et al., 2002) with the assumption that before pseudogenization, the two duplicate genes experienced the same degree of selective constraint and that when one of the duplicates became a pseudogene, the pseudogenized copy evolved neutrally (see Methods, Figure 2.7A). Based on these assumptions, the estimated timing of pseudogenization indicates that over the past 25-30 million years, pseudogenization may have occurred gradually or at a steadily

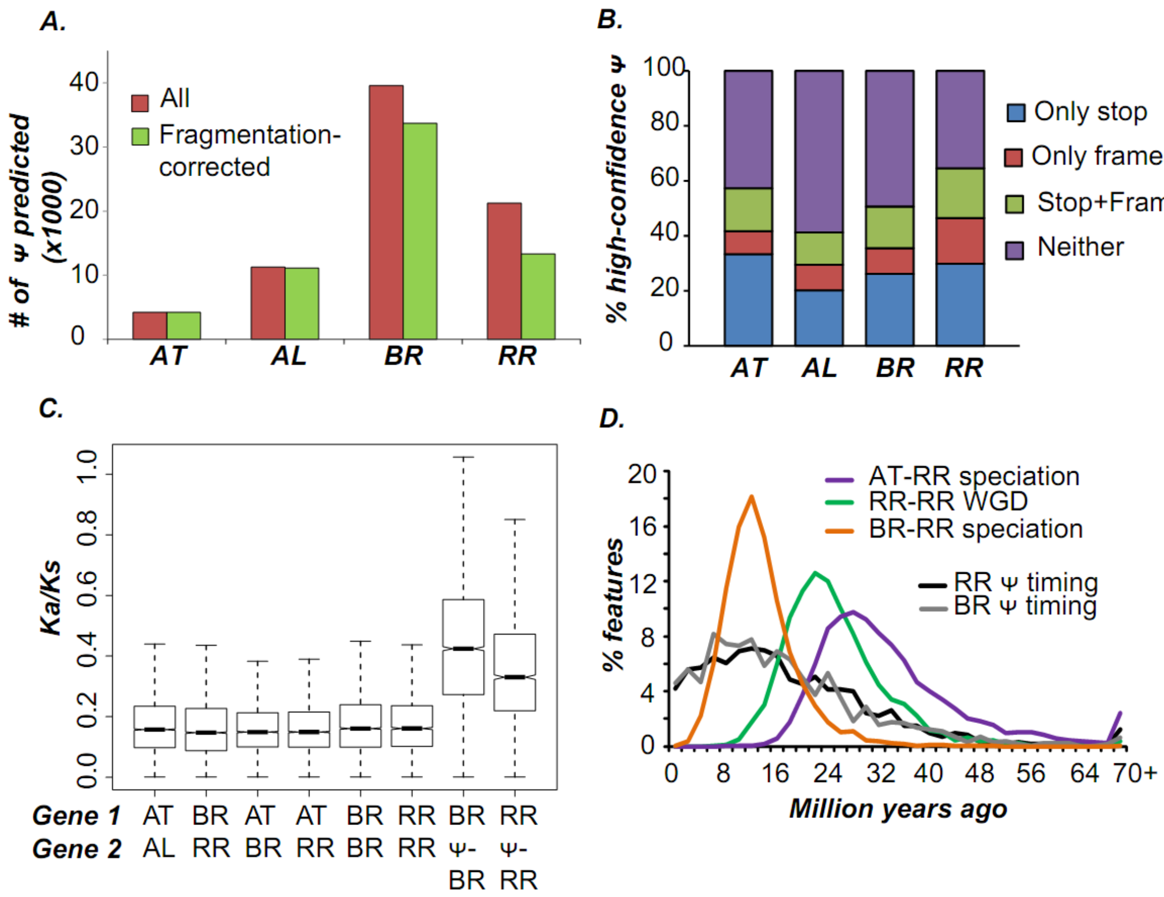


Figure 2.6: Patterns of pseudogenization in Brassicaceae species. (A): Number of pseudogenes (ψ) predicted in each species, before (red) and after (green) correcting for the fragmented nature of the genomic assemblies. (B): Comparing selective constraint between each pair of functional homologs and between functional gene- pseudogene pair shows that pseudogenes are under significantly relaxed constraint than functional genes. (C): Timing of pseudogenization (black and gray lines) compared to timing of other events. The distribution of the times when the functional ancestors of the pseudogenes were duplicated, based on *dS*, is shown as a dotted black line.

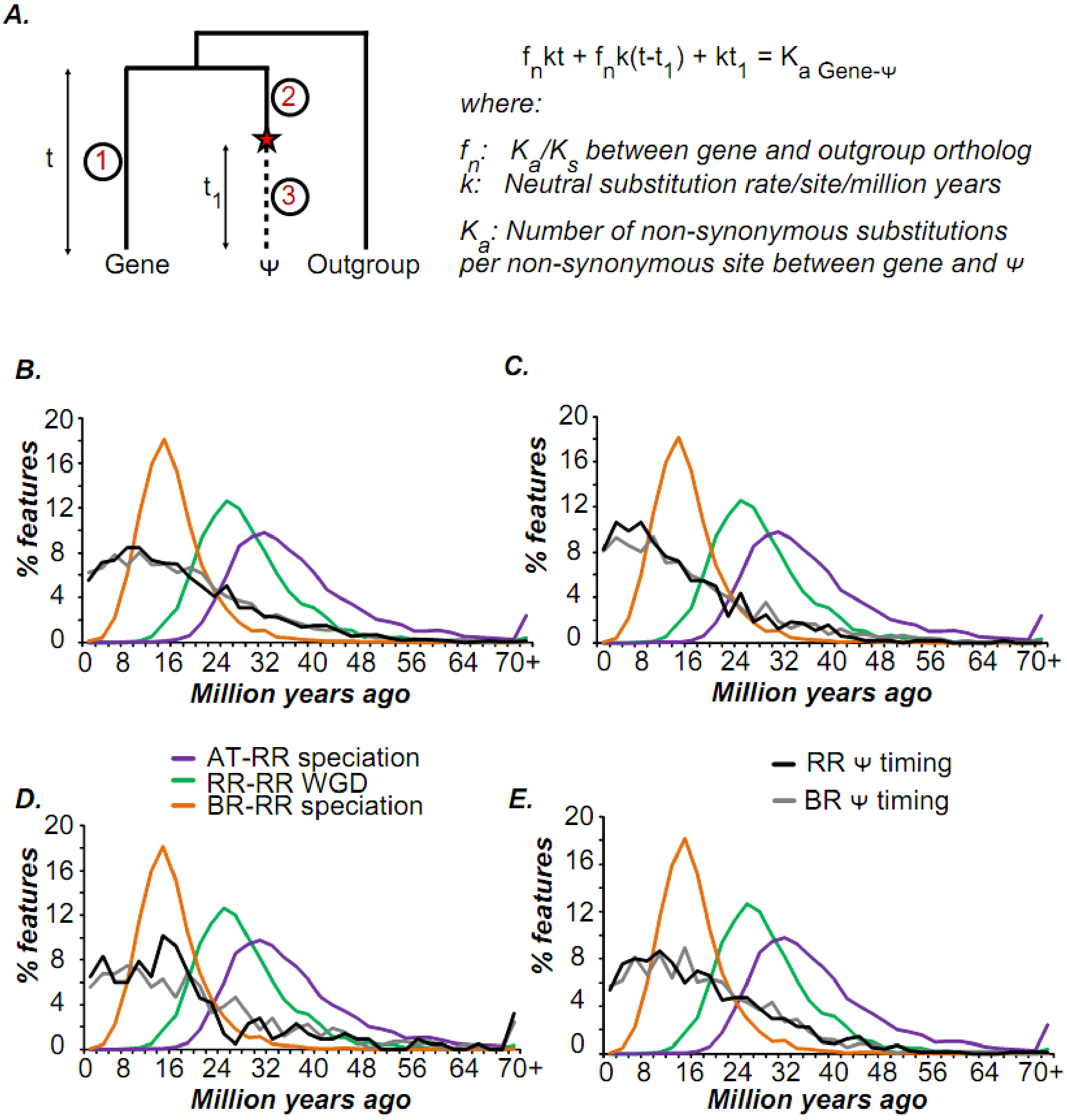


Figure 2.7: Patterns of pseudogenization in studied species. (A): Schematic representation of the formula used for estimation of timing. The red star represents the pseudogenization event. (C-F): Pseudogenization timing distributions using different

Figure 2.7 (cont'd)

criteria for a) choosing WGT derived pseudogenes and b) estimating timing. See Methods for more details.

increasing rate (Figure 2.6C). The behavior of both BR and RR pseudogenes was very similar, suggesting similar selective constraints on genes acting in both species independently. We cannot distinguish between a constant pseudogenization rate and a steadily increasing pseudogenization rate owing to the errors associated with estimation of divergence times using the molecular clock assumption. For example, the lower pseudogenization rates observed between 25-35 mya could simply be due to estimation errors as opposed to a true low initial pseudogenization rate. In addition, the choice of thresholds for defining α' derived pseudogenes was found to affect slightly alter the shape of the timing distribution (see Methods, Figure 2.7B-E). However, regardless of the thresholds used, we do not see any evidence of rapid pseudogenization immediately post α' WGT, suggesting that the triplicated gene content which escaped deletion in the neopolyploid ancestor of BR and RR may have been tolerated for some time.

Sequence divergence of duplicate genes post α' WGT

Although a large proportion of the triplicated gene content has been lost, ~15% of the duplicates are still retained. Given that ~27 million years have elapsed since the α' WGT event, these retained duplicates may sub-functionalize or neo-functionalize over time via expression or sequence divergence. Such sequence divergence may be the consequence of accelerated evolution in one gene over the other. To differentiate whether the α’ derived duplicates evolved at a similar or distinct rate, we used relative rates test (Goldman and Yang, 1994) to compare BR and RR paralogs using their AT ortholog as an outgroup (see Methods). We found that the rate of neutral evolution based on synonymous sites at the third codon position was similar between almost all duplicate gene pairs, with only ~4% gene pairs evolving asymmetrically (Figure 2.8A). On the other hand, when considering coding and amino acid sequences, 10-13% BR and 11-18% RR gene pairs evolved asymmetrically, respectively. Most α' duplicates appear to evolve at uniform rates (~83-87% for amino-acids) (Figure 2.8A). We also found that distributions of fold-difference between branch-wise *dN/dS* for the asymmetrically evolving pairs was significantly greater than that for the symmetrically evolving pairs (KS test p<1e-15) (Figure 2.8B), suggesting that the asymmetry in amino acid and nucleotide substitution rates is associated with a significant relaxation of selection on one of the branches, similar to observations in yeast (Fares et al., 2006). We did not find any functional bias (based on GO biological process categories) among gene pairs evolving asymmetrically. However, of the 443 and 491 OGs to which the BR and RR asymmetric duplicates belonged to, 159 (36.9%) OGs were the same. The overlap was highly statistically significant, since a simulation performed for 100,000 iterations by randomly picking 443 and 491 OGs could only find a maximum of 16.5% overlap across all iterations. This finding suggests that a significantly high number of OGs underwent parallel instances of asymmetric evolution in BR and RR, indicating a potential bias for some duplicates to evolve asymmetrically. In addition, some gene pairs may have undergone asymmetric evolution in the shared lineage of BR and RR.

In cotton, comparison of 16 duplicated genes between allotetraploid cotton and its diploid progenitors found no evidence for aymmetric evolution between duplicates post polyploidization (Cronn et al., 1999). Analyses of protein evolution rates in AT duplicates derived from the α duplication event also indicated that <20% of the 833 pairs analyzed were evolving asymmetrically from each other at the sequence level

(Blanc and Wolfe, 2004). In contrast, another study in yeast found ~60% of the WGD derived duplicates experienced asymmetric rates (Byrne and Wolfe, 2007). Our findings that only 15-20% of the α' WGT duplicates evolved asymmetrically are more consistent with study in cotton and AT. Such asymmetry coupled with a relaxation of selection on one branch may lead to eventual pseudogenization of the gene on the leaf node. However, neo-functionalization or, in some cases, sub-functionalization may also be the likely fates (He and Zhang, 2005; Hahn, 2009). Which of these scenarios predominate among duplicates evolving asymmetrically remains to be understood. On the other hand, the >80% of the duplicate gene pairs which did not show asymmetry appear to be constrained to a similar extent at the sequence level (Figure 2.8B). These pairs may still diverge from each other via accumulation of distinct mutations and/or via expression divergence. The fact that they appear to be under more selective constraint than asymmetric pairs makes sub-functionalization the most likely scenario for retained duplicates, at least at the sequence level, although neo-functionalization or a mixture of neo and sub-functionalization cannot be completely ruled out. Also, given the plasticity of gene expression, regulatory variation may play a bigger role in functional divergence among WGD duplicates than sequence divergence (Blanc and Wolfe, 2004).

Our comparative analyses of BR and RR gene content indicates that retained duplicates in both genomes experienced similar patterns of asymmetric evolution post WGD. In addition, the timings of pseudogenization of BR and RR pseudogenes also followed similar distributions. Given BR and RR have been evolving independently since the past 13-19 mya, these results suggest that process of duplicate gene retention and loss may have exhibited similar biases in both BR and RR lineages.

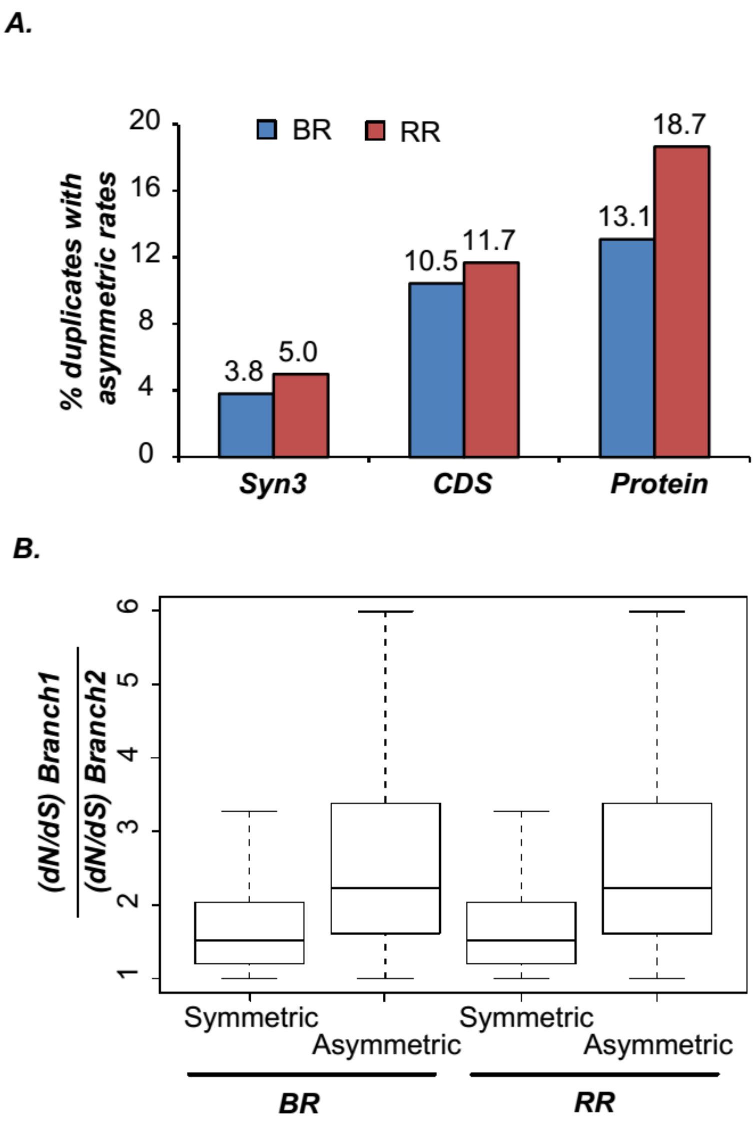


Figure 2.8: Relative rate of evolution of α duplicates. (A): Results of relative rates test between α' duplicates. Syn3 corresponds to synonymous sites at 3rd codon position. See Methods for more details. (B): Distributions of fold differences between degrees of constraint on the two branches leading to duplicate genes.

Expression divergence between α' duplicates

To understand the extent of expression divergence among duplicates, we focused on expression level in a single developmental stage (flower) and asked if, compared to their AT orthologs, RR genes showed signatures of expression level divergence and whether the patterns were different between 1:1, 1:2 and 1:3 AT:RR OGs. Based on the quintiles of the expression distribution of all AT and RR genes, we partitioned their expression levels into five states – Very Low (VL), Low (LO), Medium (MD), High (HI) and Very High (VH) – as well as a sixth Not Expressed (NE) state, and examined transitions between states for pairwise AT:RR comparisons, with the assumption that the AT gene expression level represented the ancestral expression state of the RR duplicates (see Methods). Our results indicate that for 1:1 OGs, expression level is significantly conserved – ~30% of all RR genes show conservation of the same state as the AT gene (10<z-score<35; Figure 2.9A). On the other hand, the significance of enrichment drops substantially as we move to 1:2 and 1:3 OGs, where transitions are more frequent, with more transitions occurring to lower levels of expression (Figure 2.9B). These results suggest that RR genes in 1:2 and 1:3 OGs have experienced expression level divergence since the WGT event while those in 1:1 OGs tend to conserve their expression level in floral tissues.

Overall, we find that most instances of expression divergence in 1:2 and 1:3 OGs occur via expression loss in one of the branches (Figure 2.9A,B). Based on random expectation, expression gain and loss are equally likely, however, more expression loss is observed than random expectation (Figure 2.9A,B). Expression loss in 1:2 and 1:3 OGs is also associated with a higher proportion of RR genes with HI and VH expression states in these OGs than 1:1 OGs and random expectation (Figure 2.9C). Based on our previous observations (Figure 2.9A), the expression level in genes that lose expression does not go to zero; instead, there is just a decrease in expression level. Assuming these RR genes are still functional, the most likely explanation is that they are now expressed in some other tissue/condition, resulting in sub- or neo- functionalization. Interestingly, we find that one of the copies in 1:2 and 1:3 OGs has divergence similar to the copy in 1:1 OG, suggesting that despite a general trend towards divergence, expression in flowers may still be conserved in these OGs. Further partitioning expression conservation among branches, we find that the AT expression level is conserved on at least 1 branch in ~60% and ~70% of 1:2 and 1:3 OGs, respectively. In addition, the possibility that all existing copies in RR diverge in expression >2-fold (no branch conserved) from AT is highly under-represented in all three OG types (-35 <z-score< -7). These results suggest that although 1:2 and 1:3 OGs diverge more in expression than 1:1 OGs, one of the copies is significantly more likely maintain the ancestral expression state while the other copies are free to diverge in expression level and perhaps, tissue-specificity.

Comparing between sequence and expression evolution, we find that most divergence between duplicates occurs at the level of expression, since ~70% of the AT-RR pairs were found to have >two-fold divergence in expression level. On the other hand, asymmetric sequence divergence occurs only in ~15% of the duplicates. Although the sequence evolution pattern suggests sub-functionalization of retained duplicates in BR and RR, a broader expression sampling of multiple tissues would be needed to determine the predominant mode of duplicate evolution.

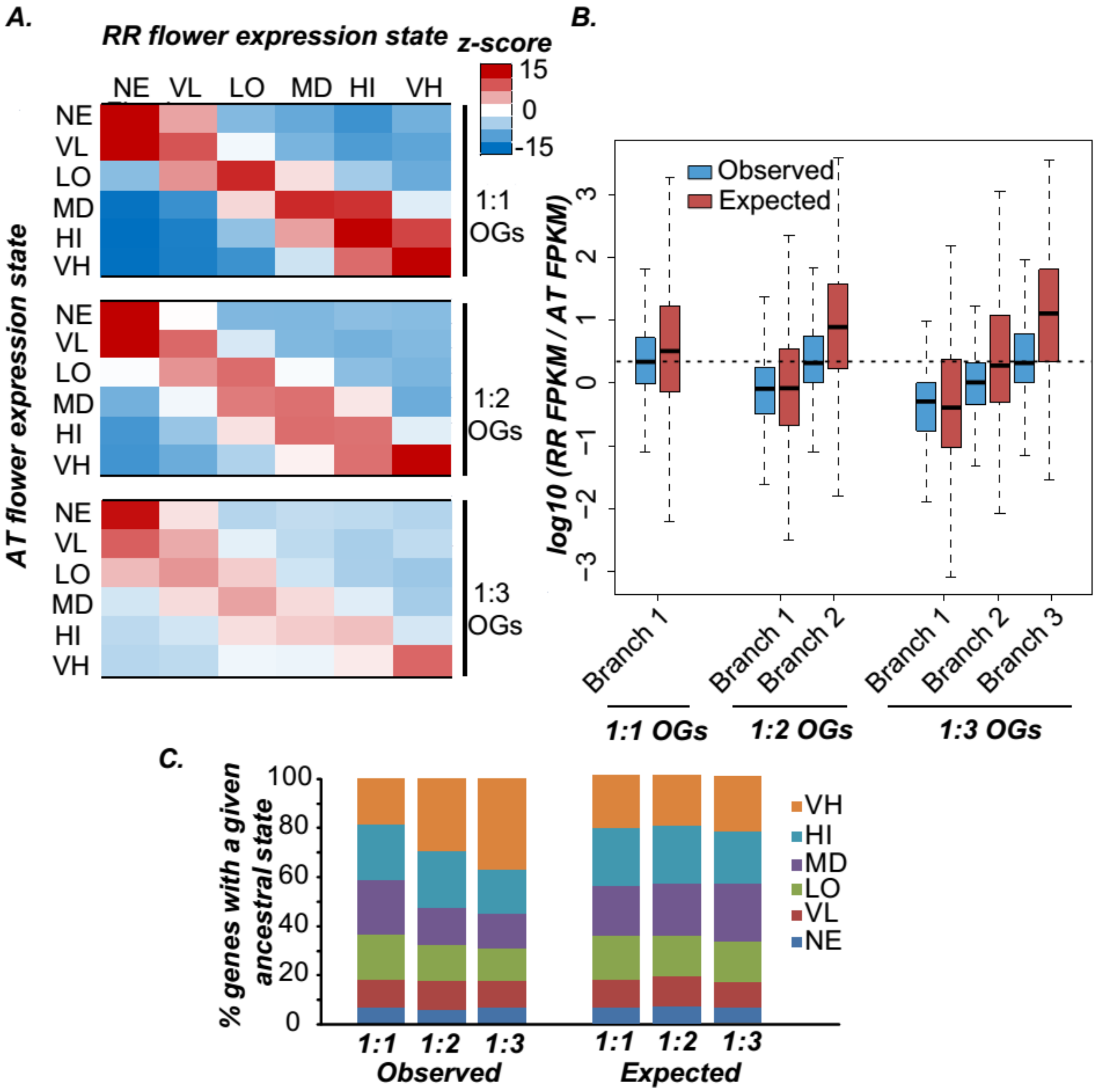


Figure 2.9: Expression divergence of α' duplicates. (A): Significance testing of % overlaps between AT and RR expression states. To obtain percentages, for each AT gene with a given expression state, the number of RR genes having each of the six expression states was determined. A distribution of random percentages for each cell in the table was obtained using 10,000 replicates of randomized data. The observed percentages were compared against the randomized distribution to obtain the z-score

Figure 2.9 (cont'd)

for each observed value. (B): Observed and expected distributions of FPKM fold change in the three OG types. The black horizontal dotted line indicates expected divergence between AT and RR expression levels based on the observed divergence in 1:1 OG type. (C): Expression states of genes in OG branches. For each OG type, RR genes along each of the branches and their ancestral expression state were defined as noted (see Methods) and were used to estimate the observed frequencies. To calculate expected frequencies, we obtained a dataset of the same size as the observed dataset, with randomized associations between AT and RR. This dataset was used to calculate the expected frequencies.

Our comparative analyses of BR and RR gene content indicate that retained duplicates in both genomes experienced similar patterns of asymmetric evolution post WGD. In addition, the timings of pseudogenization of BR and RR pseudogenes also followed similar distributions. Given BR and RR have been evolving independently during the past 13-19 mya, these results suggest that process of duplicate gene retention and loss may have exhibited similar biases in both BR and RR lineages.

Predicting duplicate gene retention

Our results so far indicate that duplicates in ~15% of the OGs may have been retained post α' WGT. Such retained duplicates may exhibit functional or other biases (Pál et al., 2001; Chapman et al., 2006; Schnable et al., 2011). However, it remains unclear whether these characteristics apply to α' WGT duplicates and whether some of these features are better predictors of duplicate retention than others. To address these questions, we examined five types of gene features including GO-Slim classification, sequence-related features, expression-related features, network-related features and conservation-related features (see Methods, Table 2.4). For each feature, we asked if the feature values of retained duplicates were significantly different from those of singletons. In addition, we compared the properties of α’ retained duplicates and singletons against those derived from the α WGD event (Bowers et al., 2003). Because the general trends in BR and RR are essentially the same, in all subsequent discussions we discuss the joint results of both species.

We found that, except from inconsistencies between some features (e.g.: protein size, gene size), most other features are consistent between the α’ WGT and the α

Table 2.4: Datasets used for enrichment and SVM analyses

No	***Feature set***	***Source***	***Comments***
GO-Slim and Sequence-related features			
1	GO-Slim categories	TAIR FTP	Only biological process categories were used.
Sequence-related features			
1	Protein size Gene size GC3 content	Custom Python scripts	Values were obtained by analyzing the FASTA and GFF files.
2	PFAM Domain size	HMMER	HMM Domains of AT, BR and RR proteins were obtained by running HMMER with the options *–cut_tc –noali* and further filtering the domains with Evalue<1e-5
Expression-related features			
1	Breadth and level of expression (NASCarray)	NASCArray	Pearson's Correlation Coefficient was calculated between NASCArray datasets using the ATH1 chips. Of the datasets with > 0.98 PCC, only 1 representative dataset was kept. Breadth and level of expression were calculated for the remaining 1779 datasets, after excluding multigene probes. Low/Medium/High expression levels and breadth were defined as <25th percentile, 25th-75th percentile and >75th percentile of the entire distribution.
2	Biotic and abiotic responsiveness	ATGenExpress	Previously published data Zou et al, 2011 was used. Genes showing more than 2X upregulation or downregulation in atleast one condition were defined as responsive to stress.
3	RNA-seq	Previously published data	Data from Moghe et al, 2013 was used for this study. Low/Medium/High expression levels and breadth were defined as <25th percentile, 25th-75th percentile and >75th percentile of the entire distribution. Low/Medium/High expression breadth was defined as expression in 0-3, 3-5 and 5-8 datasets respectively.

Table 2.4 (cont'd)

Network-related features			
1	Number of interacting partners	Aranet	Number of interactions in the integrated Aranet network inference were used.
Conservation-related features			
1	Breadth of conservation across plants	Phytozome	TBLASTN was performed between AT or BR/RR peptide sequences (Query) and the genome fasta sequence of all Phytozome species (Subject). All hits with E>1e-10 were eliminated. Number of species with significant hits was enumerated.
2	*dN/dS* values	Custom python script	*dN/dS* was calculated between orthologs using the yn00 function in the PAML package. To obtain one *dN/dS* value for each AT gene, the average *dN/dS* value between AT-BR and AT-RR orthologs was computed and used for this analysis.

WGD events (Figure 2.10A,B). For example, among biological functions, retained duplicates were most strongly enriched in GO-Slim categories related to transcriptional regulation, stress response, signal transduction and transport for both polyploidization events (Figure 2.10A,B). Duplicates retained after both polyploidization events tend to have larger gene sizes, higher GC3 content (*p*<1e-9 and *p*<1e-21, respectively), higher expression levels and broader expression profiles (*p*<1e-25 and *p*<1e-3, respectively for RNA-seq data), responsiveness to biotic and abiotic stresses (*p*<1e-7 and *p*<1e-4, respectively) and greater network connectivity (*p*<1e-21 and *p*<1e-59, respectively) than singletons. In addition, retained duplicates tend to have homologues in a higher number of land plant genomes (*p*<1e-45) and show lower *dN/dS* values with with their AT orthologs (*p*<1e-24) than singletons. It is likely that some features are correlated with each other e.g.: higher GC3 content has been shown to be correlated with stronger purifying selection, greater codon usage bias and higher frequency of DNA methylation (Elhaik and Tatarinova, 2012), and may be associated with expression-related characteristics of retained duplicates. Similarly, higher conservation among retained duplicates may be associated with their biological roles, network connectivity and expression profiles.

Overall, our enrichment analyses indicates biases amongst retained duplicates and singletons, many of which are consistent between α and α' events. However, what is the relative importance of these features, and can they successfully classify retained duplicates from singletons? To address this question, we considered all features regardless of whether their values differ significantly between retained duplicates and singletons.

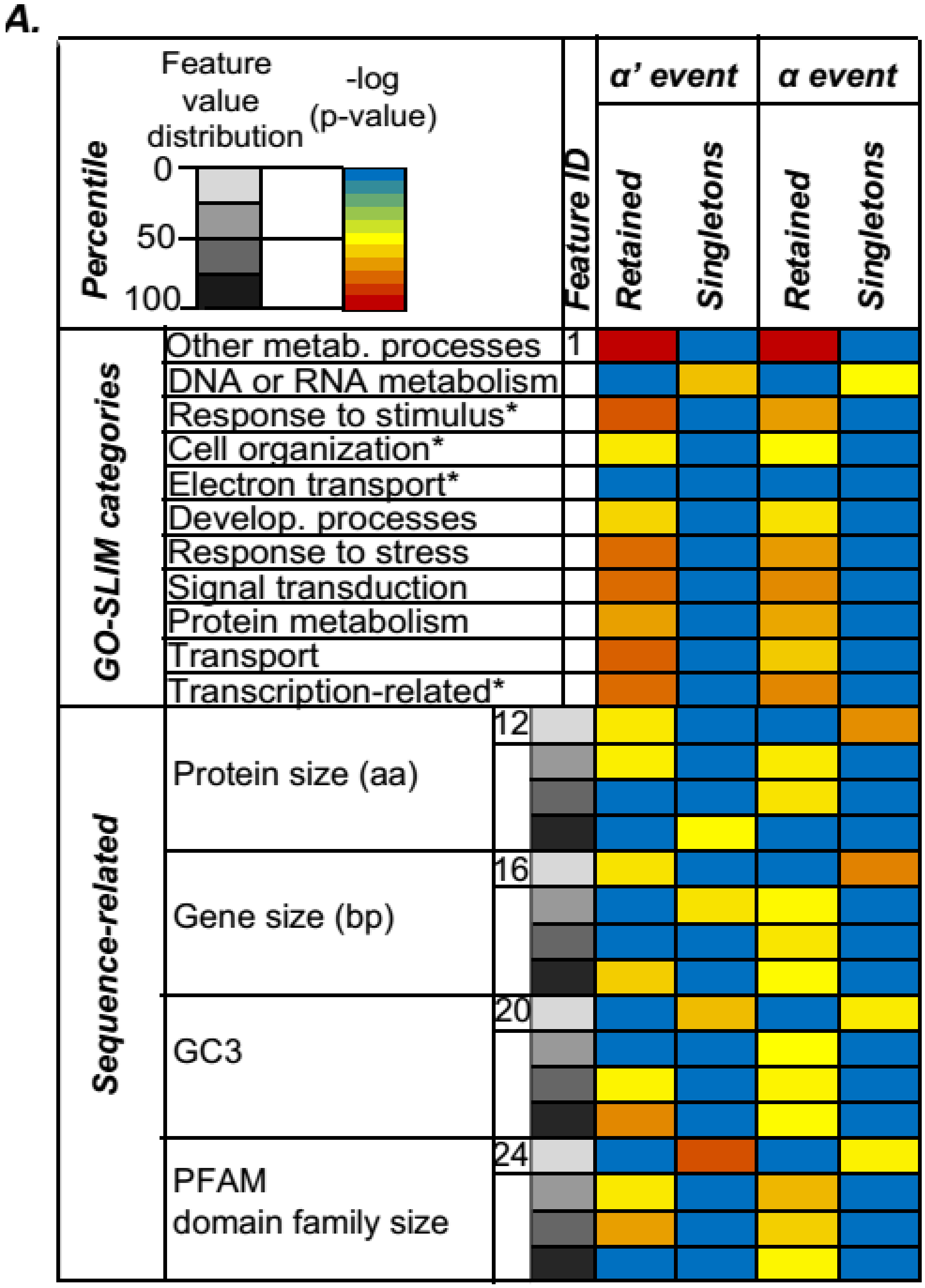
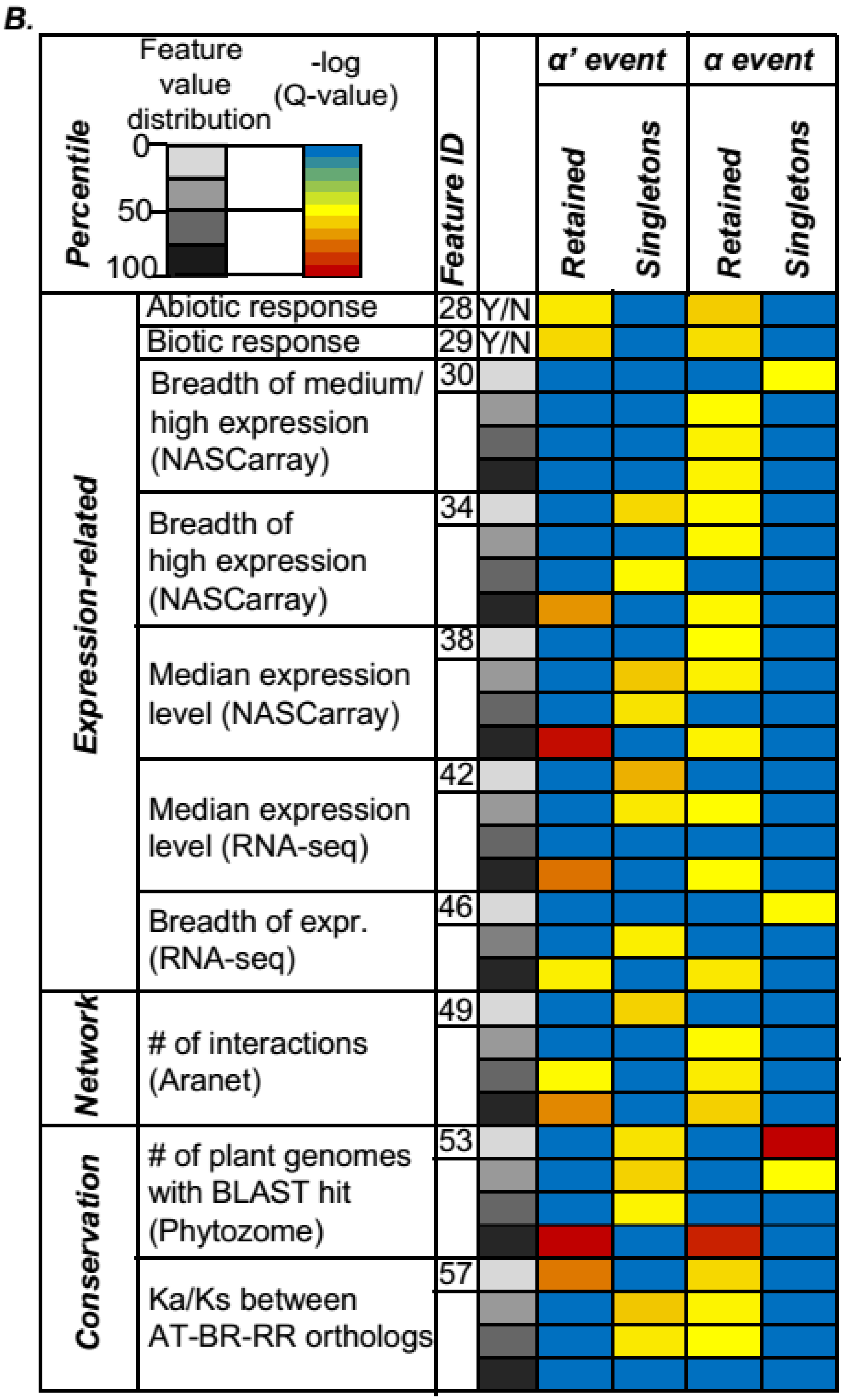


Figure 2.10: Comparison of features between retained duplicates and singletons. (A and B): Results of enrichment analysis showing various features classified based on their type. Feature IDs are sequential throughout the entire set of features (including A

Figure 2.10 (cont'd)

and B). The value distributions of each feature were divided into four bins corresponding to quartiles of the feature distribution, indicated by increasingly darker shade of gray for each feature bin in the figure. The colors represent degree of enrichment, from over-representation (red) to no enrichment (white) to under-representation (blue).

Figure 2.10 (cont'd)



and generated predictive models using a machine learning approach called Support Vector Machine. The model performance was evaluated using Area Under Curve (AUC, area under the curve of true positive rate vs. false positive rate) where a AUC of 1 indicate a perfect model while 0.5 indicate a model with no merit. Models for the α WGD and α' WGT events, as well as randomized data for each WGD event were generated.

For the model predicting α' duplicate retention using all features (the full model), the average AUC is 0.73 significantly better than the model constructed with randomized data (average AUC=0.51, Figure 2.11A,B) or using single sets of features ("the individual models", average AUC=0.56, Figure 2.12A). The results are similar for α duplicates, although, compared to random guesses, the performance in classifying α duplicates (average AUC = 0.75) is slightly better than predicting α' duplicates (Figure 2,11A,B), likely because some features such as GO-Slim, expression related features and network related features for the BR and RR genes were inferred from their AT orthologs. We also found that excluding one feature set at a time from the full model (leave one out models) did not significantly affect the model performance **(**average AUC = 0.72, Figure 2.12A). Our findings suggest that combining multiple features into a single model allows for a better classification of retained duplicates from singletons than random guesses and single features. In addition, the model trained on the α' dataset generated an average AUC of 0.61 when used to classify α duplicates, while the one trained on the α dataset generated an average AUC of 0.67 for α' duplicates. While both AUCs are better than the individual models and random guesses, it seems that model performance is reduced when tested on WGD duplicates from an event it is not trained on, possibly due to the noise introduced due to inferring feature values from orthologous genes or due to unique properties of retained duplicates associated with each WGD event.

To identify consistent and divergent properties between α and α’ events, we used the weights obtained by SVM analysis. The SVM weights for each feature correlate with how well a feature allows differentiation of retained duplicates and singletons. We found that features related to DNA/RNA metabolism (#2), electron transport or energy pathways (#5), transcriptional regulation (#11), protein and gene size (#12 and #19, respectively), GC3 content (#23) and number of plant genomes with BLAST hits (#56) showed consistency between models (Figure 2.12B). These features also are also consistent with results of enrichment analyses. On the other hand, some features (e.g.: gene size, network interactions) had higher weights in one SVM than the other or weights in the opposite direction. In addition, features related to expression level (#43 and #45) or breadth of expression (#37 and #48) were not as important in classifying both α and α’ duplicates from singletons. These observations suggest that some features (e.g.: function, structural composition) may be important during the retention process but others (e.g.: expression-related) may play a less significant role.

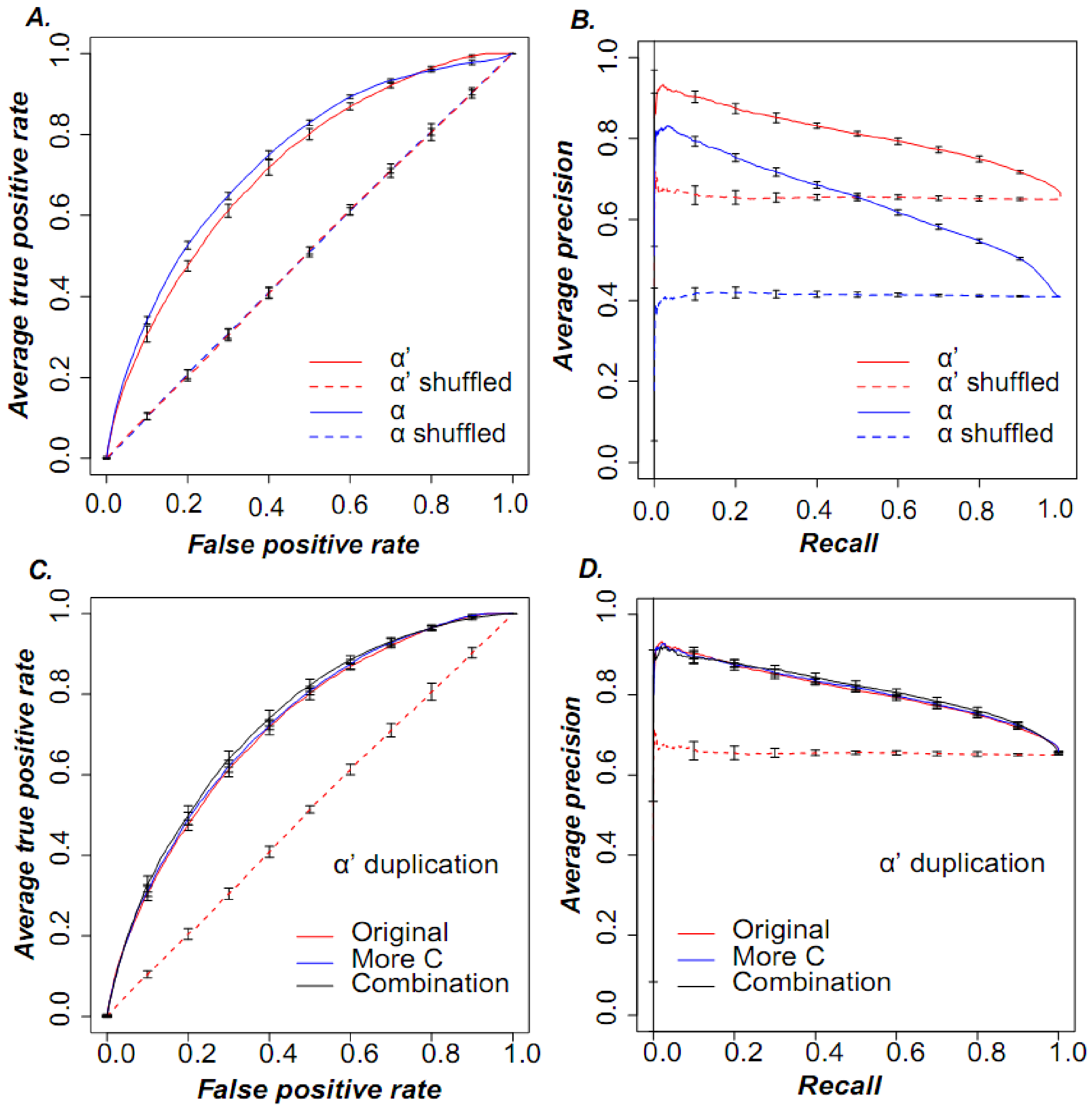


Figure 2.11: SVM analyses results (A and B): The AUC-ROC curves (A) and Precision-Recall curves (B) obtained using linear SVM model for α WGD and α' WGT events using all features (as noted in Figure 2.10A,B). See Methods for details on how the shuffling was performed. (C and D): Increasing the range of soft margin values (more C) or using a model made from pairwise combinations of all 60 features used in the original

Figure 2.11 (cont'd)

model (combination) do not increase model performance, as shown in the AUC/ROC curve (C) and the P/R curve (D).

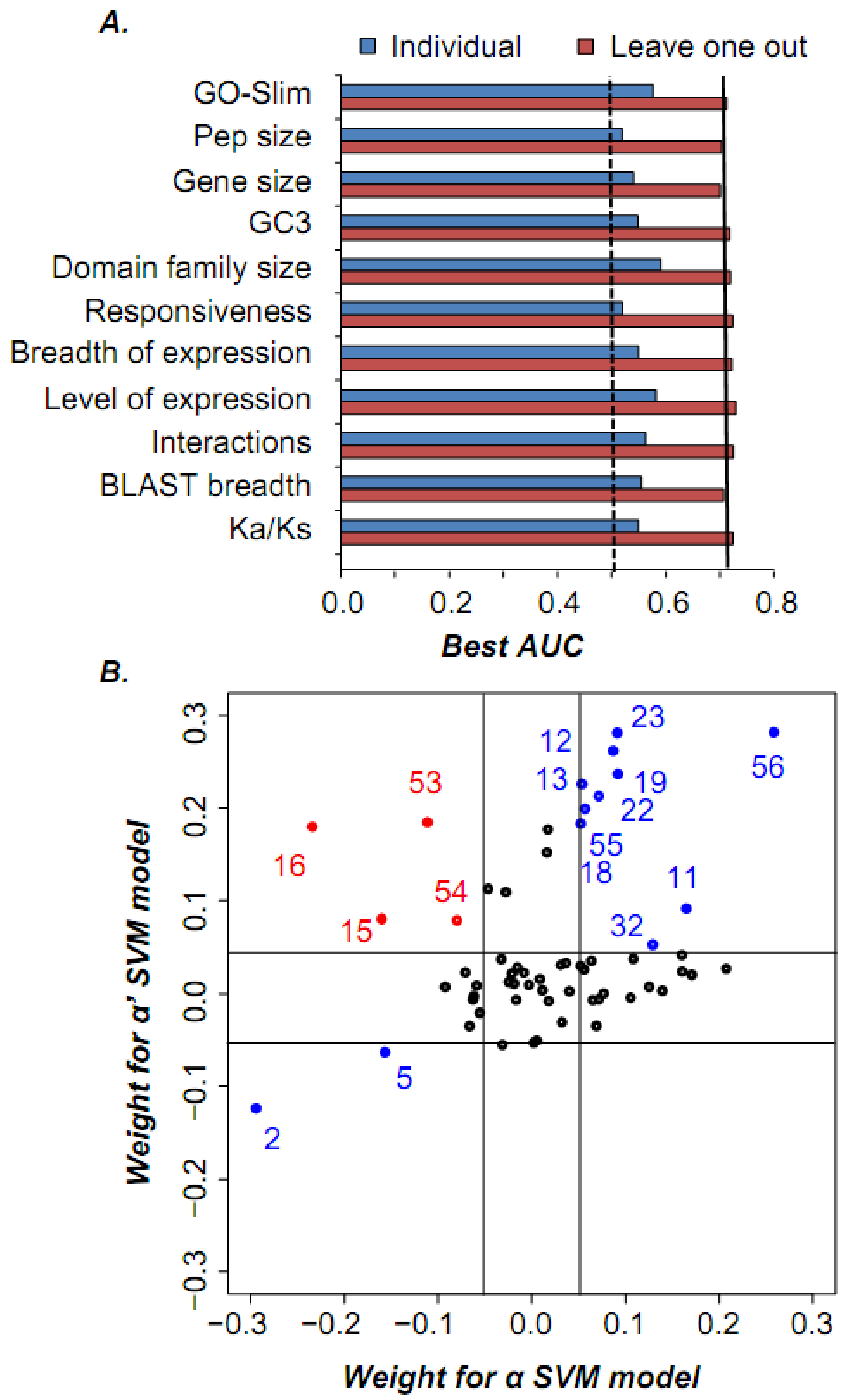


Figure 2.12: Identifying features most important for classification. (A) The "individual" model involved running SVM with only the given set of features, while the "leave one out" strategy involved running SVM after excluding only the given set of

Figure 2.12 (cont'd)

features. The dotted and the solid lines correspond to the AUCs of the randomized model and the full original model, respectively (B): Comparison of SVM weights for each feature between α WGD and α' WGT. Each dot represents values for a single feature. Features with significant and consistent weights are colored blue while those with significant yet opposite weights are colored red. Numbers in brackets correspond to Feature IDs as noted in third column of Figure 2.10A,B.

CONCLUSIONS

In this study, we have sequenced the genome of RR, a wild relative of the cultivated crops *Raphanus sativus* and BR. The 254 Mb assembly encompasses ~49% of the estimated genome size, has an N50 of 10.1kb and houses a majority (38,174) of the genes in the RR genome. We used these gene models to understand the evolution of duplicate genes and pseudogenes in BR and RR post α' WGT.

The loss of ~60% of the genes in BR and RR in a consistent fashion suggests that gene loss is the predominant fate of duplicates post WGD. Such gene loss may occur in bursts (e.g.: via deletion in early generations) or gradually (via pseudogenization). However, several thousand genes are still retained within the BR and RR genomes and may contribute to evolutionary novelty. For example, a recent study showed that circadian rhythm regulated genes are over-retained in BR (Lou et al., 2012), suggesting the possibility of phenological changes in post α' species. In our study, retained duplicates were also found to possess functions related to transcriptional regulation, stress regulation and development. Diversity in such functions may allow conquest of new ecological niches.

What are the properties of such retained duplicates? Over the past decade, several studies have taken advantage of the increased availability of genome sequence data and comparative genomic tools to analyze the evolution of WGD derived duplicate genes in multiple species, assessing features important for the loss and retention of WGD derived duplicate genes (Blanc and Wolfe, 2004; Schnable et al., 2011; Jiang et al., 2013). In this study, we started of with the features assessed as important in these studies and confirmed their degree of importance in distinguishing α' duplicates from singletons using enrichment analysis and machine learning. Our framework identifies features which were consistently important in the loss/retention across the α and α' duplicates. The fact that the performance of the full model was good but not complete suggests that although existing knowledge is useful, additional features may be important for the gene retention process.   We did not include features such as subgenome bias or random loss in our model, and further research using such additional features would be needed to increase the predictive power of the full model.

One of the results from our analyses was the high degree of conservation of retained duplicates at smaller and larger time scales. This phenomenon has been noted before in plants (Jiang et al., 2013), and may indicate similar biases in the retention process across multiple WGD events. It has been suggested that retained duplicates tend to provide a buffering function against loss of a gene copy, which erodes with time due to mutation accumulation but nevertheless contributes to a lower evolutionary rate (Chapman et al., 2006). However, given evolution cannot see into the future, this model cannot fully explain the observed pattern of lower rates after >25 million years of divergence. It is possible that the lower evolutionary rate is simply a function of its correlation to other features, such as network connectivity, expression profiles and biological function, found to be enriched among retained duplicates in our study. Because some duplicates in the neopolyploid possess these features, they tend to evolve slowly, allowing accumulation of novel, functionally important changes resulting in neo or sub functionalization and eventual retention (Sémon and Wolfe, 2007). If neo-functionalization were to be the predominant fate of retained duplicates, it would be reflected in an elevated *dN/dS* ratio or an asymmetric rate of evolution between most duplicate pairs. However, we found that a large majority (~80-85%) of duplicates are evolving symmetrically and are under selective constraint. Such a scenario may be explained by the sub-functionalization model.

Our results also suggest a complex pattern of expression evolution between retained duplicates in RR that needs to be investigated in more detail using transcriptomic data from multiple tissues/conditions. In addition, in recent years, genomic and transcriptomic data from multiple plant species, many of which have undergone recent or ancient polyploidization events, are also available. We surmise that comparative analyses of pseudogenes and duplicate genes derived via WGD events in the plant kingdom will provide a comprehensive picture of the loss/retention process in plants.

MATERIALS AND METHODS

Genomic DNA and EST Sequencing

RR is an obligate out-crosser. To reduce the amount of heterozygosity in the genome, RR subspecies raphanistrum (weedy) from the Binghamton population in New York was inbred for five generations. Total DNA was extracted from the leaves of the 5th generation inbred plants using Qiagen DNEasy Maxi kit. The extracted DNA ethanol precipitated and assessed for quality using CHEF gel electrophoresis. For 454 sequencing, DNA was sheared by Covaris sonication, size selected by gel electrophoresis and a 3 kb mate pair library constructed according to manufacturer's instructions (Roche-454). A total of 6 full plates and three half-plates were sequenced using the Titanium chemistry. DNA was further sheared and an Illumina fragment library constructed (peak 520 bp). A total of 7 lanes of 100 bp paired-end sequence was generated on an Illumina GAII sequence analyzer.

ESTs were sequenced from three *R. sativus* cultivars (convars *sativus, caudatus, oleifera*) and four *R. raphanistrum* populations (subspecies *raphanistrum* NY weedy population, *raphanistrum* Central Spain population, *maritimus* Coastal Spain population and *landra* France population). Total RNA from whole seedlings of *Raphanus raphanistrum* and *Raphanus sativus* (with 1 set of true leaves), buds, and anthers were pooled together. Double strand cDNA were synthesized from pooled RNA using SMART technology (Clontech). The prepared cDNA was normalized by cDNA denaturation/reassociation, treatment by duplex-specific nuclease (DSN) and amplification of normalized fraction by PCR. The normalized cDNA was then digested with SfiI, fractioned, directionally ligated into pDNR-LIB (Clontech) and electroporated into GC10 competent cell (Gene Choice). Sequences were generated from 5' and 3' ends of clones. A total of 185.4 Mb was sequenced. A total of 310,844 EST sequences were deposited in NCBI dbEST.

Genome assembly and quality assessment

Before assembly, Illumina reads were trimmed from the 3' end to a Phred quality score ≥20 and length ≥50. The 454 reads were split at linker sequences and only reads with mate pairs were used for assembly. The filtered Illumina and 454 reads represented a 47X and 3X coverage of the estimated 573Mb genome. To assemble the RR genome, we explored three different approaches. We first created an Illumina-only assembly using ABySS 1.2.5 (Simpson et al., 2009) with the optimal kmer length (k=39). We then split the Illumina contigs into overlapping fragments of 1998bp (maximum size allowed for input to Newbler) with 1000bp step size at a coverage of 10X per fragment. These split Illumina contig fragments and the quality-filtered 454 reads were used as input to Newbler 2.5.3 (Margulies et al., 2005) to create a hybrid assembly. The following parameters were used for the Newbler assembly: -large -mi 98 -cpu 1 -ml 80 -ud -rip -m -e 8. The Newbler assembly showed a marginal improvement in N50 and total assembly size compared to an Illumina-only assembly (Figure 2.1). In the second approach, the RR assembly was generated with the Celera Assembler using split 454 reads (sffToCA program, option "-clear 454 -trim chop") and Illumina reads trimmed (https://github.com/tanghaibao/trimReads) so the base quality is at least Phred 20. We ran Celera Assembler version 6.1 with unitigger "BOGART" with kmerSize=30 (Miller et al., 2008). Finally, we used the program Minimus2 (Sommer et al., 2007, 2) from the AMOS 3.1.0 package to merge the ABySS/Newbler and the Celera assemblies. The Minimus2 merging step was repeated three times with the merged contigs and the unmerged contigs till convergence. The final Minimus2 assembly was substantially better than the ABySS/Newbler and the Celera assemblies (Figure 2.1) and was used in all subsequent analysis .

All the Illumina and 454 reads are currently being deposited in NCBI SRA under the accession numbers PRJNA209513 (BioProject), SRX326772 and SRX326773 (Experiments).

Structural and functional annotation

The MAKER 2.10 pipeline (Cantarel et al., 2008) was used to annotate the RR assembly detailed in Figure 2.2A. The 74,568 gene models predicted were filtered based on their Annotation Edit Distance (AED) values or the presence of a protein domain as predicted by HMM-PFAM (Eddy, 2008). Two sets of gene models with different levels of accuracy were created: (1) Set I (41,122 models) consisted of gene models with AED≤1, domain E-value<1e-3 and (2) Set II (38,174 models) consisted of models with AED<0.5 or (AED>=0.5 and domain E-value<1e-5) (Figure 2.2B). All gene models possessing specific transposon-related domains over-represented in RR vs. BR (PF03732.12, PF13975.1, PF03384.9, PF03108.10, PF14392.1, PF14111.1, PF03078.10, PF00075.19, PF13966.1, PF09331.6, PF13456.1) were also discarded from Set II via manual keyword searches. All analyses were performed using Set II gene models given their higher level of agreement with evidence. Functional annotations of gene models were obtained using BLAST2GO (Conesa et al., 2005). The genome sequence and the annotation will be made available for download at the following URL: http://shiulab.plantbiology.msu.edu/.

Timing of speciation and duplication events

Previous studies have estimated the timings of the speciation and duplication events in Brassicaceae. However, many of these estimates were obtained using a now unavailable fossil pollen calibration point placed in the genus *Rorippa* in Brassicaceae, based on synonymous substitution rate derived from two individual loci (Koch et al., 2000) or assuming a constant rate of evolution across the Brassicaceae family. These issues have been reviewed exhaustively in a previous study (Beilstein et al., 2010). Based on the relative rate test (Goldman and Yang, 1994), the synonymous substitution rate at the third codon position did not increase significantly after the polyploidization event, consistent with the molecular clock assumption. Therefore, this rate can be used for determining the age of the α' WGT event and the BR-RR speciation event.

For determining timing of speciation and duplication events using multidivtime (Rutschmann F., 2005), we first assigned *Carica papaya* (CP) genes to the predicted Brassicaceae orthologous group. if a CP gene had a significant hit to ≥1 species in each of AT/AL and BR/RR and no hit to any other orthologous group. Synonymous substitution rates of singletons at the third codon positions between AT:AL:BR:RR and CP (outgroup) were used to determine the times for speciation. The rates of retained duplicates were used to estimate the timing of duplication in BR and RR, using the AT ortholog as outgroup. A synonymous site substitution rate of $7*10^{-3}$ substitutions/site/million years (Ossowski et al., 2010) was used for determining the timing of speciation and duplication, with a prior age of 36 mya between the root and the tip. Based on findings in a previous study (Beilstein et al., 2010), we fixed the lower constraint for AT-BR divergence at 30 million years.

For determining age, *dS* was calculated between pairs of singleton genes and between pairs of retained duplicates using codeml (Yang, 2007). Divergence time was obtained using the formula T= *dS*/(2*neutral rate). As expected, if dates are estimated using the previously used substitution rate of $15*10^{-3}$ substitutions/site/million years (Koch et al., 2000), the median ages of different events are almost halved (Figure 2.3B).

Prediction of pseudogenes and pseudogenization timing

A modified version of a previously defined pseudogene pipeline (Zou et al., 2009) was used to predict pseudogenes in genomes of all four species under study. These pseudogenes are exclusively derived from protein-coding genes and not from non-coding RNA genes. Specifically, we performed TBLASTN using protein coding genes as query and genomic sequences as the subject using BLAST 2.2.25. We then filtered the output using the thresholds: E-Value < 1e-5, %Identity > 40%, Match Length>30aa and Coverage > 5% of the query sequence to obtain pseudo-exon definitions. Pseudo-exons in close proximity to each other (based on the 95th percentile of the intron length distribution) and having matches to the same protein were then joined together to form putative pseudogenes based on their Smith-Waterman score. Putative pseudogenes overlapping with annotated protein coding regions were removed from the dataset. In addition, pseudogenes with significant similarity to known *Viridiplantae* repeats (Cutoff=300, Divergence=30) as determined by RepeatMasker 3.3.0 were discarded.

Finally, because of the fragmentary nature of the BR and RR genomes, there was a high false positive rate due to proteins split between contigs being counted as pseudogenes. To reduce the false positive rate, high confidence pseudogenes were determined using a custom python script. Specifically, a pseudogene is considered a high-confidence pseudogene if it contains stop codons or frame-shifts or if it passes a particular test. This test states that a protein is a high confidence pseudogene if $X_U >= Y_U + Z$ and $X_D >= Y_D + Z$, where X_U and X_D are the absolute distances between the pseudogene and the each end of the contig it is on for both sides of the pseudogene, upstream and downstream relative to the orientation of the matching protein, respectively, and where Y_U and Y_D are the absolute distances between the matching region on the protein and the end of the protein for both sides of the protein, upstream (N-terminal side) and downstream (C-terminal side), respectively, and where Z is the 95th percentile intron length for the species being tested.

The number of detectable pseudogenes is higher in post-α'-polyploidization species compared to AT/AL. For each annotated protein-coding gene in AT and AL, there are 0.15 and 0.34 pseudogenes, respectively. In contrast, there are 0.96 and 0.56 pseudogenes/annotated gene for BR and RR, respectively (or, after correcting for the fragmentary nature of the BR and RR genomes, 0.82 and 0.35, respectively). The low proportion of pseudogenes/annotated gene in RR is likely because of the incomplete RR assembly as well as overcorrection. The pseudogene numbers obtained for BR and RR are likely to be an underestimate of the actual number of pseudogenes derived from transposition events in BR and RR, given that the repetitive genomic fraction was largely missed in both the assemblies. In addition, putative pseudogenes resembling repeats – 5060 BR pseudogenes and 518 RR pseudogenes – were discarded. There are substantially fewer repeat-related pseudogenes in RR most likely because of the lower coverage of the RR genome than the BR genome.

To estimate the timing of pseudogenization, we used a published approach (Figure 2.7A) (Chou et al., 2002). All estimates ≤ 0 mya were discarded. To determine whether the timing was robust to the definition of α' pseudogenes, we used four different approaches to define and estimate pseudogenization timing of such pseudogenes: 1) definition based on *dS* only, timing using the entire pseudogene sequence (3300 BR, 2171 RR pseudogenes, Figure 2.7B), 2) definition based on *dS* only, timing using only the sequence past the first disabling mutation (1266 BR, 924 RR pseudogenes, Figure 2.7D), 3) definition based on homeology, timing using the entire pseudogene sequence (1522 BR, 652 RR pseudogenes, Figure 2.8C), and 4) definition based on homeology, timing using only the sequence past the first disabling mutation (564 BR, 215 RR pseudogenes, Figure 2.7D). Our results are slightly biased towards more recent pseudogenization timings if we only use the pseudogenes lying in homeologous segments with their paralogs (Figure 2.7C,D).

Thresholds for *dS* with respect to a pseudogene's functional paralog were set at 0.2 ≤ *dS* ≤ 0.6. The lower and upper bounds were based on the 25th and 75th percentiles of the duplicate gene *dS* distribution (Figure 2.4A). Changing the *dS* threshold to a more stringent one (0.30 ≤ *dS* ≤ 0.42) for identifying WGD derived pseudogenes did not influence the estimates significantly (Figure 2.7E).

To determine whether our findings are robust to the estimate of the duplication time in the timing formula, we defined duplication times using three methods: 1) a fixed duplication time of 25 mya, 2) random sampling from a Gaussian distribution with mean=25 and sd=7 (based on the functional duplicate gene *dS* distribution) and 3) Calculating the duplication time based on the *dS* between pseudogene and the parent gene. In all cases, the distributions obtained for pseudogenization timing were very similar and do not affect our interpretations (data not shown). All timing estimates ≤0 were discarded.

Classifying retained duplicates and singletons with machine learning

We used the Support Vector Machine (SVM) approach to generate classifiers that allow distinguishing retained duplicates and singletons. The feature sets used in this study are detailed in Table 2.4 (Kilian et al., 2007; Lee et al., 2010; Goodstein et al., 2012; Moghe et al., 2013). If the feature values could not be obtained from BR/RR directly, values were inferred from the AT orthologs of the BR/RR genes.

For all quantitative features, we binned the values into four quartiles based on the feature value distribution across all genes. All other features (GO-Slim categories and responsiveness to biotic or abiotic stress) were treated as discrete categories. The 4702 retained duplicates and 2533 singletons were assigned roughly equally and randomly to the training and the test dataset. The random split was repeated ten times. SVM-Light (Joachims, 1999) was used to generate classifiers and feature weights. A grid search was performed to determine the optimal SVM parameters. Increasing the C sampled from 1e-06 to 1000, with 10-fold change or using pairwise combinations of all features did not result in any improvement in the AUC and Precision/Recall curves (Figure 2.11C,D). Using a radial basis function with varying gamma values from 1e-06 to 1, with 100-fold change for the next value, also did not result in improved model performance (data not shown). This suggests that although the full model performs better than random guesses, its performance is reduced when tested on WGD duplicates from an event it is not trained on, possibly due to the uniqueness of each WGD event or the noise introduced via inferring feature values from orthologous genes.

ACKNOWLEDGEMENTS

We thank Melissa Lehti-Shiu for maintaining the radish plants, Barry Williams for discussion, Alicia Massa for helping isolate genomic DNA samples, Robin Buell for discussion on sequencing strategies, John Johnston for computational support, Guangxi Wu for assistance with orthology prediction and Alex Seddon for assistance with SVM analyses.

CHAPTER THREE

Characteristics and significance of intergenic PolyA RNA transcription in *Arabidopsis thaliana*[1]

[1] The work described in this chapter was published in the following manuscript:

Gaurav Moghe, Melissa Lehti-Shiu, Alex Seddon, Shan Yin, Yani Chen, Piyada Juntawong, et al (2013) Characteristics and significance of intergenic polyA transcription in *Arabidopsis thaliana*. *Plant Physiology*, 161(1):210-224

ABSTRACT

The *Arabidopsis thaliana* genome is the best annotated plant genome. However, transcriptome sequencing in *A. thaliana* continues to suggest the presence of polyA transcripts originating from presumed intergenic regions. It is not clear whether these transcripts represent novel non-coding or protein coding genes. To understand the nature of intergenic polyA transcription, we first assessed its abundance using multiple mRNA-sequencing datasets. We found 6,545 Intergenic Transcribed Fragments (ITFs) occupying 3.6% of *A. thaliana* intergenic space. In contrast to transcribed fragments that map to protein coding and RNA genes, most ITFs are significantly shorter, are expressed at significantly lower levels and tend to be more dataset-specific. A surprisingly large number of ITFs (32.1%) may be protein coding based on evidence of translation. However, our results indicate that these "translated" ITFs tend to be close to and are likely associated with known genes. To investigate if ITFs are under selection and are functional, we assessed ITF conservation through cross-species as well as within-species comparisons. Our analysis reveals that 237 ITFs, including 49 with translation evidence, are under strong selective constraint and relatively distant from annotated features. These ITFs are likely parts of novel genes. However, the selective pressure imposed on most ITFs is similar to that of randomly selected, untranscribed intergenic sequences. Our findings indicate that despite the prevalence of ITFs, apart from the possibility of genomic contamination, many may be background or noisy transcripts derived from "junk" DNA whose production may be inherent to the process of transcription and which, on rare occasions, may act as catalysts for creation of novel genes.

INTRODUCTION

The advent of tiling arrays and high-throughput sequencing has led to the discovery of a complex transcriptional landscape in eukaryotic genomes. Studies in yeast (David et al., 2006), animals (Bertone et al., 2004b; Carninci et al., 2005b) and plants (Yamada et al., 2003b; Matsui et al., 2008; Li et al., 2007b) have revealed the presence of a large number of unannotated, novel transcripts. These novel transcripts may represent alternatively spliced forms of known genes (Filichkin et al., 2010), products of antisense (Yamada et al., 2003b) or bidirectional (Xu et al., 2009) transcription, retained introns (Ner-Gaon et al., 2004; Filichkin et al., 2010), transcript fusions (Ruan et al., 2007) or intergenic transcriptional units (referred to hereafter as Intergenic Transcribed Fragments or ITFs). Among these novel transcripts, ITFs are unique in that they do not overlap with known genomic features and may represent novel genic sequences. The prevalence of intergenic transcription raises the possibility that there are many more functional genes yet to be discovered. However, there are two outstanding questions regarding ITFs. First, it is not clear what proportion of ITFs code for proteins. Secondly, whether or not most ITFs are functional is under debate (Mattick, 2009; Ponting and Belgard, 2010).

After ITFs are identified with whole genome tiling arrays or high-throughput sequencing, computational methods are used to determine if they display characteristics of non-coding RNA (ncRNA) (Li et al., 2007a; Fahlgren et al., 2007; Gregory et al., 2008). These methods rely on secondary structure prediction, similarity to known ncRNA and conservation between species. The protein coding potential of ITFs, on the other hand, is determined based on *ab initio* gene prediction, open reading frame (ORF) length, evolutionary conservation measures, pairwise alignment scores, predicted secondary structure and entropy (Dinger et al., 2008; Nekrutenko et al., 2002; Liu et al., 2006). For example, in a global gene expression study in *A. thaliana*, a 50 amino acid length threshold was used to define potential protein coding intergenic transcripts (Stolc et al., 2005). Similarly, the FANTOM consortium defined putative protein coding mRNAs using an open reading frame (ORF) length cutoff of 300 nt (Okazaki et al., 2002). Reliance on length cutoffs can result in longer random ORFs being falsely annotated as protein-coding and will also lead to the exclusion of true small ORFs such as those that have been identified in yeast, humans and *A. thaliana* (Basrai et al., 1997; Pruitt et al., 2007; Hanada et al., 2007). Proteomics and polyribosome immuno-precipitation (Zanetti et al., 2005; Sparkes et al., 2006) allow more direct identification of potentially protein-coding ITFs than computational approaches. Currently, there has yet to be a systematic assessment of ITF protein coding potential based on a combination of computational and experimental approaches.

In addition to the question of whether ITFs code for proteins or not, the functional relevance of intergenic transcription is not well understood. One hypothesis is that most transcripts simply represent transcriptional noise. For example, based on the genome-wide distribution of RNA polymerase II and TATA-Box binding protein in yeast, ~90% of RNA polymerase II transcriptional initiation events were estimated to be the result of low polymerase fidelity and may represent transcriptional noise (Struhl, 2007). Consistent with the “noise” hypothesis, several studies have shown that ITFs tend to have significantly higher evolutionary rates than known genes. For example, the ENCODE consortium found that 93% of the unannotated transcribed regions in the human genome show no clear evidence of evolutionary constraint (Birney et al., 2007). The alternative hypothesis is that most ITFs are functional (Dinger et al., 2009). Differential expression, alternative splicing and/or association with chromatin modification marks have been cited as evidence for ITF functionality (Hiller et al., 2009; Guttman et al., 2009). In addition, the functions of a growing number of novel transcripts have been experimentally determined. Examples include *Xist, RepA, Air* and *Hotair*, which regulate recruitment of Polycomb proteins onto DNA (Mercer et al., 2009) as well as a recently discovered long non-coding RNA called *COLDAIR* shown to be important in regulating vernalization responses in *A. thaliana* (Heo and Sung, 2011). Based on these studies, it is clear that some ITFs are functional. The main question is the abundance of functional ITFs relative to those derived from noisy transcription.

To date, most studies of intergenic transcription have focused on the presumably non-coding fraction of the transcriptome. In addition, currently there is no published study assessing the evolutionary significance of plant intergenic transcription. In this study, we focused on intergenic polyA RNA transcripts to gain more insight into the nature of plant intergenic transcription by RNA polymerase II. We first analyzed eight different *A. thaliana* messenger RNA-Sequencing (mRNA-seq) datasets from this study and two other sources (Jiao and Meyerowitz, 2010; Filichkin et al., 2010) to determine the extent of intergenic polyA transcription. We then investigated whether ITFs are likely protein coding using (1) ribosome immuno-precipitation data generated in this study as well as public datasets (Jiao and Meyerowitz, 2010), (2) proteomics data (Castellana et al., 2008; Baerenfaller et al., 2008), and (3) fusion protein expression studies on selected targets. Finally, making use of the polymorphism data from 80 different *A. thaliana* accessions (Cao et al., 2011) and protein coding genes and genome sequences of other plants, we explored whether ITFs, especially those that may code for proteins, are likely functional based on within and cross-species conservation.

RESULTS AND DISCUSSION

Defining transcribed regions in the *A. thaliana* genome

To explore the functional significance of intergenic transcription further, a rigorous definition of transcribed regions within the *A. thaliana* genome is necessary. To this end, we analyzed mRNA-seq data from three different sources: (1) 7-day old seedlings generated in this study, (2) whole flower (Jiao and Meyerowitz, 2010), and (3) 12-day old seedlings grown under six environmental conditions (Filichkin et al., 2010) (Table 3.1). We assembled transcript fragments (TxFrags) using two approaches (see Methods). In the first approach, contiguous regions in *A. thaliana* occupied by mapped mRNA-seq reads were defined as expressed (Set 1 TxFrags). In the second, more stringent, approach, we assembled TxFrags using the transcript assembly program Cufflinks (Set 2 TxFrags).

We first compared the characteristics of Set 1 TxFrags among annotated features including protein coding genes, RNA genes, pseudogenes, and transposons. Regardless of the genomic feature and dataset, the Set 1 TxFrag length distributions are bimodal with the first peaks located near the mRNA-seq single read length, indicating most Set 1 TxFrags consist of a single read (Figure 3.1A). Next, the Fragments Per Kilobase of exon model per Million mapped reads (FPKM) measure was used to assess Set 1 TxFrag expression level. Similar to length distributions, the Set 1 TxFrag FPKM distributions are bimodal with the first peaks at very low FPKM, mostly consisting of single read TxFrags (Figure 3.1B). The likely sources of low FPKM TxFrags are: (1) genes with very low level or highly specific expression, (2)

Table 3.1. Description of datasets for transcriptome analyses

Source	Stage / condition	Reads mapped[2]	Intergenic reads[3]	Set 1 TxFrag[4]	Set 1 ITF[5]	Set 2 TxFrag[4]	Set 2 ITF[5]
This study	7d old seedlings	4,783,510	19,963 (0.4%)	228,968	5,203	28,611	77
This study	T87 cells	30,304,063	272,063 (0.9%)	NA	NA	30,987	1289
Jiao & Meyerowitz, 2010	Stage 4 flowers	19,793,325	146,951 (0.7%)	223,377	20,556	46,276	334
Filichkin et al. 2010[1]	Control	7,841,527	165,976 (2.1%)	354,199	56,186	45,681	815
	Cold	5,653,569	151,179 (2.6%)	356,418	59,063	45,279	1046
	Salt	4,166,706	208,340 (5.0%)	364,001	66,015	32,780	1756
	Heat	5,688,184	84,280 (1.5%)	280,182	35,625	30,400	323
	Drought	4,830,498	181,449 (3.8%)	364,001	68,429	44,343	2194
	High light	6,645,853	549,639 (8.2%)	402,608	91,882	59,905	9363

[1] 12-day old seedlings.

[2]. To the nuclear genome assembly of Arabidopsis thaliana, TAIR 10 release.

[3]. In parenthesis: percent reads mapped to the genome that are intergenic.

[4] TxFrags: transcribed fragments. Set 1 and Set 2 TxFrags were generated with two transcript assembly methods without FPKM threshold, see Methods for details.

[5] ITF: Intergenic TxFrags without FPKM threshold.

“transcriptional noise” representing background genome transcription (Struhl, 2007), or (3) low level genomic DNA contamination in the sequenced mRNA sample. If the presence of ≥1 Set 1 TxFrags is considered evidence of expression, 78-94% of protein coding genes are expressed. However, 19-68% of pseudogenes and 5-69% of transposons would be considered expressed based on the same criterion. Given that *A. thaliana* transposons have been documented to be under-expressed (Schmid et al., 2005) and subject to strong post-transcriptional silencing through DNA methylation (Zilberman et al., 2007; Zhang et al., 2006), transposon expression was used as a conservative error estimate of expression calls.

To stringently control for false positives arising from background transcription and/or low level genomic contamination, we applied multiple FPKM thresholds defined according to the percentage of transposon TxFrags considered expressed (Figure 3.1C). Comparing percent transposons expressed, we found that the FPKM thresholds have significantly different impacts on datasets (Figure 3.1C). For example, an FPKM threshold based on the 90th percentile of the transposon expression distribution results in a 57.4% reduction of transposon expression in the 12-day seedling drought stress dataset compared to no FPKM threshold, but no reduction in the 7-day data (Figure 3.1C). This difference in the degree of transposon expression due to FPKM threshold choice is not simply due to differences in sequencing depth as the numbers of mapped reads are both ~4.8×10^6 (Table 2.1). In addition, this difference cannot be attributed to stress treatments as degrees of transposon expression in the stress treatment and control samples are similarly regardless of FPKM thresholds (green, Figure 3.1C). We note that only 22-38% of reads from the 12-day datasets can be mapped to the *A.*

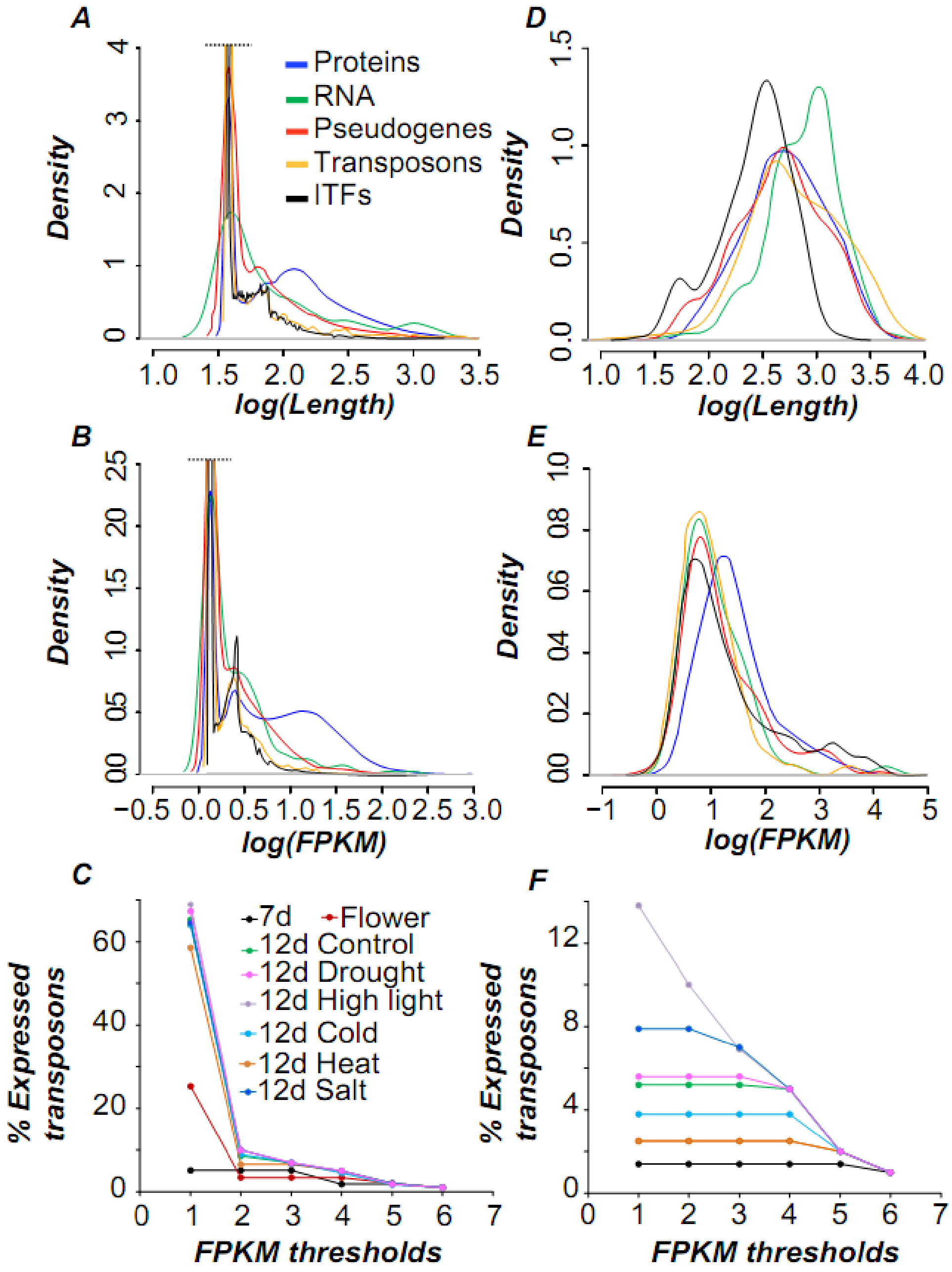


Figure 3.1: Characteristics of Set 1 and Set 2 TxFrags. (A) Length and (B) expression level distributions of various genomic features - proteins (blue), RNA (green), pseudogenes (red), transposons (orange) and ITFs (black) - based on Set 1 TxFrags identified across all eight RNA-seq datasets. Both axes are logarithmically scaled with base 10. To emphasize the lower peaks, curves beyond the black dashed

Figure 3.1 (cont'd)

line are truncated. (C) Percent transposons considered expressed based on Set 1 TxFrags identified from eight datasets at various FPKM thresholds. (D) Length and (E) expression level distributions for Set 2 TxFrags. (F) Percent transposons considered expressed based on Set 2 TxFrags. Please note that the color legends for (C) and (F) are different from those of (A), (B), (D) and (E).

thaliana genome compared to 71% and 75% for 7-day old and flower data, respectively. Thus, data quality may significantly impact gene expression calls, even after quality filtering and mapping the reads to the genome. For comparison, we applied a second, more stringent transcript assembly approach using Cufflinks with bias corrections of transcript-models based on sequences, positions, and abundance (Trapnell et al., 2010) to generate Set 2 TxFrags. Compared to Set 1 TxFrags, Set 2 TxFrags are significantly longer (Figure 3.1D, Kolmogrov-Smirnov (KS) test p<2.2e-16) and have significantly higher FPKM values (Figure 3.1E, KS test p<2.2e-16). In addition, Set 2 TxFrags length and coverage distributions overlap with the right tails of Set 1 TxFrags (Figure 3.1A,B,D,E), indicating the main difference between these two sets is enrichment for longer and more abundant transcripts in Set 2 TxFrags. Increasingly stringent FPKM thresholds still have a significant effect on the numbers of transposons considered expressed for several 12-day datasets (Figure 3.1F). Nonetheless, the second approach (Set II TxFrags) allows for better control in calling transposon expression, which we considered to be mostly false positive, than the first, simpler approach.

Pervasiveness of intergenic transcription in *A. thaliana*

Previous microarray-based studies in *A. thaliana* have shown that a large number of polyA transcripts are produced from the intergenic regions of the genome (Yamada et al., 2003b; Matsui et al., 2008). Considering the advantages of RNA-seq over microarrays for expression studies (Agarwal et al., 2010), we re-assessed the preponderance of intergenic transcription using RNA-seq datasets. Here, TxFrags located within intergenic regions are referred to as Intergenic TxFrags (ITFs). We found that the analysis method (Set 1 vs. Set 2), FPKM threshold, and dataset significantly

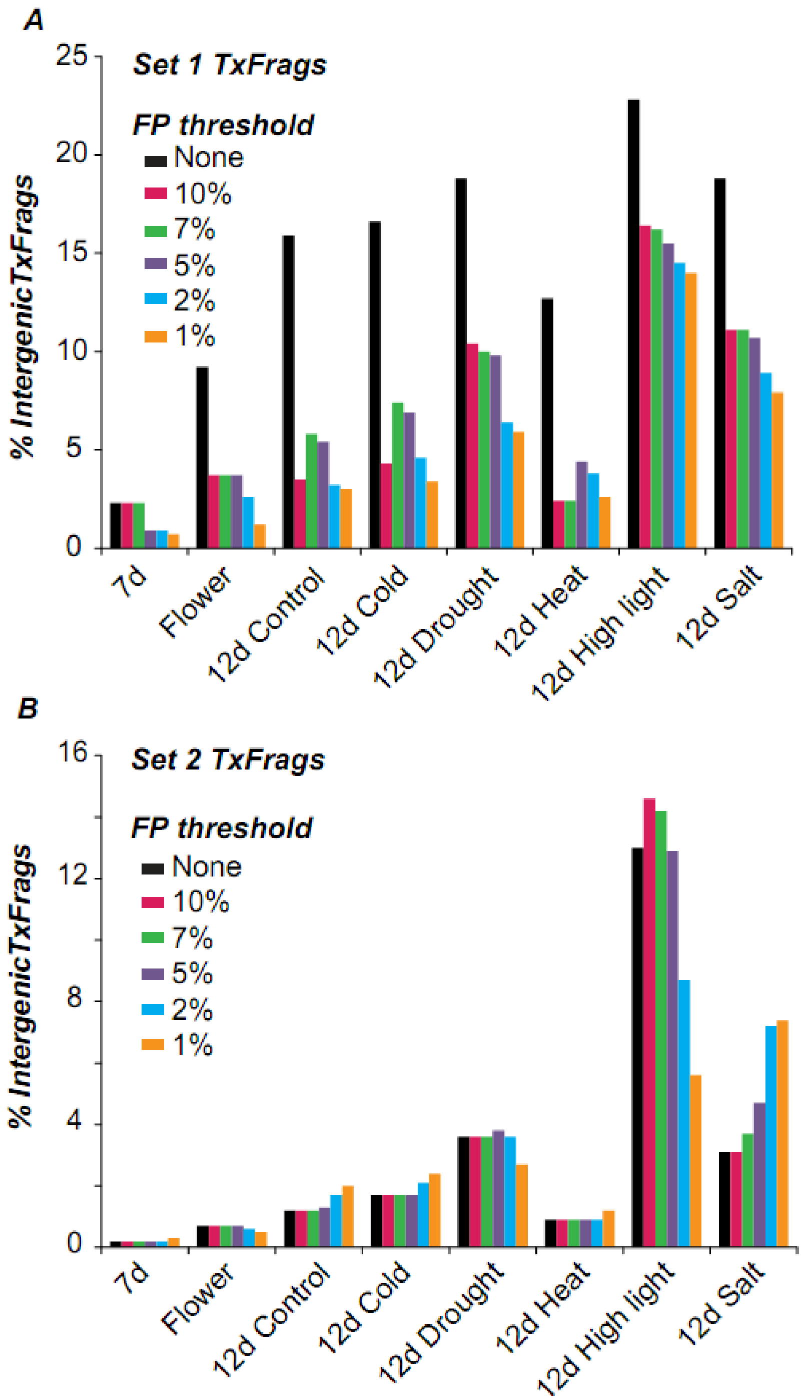


Figure 3.2: Percent TxFrags defined as intergenic at different FPKM thresholds among datasets. Set 1 TxFrags (A) and Set 2 TxFrags (B) identified as intergenic without an FPKM threshold (black) and at progressively more stringent FPKM thresholds according to transposon-based False Positive (FP) rates of 1%, 2%, 5%, 7%

Figure 3.2 (cont'd)

and 10%. The X-axis indicates the datasets used to identify TxFrags. The Y-axis represents percent true positive TxFrags that are intergenic at each FP threshold. Note that the percentage did not monotonically decrease because some TxFrags overlapping with annotated features also were filtered out when FP thresholds were applied.

influence estimates of ITF abundance (Figure 3.2A,B). For example, 9.2% Set 1 TxFrags from the flower data are considered ITFs when no FPKM threshold is applied, but this proportion drops to 3.7% with an FPKM threshold of 1.33, which corresponds to a 10% false positive rate (Figure 3.2A). Comparing between datasets by allowing a 10% false positive rate, ITF estimates differ by 7 (2.3-16.4%) and 73 (0.2-14.6%) fold based on Set 1 and Set 2 TxFrags, respectively. Despite these differences, there are two consistent ITF characteristics among datasets that separate Set 1 and Set 2 ITFs. Set 1 ITFs tend to be significantly shorter than Set 2 ITFs (Figure 3.1AD; KS test p<2.2e-16). In addition, Set 1 ITFs expression levels are not significantly different from those of Set 1 transposon TxFrags (Figure 3.1B, KS test p=0.28) but are significantly lower than protein coding gene TxFrags (KS test p<2.2e-16). Set 2 ITFs have significantly lower expression levels than protein coding gene TxFrags as well (Figure 3.1E, KS test p<1e-2), although the pattern is not as pronounced as for Set 1 ITFs, presumably due to the bias corrections applied on the dataset by Cufflinks. Our findings are consistent with earlier studies in *A. thaliana* (Matsui et al., 2008; Hanada et al., 2007) and mammals (van Bakel et al., 2010; Wang et al., 2004) which found that intergenic sequences tend to be lowly expressed.

We next focused on Set 2 TxFrags, which represent a more stringently defined set of transcripts. Across datasets, 0.2-14.6% of TxFrags are potentially derived from intergenic transcription based on a 5% false positive rate (Figure 3.2B). This proportion corresponds to 10,511 ITFs across eight RNA-seq datasets, together representing 6,545 non-overlapping intergenic transcribed genomic regions and spanning 3.6% of the assembled intergenic region in *A. thaliana*. Our ITF estimate is comparable to an earlier tiling array based study in *A. thaliana* where 7,719 un-annotated transcriptional units were defined as novel, non-protein coding RNAs (Matsui et al., 2008). Other studies have provided more conservative estimates of *A. thaliana* intergenic expressed regions - from 104 (Stolc et al., 2005) to 2,397 (Yamada et al., 2003b). In mammals, however, the ENCODE project as well as other studies have reported significantly more pervasive intergenic transcription (Bertone et al., 2004b; Birney et al., 2007; Kapranov et al., 2007). The ENCODE project reported that 488,906 (22.6%) TxFrags lie in intergenic regions and that 93% of the ENCODE bases have transcription evidence (Birney et al., 2007). Compared to *A. thaliana* (3.6%)*,* a significantly larger proportion of the ENCODE region is transcribed, even if we consider Set 1 TxFrags (13.7% at a 5% false positive rate) that are not as rigorously defined as Set 2 TxFrags.

There are several possible explanations for the differences in ITF pervasiveness between plants and humans. First, the ENCODE study analyzed transcripts obtained from 31 different cell lines and tissues, which represents a much broader sampling of the transcriptome than our study. Second, known issues with tiling arrays used in the ENCODE study, particularly cross-hybridization (van Bakel et al., 2010; Agarwal et al., 2010) may lead to an over estimation of ITFs. Consistent with this possibility, a recent RNA-seq study of human 293T cell total RNA found that only ~4% of reads were intergenic (van Bakel et al., 2011), similar to our *A. thaliana* estimate. Third, the intergenic space in *A. thaliana* comprises only ~40% of the genome, compared to ~99% in the human genome. If intergenic transcripts are largely derived from noisy transcription or genomic contamination, species with larger genomes may have more mRNA-seq reads from intergenic space. The fourth reason may be that larger genomes have more functional elements. However, variation in genome sizes can be due to extreme proliferation of transposable elements (Hawkins et al., 2006; Piegu et al., 2006). Thus larger genomes do not necessarily contain more genes. Finally, elements of our experimental design, such as the use of tissue samples with multiple cell types or insufficient coverage, may lead to an underestimate of ITFs. To address some of the issues concerning our study design, we analyzed cell-type specific transcriptome data obtained using directional Illumina sequencing.

Factors affecting ITF estimates

The datasets we analyzed have the following limitations that may affect estimates of ITF abundance (Clark et al., 2011). First, all datasets were generated using complex tissue samples that may render cell-type specific ITFs undetectable. Second, the sequencing was performed using single reads without directionality information, which may result in mis-assembly of ITFs. Third, the read length and coverage may be insufficient for detecting ITFs expressed at low levels. To address these issues, we directionally sequenced polyA-selected RNA from T87 suspension culture cells with longer reads (72bp) and greater depth (2-9 times more sequenced bases; ~2.3Gb, ~$3x10^7$ reads). We found that 0.9% of reads and 4.2% of TxFrags (identified using the same criteria as Set 2 TxFrags) from the suspension culture data are intergenic (Figure 3.3), consistent with the proportions of intergenic reads and TxFrags identified from more complicated tissues (Table 2.1). In addition, in the suspension cell dataset, 3,052 (9.8%) TxFrags and 170 (13.1%) ITFs overlap with ≥1 other TxFrags and ITFs, respectively, that are in the opposite orientation. Thus, lack of read directionality

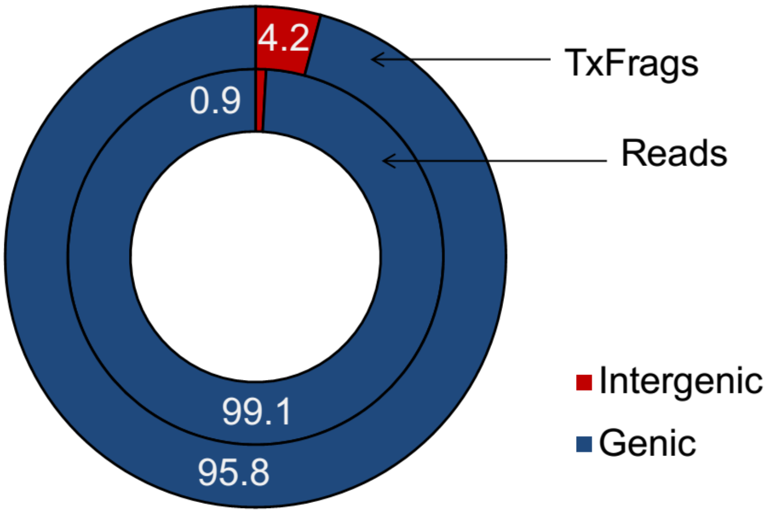


Figure 3.3: Directional sequencing of mRNA from T87 cells. Shown are the percentages of total reads mapping to the annotated vs intergenic portions of the *A. thaliana* TAIR v10 nuclear genome.

information and mis-assembly can lead to an ~13% under-estimate of ITFs that overlap in opposite orientations.

Another factor affecting the estimate of ITFs is that the datasets we have analyzed so far are derived from polyA RNA. Non-polyA RNA may comprise the bulk of the transcriptome and significantly contribute to intergenic expression (Xu et al., 2010; Armour et al., 2009; Cheng et al., 2005). However, an earlier study focusing on both polyA and non-polyA RNAs in *A. thaliana* found that 3.5% reads are intergenic (Lister et al., 2008), which is comparable to the 0.4-8.2% reads that are intergenic in the mRNA-seq datasets we analyzed (Table 2.1). In addition, a study of human 293T cell rRNA-depleted total RNA revealed that ~4% reads were intergenic (van Bakel et al., 2011). This suggests that our estimates of intergenic transcription in *A. thaliana* based on polyA RNA sequencing are reasonable. Nonetheless, detailed studies of non-polyA ITFs will be necessary to estimate the contribution of non-polyA transcripts to intergenic transcription.

Taken together, we have identified 6,545 ITFs (5% false positive rate) that are likely novel transcriptional units not previously defined in the *A. thaliana* genome. Two outstanding questions remain. First, because these ITFs are derived from polyA RNAs, are they parts of novel protein-coding or non-coding RNA genes? Second, do some of these ITFs have clear evidence of selection, therefore suggesting their functionality? To address the first question, we assessed the protein coding potential of ITFs by analyzing ribosome-associated transcripts and shotgun proteomics datasets.

Distinguishing coding from non-coding intergenic transcripts based on ribosome association

Translation initiation is the rate-limiting step in protein translation; therefore, transcripts associated with the ribosome are more likely to be translated (Kawaguchi and Bailey-Serres, 2002; Bailey-Serres et al., 2009). Studies in *A. thaliana* (Jiao and Meyerowitz, 2010; Branco-Price et al., 2008), mouse (Doyle et al., 2008) and yeast (Ingolia et al., 2009) have taken advantage of this property to globally investigate translational regulation. To assess whether ribosome association of intergenic transcripts is a good measure of their translation potential, we first sequenced ribosome-associated transcripts from 7-day old seedlings. After identifying the ribosome-associated TxFrags (R-TxFrags), we selected eight genomic regions with evidence of ribosome-association and seven without for *in vivo* translation studies. These regions overlap with putative small ORF (sORF) genes that were originally computationally predicted from intergenic regions (Hanada et al., 2007). Several of these regions have since been annotated based solely on computational predictions and/or cDNA evidence. The 5' UTRs and coding sequences of the sORFs were fused in frame to a yellow fluorescence protein (YFP) reporter that lacks a translational start codon, and the translation of these sequences in transiently transformed tobacco leaf epidermal cells was evaluated (see Methods).

Of the eight genomic regions with R-TxFrag evidence, five were translated in tobacco while only one of the seven regions without R-TxFrag support was translated. Thus, there was a significant enrichment of sORFs with R-TxFrag evidence among those translated *in vivo* (Fisher Exact Test, $p < 0.05$). The observed localization patterns of the protein fusions were largely consistent with signal peptide predictions, indicating that the fusion proteins were likely correctly translated and targeted in tobacco. However, three sORFs with ribosome association evidence do not appear to be translated in the transient expression assay. These sORFs may not be translated or this may be an artifact due to the use of a heterologous system (tobacco). One translated sORF is an annotated “Other RNA” gene (At1g31935; Table 2.2). In addition, a number of annotated "Other RNA" genes have either ribosome association or proteomics evidence, which highlights the importance of experimentally evaluating protein coding potential. Overall, based on the findings of our *in vivo* translation assays, we conclude that features with evidence of ribosome association are more likely to be translated than those without.

Translation evidence for ITFs

Given that ribosome association is a good indicator of translation potential of intergenic sequences, we further analyzed R-TxFrags from the 7-day-old seedling data to estimate the proportion of ITFs likely to be parts of coding genes. To address potential issues due to sequencing coverage or tissue-specific expression and translation, R-TxFrags were also identified using ribosome associated transcript data of whole flowers and specific floral domains expressing three homeotic genes (Jiao and Meyerowitz, 2010). For comparison, we also incorporated shotgun proteomics data from two studies examining protein expression in multiple tissues and developmental stages (Castellana et al., 2008; Baerenfaller et al., 2008). These data are collectively referred to as “translation datasets” and are summarized in Table 3..2.

Table 3.2. Description of datasets used for translation analyses

Source	**Stage or condition**	**Data type**[3]	**Read length (bp)**	**Reads mapped**[4]	**TxFrags**[5]
This study[1]	7-day old seedlings	R	36	12,077,020	31,230
Jiao and Meyerowitz, 2010[2]	AG domain (Stages 4, 6-7)	R	38	39,325,787	44,686
	AP1 domain (Stages 4, 6-7)	R	38	19,020,560	52,544
	AP3 domain (Stages 4, 6-7)	R	38	42,960,909	46,973
	Flower (Stage 4)	R	38	19,814,409	35,280
Castellana et al. 2008	Multiple tissues and developmental stages	P	NA	176,880	NA
Baerenfaller et al. 2008	Multiple tissues and developmental stages	P	NA	85,790	NA

[1] *Reads from three lanes of sequencing of technical replicates were pooled together for transcript assembly*

[2] *Reads from Stages 4 and Stages 6-7 were pooled together for transcript assembly for each domain*

[3] *R: ribosome associated transcript. P: proteomics.*

[4] *To the nuclear genome assembly of Arabidopsis thaliana, TAIR 10 release.*

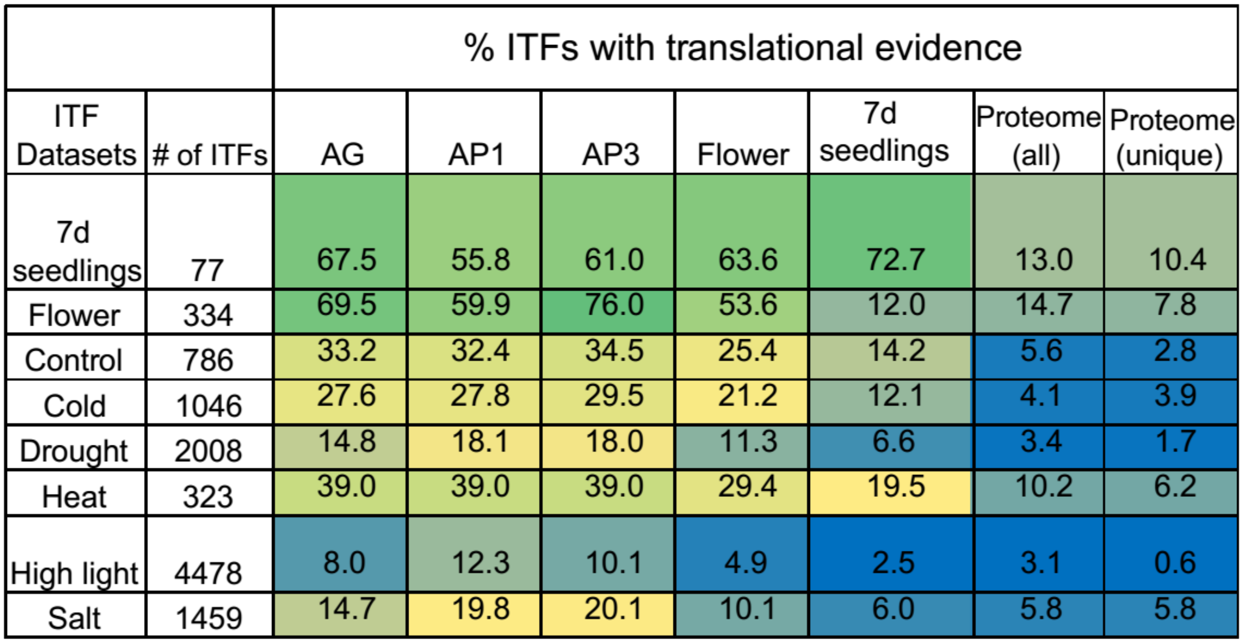
[5] *TxFrags generated using Cufflinks.*

As with mRNA-seq, most R-TxFrags and proteomics tags mapped to previously annotated regions in the genome, particularly protein-coding genes. Among ribosome immuno-precipitation datasets, 63-73% of protein coding genes have ≥1 R-TxFrags (Figure 3.4; Figure 3.5). Similarly, 74% of protein coding genes have ≥1 proteomics tags (Figure 3.5). In addition, 62-67% of the R-TxFrags overlap with ≥1 proteomic tags. These findings demonstrate that ribosome associated transcripts tend to be translated, consistent with our *in vivo* translation studies. On the other hand, 5-23% of annotated ncRNA genes and 7-15% of pseudogenes have uniquely mapped R-TxFrags and/or proteomics tags. If all annotated RNA genes are truly non-coding, calling a feature translated based on a corresponding R-TxFrag and/or proteomics tag can have a 5-23% false positive rate depending on the dataset (RNA, Figure 3.5). One anomaly is that 34.0% of transposons have proteomic tags, although only 1.9-3.4% have R-TxFrags (Figure 3.5). This discrepancy is in sharp contrast to our finding that the proportions of protein coding genes possessing R-TxFrags and proteomics tags are both ~70% (Figure 3.5). This observation, also noted in the original study (Castellana et al., 2008), is inconsistent with studies demonstrating reduced transcription of transposons (Schmid et al., 2005) and their extensive methylation (Zilberman et al., 2007; Zhang et al., 2006). Using number of proteomic tags as a proxy of protein expression level, transposons with proteomics evidence tend to have significantly fewer tags than protein coding genes (Figure 3.6, KS test, p<2.2e-16). In addition, 67.6% of transposons with proteomics evidence have only one tag compared with 26.6% of protein-coding genes suggesting that if the transposons are expressed and translated, it happens at significantly lower levels than protein-coding genes.

How many ITFs have evidence of translation? Among the 6,545 non-overlapping ITFs identified from eight mRNA-seq datasets, 2,107 (32.2%) have ≥1 R-TxFrags from ≥1 of the translation datasets analyzed (Figure 3.4, Figure 3.5). Unlike protein coding genes, there is substantially stronger support for ITF translation from ribosome association than from proteomics data (Figure 3.5). Some of the ribosome associated ITFs may contain protein coding regions even though there is no proteomics support, partly due to the fact that proteomics data tend to be biased toward more abundantly translated proteins (Baerenfaller et al., 2008). It is also likely that a significant number of ribosome associated ITFs are derived from the un-translated regions (UTRs) of protein coding transcripts. Taken together, even if the false positive rate is 23%, ~1,622 ITFs are likely parts of transcripts destined to be translated after eliminating potential false positives. Thus, a significant number of intergenic transcripts may be part of larger protein-coding genes, either as coding sequences or as UTRs. Our finding highlights the importance of assessing translational potential of polyA intergenic transcripts before defining them as sequences that function solely at the RNA level.

Relationship between ITFs and neighboring, annotated genes

Based on analysis of mRNA sequencing data, we uncovered thousands of short, low abundance transcripts from intergenic regions. In addition, many of these ITFs are supported by translation evidence. One immediate question is whether these ITFs, translated or not, are extensions of previously annotated or novel protein-coding genes. To address this question, we assessed whether there is a significant bias in where ITFs are located within the *A. thaliana* genome. Using ITFs identified from eight RNA-seq datasets (Table 2.1), we calculated the distance between each ITF and its closest



		% ITFs with translational evidence						
ITF Datasets	# of ITFs	AG	AP1	AP3	Flower	7d seedlings	Proteome (all)	Proteome (unique)
7d seedlings	77	67.5	55.8	61.0	63.6	72.7	13.0	10.4
Flower	334	69.5	59.9	76.0	53.6	12.0	14.7	7.8
Control	786	33.2	32.4	34.5	25.4	14.2	5.6	2.8
Cold	1046	27.6	27.8	29.5	21.2	12.1	4.1	3.9
Drought	2008	14.8	18.1	18.0	11.3	6.6	3.4	1.7
Heat	323	39.0	39.0	39.0	29.4	19.5	10.2	6.2
High light	4478	8.0	12.3	10.1	4.9	2.5	3.1	0.6
Salt	1459	14.7	19.8	20.1	10.1	6.0	5.8	5.8

Figure 3.4: Pairwise proportions of ITFs with translation evidence across different datasets. Each row represents each mRNA-Seq transcriptome dataset while each column represents a unique translation dataset - 5 translatome and 1 proteome. The percentage of ITFs in each transcriptome dataset having an overlapping piece of translation evidence is noted in the cells. A range of colors from green to blue is used to represent high to low percentages. Only uniquely mapping peptide tags were considered for overlap analysis when estimating percentages in the last column.

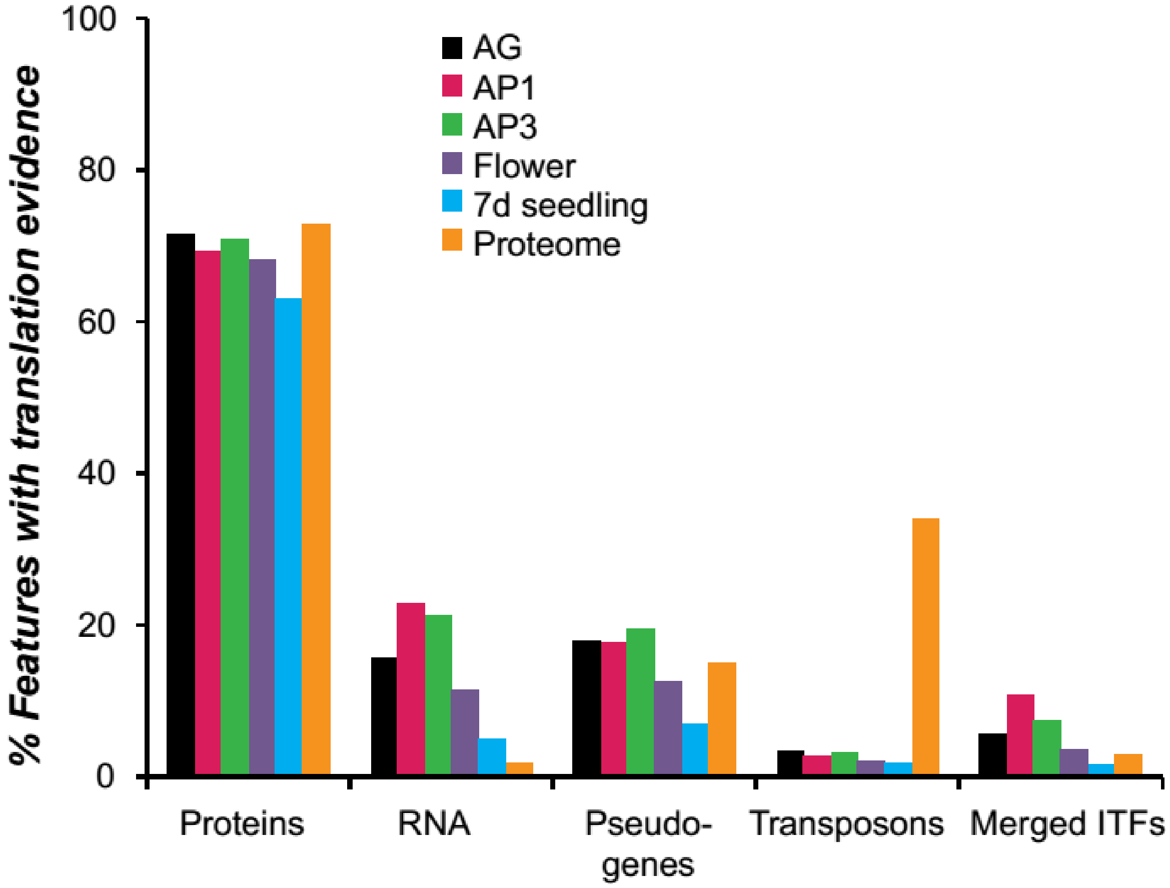


Figure 3.5: Percent total features with translation evidence. Percent features with overlapping translation evidence was calculated for protein-coding genes, RNA genes (excluding Other RNA), pseudogenes, transposons and ITFs obtained from the 7d seedling and flower transcriptomes. Ribosome immuno-precipitation data: AG, AP1, AP3, flower, and 7-day seedling. Proteomics data: combined data from two studies. Only uniquely mapping R-TxFrags and proteomics tags were used as evidence.

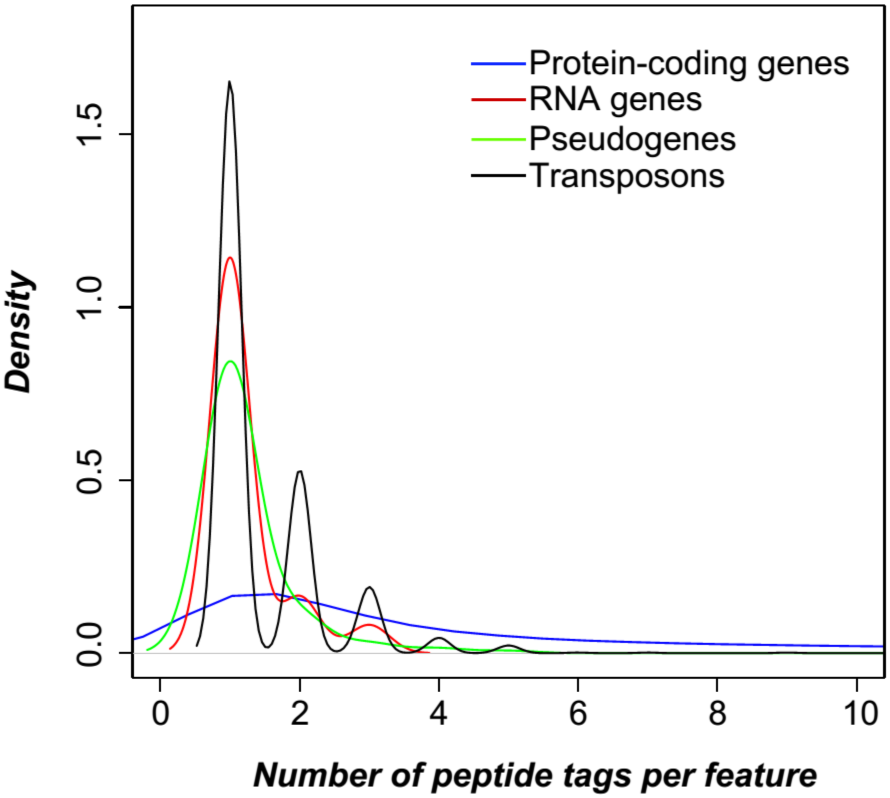


Figure 3.6: Number of peptides per feature. Distributions of the number of peptide tags for protein-coding genes (blue), RNA genes (red), pseudogenes (green) and transposons (black).

annotated protein coding gene. We found that although a substantial number of ITFs are closer to genes, they are not any closer than intergenic sequences sampled randomly based on ITF number and size (Figure 3.7B). This is contrary to the expectation that ITFs are predominantly extensions of existing genes.

Given that ITFs in general are not closer to neighboring genes than randomly selected intergenic sequences, do ITFs with translation evidence behave similarly? Firstly, we found that translation evidence (proteomic tags and R-TxFrags) tends to lie farther away from protein coding genes than random expectation (Figure 3.8A). However, ITFs with translation evidence tend to lie closer to genes than ITFs without translation evidence (Figure 3.8B), suggesting that most ITFs with translation evidence may be parts of neighboring protein-coding genes. If translated ITFs are indeed missing parts of annotated genes, ITFs closer to genes with transcription evidence should be enriched in the translated set compared to ITFs closer to non-transcribed genes. Consistent with this expectation, among the 4,942 ITFs closest to a transcribed annotated protein, 37.9% have translation evidence while among the 563 ITFs closest to a non-transcribed annotated protein, only 14.2% have translation evidence (Fisher Exact Test, p<2.2e-16, Figure 3.7C). These observations suggest that most ITFs with translation evidence that are close to annotated genes may be missing parts of those genes or associated with the transcription of those genes via an unknown mechanism.

Taken together, we have demonstrated the presence of ITFs from >6,000 intergenic regions in *A. thaliana* from multiple RNA-sequencing datasets. More than 20% of these ITFs are likely translated or are part of protein coding transcripts. Among the 6,545 ITFs, 59.4% are located >300bp away from an annotated gene. Of these, 847

(21.7%) have translation evidence. Considering that 300bp is ~90th percentile of both *A. thaliana* intron and UTR lengths, these relatively distant ITFs may be parts of novel transcriptional units. However, ITFs, in general, tend to be significantly shorter and expressed at lower levels than protein coding genes. We also find that ITFs, in general, tend to be expressed narrowly, in a dataset-specific manner, while TxFrags corresponding to annotated features are present in multiple datasets (Figure 3.7A). The translation of ~32.2% ITFs is supported by ≥1 ribosome immuno-precipitation and/or proteomics datasets, compared to 88.0%, 44.6%, and 36.9% for protein coding genes, pseudogenes, and transposons, respectively. In terms of translation, ITFs behave similarly to pseudogenes and transposons. Previous studies have suggested that breadth of expression as well as level of expression can be considered as proxy indicators of functionality (Nuzhdin et al., 2004; Subramanian and Kumar, 2004; Movahedi et al., 2011). However, genes can have highly specific expression and/or low expression levels. Thus, one remaining question is whether these ITFs are parts of functional sequences with clear evidence of selection.

Evidence of natural selection on ITFs at the nucleotide level

Intergenic transcripts that are independent transcriptional units may be derived from noisy, background transcription or unannotated genes that are functional (Struhl, 2007; Dinger et al., 2009; van Bakel et al., 2010; Clark et al., 2011; van Bakel et al., 2011). Transcripts not important to cellular function are expected to accumulate mutations much like neutrally evolving sequences. In contrast, functional ITFs should be selected for and show signs of non-neutral evolution. To assess whether there is a clear signature of natural selection that is indicative of functionality, the ITF nucleotide

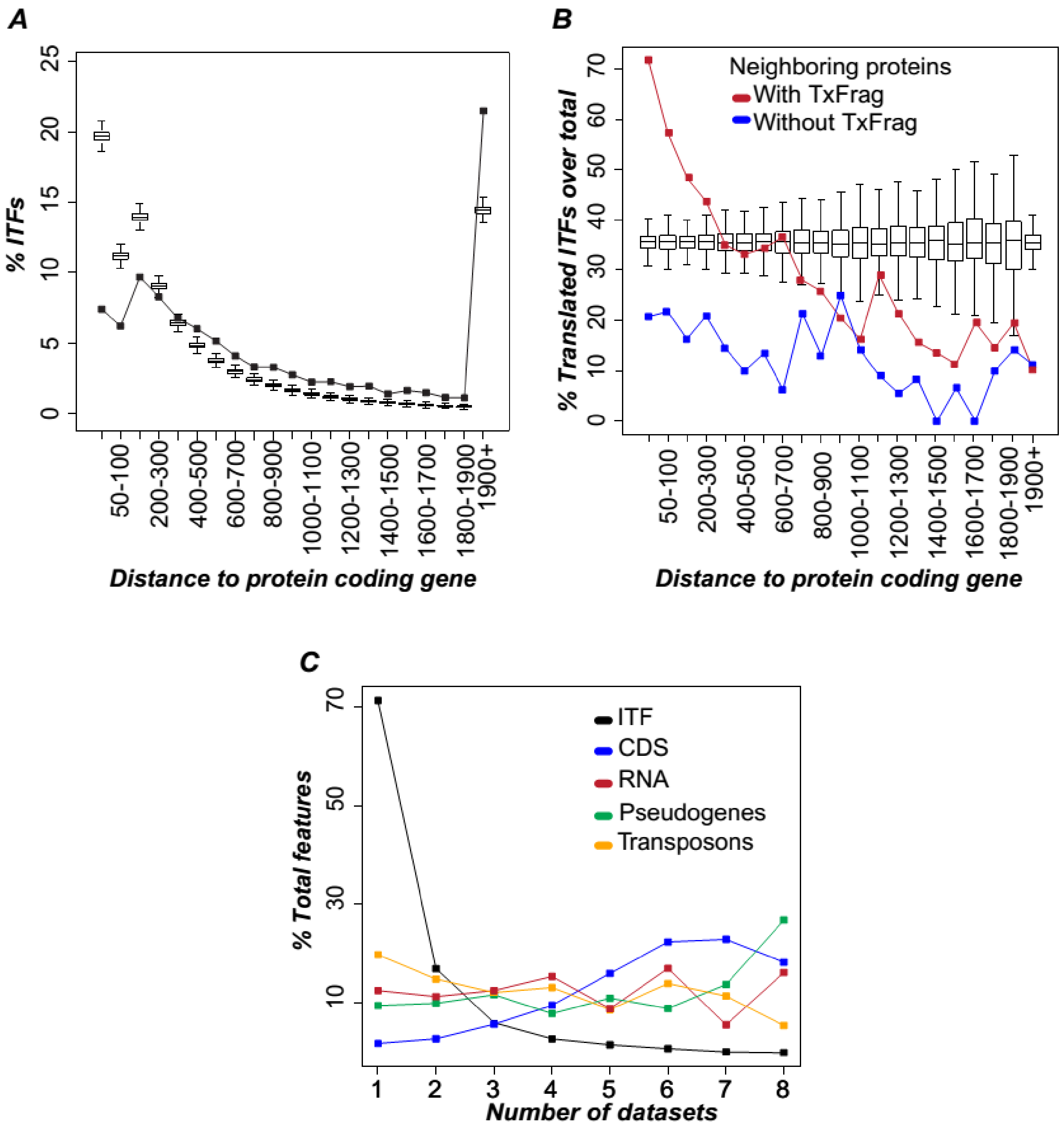


Figure 3.7: Proximity of ITFs to neighboring genes. (A) Distance distribution of ITFs to their nearest protein-coding genes. The boxplots depict distance distributions between 10,000 sets of randomly sampled intergenic sequences and their nearest protein-coding genes. (B) Percent translated ITFs over all ITFs in the same distance bin is shown as a function of distance to the nearest protein coding gene. ITFs neighboring

Figure 3.7 (cont'd)

proteins with and without transcript evidence are represented by red and blue lines, respectively. Boxplots represent the randomly expected proportions in each distance bin obtained by permuting the association between distance and presence/absence of translation evidence. The medians of random expectations are ~35% because ~35% ITFs have ≥1 translation evidence. (C) "Breadth of expression" (as indicated by the number of datasets where a feature can be found) of ITFs (black) and TxFrags mapped to protein-coding genes (blue), RNA genes (red), pseudogenes (green) and transposons(orange).

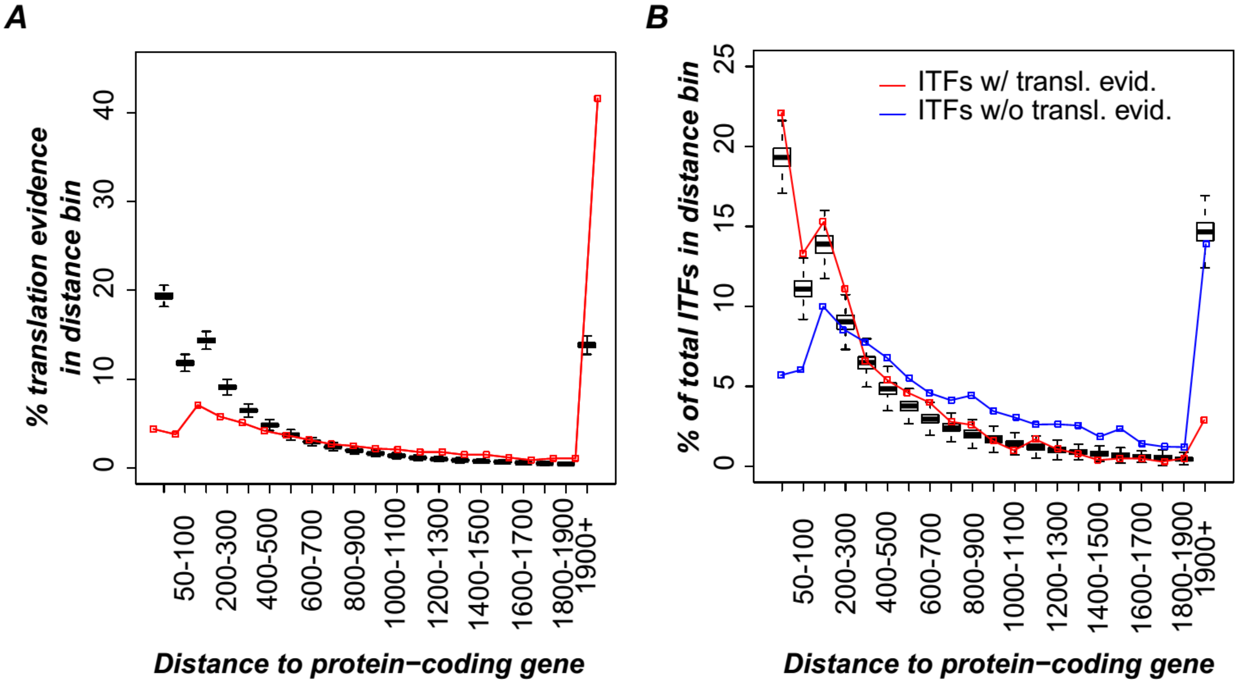


Figure 3.8: Distances of translation evidence from annotated genes. (A) % of total translation evidences (R-TxFrags + peptide tags) (Y-axis) and (B) % of ITFs with (red) and without (blue) translation evidence as a function of distance from nearest gene (X-axis). The boxplots represent median % distribution obtained through a simulation

Figure 3.8 (cont'd)

involving randomly sampled intergenic sequences. The lines represents observed % of translation evidence in each distance bin.

substitution rates were estimated using syntenic genomic regions of *A. thaliana* and *A. lyrata*, which diverged from their common ancestor ~10 million years ago (Hu et al., 2011). We also estimated the four-fold degenerate site substitution rates of protein-coding orthologs as proxies for neutral evolution rates. Substitution rates for protein coding genes, RNA genes and randomly chosen intergenic regions not overlapping with ITFs were also estimated for comparison.

Among 6,545 ITFs, only 1,238 (18.9%) have identifiable syntenic regions for substitution rate estimation between the two *Arabidopsis* species. The proportion of syntenic ITFs is significantly lower than those of protein coding genes (90.9%, Fisher Exact Test p<2.2e-16), RNA genes (35.7%, p<9.1e-10), and pseudogenes (33.4%, p<2.2e-16) but significantly higher than transposons (10.9%, p<1.5e-16). Thus, many "orphan" ITFs without putative orthologs likely evolved rapidly with little or no selective constraint. For ITFs found within syntenic regions, substitution rates are significantly higher than not only those of annotated protein coding genes (KS test, p<2.2e-16; Figure 3.9A). On the other hand, ITF substitution rates in general are significantly lower than those of four-fold degenerate sites (KS test, p<2.2e-16; Figure 3.9A). These observations suggest that ITFs may constitute a mixed population; the first population under strong selective constraint and the second one evolving neutrally. Using the 5th percentile of the four-fold degenerate site rate distribution (rate=0.07) as a threshold, only 6.4% of the 6,545 ITFs are likely under strong purifying selection. The remaining 93.6 % are likely under little or no purifying selection. To control for local rate variation, we compared the rate of each ITF to the four-fold site rates of neighboring genes.

Based on this approach, a much smaller percentage, 2.7%, of ITFs were found to be under selection (Figure 3.9B).

One issue in comparing any sequence feature to four-fold sites is that there can be significant alignment bias. This is because a sequence feature, e.g. an ITF, is aligned to its putative ortholog at the nucleotide level whereas four-fold sites are identified from nucleotide sites originally aligned based on protein sequences. Given that the alignment process involves finding an alignment with the best score, regardless of whether the sites are homologous or not, it will tend to make a nucleotide-based alignment look more similar than it really is. Thus, the lower substitution rate among sequence features compared to four-fold sites can simply be due to this artifact. To account for this, we selected random intergenic regions that have no evidence of expression and calculated their substitution rates. Similar to ITFs (6.4%), 7.5% of the random intergenic region samples are under strong purifying selection. More importantly, there is a small but statistically significant difference in the substitution rate distributions between ITFs and random intergenic sequence (Median rates: 0.09 and 0.08, respectively, KS test, p<2e-05). Therefore, after accounting for potential alignment bias, potentially even fewer than 6.4% of ITFs are under significant selective constraints. The implication is that the majority of ITFs appear to evolve similar to presumably non-functional, non-expressed random intergenic regions between species.

To assess the possibility that that some ITFs may have a species-specific function in *A. thaliana* making the signature of selection only obvious at the intra-specific level, we analyzed genomic sequences of 80 accessions of *A. thaliana* (Cao et al., 2011) and estimated the Nucleotide diversity (π) for ITFs, annotated sequence features

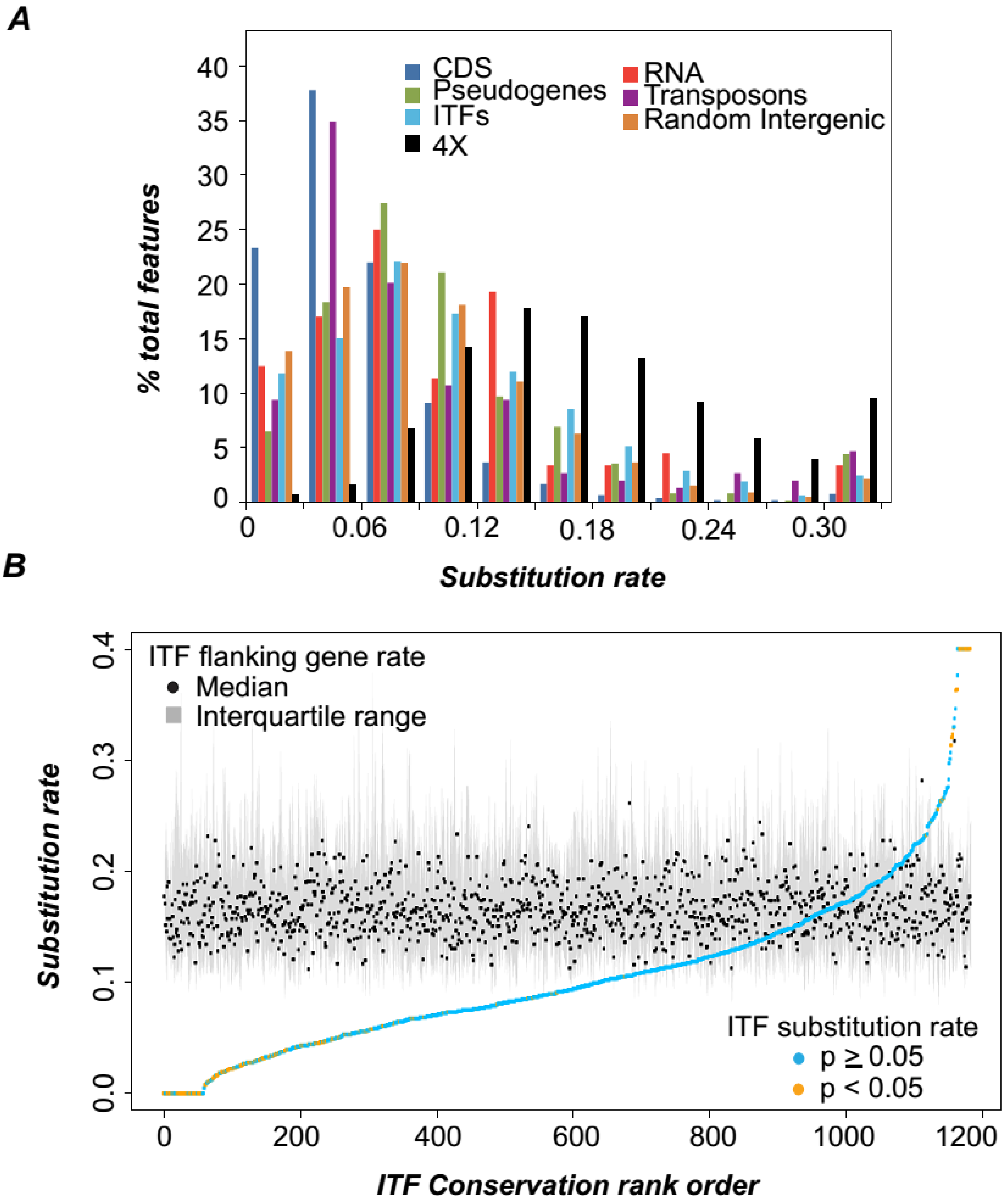


Figure 3.9: Evolutionary conservation of ITF sequences. (A) Between species nucleotide substitution rate distributions of different features and four-fold degenerate sites (4x). (B) Substitution rates of ITFs compared with local substitution rates of 4x

Figure 3.9 (cont'd)

sites. Fourfold degenerate sites of up to 60 neighboring protein-coding genes were used to determine distributions of local substitution rates. Black circles indicate medians of the distributions, gray lines define the interquartile ranges and each filled orange or blue circle indicates the substitution rate of the ITF in the given region. The ITFs are arranged from low to high z-scores. A filled orange circle indicates a significant z-score at p<0.05, while a filled blue circle indicates p≥0.05.

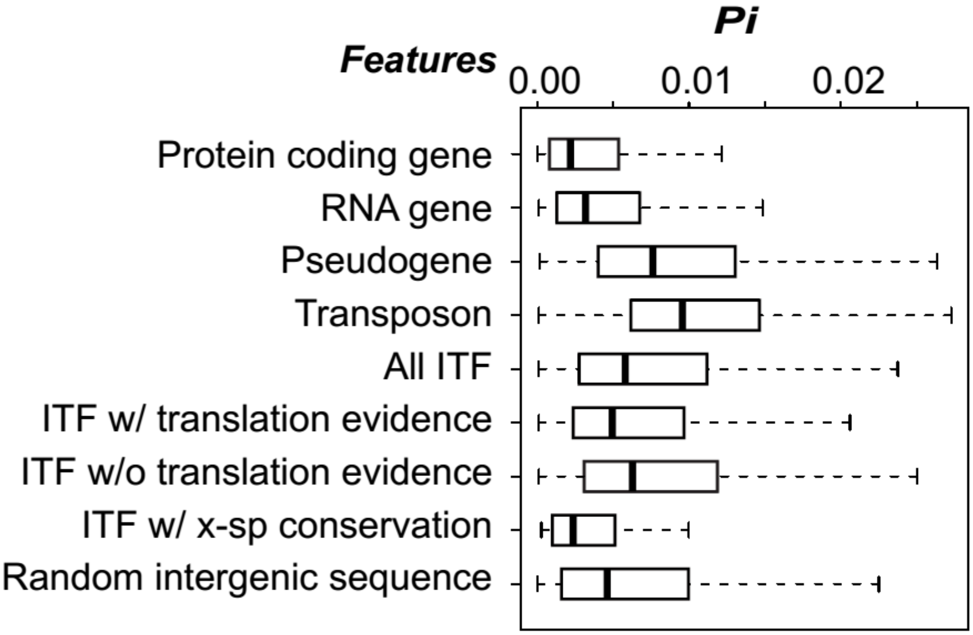


Figure 3.10: Distribution of π values for genomic features. The π values were calculated using population genomic data of 80 *A. thaliana* accessions. X-sp: cross-species. Random intergenic sequences were selected from regions without transcript support.

and randomly selected intergenic sequences not overlapping with ITFs. The π value allows us to assess genetic variability of different genomic features among *A. thaliana* populations (Li, 1997). Our findings suggest that ITFs have π values significantly higher than coding sequences and RNA genes (KS tests, both p<2.2e-16), but similar to random intergenic sequences not overlapping with ITFs (Figure 3.10). We also estimated Tajima's D, Fu and Li's D and Fay and Wu's H based on site-frequency spectrum to assess if ITFs are under selection. The distributions of all three statistics were comparable between ITFs and randomly sampled, unexpressed intergenic sequences (KS tests, all p>0.1) but significantly different from those of protein coding genes and RNA genes (KS tests, all p<0.001). These findings suggest that there is much more relaxed selection within species on ITFs compared to protein coding genes. Furthermore, the intensity of selection on ITFs is similar to that on random intergenic regions, which are likely largely non-functional and evolve neutrally.

Selection on ITFs with translation evidence

Our findings indicate that some ITFs are under strong selective constraint and may be functional. However, a much larger number of ITFs do not have clear signatures of selection. One immediate question is whether ITFs under strong selective constraint tend to be those that are translated given that ITFs with translation evidence tend to be located closer to neighboring genes. We performed a similarity search between ITF sequences and the genomes of fifteen land plants ranging from a bryophyte to angiosperms. For comparison and to address potential annotation issues, we also analyzed TxFrags mapping to protein-coding genes and RNA genes. ITFs with evidence of translation were slightly more conserved over ITFs with no evidence of translation (compare Figure 3.11A, B), but most ITFs have significant similarities only between *A. thaliana* and *A. lyrata*, with sequence similarity rapidly declining beyond the *Arabidopsis* genus. On the other hand, there is a significantly higher degree of cross-species similarity between protein-coding genes – 6,895 of the 10,000 randomly selected protein sequences had E-values <1e-5 in >1 species (Figure 3.11C). Even RNA genes, which are not expected to be translated, have higher sequence similarities than ITFs (Figure 3.11D). Thus, at both the nucleotide and amino acid sequence level, relatively few ITFs are under selection based on cross-species comparisons.

Of the 847 ITFs with translation evidence that are located >300bp away from genes, 799 did not show significant conservation. Considering that these ITFs tend to be expressed at low levels, such sequences may represent translational noise. But we cannot rule out the possibility that they are lineage-specific coding sequences. Of the 49 ITFs that did show conservation, 16 had similar sequences present in >1 species at the amino acid level (E-value<1e-5) and 10 showed overlap with computationally predicted small ORFs with high protein coding potential (Hanada et al., 2010). The π value distribution of these 49 ITFs is statistically indistinguishable from that of TxFrags mapping to protein coding sequences (KS test *p*=0.2; Figure 3.9B), suggesting that these ITFs are under a similar selective constraint as protein coding sequences. These ITFs may thus represent novel functional genes. Nonetheless, only ~5% ITFs with translational evidence are subject to strong purifying selection among *A. thaliana* ecotypes, reinforce the notion that most of them are products of noisy transcription.

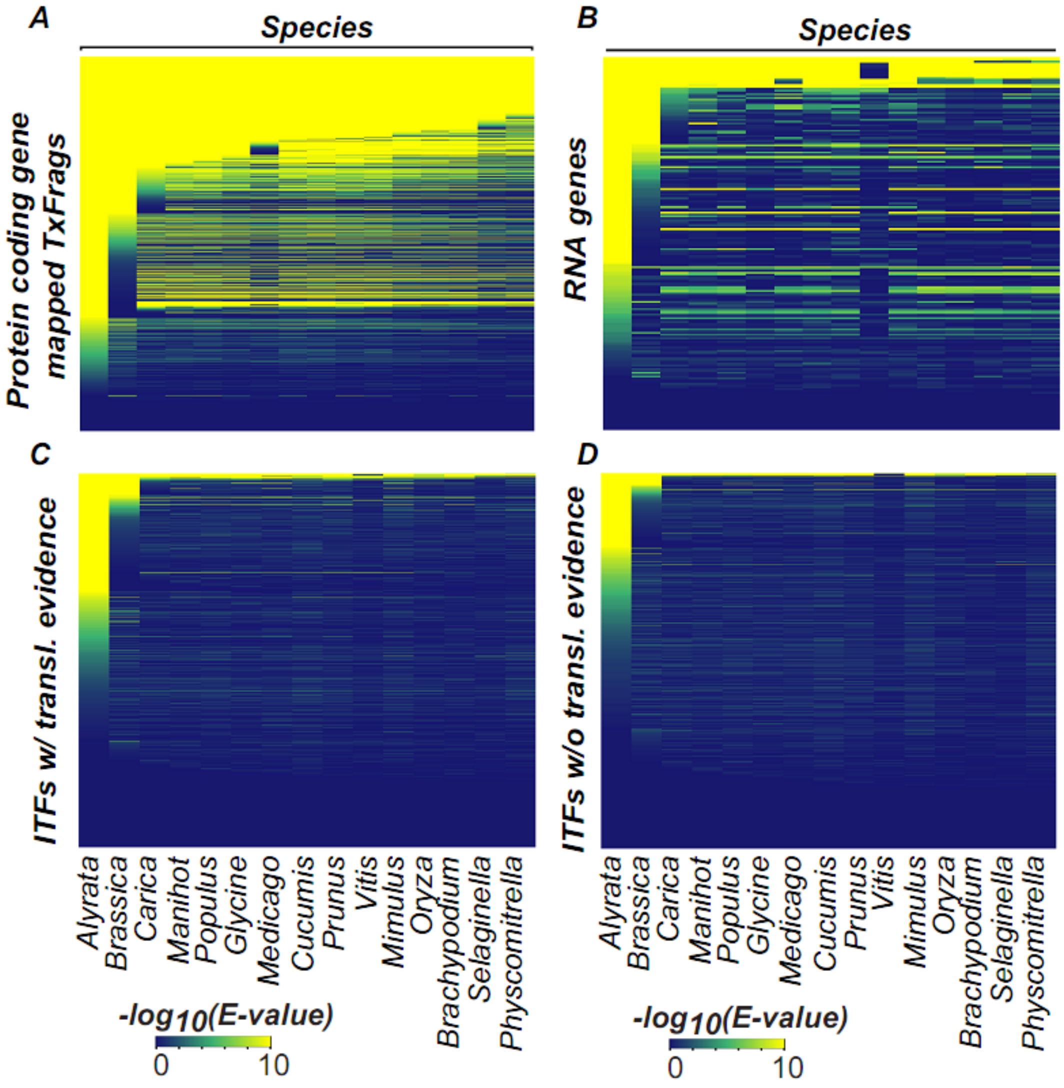


Figure 3.11: Evolutionary conservation of ITF sequences. (A-D) Heat maps indicate degree of cross-species similarity of ITFs with translation evidence (A) 10,000 randomly selected TxFrags mapped to protein coding genes ITFs without translation evidence (B) annotated RNA genes, (C) ITFs with translation evidence and (D), ITFs without translation evidence. Rows represent features and columns represent the subject species for similarity search. The Expect (E)-values were converted to a negative logarithmic scale and adjusted to be between 0 to 10, with 0 (blue) indicating E-value ≥1 and 10 (yellow) indicating an E-value ≤1e-10.

CONCLUSIONS

In this study, we analyzed the intergenic polyA transcriptome of *A. thaliana* to address the issues of abundance, coding/non-coding nature and functional relevance of intergenic polyA transcripts. Our results indicate that ~5% of the TxFrags in the *A. thaliana* transcriptome can be reliably called intergenic. One limitation of our analyses, as we have noted before, is our focus on the polyA fraction of the transcriptome. It is likely that the non-polyA fraction of the transcriptome may harbor additional novel non-coding genes that need to be further investigated. Another limitation is that the read lengths of the RNA-seq data we used are short. It is possible that some ITFs belong to the same transcriptional units, making the number of ITFs an overestimate.

Our results indicate that ~3.6% of the intergenic space in *A. thaliana* is transcribed by RNA Pol II, and ~40% of what is transcribed tends to lie within 300bp of annotated genes. Around one third of ITFs have translation evidence, and we find a significant bias in their distribution; they tend to be closer to transcribed protein-coding genes, raising the possibility that some ITFs may in fact be unannotated extensions of known genes. Our primary sequence-level evolutionary analysis indicates that a relatively low fraction (~5%) of the ITFs have experienced strong purifying selection either within-species or between-species. We should emphasize that our criteria for evaluating selection is stringent. Furthermore, some ITFs may be more strongly constrained at the secondary structural level similar to non-coding RNA genes (Washietl et al., 2005) i.e. sequence level changes may be tolerated as long as a functional structure is maintained. In addition, some long non-coding RNAs such as *Air* and *Xist* are poorly conserved (Pang et al., 2006; Ponting et al., 2009), indicating that lack of conservation may not always mean lack of function. Thus, it is likely we will miss some ITFs that are under selection or are functional. However, most ITFs are short, and unlike the long non-coding RNAs, tend to be dataset specific, and are expressed at a very low level compared to annotated genes. In addition, most ITFs have characteristics more similar to pseudogenes and transposons than to protein coding and RNA genes. Even compared to intergenic sequences without evidence of expression, ITFs tend to evolve faster. Taken together, most ITFs bear the hallmarks of neutrally evolving sequences, suggesting they are products of noisy transcription as proposed earlier (Struhl, 2007).

The idea of transcriptional noise has been intensely debated over the past few years. Some studies support the theory that the transcriptional machinery might be error-prone and that many transcripts may be the result of false starts and/or stops (Struhl, 2007; Li et al., 2007a; Xu et al., 2009; van Bakel et al., 2010). Such errors may occur because it is not possible to regulate any biological process to the point that there is no error; noisy transcription may exist simply because it incurs little fitness cost. Considering that the vast majority of mutations are neutral or nearly neutral (Ohta, 1992), this paradigm for gene evolution may also apply to other molecular events, including transcription. As has been postulated before, the target of natural selection may be the effects of error-prone transcription rather than the transcriptional process itself (Hurst, 2009). Another possibility is that the effect of genetic drift, particularly on organisms with smaller effective population sizes, may render selection against erroneous transcript production ineffective. In either case, the hypothesis is that transcriptional errors may not always be subjected to purifying selection. Based on our findings, it would seem that much of the intergenic transcription falls into this category. We should emphasize that, for most ITFs, there is little or no evidence to reject the null hypothesis that ITFs are non-functional. The reason to consider non-functionality as null is simply because only functionality can be experimentally tested (van Bakel et al., 2010).

In a recently published series of papers by the ENCODE consortium, 80.4% of the human genome was found to have evidence of functionality (Bernstein et al., 2012). Functionality of a sequence, in this case, was defined at the biochemical level. That is, a functional sequence is presumed to have at least one RNA and/or chromatin-associated event, such as transcription factor binding, nucleosome binding, DNA methylation, in at least one cell type. However, it remains unclear to what extent these biochemically functional sequences may have physiological function, given these biochemical events can also be due to noise. As an example, among the novel long intergenic polyA TxFrags obtained in the ENCODE study, only 4% of the bases were conserved between humans and macaques. Among the 96% bases not conserved, only 6-11% showed evidence of lineage-specific constraint in humans, comparable to what we find in *A. thaliana* (Ward and Kellis, 2012). In addition, intergenic TxFrags found in this study were found to be present at <0.1 copies/cell (Djebali et al., 2012), consistent with our finding that most of the *A. thaliana* ITFs have a very low expression level.

Considering the dramatically increased sensitivity and throughput in sequencing, noisy transcription and even contaminating sequences, such as trace amount of genomic DNA, can be readily detected. Thus, we propose that a transcription event as detected by sequencing may not be considered functional by default. Evolutionary constraint acting on novel sequences or other evidence should be demonstrated prior to their annotation. The increasing availability of population-wide polymorphism datasets and genome sequences of related species provide more robust tools for such evolutionary studies, especially those focusing on lineage-specific selection (Ward and Kellis, 2012). In addition to the question of functionality, we show that a significant number of ITFs are associated with ribosomes and a smaller fraction of them have proteomics tags. Thus, novel transcripts should not be regarded as non-coding by default without rigorous experimental analysis of their coding potential. Our results also suggest the need to have a clearer understanding of the mechanistic aspects of RNA polymerase action on how noisy transcription may arise. Use of an integrated approach to validate novel RNA predictions and their functionality would be important in this regard. For example, a recent study in mouse used an array of approaches including identification of conserved histone modification marks, evolutionary analyses of promoter regions, gene set enrichment analysis, transcription factor chromatin immuno-precipitation and RNA interference assays to identify putative functional long non-coding RNA (Guttman et al., 2009). We surmise that such an approach will allow us to explore more deeply the mechanistic and evolutionary aspects of the transcriptional process in plants.

MATERIALS AND METHODS

Plant material and RNA isolation

For transcriptome and ribosome immuno-precipitation studies, transgenic *A. thaliana* Columbia seeds expressing a His_6FLAG-tagged version of the ribosomal large subunit protein L18B (*35S:HF-RPL18B*), were surface sterilized, stratified for three days, and sown on 0.5X Murashige and Skoog media containing 1% (w/v) sucrose and 0.4% phytagel. Seedlings were grown vertically under a 16h day (125 μE m^{-2} s^{-1} photosynthetically active radiation)/8h night cycle for seven days as described previously (Branco-Price et al., 2008). Seven-day old seedlings were harvested at the end of the light period. Total RNA extraction and ribosome immunoprecipitation were done as previously described (Branco-Price et al., 2008) for three biological replicates except that the RNeasy Plant Mini Kit purification step was omitted. Total RNA and ribosome-immunoprecipitated RNA was quantified using a Nanodrop spectrophotometer (Nanodrop Technologies, USA) and RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

Illumina RNA-seq and data analysis

For each in-house (7-day seedlings) RNA sample, cDNA libraries were constructed by first isolating poly-adenylated RNA from three µg of total or ribosome-associated RNA. Libraries for RNA-seq were prepared using the Illumina mRNA-seq Sample Prep Kit. Briefly, polyA RNA was fragmented and reverse transcribed using random primers. Adapters were ligated to double stranded cDNA, and fragments from 175-225 bp were gel-purified. After PCR amplification, the cDNA libraries were sequenced on an Illumina Genome Analyzer. Each library was loaded onto at least two lanes; however, usable sequence was only obtained for one polyA RNA library and two ribosome-associated RNA libraries (four lanes total). Three lanes of ribosome-associated RNA sequencing, corresponding to two biological replicates, were combined to give 19,818,643 reads and one lane of polyA RNA yielding 7,028,772 36-bp reads was obtained. The original sequencing reads were deposited in NCBI Short Read Archive under the accession number SRA053376 (http://www.ncbi.nlm.nih.gov/Traces/sra). All public datasets (Tables 2.1, 2.2) were downloaded from NCBI Short Read Archive. The following procedure was commonly performed on both the in-house and the public datasets.

The short reads, after quality trimming, were mapped to the TAIR10 *A. thaliana* genome using Bowtie v 0.12.7 (Langmead et al., 2009) and TopHat v 1.2.0 (Trapnell et al., 2009). The default settings were used except that the maximum combined intron size was set at 5,000 bp. The mapped reads were assembled with two approaches. In the first approach, reads with overlapping genomic locations were merged into TxFrags (Set 1 TxFrags) without considering the possibility that neighboring TxFrags may be derived from the same transcriptional units. In the second approach, Cufflinks 0.9.3 (Trapnell et al., 2010) was used with default parameters except a maximum combined intron size was set at 5,000bp (Set 2 TxFrags). All TxFrags overlapping with annotated features by ≥1 bp, including introns or UTRs, were flagged as genic transcripts.

Estimating level and breadth of expression

For estimating expression level, the Fragments Per Kilobase of exon model per Million mapped reads (FPKM) measure was used. Since Set 1 TxFrags represent a set of unique, non-overlapping TxFrags, the entire TxFrag was considered an exon for the purpose of FPKM estimation. For Set 2 TxFrags, FPKM values were estimated by Cufflinks. The breadth of expression was calculated for the 6,545 merged ITFs. For comparison, we also measured the expression breadth of TxFrags mapping to annotated features. We used the number of datasets in which a particular feature had expression evidence (≥ 1 overlapping TxFrag) as a measure of the breadth of expression of that feature.

5’ RACE and transient expression of YFP fusion proteins in tobacco

These experiments were performed by Dr. Melissa Lehti-Shiu in collaboration with Yanni Sun and Dr. Federica Brandizzi.

The 5’ UTRs of putative coding sORF sequences were identified from publicly available cDNA sequences (Aubourg et al., 2007) or were amplified by 5’ RACE (RLM-RACE kit, Ambion or SMART RACE cDNA amplification kit, Clontech). The 5’UTRs and coding sequences of each sORF were amplified from genomic DNA and cloned into the TOPO-TA entry vector (Invitrogen). The sequences were then transferred by recombination mediated by LR clonase (Invitrogen) into a modified pMDC83 destination vector (Curtis and Grossniklaus, 2003), containing the enhanced YFP sequence (Clontech), lacking a translational start codon, under the control of the 35S promoter. Constructs containing sORFs fused in frame with YFP were transformed into *Agrobacterium tumefaciens* GV3101. Transient transformation was performed to express sORF-YFP fusions in tobacco (*Nicotiana tabacum*) cells (Sparkes et al., 2006). Transgenic *A. tumefaciens* cells were cultivated overnight, and 200ul of the culture (OD A600~1-2) was pelleted and resuspended with sterile water to 0.1 OD*. A. tumefaciens* cells were infiltrated into tobacco leaves, and the infiltrated tobacco was kept under constant light for 72 hours. Infiltrated areas of tobacco leaves were detached, and observed under an inverted laser scanning confocal microscope (Olympus Spectral FV 1000). YFP signals were detected with the 514nm argon laser excitation line with a band pass emission filter of 517.5 to 542.5 nm. For visualization of AT_1|-|2|5786755-5786853 and ERD2 colocalization, equal volumes of *A. tumefaciens* cultures were mixed prior to infiltration. Fluorescence was visualized after three days with a Meta Zeiss confocal using the Argon laser excitation lines of 458 nm and 514 nm, and bandpass emission filters 475 to 525 nm and 530 to 600nm for blue shifted GFP and YFP, respectively.

Evolutionary conservation analyses

To identify ITFs under selection at the nucleotide level, the *A. thaliana* ITFs were first mapped to the *A. lyrata* genome using GMAP version 2007-09-28 (Wu and Watanabe, 2005) with default settings. Putative ITF orthologs were defined as pairs of similar sequences (≥80% coverage, ≥80% identity, ≥40 bp match length) between *A. thaliana* and *A. lyrata* flanked by ≥1 putative orthologous genes among 10 protein-coding genes on either side of the ITF. Putative orthologs between these two species were identified based on reciprocal best match and synteny information. The orthologous ITFs were aligned using Clustal 2.1 (Thompson et al., 1994) and the nucleotide substitution rate was calculated using baseml with the HKY substitution model in PAML (Yang, 2007). ITFs with a substitution rate lower than the 95th percentile (HKY distance ≤ 0.07) of the fourfold-degenerate site substitution rates of all protein-coding orthologs were deemed to be evolving under strong purifying selection. To control for genome wide variation in local substitution rates, the fourfold-degenerate site substitution rates of up to 60 protein coding genes in the vicinity of the ITFs were used to determine the 5% significance level using a z test. We did not conduct a *Ka/Ks* analysis for ITFs because (1) most ITFs are short thus the variance of Ka and Ks estimates for short sequences tend to be high and (2) it is not clear what the correct reading frame is, if these ITFs are translated. Instead, to compare the levels of conservation at the coding sequence level, we performed a translated BLAST search between ITF/TxFrag sequences and the draft assemblies of fourteen plant species in Phytozome 5.0 (http://www.phytozome.org/). The negative logarithm of the E-value of the top match in each species was used to plot a heatmap. All negative log values ≥10 or ≤0 were set to 10 and 0 respectively.

For conservation analyses within species, we used polymorphism data in the form of a genome matrix file from 80 different *A. thaliana* accessions (Cao et al., 2011). For each genomic feature type, we reconstructed the aligned sequences based on the genome matrix file. The aligned sequences were analyzed for π, Tajima's D and Fu and Li's D using Variscan (Vilella et al., 2005) with the following parameters: *RefPos=1, Outgroup=none, RunMode=12, UseMuts=0, CompleteDeletion=0, FixNum=1, NumNuc=60*. For Fay and Wu's H, we used the orthologs in *A. lyrata* as outgroups with RunMode=22 (Figure 3.12). For comparison, π values for 10,000 randomly chosen protein coding genes, RNA genes, transposons, and pseudogenes were also calculated. For features with <10,000 sequences, bootstrap samples were used. To determine the background π values, 10,000 random intergenic sequences were sampled based on the size distribution of ITFs. Only those intergenic sequences not overlapping with any TxFrags were used for analysis. For each sequence in each feature type, a π value was estimated. The π distributions were then compared statistically.

Presence of ambiguous nucleotides or the short size of the ITFs can affect the error margins associated with π estimates. To assess whether these factors influence our findings, we conducted additional analysis by changing the minimum number of sites analyzed (MinLength) and proportion of the aligned length with non-ambiguous bases (coverage). We sampled a range of MinLength (0,50,100,150,200) at no coverage threshold and a range of coverage (0,0.25,0.50,0.75, and 1) at no MinLength threshold. Our analyses suggested that the trend observed in Figure 3.9B is not affected by presence of ambiguous nucleotides or the short length of the ITF (data not shown).

ACKNOWLEDGEMENTS

We are grateful to Kevin Carr, Shari Tjugum-Holland, and Jeff Landgraff for help with sequencing and data pre-processing, Melinda Frame for help with Confocal Laser Scanning Microscopy, John Johnston for computational support and Andy Lin for his contributions towards preliminary studies.

CHAPTER FOUR

Evolution of genic and intergenic expression patterns in Poaceae species[1]

[1] A portion of the work described in this chapter was published in the following manuscript:

Rebecca Davidson, Malali Gowda, **Gaurav Moghe**, Haining Lin, Brieanne Vaillancourt, Shin-Han Shiu et al (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant Journal,* 71(3):492-502.

ABSTRACT

The Poaceae family, also known as the grasses, includes agronomically important cereal crops such as rice, maize, sorghum, and wheat. Previous comparative studies have shown that much of the gene content is shared among the grasses; however, functional conservation of orthologous genes as well as the extent and characteristics of intergenic transcription in these species have yet to be explored. To gain an understanding of the genome-wide patterns of expression evolution, we employed a sequence-based approach to compare analogous transcriptomes in species representing three Poaceae species including *Brachypodium distachyon* (purple false brome), *Sorghum bicolor* (sorghum) and *Oryza sativa* (rice). For analyzing intergenic expression, we also analyzed *Zea mays* (maize). Our transcriptome analyses reveal that evolution of gene expression profiles and coding sequences in the grasses may be linked. Genes that are highly and broadly expressed tend to be conserved at the coding sequence level while genes with narrow expression patterns show accelerated rates of sequence evolution. Analyzing patterns of intergenic expression, we show that despite the availability of 120-250 Mb of intergenic space, only ~5 Mb is transcribed in all four species, ~30% of which occurs near genes. The transcription of intergenic regions is more significantly correlated with the transcription of the nearest gene than random expectation. In addition, we find that intergenic regions that are expressed tend to be more highly associated with open chromatin marks and less associated with repressive chromatin marks, compared to untranscribed intergenic regions. These results suggest that transcription of intergenic regions may occur due to regulatory influence of neighboring genes or due to presence of an open chromatin architecture. Additional

analyses would be needed to identify intergenic transcripts with signatures of functionality. Overall, our analyses help reveal patterns of genic and intergenic expression evolution in Poaceae.

INTRODUCTION

The Poaceae family of grasses comprises over 600 genera and more than 10,000 plant species that belong to three major subfamilies, the Pooideae (wheat, barley, oat), the Panicoideae (maize, sorghum, sugarcane, switchgrass, and millets), and the Ehrhartoideae (rice, cut grass, and veldt grass) (Bolot et al., 2009; Kellogg and Buell, 2009). The Poaceae is an attractive group for comparative genomics because it includes many agriculturally important cereal crops with diverse native distributions and at least 35-fold variation in genome size (e.g. *Brachypodium distachyon* (hereafter BD) = 270 Mb; *Triticum aestivum* (wheat) = 16,000 Mb]. Although multiple Poaceae genomes have been sequenced (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Schnable et al., 2009; International Brachypodium Initiative, 2010), comparative expression analyses have yet to be performed in annotated cereal genomes, thereby limiting knowledge of the evolution and regulation of the core Poaceae transcriptome as well as the proportions of genes that are lineage specific. In addition, although the Poaceae genomes vary to a great extent in size, the number of genes in each genome are comparable to each other, indicating a large variation in the sizes of the intergenic regions in these species. Whether the amount of intergenic space transcribed increase with genome size, whether the intergenic transcripts lie close to genes as observed in *Arabidopsis thaliana* and whether any of these transcripts constitute novel genes is poorly understood.

Next-generation transcriptome sequencing offers advantages over previous expression profiling methods and provides the opportunity to fully evaluate global gene expression patterns (Marioni et al., 2008; Agarwal et al., 2010; Bernstein et al., 2012). In this study, I analyzed transcriptome data from eight related developmental stages of flower and seed, as well as leaves of four Poaceae species including BD, rice (*Oryza sativa*, OS), sorghum (*Sorghum bicolor*, SB) and maize (*Zea mays*, ZM) to understand the evolution of genic and intergenic expression patterns in Poaceae. These species represent three subfamilies within the Poaceae in which BD and OS share the most recent common ancestor (~45 mya), SB and ZM share a common ancestor (~25 mya) and the last common ancestor of all four species dates to 45–60 Ma (Bowers et al., 2005; Bennetzen, 2007; Paterson et al., 2009; International Brachypodium Initiative, 2010). These evolutionary relationships were utilized to address the following questions regarding genic and intergenic expression. First, what are the patterns of evolution of orthologous genes across these four species? Second, is there a relationship between sequence divergence and expression divergence? Third, how pervasive is intergenic transcription in these species? And finally, what could be the probable mechanistic explanations for intergenic transcription?

RESULTS AND DISCUSSION

Transcriptome sequencing and expression clustering

The tissues evaluated in this study include eight structures across flower and seed development and one vegetative stage from BD (Pooideae), OS (Ehrhartoideae), and SB (Panicoideae). Despite differences in flowering times, floral tissues were harvested at visually similar developmental stages including floral pre-emergence from the flag leaf (inflorescence-1), post-emergence from the flag leaf (inflorescence-2), and anthesis (anther and pistil); leaf and developing seed tissues were harvested at prescribed days after sowing and pollination, respectively. ZM (Panicoideae) was not used for comparative analysis of orthologous gene expression since its developmental stages could not be matched with BS, OS and SB. Reliable expression levels in units of fragments per kilobase of exon model per million fragments mapped (FPKM) could be estimated for 86, 66, and 73% of annotated genes in BD, OS and SB, respectively. For additional details regarding sequencing and transcript assembly, please refer to the original publication (Davidson et al., 2012).

The normalized transcript abundances for these 27 samples (nine tissues x three species) were used for expression clustering using k-means clustering, which detected eight co-expression clusters, and an additional cluster with low levels of expression ($\log_2$FPKM ≤ 2) across all tissues. To assess the functional significance of the genes in these clusters, we determined if any Gene Ontology (GO) terms were enriched in co-expression clusters. Many of the enriched GO categories were consistent with the known physiological processes in tissues including Cluster 1 genes that were upregulated in anthers and were enriched for the terms ‘sexual reproduction’ and ‘cell wall modification’. Cluster 6 genes exhibited elevated leaf expression and showed enrichments in ‘photosynthesis’, ‘cell iron-sulfur cluster’, and ‘cell redox homeostasis’ similar to previous observations in maize (Li et al., 2010). Genes in cluster 8 showed constitutive expression across tissues and were enriched for GO terms associated with primary metabolism such as ‘glycolysis’, ‘ATP synthesis’, and ‘fatty acid biosynthesis’ as well as various nucleotide metabolic processes suggestive of roles in core metabolic functions(Davidson et al., 2012).

The last common ancestor of the three species under study was ~45-60 mya. Since then, orthologous genes in these species will have experienced sequence as well as expression divergence. We first asked if these modes of divergence were coupled over the 45-60 million years of evolution in Poaceae.

Conservation and diversification in Poaceae orthologs

To determine if the evolution of orthologous gene expression is coupled to the evolution of their coding sequences, we estimated the non-synonymous (Ka) and synonymous (Ks) substitution rates of protein-coding orthologous pairs between the three species. There was no significant difference in the Ka/Ks ratio distributions between orthologs in all three species pairs, suggesting a similar level of selection pressure on protein-coding genes in all three species (Figure 4.1A). However, significant differences were observed between OrthoMCL categories, with genes in the 2xN category having higher Ka/Ks ratios (Figure 4.1A) than genes in other orthologous categories. To understand which biological processes were over-represented in the 2xN category, we performed a GO enrichment analyses for the 932 2xN genes for which GOs could be confidently assigned. In general, 2xN genes with GO annotation tend to be involved in stress-related functions (‘response to biotic stimulus’, ‘defense response’, ‘apoptosis’), lipid transport, secretion (‘exocytosis’), and general oxidation–reduction reactions. Regarding the enrichment in stress-related functions, the observation is consistent with previous studies in plants that have shown that genes that are responsive to stress tend to experience a higher degree of lineage-specific duplications (Hanada et al., 2008). This is also observed in the 1xN category which is similar to 2xN in that there are gene losses in one or two species with lineage-specific duplicates.

Stress-related GO categories were also enriched in the 3xN category, although to a lesser extent. In the 3xN category, several core metabolic functions such as ‘translation’ (404 genes), ‘ATP biosynthesis’ (101 genes), ‘nucleosome assembly’ (103 genes), and ‘biosynthetic process’ (119 genes) were found to be overrepresented, along with processes such as ‘oxidation–reduction’ (693 genes), ‘response to wounding’ (27 genes), and ‘sexual reproduction’ (28 genes). The 3xN category contains genes that are present in all three species but with different degrees of lineage-specific gains. Some will have very limited expansion (e.g. 1:1:2) while the others may have dramatic differences (e.g. 1:1:100). Thus, this category includes genes with both essential functions as well as genes involved in processes that are known to evolve quickly, which may explain why these genes have higher substitution rates than genes in the 3x3 category, but lower rates than genes in 2xN (Figure 4.1A). Out of 12,497 3x3 genes, the largest enriched GO categories included essential functions such as ‘regulation of transcription’ (>1000 genes), ‘protein folding’ (253 genes), ‘intracellular protein transport’ (123 genes), and ‘glycolysis’ (91 genes)(Davidson et al., 2012). In the 2x2 category, only three GO terms were found to be enriched – ‘protein amino acid phosphorylation' (295 genes), 'regulation of transcription' (168 genes) and 'response to oxidative stress' (64 genes). These GO terms are suggestive of the roles of 2x2 genes in transcriptional and protein level regulatory functions.

The evolutionary trend visible for the orthologous groups at the coding sequence level is mirrored at the expression level. Using Pearson correlation coefficients (PCC) as a measure of expression correlation, we found that the 3x3 group has a higher proportion of gene pairs (47.8%) with correlated expression (PCC ≥ 0.6) compared with the 2x2 (44.1%), 3xN (37.8%), and 2xN (30.7%) groups (Figure 4.1B). These results suggest that the 3x3 single-copy genes, which are associated with core metabolic functions, have experienced stronger purifying selection in the Poaceae family in contrast to the 3xN and 2xN multicopy genes that are undergoing relatively rapid diversification not only at the coding sequence level but also at the level of gene expression. After examining PCC distributions among species pairs further, we found the highest proportions of correlated (PCC ≥ 0.6) gene pairs to be between OS and SB in all categories, indicating that these two species are the most transcriptionally similar among the three Poaceae species (Figure 4.1B). Interestingly, we observed higher proportions of correlated (PCC ≥ 0.6) gene pairs for BD–SB compared with BD–OS in all orthologous groups. These results are inconsistent with the phylogenetic relationships between these species (Kellogg, 2001; International Brachypodium Initiative, 2010) and therefore could reflect developmental differences in flower and seed morphologies including BD's unique floral branching (spike versus panicle), OS's lack of

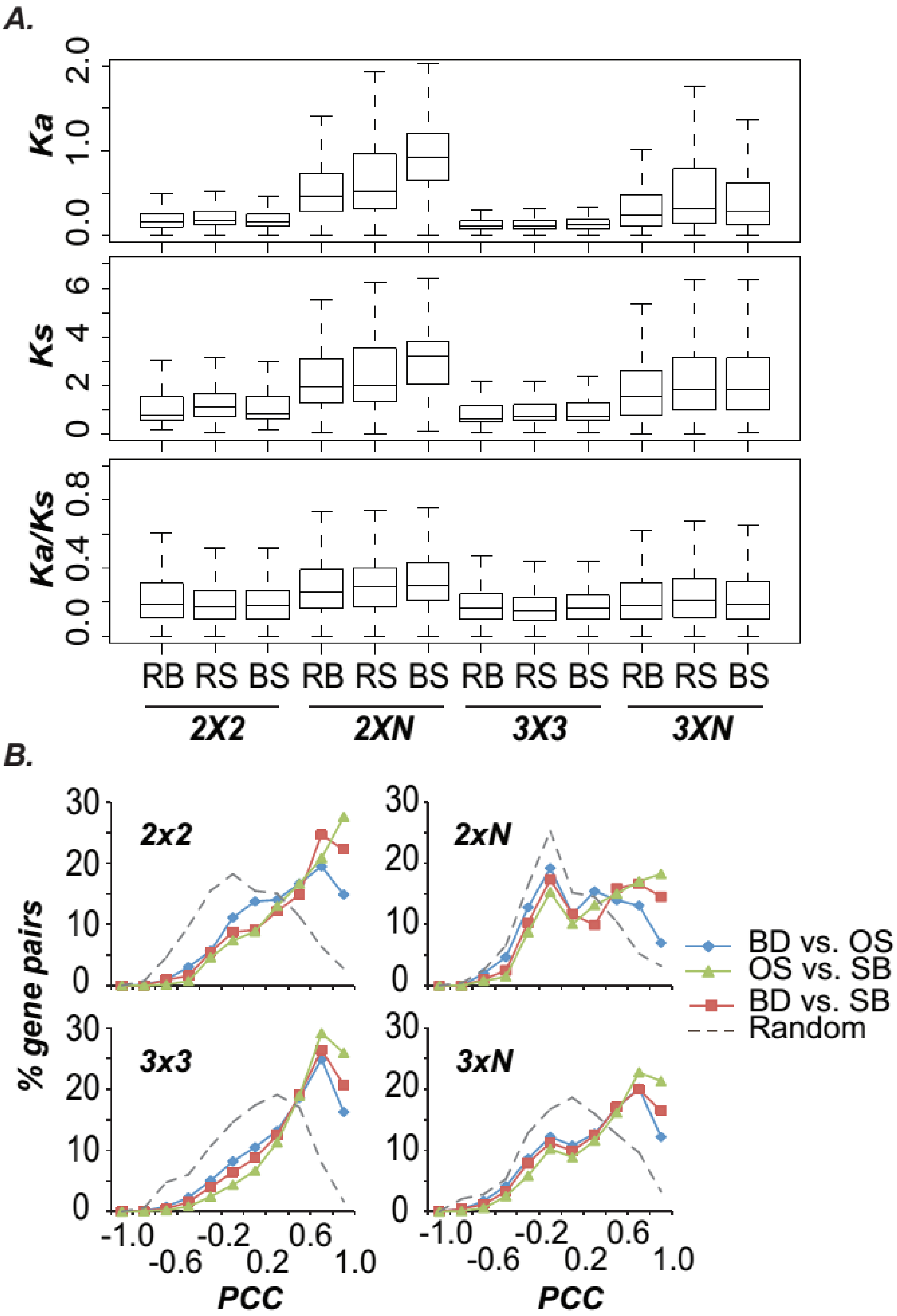


Figure 4.1: Conservation and diversification of OrthoMCL groups. (A): Coding sequence divergence. For each OrthoMCL group, non-synonymous (Ka) and synonymous (Ks) rates and their ratios (Ka/Ks) were calculated between cross-species

Figure 4.1 (cont'd)

gene pairs. Rate distributions are shown for Rice–Brachypodium (RB), Rice–Sorghum (RS) and Brachypodium–Sorghum (BS) comparisons. (B): Relationship between Pearson Correlation Coefficient (PCC), a measure of expression similarity, and the OrthoMCL groups.

awn production, and BD and SB's shared number of anthers (Kellogg, 2001). These trends may also correspond to environmental adaptations of tropical grasses (OS and SB) versus a temperate grass (BD).

To better understand the link between coding sequence and expression divergence of orthologs, we asked if gene pairs under stronger selective constraint have a higher correlation in expression. With conflicting results, previous studies have explored this question in yeast (Tirosh and Barkai, 2008), Drosophila (Nuzhdin et al., 2004) and mammals (Jordan et al., 2005; Khaitovich et al., 2005; Liao and Zhang, 2006). In our dataset, we found a statistically significant enrichment only in highly conserved gene pairs (Ka/Ks<0.1; z-test P-value < 0.05) in all orthologous groups; however, the effect size was not large and no enrichment was observed at higher Ka/Ks (Figure 4.2). This result, similar to that observed for *A. thaliana* and OS comparisons in a recent study (Movahedi et al., 2011), suggests, at best, only a weak relationship between coding sequence and expression divergence with the exception of highly conserved proteins.

Although expression divergence and coding sequence divergence are not highly associated with each other, genes which are highly expressed or broadly expressed may exhibit some association with sequence conservation. We tested this possibility using the expression data and orthlogy inference from this study.

Coding sequence evolution is related to the level and breadth of expression

Previous studies in yeast, vertebrates and *A. thaliana* have suggested that genes with higher expression levels and/or wider breadth of expression are more similar at the coding sequence level (Pál et al., 2001; Nuzhdin et al., 2004; Subramanian and Kumar, 2004; Wright et al., 2004; Movahedi et al., 2011). We thus asked whether, in the three Poaceae species, expression level and/or breadth of expression distinguish highly conserved genes from fast evolving genes. Expression levels of orthologous gene pairs were categorized into five groups (high, intermediate, low, not expressed, or divergent) based on the median FPKM value of each gene. The divergent category was the most highly variable between all four groups (Figure 4.3A)*.* We found that 60% (3x3) and 38% (3xN) of the gene pairs with high expression levels were under strong purifying selection (Ka/Ks <0.2) compared with all other groups (Figure 4.3B). In addition, pairs in which both genes are not expressed tend to evolve faster than pairs in which both genes are lowly expressed (KS test, P<1e-16). Among the multi-taxa orthologous groups, the 2xN group had the most gene pairs classified as ‘not expressed’ (Figure 4.3B). Given that genes in the 2xN group also tend to have high Ka/Ks values, it is possible that these genes may be undergoing pseudogenization. A recent study in *A. thaliana* found that lowly expressed genes tend to have a higher Ka and an excess of mutations in their promoters as compared with highly expressed genes, a finding that may also apply to the 2xN genes (Yang et al., 2011). Other studies have also found a link between expression level, Ka/Ks and pseudogenization (Frith et al., 2006; Zou et al., 2009).

We then explored the relationship between breadth of expression and evolutionary rate. We defined breadth of expression into four categories (broad, narrow, not expressed, or divergent) based on the number of tissues in which a given gene pair is co-expressed. Our results indicate that genes with a broader expression pattern (Figure 4.4A) are under higher evolutionary constraint than all other categories of

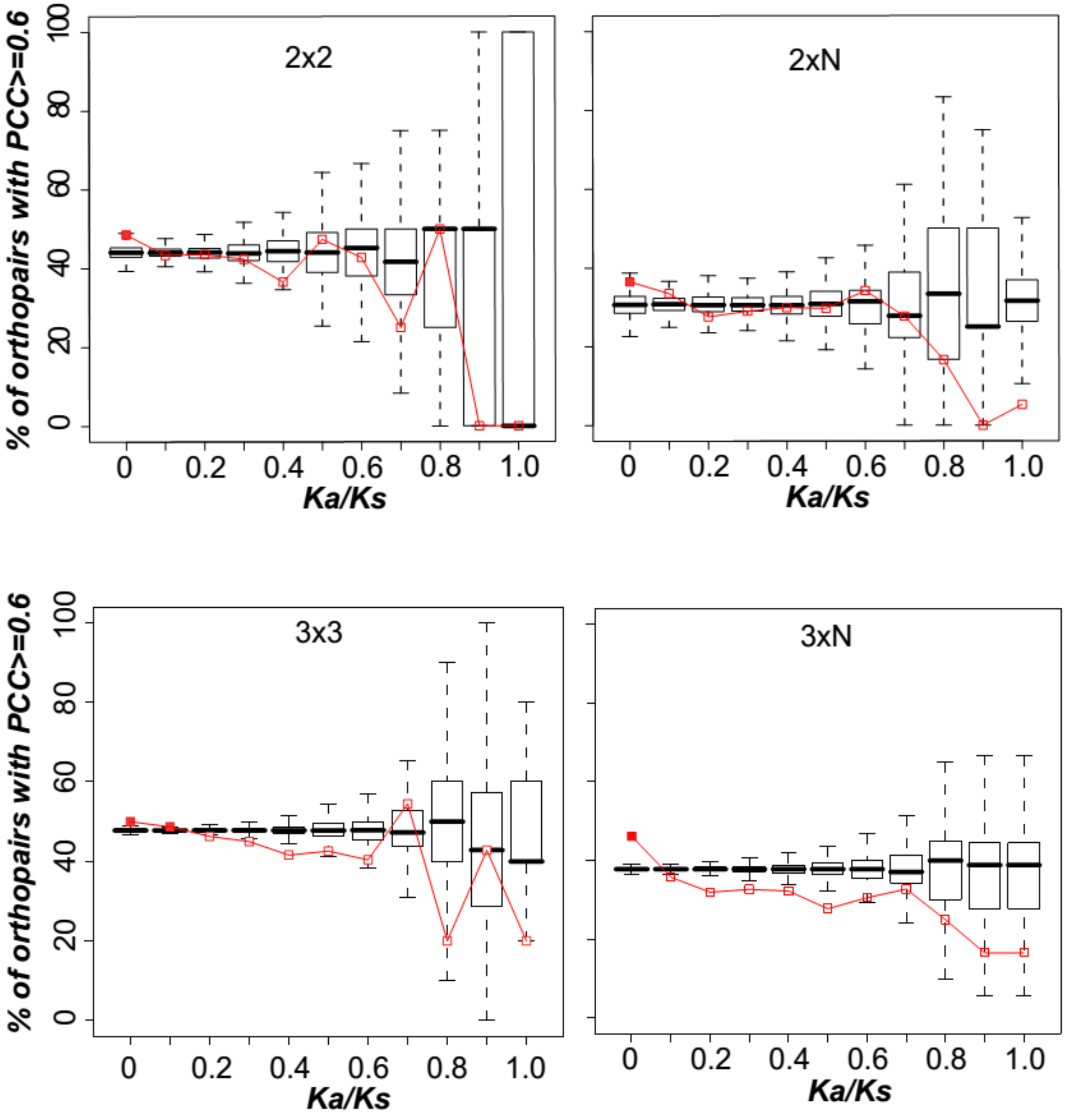


Figure 4.2: Gene pairs under strong evolutionary constraint (Ka/Ks<0.1) tend to be correlated in expression. The boxplots indicate random expectation while the red line indicates observed percentages.

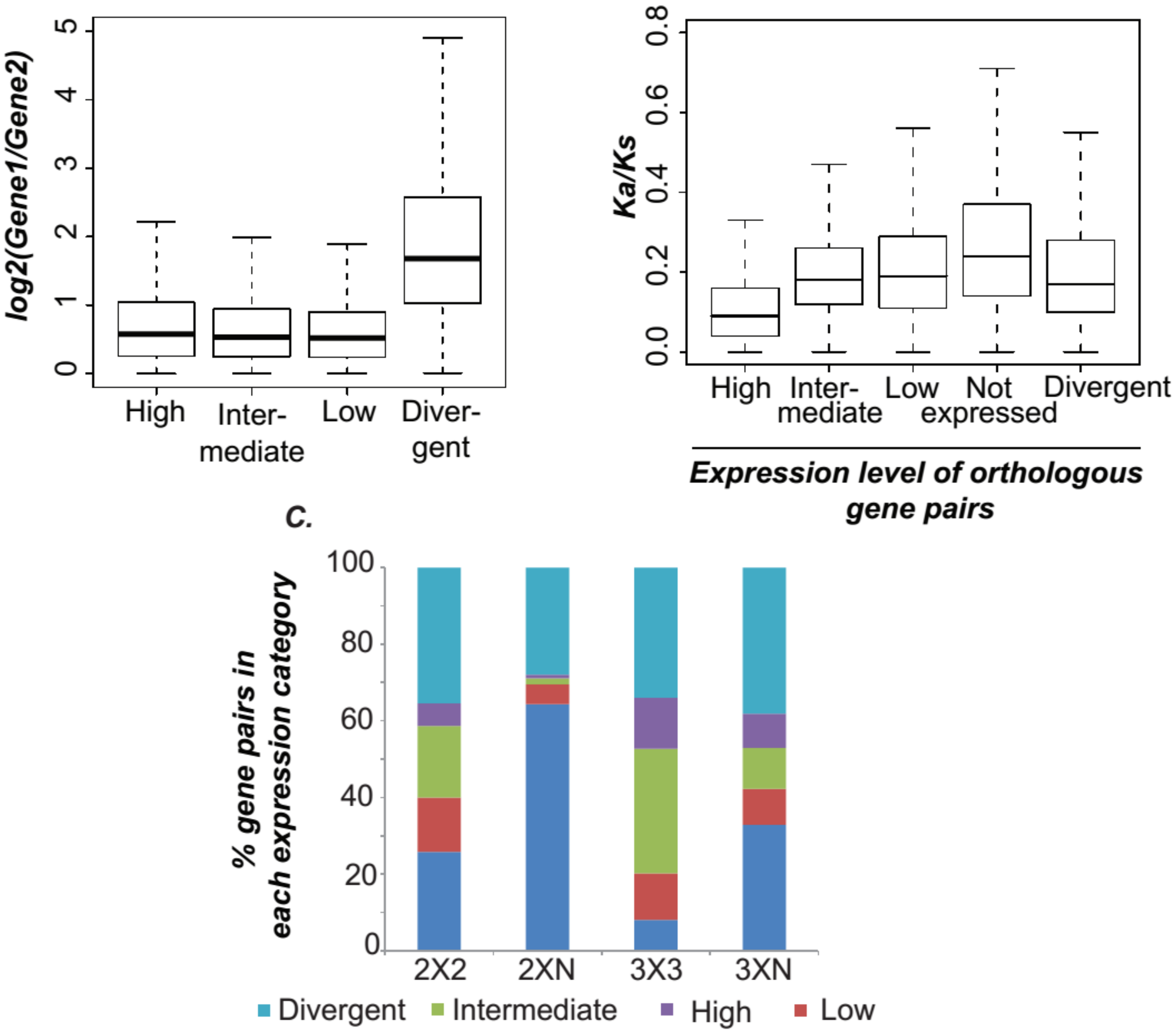


Figure 4.3: Coding sequence evolution vs gene expression level. (A): Divergently expressed gene pairs have significantly higher fold-change difference than gene pairs in other categories. (B): Rate of evolution as a function of the expression level categories. All categories had significantly different Ka/Ks distributions from all other categories (KS test P<1e-15). (C): The percentage of gene pairs in each of the expression level categories for each OrthoMCL category.

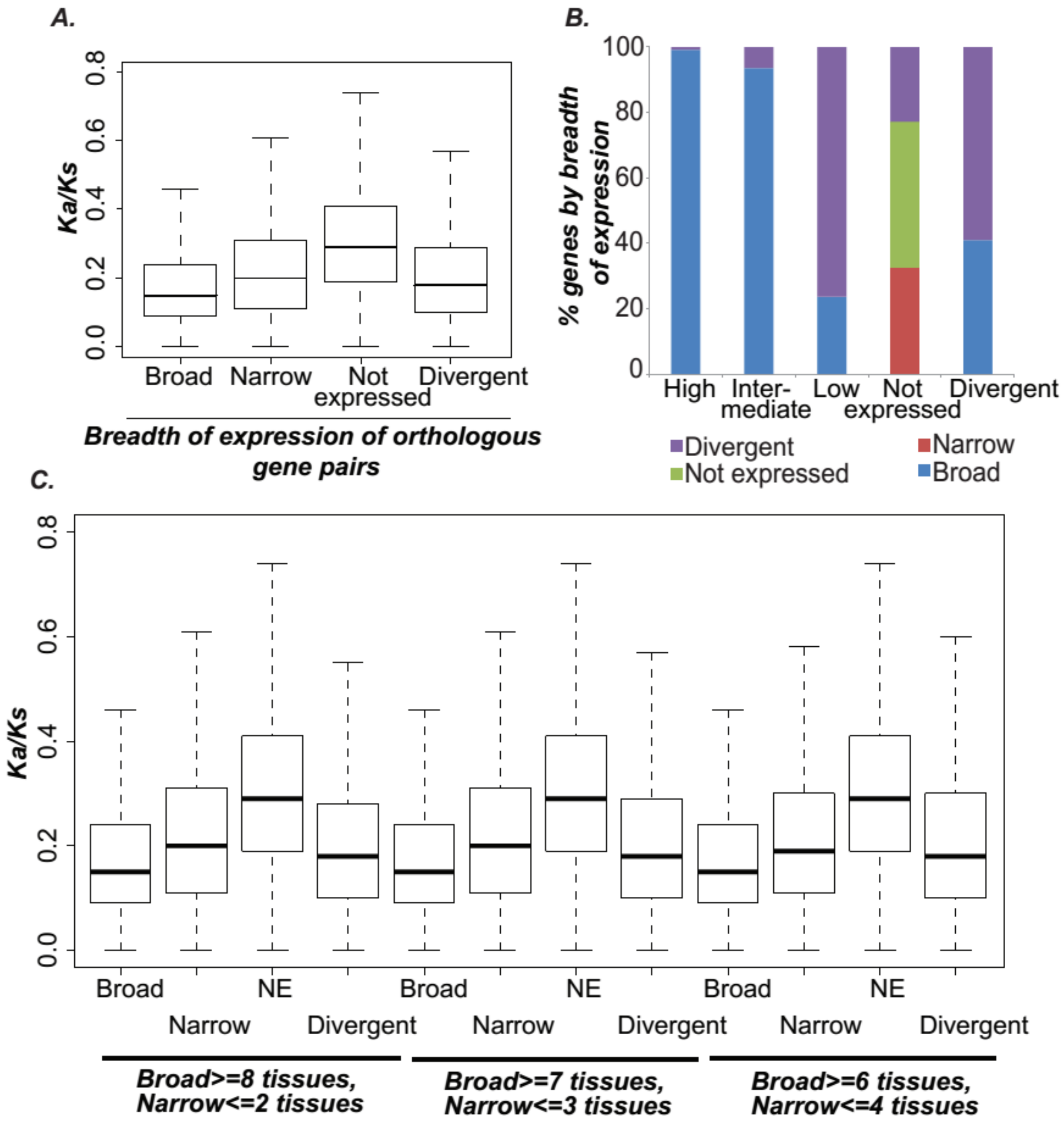


Figure 4.4: Coding sequence evolution vs gene expression breadth. (A): Rate of evolution as a function of the breadth of gene expression. All categories had significantly different distributions from all other categories (KS test P<1e-15)). (B): Classification of genes based on expression level and breadth of expression. (C): Changing thresholds for defining breadth do not influence the observed patterns.

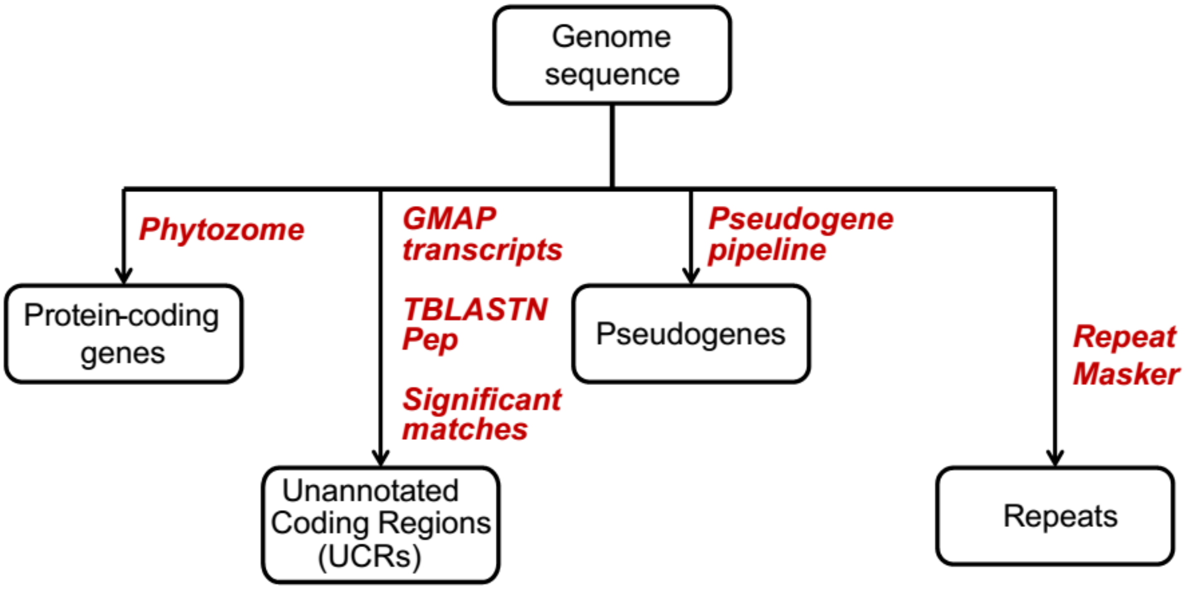
genes (KS test, P<1e-15). We also observed that gene pairs in which both genes are defined as 'not expressed' evolve at significantly higher rates than other defined categories (Figure 4.4A; KS test, P-value < 1e-16). The observed patterns were robust to changing the thresholds for defining the breadth of expression categories (Figure 4.4B). It is possible that breadth and level of expression are not independent; 99.2% of highly expressed genes have broad expression compared with 23.6% of the lowly expressed genes (Figure 4.4C).

In addition to understanding genic conservation, the global transcriptome profiles in BD, OS and SB can also be used to ask questions regarding pervasiveness and characteristics of intergenic transcripts. Alongwith these species, we also included RNA-seq data from ZM. Given its large genome size, analysis of ZM transcriptome may provide novel insights into how the characteristics of intergenic expression vary with genome size.

Reannotating the Poaceae genomes

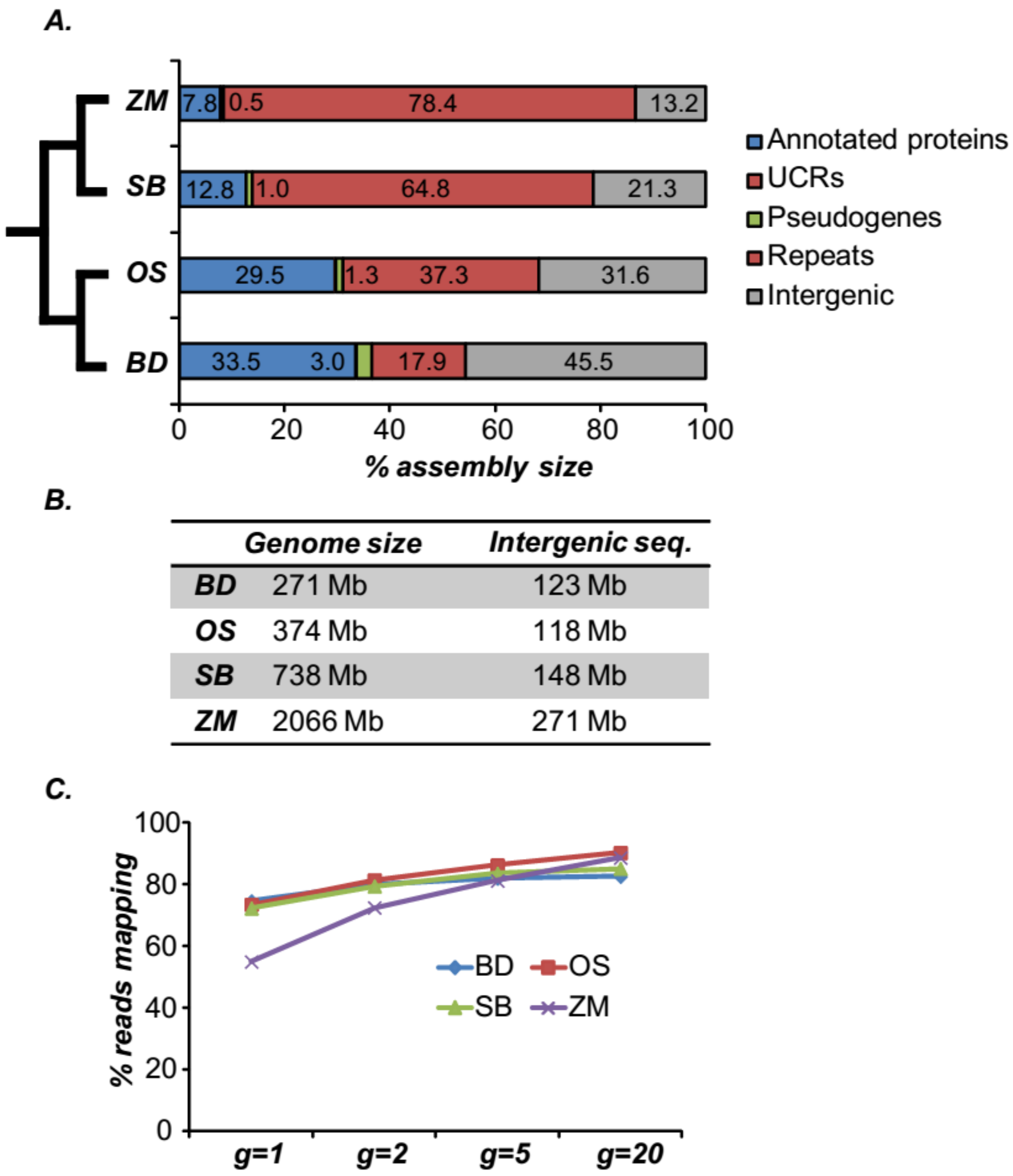
The genomes of BD, OS, SB and ZM have been predicted to have 26,552, 39,049, 27,608 and 39,656 genes respectively. Since these gene models were obtained using different strategies, it is likely that some models in one species may escape annotation in another. Such absent gene predictions can affect the estimates of intergenic transcription. Hence, in order to exclude the possibility of incomplete annotation, we predicted potentially missed Unannotated Coding Regions (UCRs) present in the intergenic regions using a combination of nucleotide-based and protein-based similarity searches (see Methods). We also predicted pseudogenes as well as repetitive sequences within the Poaceae genomes using previously published strategies (see Methods, Figure 4.5). For the purposes of this study, we defined intergenic regions as genomic regions not overlapping with protein-coding genes, UCRs, pseudogenes as well as repeats.

Our re-annotation revealed that most coding regions in each genome have already been successfully annotated (Figure 4.6A*)*. The UCRs accounted for only ~0.2% of the genome of each species. Repetitive regions occupy a significant proportion of the genomic space, ranging from 17.9% in BD to 78.4% in ZM. The percentage of intergenic regions based on the re-annotated assemblies in the Poaceae genome varied from 45.5% in the BD genome to 13.2% in the ZM genome (Figure 4.6A). The amount of intergenic space varied from ~120 Mb in BD and OS to ~150 Mb in SB and ~270 Mb in ZM (Figure 4.6B). Thus, there is more space available for intergenic transcription in ZM as compared to BD and more space may result in more intergenic bases transcribed. However, we also find that intergenic region size does not increase linearly with genome size. In addition, whether the percentage of the intergenic space transcribed would also increase, whether such transcribed regions are dispersed across the genome or clustered near genes and whether some of these intergenic TxFrags (ITFs) represent novel gene transcripts is not clear. We addressed these issues using RNA-seq expression data obtained from multiple developmental stages in all three species.



	Genome size	Ann.coding	UCRs	Pseudogenes	Repeats
BD	271 Mb	26,552	1194	18,338	68,447
OS	374 Mb	39,049	331	20,381	265,371
SB	738 Mb	27,608	495	18,548	418,497
ZM	2066 Mb	39,656	2316	39,118	963,412

Figure 4.5: Overview of the pipeline for reannotation of the Poaceae genomes. See Methods for more details. The table shows the number of features of each type identified in each species, alongwith the genome size.



	Genome size	Intergenic seq.
BD	271 Mb	123 Mb
OS	374 Mb	118 Mb
SB	738 Mb	148 Mb
ZM	2066 Mb	271 Mb

Figure 4.6: Genome composition after reannotation. (A): % of bases in the genome assemblies occupied by each feature type. UCRs occupy ~0.5% and are not visible. (B): Size of intergenic space in each species (C): Number of reads mapped by relaxing the

Figure 4.6 (cont'd)

"number of hits per read" threshold from 1 (unique-mapping) to 20 (mapping to max. 20 regions in the genome). Analysis was done only for the anther dataset.

Extent of transcription for annotated features and intergenic regions

In a previous study, we obtained polyA RNA-seq data from nine developmental stage matched tissues in BD, OS and SB (Davidson et al., 2012) as well as from multiple tissues in ZM (Davidson et al., 2011). Given the abundance of repeats in these genomes, we first asked what proportion of the transcriptomic reads arise from repetitive regions using the anther dataset as an example. We found that as we relaxed the "number of hits" threshold, more reads mapped, and this proportion increased faster for SB and ZM (Figure 4.6C). This suggests that a larger percentage of the transcriptome is derived from repetitive sequences in the SB and ZM genome. For further analysis, however, we employed a conservative approach and only used reads that mapped to a unique location in the genome. The choice of unique mapping reads may influence the results for repetitive regions to a greater extent than genes/pseudogenes/intergenic sequences.

The uniquely mapping RNA-seq reads were mapped to their cognate genomes and assembled into transcript fragments (see Methods). Our previous study indicated that TxFrags obtained using Cufflinks, although not representative of all the reads in the transcriptome, represent a set of high-confidence TxFrags (Moghe et al., 2013). For these TxFrags, we estimated the expression level in terms of Fragment Per Kilobase of exon model per Million mapped reads (FPKM) values. For TxFrags lying in the intergenic regions (called Intergenic TxFrags, ITFs), we also assumed that ITFs closer than 150bp to each other belong to the same transcriptional unit and joined such TxFrags across all developmental datasets, together, denoting the corresponding genomic regions as Intergenic Transcribed Fragment Regions (ITFRs). The 150bp threshold was chosen as the 95th percentile of the distance distribution between TxFrags which mapped to the same protein-coding exon. In other words, there is only a 5% chance that TxFrags mapping <150bp away from each other are likely part of different coding transcripts, in a given genome at a given sequencing coverage.

Based on information from ITFRs, we found that ~5% of the intergenic space (5-8 Mb) had evidence of transcription, regardless of the genome size or the size of the intergenic region of the species (Figure 4.7A,B), suggesting that species with larger genomes or larger intergenic spaces do not have more intergenic transcription. These ITFRs have been stringently defined using Cufflinks, which is reflected in the their FPKM distributions which are comparable to those of genes (Fig. 4.7C). However, despite their high expression levels, ~50% of the ITFRs were expressed in only one tissue (Figure 4.7D). In contrast, we found that on an average, 10.6%, 20% and 36% of the proteins, UCRs and pseudogenes, respectively, were expressed in only one tissue (Figure 4.7D).

The low proportion of intergenic space being transcribed suggests absence of pervasive intergenic polyA transcription in all Poaceae species. In addition, the fact that only ~5 Mb of the intergenic space is transcribed despite there being ~200 Mb of intergenic bases available for transcription may indicate some biases in the transcriptional process. We tested two hypotheses regarding the identity of ITFRs. First, we asked whether the transcription of ITFRs tends to be associated with transcription of annotated genes and features. Second, we investigated whether ITFRs tend to lie preferentially in areas of open chromatin.

A significant proportion of ITFRs lie near genes

In a previous study, we had found that >30% of the ITFRs lie close to annotated genes, however, the proportion was comparable to random expectation (Moghe et al., 2013). To understand whether a similar trend holds in genomes with larger intergenic regions, we calculated the distance of ITFRs from protein-coding genes and all features (protein-coding genes + UCRs + pseudogenes + repeats). We also obtained a random expectation of distance by sampling random sequences from the intergenic regions in each of the four species (see Methods) and determined whether the observed percentage of ITFRs occuring in the neighborhood of a gene was more than the percentage randomly expectated.

Our results suggest that ~30% of the ITFRs in all species are located <400bp away from protein-coding genes (Figure 4.8A). This proportion is much more than randomly expected e.g.: in ZM, there are 6X higher number of ITFRs in the 0-200 bp neighborhood than expected by chance (Figure 4.8B). On the other hand, although 35-52% of the ITFRs are located >2000bp away from protein-coding genes, we find that such ITFRs tend to be in close proximity to other features (Figure 4.8C). Given the wide spread of repeats in the non-genic space in the Poaceae genomes, the pattern observed in Figure 4.8C may be expected (Figure 4.8D). Nonetheless, these results suggest that there may likely be an association between transcription of the protein-coding genes and transcription of the intergenic region in their neighborhood.

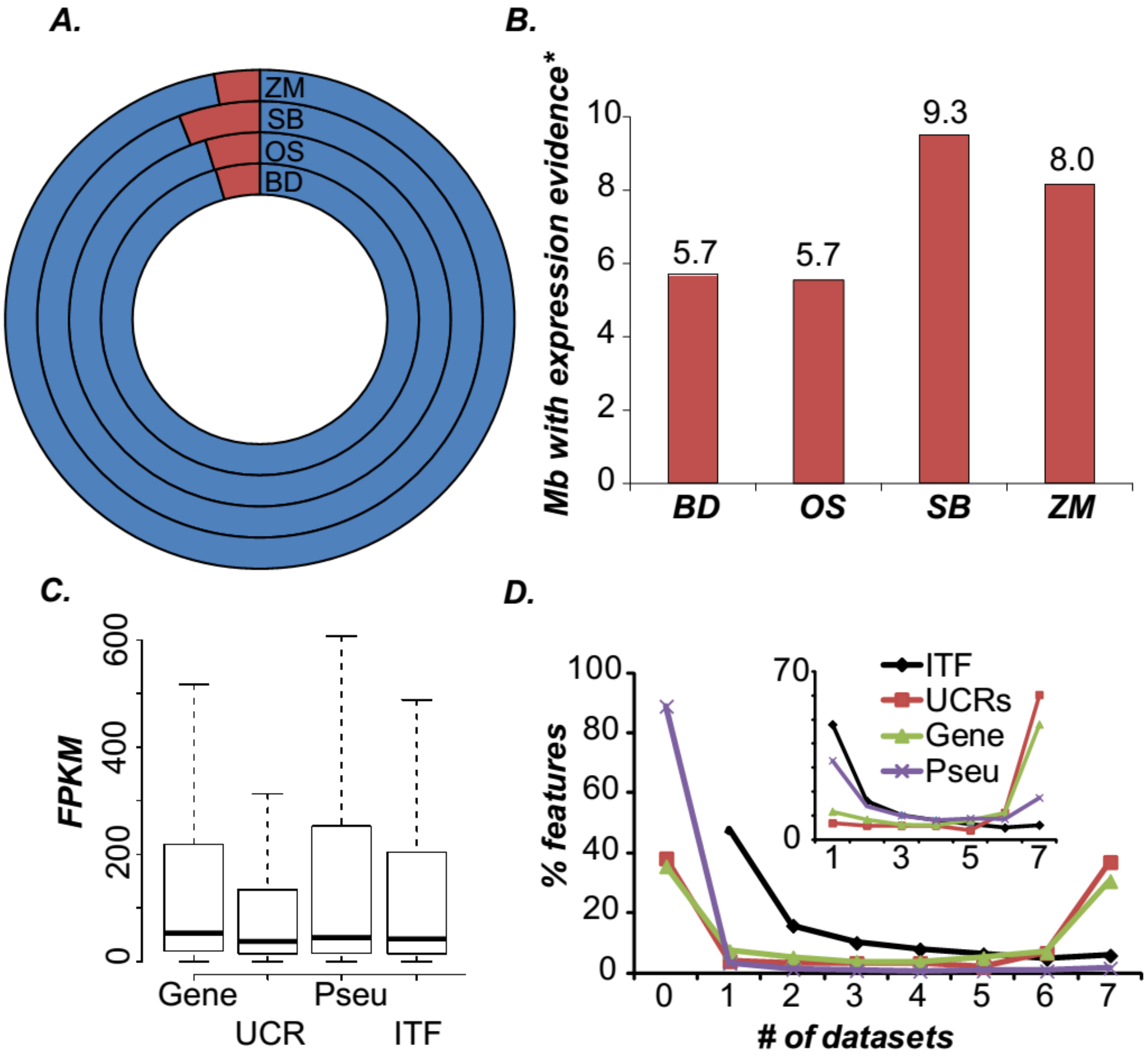


Figure 4.7: Characteristics of intergenic transcripts. (A and B): Proportion (A) and total size (B) of intergenic space with evidence of transcription. The observed proportions were scaled by factors of 1.1 (BD), 1.2 (OS, SB) and 1.6 (ZM) based on the fold increase observed in number of reads mapped to the genome after increasing the g threshold in bowtie (Figure 4.6C), to account for missed intergenic reads due to duplicate hits. (C): FPKM levels of different features. Pseu=pseudogenes (D): Breadth of expression of different features. Inset shows breadth of only those features expressed in ≥1 datasets.

There are three possibilities regarding ITFR-neighboring gene association. First, the ITFs could be unannotated regions of a gene transcript. Second, the ITFs could be produced as a result of transcription of the neighboring gene but are not physically linked. Third, ITFs could be completely independent transcripts, not associated with the neighboring gene at all. All three scenarios differentiate the mode of generation of the ITFs but cannot distinguish functional ITFs from noise.

Nevertheless, if the transcription of the ITFRs was occurring through the regulatory influence of the nearest protein-coding gene (first two possibilities), we can expect ITFRs closer to genes to have a higher correlation with the neighboring gene's expression than random expectation. To test this hypothesis, we estimated correlations between ITFR expression and that of their neighboring genes for ITFRs expressed in >50% of the tissues. Indeed, we found that ITFRs have a higher correlation to their nearest protein-coding gene than by chance (KS test p<1e-15, Figure 4.9A). In contrast, this influence does not extend to the gene on the other side of the ITFR, which were found to behave in a similar fashion as randomly picked pairs of genes (KS test p>0.1 for all pairs, Figure 4.9B). We also do not see any decay of correlation with distance from the gene to at least as far as 1kb (Figure 4.9C). These observations provide support to the hypothesis that the nature of association of ITFR expression to its neighboring gene expression is stronger than would be observed by pure chance or simply due to proximity.

Based on these observations, it is possible that the genomic loci where ITFRs are located tend to be commonly regulated by some upstream regulator. Such co-regulation may not only lead to expression of neighboring gene but also transcription of

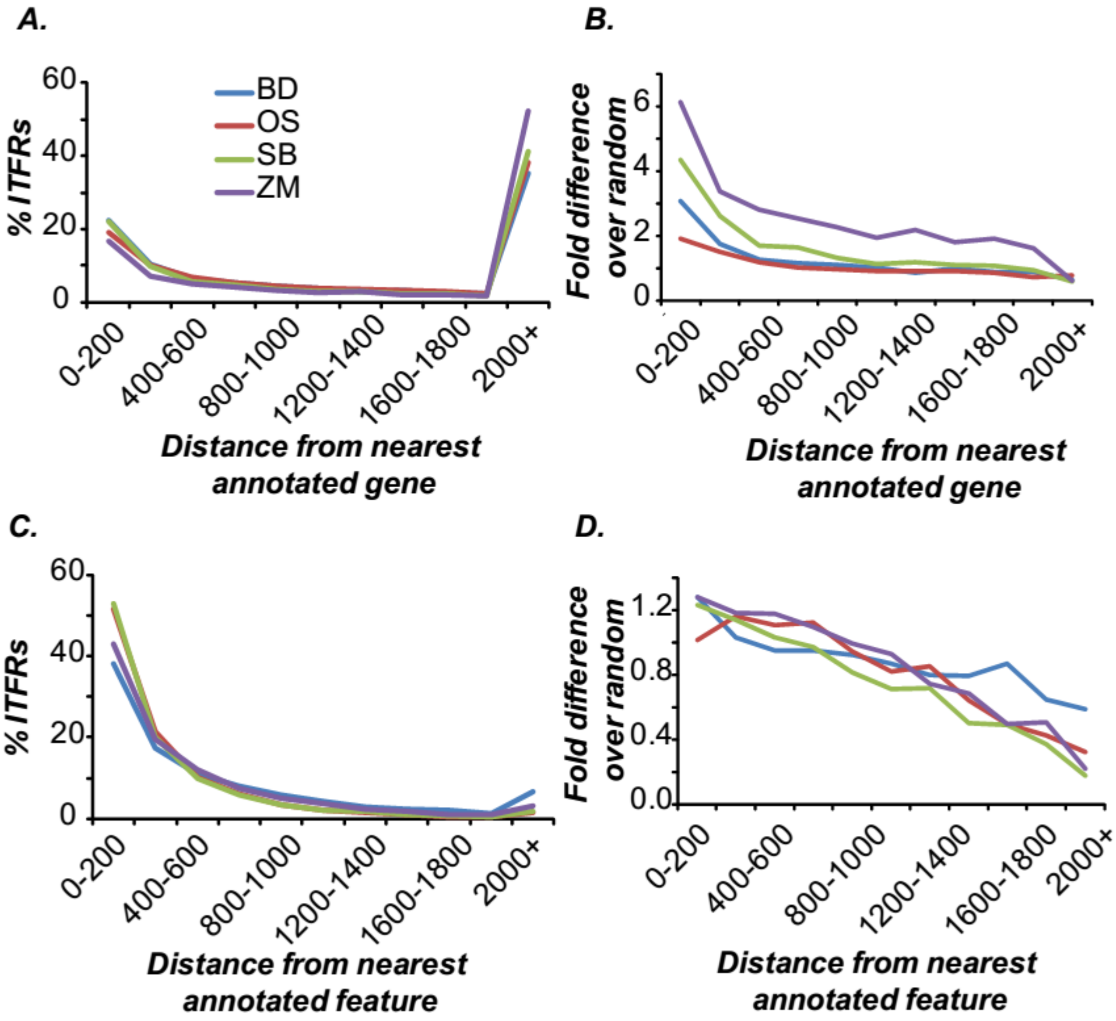


Figure 4.8: Distance of ITFRs from genes and other features (A): For each species, ITFR distance from the nearest gene is shown. (B): Compared to 10,000 randomly picked intergenic sequences, ITFRs tend to be much closer to genes (C): ITFR distance from the nearest annotated feature, including UCRs, pseudogenes and repeats. (D): Fold difference over random expectation

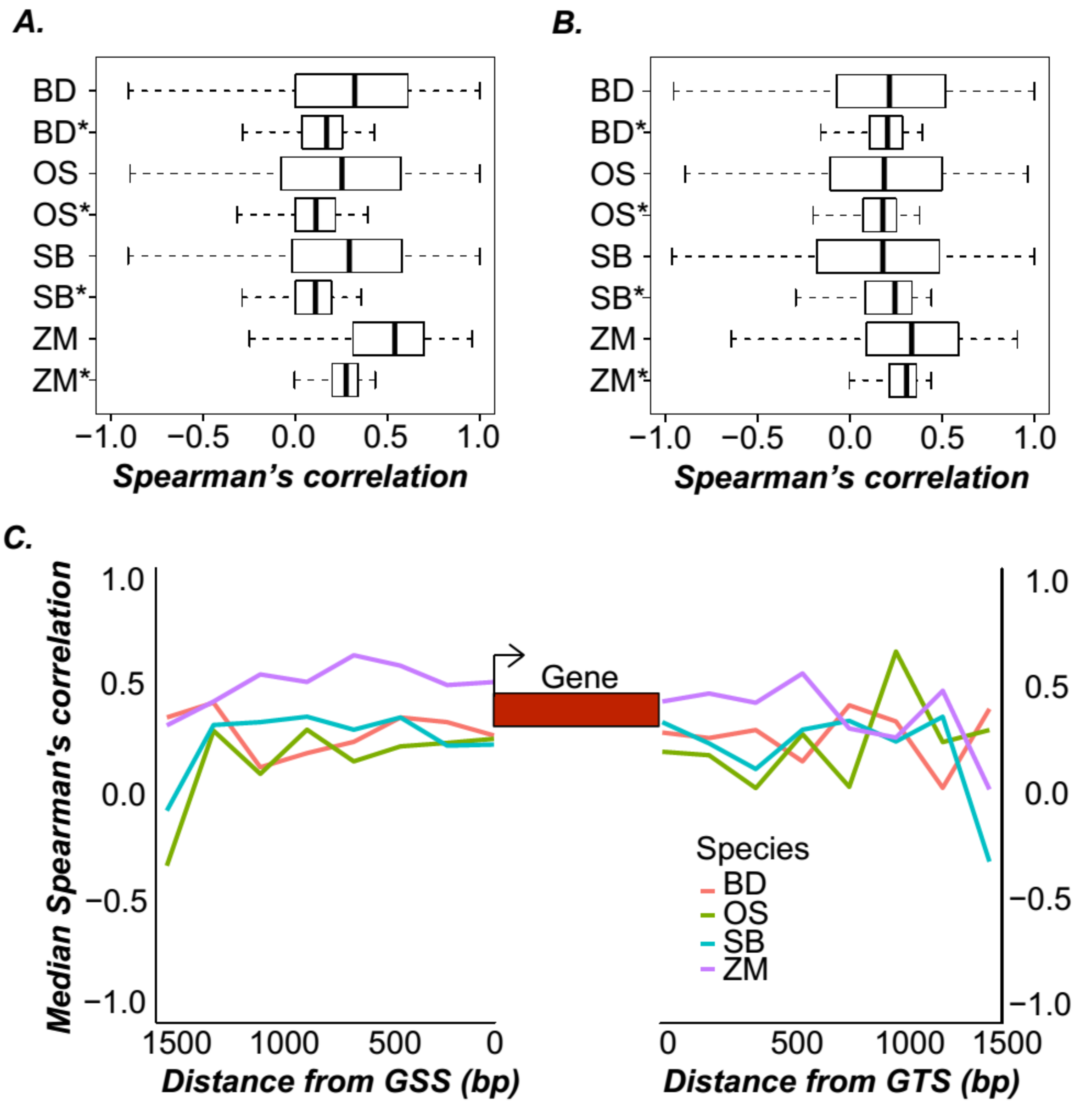


Figure 4.9: Expression correlations of ITFs. (A and B): Spearman's correlation coefficient (SCC) distributions for ITFR-Neighboring gene (A) and Gene-Neighboring gene (B) expression. * represents background expectation obtained by randomly pairing ITFRs with genes (A) or gene with gene (B). (C): Expression correlation does not decay with expression up to 1.5kb. SCC for ITFRs lying in each 200bp distance bin was calculated. For all figures, only ITFRs expressed in >50% of the tissues were used for SCC estimation. GSS=Gene Start Site, GTS=Gene Termination Site

some bases in the intergenic region between the two genes, creating ITFRs. It is possible that ITFR transcription is merely a side-effect of gene transcription, serving no real purpose. Alternatively, such transcription may have a functional role in gene regulation at the transcriptional or translational level.

ITFRs are associated with nucleosome free regions

While regulatory influence may explain some of the intergenic transcription, the fact that up to 50% of the ITFRs lie >2000bp away from genes indicates that there may be other explanations too. One hypothesis is that ITFs are produced from intergenic regions possessing an open chromatin architecture. A previous study found that ITFs were produced from nucleosome free regions in mouse and humans (van Bakel et al., 2010). To understand whether this hypothesis was true and to assess the degree of association of other chromatin marks with ITFRs, we obtained several indicators of chromatin structure – DNAse hypersensitive sites (DHS), H3K4me3, H3K9ac and H3K27me3 – from previously published datasets in OS seedlings (He et al., 2010; Zhang et al., 2012) and tested whether ITFs obtained from OS seedling RNA-seq data are significantly associated with any of these chromatin marks than randomly picked, untranscribed intergenic regions.

Our results suggest that Seedling ITFRs tend to have a higher proportion of DHS as well as H3K4me3 and H3K9ac marks than random intergenic sequences (KS test p<1e-15; Figure 4.10). While DNAse hypersensitivity correlates well with open chromatin structure, H3K4me3 and H3K9ac marks tend to be associated with transcribed sequences. We also find that Seedling ITFRs tend to have higher levels of H3K27me3 modification than random intergenic sequences(KS test p<1e-15, Figure

4.10). H3K27me3 is a mark associated with transcriptional repression, and a previous study in rice showed that the ratio between H3K4me3 and H3K27me3 is positively correlated with gene expression level (He et al., 2010). In our dataset, we see that ITFRs tend to possess higher ratios than random intergenic regions, suggesting that their transcription is influenced by chromatin architecture in the region

Interestingly, when we analyzed the scale of chromatin marks and DHS for the original set of ITFRs, we found similar patterns as with Seedling ITFRs (Figure 4.10) i.e. they are more significantly associated with DHS and other chromatin marks than random intergenic sequences. Epigenetic marks tend to be tissue specific, hence, it is surprising that we see the same trend between ITFRs derived from seedling as well as other developmental stages. The most likely explanation for this observation is that ITFRs are associated with chromatin regions that tend to be largely open over the tested developmental conditions. Additional analyses will, however, be required to determine whether the ITF-producing regions are constitutively open.

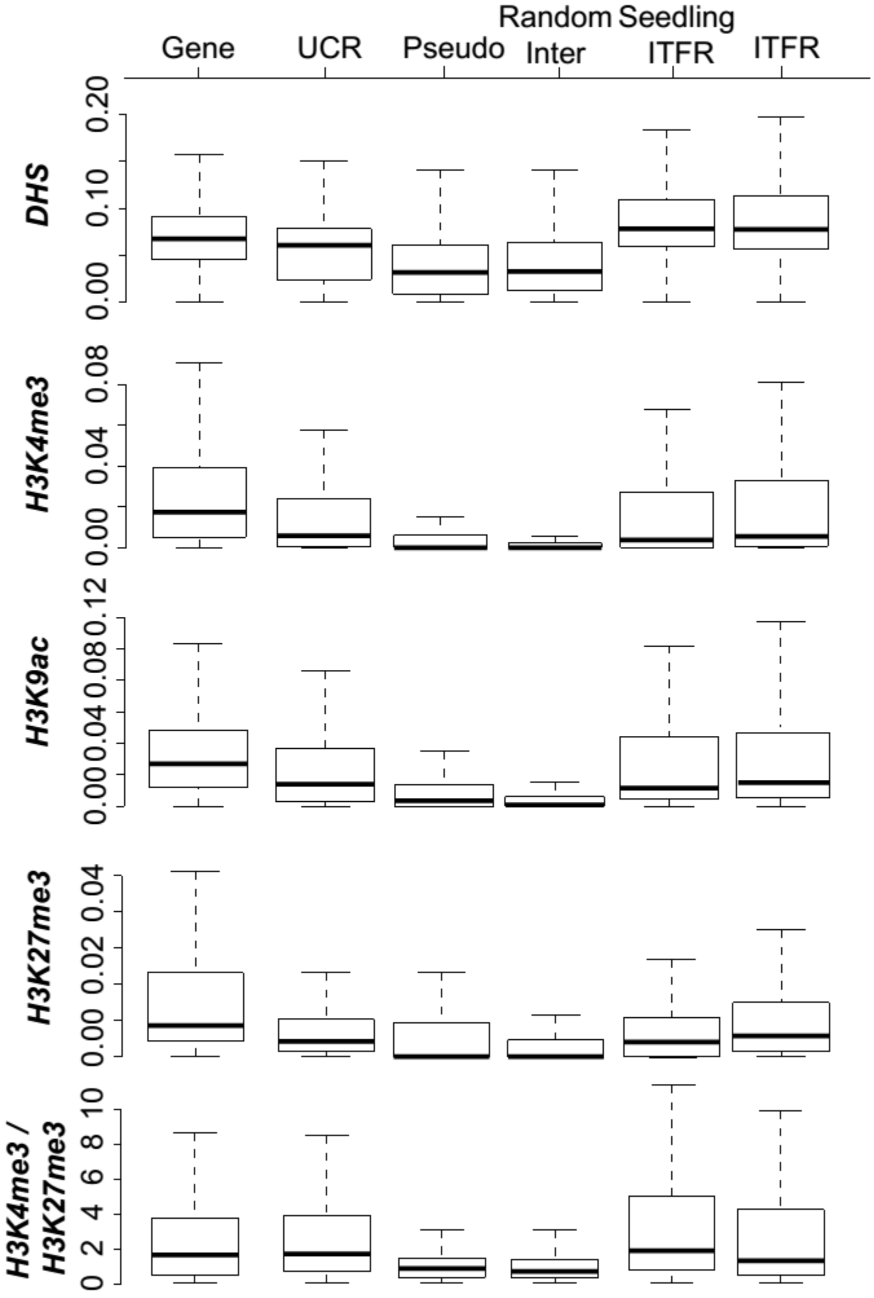


Figure 4.10: Chromatin mark association for each feature type. Y-axis represents number of reads mapping to a gene per bp of the gene. DHS=DNAse Hypersensitive Sites. Feature labeled only as "ITFR" refers to the original ITFRs discovered by combining all RNA-seq datasets in each species.

CONCLUSIONS

In this study, we asked questions regarding the evolution of genic and intergenic expression patterns in Poaceae species. Based on our analyses of expression profiles of orthologous genes between BD, OS and SB, we infer that although sequence and expression divergence only show a weak relationship with each other, there are other aspects of gene expression, namely level of expression and breadth, which show an association with sequence evolution. We cannot infer a cause-effect relationship between sequence and expression evolution; it is likely that other properties of a gene such as its network connectivity, biological function and extent of redundancy may influence both sequence as well as expression evolution. Indeed, from our analyses of GO enrichment among ortholog group types (Davidson et al., 2012) as well as those from previous studies (He et al., 2010; Zhang et al., 2012), sequence and expression conservation may very well be influenced by such properties.

In comparative studies between BD, OS, SB and ZM, we focused on understanding two questions related to intergenic polyA expression. First, whether the genome size of a species and the total size of its intergenic regions influences the extent of intergenic transcription. Second, what is the mechanistic basis of intergenic transcription? With regards to the first question, we found that ~5% of the intergenic region in all species was transcribed, regardless of the genome size or the total intergenic space. Around 5Mb of the intergenic space had any evidence of transcription in each species, despite there being 25X to 50X more space available for intergenic transcription. This result suggested that there may be some mechanistic biases in the way intergenic regions are transcribed.

We investigated two possibilities: 1) the ITFs were produced due to the regulatory influence of their neighboring genes and 2) that they were produced from regions with open chromatin configuration. Our results suggested that there was a highly significant influence of the transcription of the neighboring gene on intergenic transcription. It is possible that once in a while, the RNA-polymerase extends beyond the 3' end or there is antisense transcription from the 5' end of the gene. Previous studies in yeast and mammalian cells have pointed to presence of "ripple transcription" (Ebisuya et al., 2008), wherein the act of transcription of a gene leads to transcription of bases, sometimes, up to 10kb away. Such a phenomenon may be occurring in plant genomes too. Thus, ITFRs may either be unannotated parts of annotated transcripts or they are produced due to the transcriptional activity in their vicinity.

We also find that ITFRs tend to lie in open chromatin regions more frequently than expected by chance. Studies in Drosophila suggest that RNA polymerase is poised for transcription at the promoter of many genes, especially those involved in stress response and development, only to be fully activated by additional transcriptional factors at the right time (Muse et al., 2007). Such binding need not be completely inactive and spurious activation of RNA polymerase may occur once in a while producing ITFs. In addition, regions close to the 5' and 3' ends of a gene tend to be nucleosome free (He et al. 2010), possibly enabling spurious transcription by RNA polymerases. We also find that ITFRs are expressed in a very tissue-specific manner, despite lying in areas of open chromatin. This suggests that open chromatin is not sufficient for ITFR transcription and there may be additional factors, such as regulatory influence of neighboring gene or in some cases, the promoter region of the ITFR itself, that may cause their transcription.

The question that remains largely unanswered in our analyses is whether these ITFs are functional transcripts. Preliminary analyses suggested that there may be some conservation signal despite millions of years of divergence between the species. Specifically, we find that ~25% of the genes that have ITFs in their neighborhood in one species tend to have ITFs in the neighborhood of their orthologs in at least one other species, with ~200 genes having ITFs in their neighborhood all four species. These observations could indicate a propensity for certain genes to produce ITFs in their neighborhood, either due to their expression profiles or due to the chromatin structure around them. Alternatively, these ITFs could serve some regulatory function. We have yet to analyze sequence level conservation of these ITFs; it is a question that will be addressed in the future. Taken together, our analysis of intergenic and genic transcription provides a picture of a complex transcriptional landscape of plant genomes and identifies some of the players that influence expression evolution. The principles regarding intergenic transcription learnt from this study can not only be used as a baseline to identify novel, non-canonical genes but also to understand the nature of gene boundary transcription that may occur in all plant genomes.

MATERIALS AND METHODS

Biological samples, sequencing strategy and GO enrichment analyses

For more details regarding samples and sequencing, please refer to the published manuscript (Davidson et al., 2012).

We used GO to identify functions enriched within a given k-means cluster and OrthoMCL category. For performing enrichment analyses, we obtained the GO definitions for each of the three species from the MSU rice annotation website (ftp://ftp.plantbiology.msu.edu/pub/data/BFGR/release_3/), which were determined from a subset of InterProScan analyses (FPrintScan, HMMPfam, Gene3D, HMMPanther, ProfileScan, HMMSmart and superfamily) (Childs et al., 2012). Only the 'biological process' GOs were used for further analyses. For finding enrichments within different k-means clusters, only the genes with orthologs were considered. For enrichment tests within different OrthoMCL categories all the genes, including the single-taxa genes were used. If a gene did not have a GO definition in the three source files, it was excluded from further analyses. To determine whether a particular GO category was enriched within a given k-means cluster/OrthoMCL category, a Fisher exact test was performed with multiple testing correction as defined by Q-value (p≤0.05)(Storey, 2002).

Evolutionary rate calculations

We estimated the synonymous rate (Ks), non-synonymous rate (Ka),and evolutionary constraint (Ka/Ks) between pairs of orthologous genes, using the yn00 package in PAML(Yang, 2007). Only orthologs were used for these comparisons.

Definitions of level and breadth of expression

To define the level of expression, gene pairs in which both genes had a median FPKM≤1 were defined as 'not expressed'. The following categories were then defined: (i) high-expressed (median FPKM of both orthologs ≥ 24), (ii) intermediate-expressed (median FPKM of both orthologs between 4 and 24), and (iii) low-expressed (median FPKM of both orthologs ≤ 4 and at least one is expressed). If both genes of a pair were placed in different categories, they were deemed as divergent. The FPKM thresholds were defined based on the 25th percentile (FPKM 4.0) and 75th percentile (FPKM 24.0) of all expressed orthologous genes in the dataset. Although some gene pairs in the divergent category will have relatively small differences, we should emphasize that the median fold FPKM difference for the 'divergent' gene pairs is 3.2, which is significantly higher than all other categories (KS test, all P-values < 1e-16; ***Figure 4.5***). Thus, gene pairs in the 'divergent' category have significantly larger expression differences than gene pairs within each category. The differences between the Ka/Ks distributions of highly expressed versus all other sets of genes were statistically significant (KS test, P<1e-16).

To define the breadth of expression, all genes with their median FPKM≤1 were considered to be not expressed/trace. Broadly expressed gene pairs were defined as pairs with both genes expressed in seven or more tissues while narrowly expressed gene pairs had both genes expressed in three tissues or fewer. If both genes within a gene pair were placed in different categories, they were deemed as 'divergent'. The differences in the medians of Ka/Ks values of broadly expressed versus all other sets of genes were statistically significant (KS test, P<1e-16). Changing the definition thresholds to ≥6/≤4 or ≥8/≤2 for broadly expressed versus narrowly expressed did not affect the observed trends (Figure 4.7).

Reannotating the Poaceae genomes and identifying ITFs

We used RepeatMasker v open-4.0.0 to identify repeats in the Poaceae genomes *(--cutoff 225 –divergence 30*). The predictions of "simple repeats" and "low complexity sequences" larger than 500bp were included in the final set. A previously published pipeline was used for identification of pseudogenes (Zou et al., 2009).

To identify the UCRs, we first compiled amino acid and transcript sequences from all four species. The amino acid sequences were searched against the genome sequences of all four species using TBLASTN. Only the hits with E-value ≤ 1e-20, coverage ≥ 10% of the query, hit length ≥ 30aa and Identity ≥ 50% were considered significant. The transcripts were mapped to the genome using GMAP (Wu and Watanabe, 2005) and only the hits with Identity ≥ 70% and coverage ≥ 70% of the query were considered significant matches. We then determined the overlap between the predicted coding regions (combined predictions of transcript + aa matching) and annotated coding regions and found that 100% of the annotated coding regions in each species were identified by the predicted coding regions. We discarded all unannotated coding regions overlapping with pseudogene or repeat predictions. The final set was termed UCRs.

All features – protein-coding genes, pseudogenes, repeats and UCRs – were then compared against each other using their genomic locations to determine whether there were any overlaps between them. If there were overlapping features, they were retained based on the following order of preference: Protein-coding genes > Repeats > Pseudogenes > UCRs.

The SRA files downloaded from GenBank GEO database were converted to fastq format using SRAToolkit. The fastq sequences were filtered by quality (Q≥20) and length (L≥20) using FASTX toolkit. The processed reads were mapped to the genome using TopHat v1.4.1 and transcript fragments (TxFrags) were obtained using Cufflinks v2.1.1. Minimum and maximum intron sizes of 5000 and 50,000 bp were used for both TopHat and Cufflinks. Since Cufflinks inflates the FPKM values of TxFrags < 200bp, the *--frag-len-mean* option was changed from the default 200 to 150 to reduce the extent of FPKM inflation for smaller TxFrags.

TxFrags generated from seven tissues for each species were then compared against annotated coding regions, UCRs, pseudogenes and repeat predictions to identify feature-specific and intergenic TxFrags. The intergenic TxFrags from multiple tissues overlapping with each other were collapsed together into Intergenic Transcribed Fragment (ITF) predictions. The FPKM levels of all features in a given tissue was defined as the average FPKM of all TxFrags mapping to that feature in that tissue.

Orthology prediction and chromatin analysis

We first performed a protein BLAST (blastp) between protein sequences from *A. thaliana*, *Populus trichocarpa*, *Vitis vinifera* and the four Poaceae species. The BLAST results were filtered using an E-value < 1e-5 and Identity > 50% and orthologs and paralogs were predicted using OrthoMCL (Li et al., 2003). For identifying conserved ITFs, we used GMAP to map the nucleotide sequences of the ITFs on the genome of other Poaceae species using thresholds mentioned above. If there were multiple hits, only the longest hits were chosen.

To assess the tendency of ITFs to be associated with certain chromatin marks, we obtained data from two previous publications (He et al., 2010; Zhang et al., 2012). Sequence files were downloaded from NCBI SRA, trimmed based on length (L≥20), and mapped to the OS genome using MAQ (Li et al., 2008). Only uniquely mapping reads were used for further analyses. A custom Python script was used to determine number of reads per base of a feature per million mapped reads of a given feature.

ACKNOWLEDGEMENTS

I would like to thank Rebecca Davidson for leading the work on the published manuscript. I would also like to thank Alex Seddon and Ming-Jung Liu for assistance in analyzing the intergenic expression part of the chapter. Funding for this work was provided by the NSF (DEB-0919452 and MCB-1119778) to SHS.

CHAPTER FIVE

Conclusions and future perspectives

In my research, I have analyzed two mechanisms that contribute to gene content evolution in plants using comparative genomic and transcriptomic approaches. In this chapter, I will discuss some of the results from my work and describe avenues for future research.

EVOLUTION OF GENOMES POST WHOLE GENOME DUPLICATION

In Chapter 2, we performed sequencing, assembly and annotation of the wild radish genome and addressed questions regarding the evolution of duplicates derived from whole genome duplication. Briefly, we analyzed the patterns of pseudogenization, sequence evolution and expression divergence of retained duplicates and arrived at the following conclusions: 1) Pseudogenization of duplicate genes does not occur immediately post WGT. Instead, it occurs gradually over several million years. 2) At the sequence level, most duplicate genes diverge from each other at uniform rates and are under a greater level of constraint than those that evolve asymmetrically. 3) Tissue-specific divergence of expression among retained duplicates occurs mostly via one of the copies maintaining its ancestral state of expression and the other copies diverging via expression reduction. Thus, the original function of the gene in the tissue is preserved, and new functions are created due to sub/neo-functionalization. 4) Duplicates that were retained tend to possess specific structural, expression and functional characteristics different from genes whose duplicates were lost. Such difference in features enabled us to generate a predictive model of duplicate gene retention post WGD in plants. In addition to these findings, some questions regarding α' WGT and other WGD events remained unanswered in our study, which will be discussed here.

The wild radish assembly which we generated had ~68,000 contigs >100bp and an N50 of 10.1 kb making it difficult to construct scaffolds representative of chromosomes and to identify the order of genes in the genome. The major difficulty in creating scaffolds was due to: 1) low coverage of 454 mate pair sequencing and 2) only one insert size in Illumina paired end sequencing. Thus, this genome assembly could be made better by additional Illumina sequencing, specifically by generating paired end reads with larger insert sizes, e.g.: >1kb in length, so that repetitive regions in the genome could be assembled with higher confidence. The availability of radish scaffolds will significantly aid the ordering of sequence-based markers such as short sequence repeats.

Studies in *Brassica rapa*, which has a significantly better genome assembly, have suggested that the α' WGT event was, most probably, a two-step event, which has resulted in a very biased pattern of gene loss in the chromosomes of the post WGT species(Tang et al., 2012). In addition, there is evidence that subgenome bias exists in *B. rapa*, where one gene among the duplicated copies shows the dominant expression pattern, mostly in a parent-of-origin manner (Cheng et al., 2012). If the order of radish genes could be determined, fractionation bias between independent lineages could be studied in more detail. To determine such gene order, additional sequencing of the radish genomes with multiple large insert size libraries will be necessary. Additionally, availability of transcriptome sequencing data from multiple tissues/conditions may allow us to study the modes of divergence of duplicate genes i.e. sub-functionalization/neo-functionalization in the context of subgenome bias.

In our study, we also analyzed the pseudogene content in Brassicaceae genomes. However, we could only identify ~4000 pseudogenes that were derived from WGD in Brassica and radish. Assuming the neopolyploid ancestor of the two species had ~90,000 genes, given both species have ~40,000 genes today, where are the remaining 50,000 genes? One possibility is that most duplicates were lost through deletion of the gene segment. Another possibility is that the remaining duplicate genes were pseudogenized by insertion of transposable elements. Most of these pseudogenes would have been discarded in the pseudogene identification pipeline. The remaining duplicates, which were not deleted or do not contain transposon insertions, are what we ended up analyzing. Thus, in our study, the extent of pseudogenization occurring due to transposon insertion has remained unaddressed.

More broadly, one question that needs to be asked is, how does the repeat content in the genome behave after polyploidization? How much influence does it have in causing double stranded breaks, chromosomal translocations, genomic rearrangements and gene loss in polyploids? Does the repeat content increase after polyploidization? These questions can be studied in greater detail in synthetic neopolyploids, in recently created polyploids belonging to the *Tragopogon*, *Senecio* and *Spartina* genera or in genomes of species that recently underwent polyploidization such as cotton and maize.

Finally, during the process of diploidization of a polyploid genome, several genes are deleted. Our study showed that although most duplicates are lost, there is also some preferential retention and preferential loss, which is biased by the properties of the genes themselves. How does gene loss/retention affect biological networks? There are a few studies suggesting processes such as metabolism and circadian rhythm are affected. For example, two gene families encoding transcription factors playing central roles in circadian rhythm regulation such as *CIRCADIAN CLOCK ASSOCIATED1 (CCA1)* and *LATE ELONGATED HYPOCOTYL (LHY)* were found be preferentially retained post α' WGT (Lou et al., 2012). Documented phenotypic and fitness changes have occurred in recent polyploids such as *Spartina anglica*, *Chamerios angustifolium* and some polyploid accessions of *A. thaliana* (Ainouche et al., 2009; Baldwin and Husband, 2011; Chao et al., 2013). However, we have very little understanding of how network remodeling plays a role in such phenotypic transitions and local adaptation. An integration of molecular technologies and field biology is needed to address such questions.

FINDING NOVEL GENES AMONG INTERGENIC TRANSCRIPTS

In Chapters 3 and 4, I analyzed the transcriptomes of *A. thaliana*, *B. distachyon*, *O. sativa*, *S. bicolor* and *Z. mays*, whose genomes range in size from 140 Mb (*A. thaliana*) to 2000 Mb (*Z. mays*). The transcriptome datasets analyzed were from multiple tissues and/or conditions and thus represented a fairly large proportion of the transcriptomic space in each species. Contrary to results from some mammalian studies (Birney et al., 2007; Clark et al., 2011), we found no evidence of pervasive transcription in the plant genomes. The conflicting result may partly be explained by the fact that despite being a broad representation, our RNA-seq datasets might still incomplete, and that sampling of additional RNA-seq datasets or non-polyA and directional sequencing may yield evidence for more transcription. In addition, we were very aggressive in filtering out reads that may constitute genomic contamination and hence, we may have missed certain true yet lowly expressed intergenic transcripts. If more RNA-seq datasets were sampled and if one were very relaxed in defining intergenic transcripts, one would indeed be able to identify additional intergenic transcripts. But the question is, are these transcripts useful to the cell in any way?

To address this question, we characterized the nature of intergenic transcripts in our five study species. Our results across all species indicate that intergenic transcripts are expressed at very low levels and in a very tissue-specific manner. We also found that a significant proportion of them lie very close to genes, and as the genome size increases they tend to lie closer to genes more often than expected randomly. In other words, although there is more space available for intergenic transcription, it still tends to be enriched near genes. Interestingly, we also find that for genes that tend to have intergenic transcription in their neighborhood, even their orthologs in other species tend to have neighborhood transcription, much more than what would be expected by chance. Our results suggest two explanations for why intergenic transcription persists, especially near genes: 1) intergenic transcripts may be produced due to the regulatory influence of the transcription of their neighboring gene or feature, and some genes may produce a greater “ripple” of transcription (Ebisuya et al., 2008) than others and 2) intergenic open chromatin regions may be more prone to transcription than other regions. However, we only found an association with neighboring gene transcription and open chromatin regions and not a cause-effect relationship. So we cannot conclusively say whether intergenic transcripts constitute noise or not based on these data.

Our inference of noise is primarily based on the fact that <5% of the intergenic transcripts are conserved within or between species, suggesting that such transcripts are transient, are lost quickly through time and may be inconsequential to the fitness of the organism. Based on these observations, it seems that most of the intergenic transcription is not purposeful and may be occurring in a spurious manner as an indirect effect of other molecular phenomena occurring on the DNA molecule.

What does this mean for annotation of novel genes in currently defined intergenic regions? RNA-seq studies continue to find thousands of intergenic transcripts. While my studies indicate most of these transcripts are a result of noise, functionally important transcripts will also be present. Using machine learning approaches such as support vector machines, it may be possible to predict which intergenic transcripts are likely novel genes based on their properties such as expression level, breadth of expression, length, distance from genes, evolutionary conservation as well as properties of the genomic neighborhood such as nucleosome occupancy, chromatin marks and expression states of neighboring genes. For example, my preliminary studies suggested that there is a direct relationship between length and nucleotide diversity in *A. thaliana* intergenic transcripts. Such measures could aid in finding the proverbial “needle in the haystack” of functional among nonfunctional intergenic transcripts.

GENE CONTENT EVOLUTION IN PLANTS: A BROADER PERSPECTIVE

In my research, I focused only on genes created via whole genome duplication and on intergenic transcripts. However, as outlined in Chapter 1, there are other modes of gene origination too, namely retroposition, exon shuffling and *trans* splicing. The relative contributions of these processes to gene content evolution are unknown. Advances in sequencing technologies, however, can help shed light on these processes. For example, the ENCODE study found that 74% of the intronic bases in the human genome could be assigned a reproducible primary transcript, suggesting presence of alternative splice forms, antisense transcripts or intronic genes (Djebali et al., 2012). In addition, we only analyzed the polyA fraction of the transcriptome and hence, additional non-coding genes not transcribed by RNA polymerase II would not occur in this fraction. A meta analysis of several polyA + non-polyA transcriptome datasets in *A. thaliana* can help in addressing the contribution of other modes of gene origination.

In addition to identifying such features and understanding their characteristics, the availability of sequence data from populations also allows us to explore their conservation within a species. Several functional RNAs may show functionality for only short time intervals and such features can be identified using polymorphism data and population genetic tests. Genome data from populations as well as data from multiple plant species are now being made available. At the time of this writing, there were at least forty plant species with draft, partial or complete genomes available in a centralized plant genome data repository (Goodstein et al., 2012) and more on the way. Comparative analyses of gene content across these species can help us understand how different novel characteristics came to be present in different lineages in the plant world.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L.W., Sasidharan, R., Reinke, V., Waterston, R.H., and Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics 11: 383.

Ainouche, M.L., Fortune, P.M., Salmon, A., Parisod, C., Grandbastien, M.-A., Fukunaga, K., Ricou, M., and Misset, M.-T. (2009). Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). Biol Invasions 11: 1159–1173.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Arias, T. and Pires, J.C. (2012). A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. Taxon 61: 980–988.

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M., and Raymond, C.K. (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat. Methods 6: 647–649.

Aubourg, S., Martin-Magniette, M.-L., Brunaud, V., Taconnat, L., Bitton, F., Balzergue, S., Jullien, P.E., Ingouff, M., Thareau, V., Schiex, T., Lecharny, A., and Renou, J.-P. (2007). Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. BMC Genomics 8: 401.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 320: 938–941.

Bailey-Serres, J., Sorenson, R., and Juntawong, P. (2009). Getting the message across: cytoplasmic ribonucleoprotein complexes. Trends Plant Sci 14: 443–453.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most “dark matter” transcripts are associated with known genes. PLoS Biol 8: e1000371.

Van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2011). Response to “The Reality of Pervasive Transcription.” PLoS Biol 9: e1001102.

Baldwin, S.J. and Husband, B.C. (2011). Genome duplication and the evolution of conspecific pollen precedence. Proc. R. Soc. B 278: 2011–2017.

Basrai, M.A., Hieter, P., and Boeke, J.D. (1997). Small Open Reading Frames: beautiful needles in the haystack. Genome Research 7: 768 –771.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. 107: 18724–18728.

Bennetzen, J.L. (2007). Patterns in grass genome evolution. Curr. Opin. Plant Biol. 10: 176–181.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004a). Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004b). Global identification of human transcribed sequences with genome tiling arrays. Science (New York, N.Y.) 306: 2242–6.

Birney, E. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

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

Bolot, S., Abrouk, M., Masood-Quraishi, U., Stein, N., Messing, J., Feuillet, C., and Salse, J. (2009). The “inner circle” of the cereal genomes. Curr. Opin. Plant Biol. 12: 119–125.

Bowers, J.E. et al. (2005). Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc. Natl. Acad. Sci. U.S.A. 102: 13206–13211.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–438.

Branco-Price, C., Kaiser, K.A., Jang, C.J.H., Larive, C.K., and Bailey-Serres, J. (2008). Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in *Arabidopsis thaliana*. Plant J 56: 743–755.

Bridges, C.B., Skoog, E.N., and Li, J.-C. (1936). Genetical and cytological studies of a deficiency (Notopleural) in the second chromosome of *Drosophila melanogaster*. Genetics 21: 788–795.

Byrne, K.P. and Wolfe, K.H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175: 1341–1350.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18: 188–196.

Cao, J. et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963.

Carninci, P. et al. (2005a). The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.

Carninci, P. et al. (2005b). The transcriptional landscape of the mammalian genome. Science (New York, N.Y.) 309: 1559–63.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. U.S.A 105: 21034–21038.

Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B., and Salt, D.E. (2013). Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. Science 341: 658–659.

Chapman, B.A., Bowers, J.E., Feltus, F.A., and Paterson, A.H. (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A 103: 2730–2735.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One 7.

Cheng, J. et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308: 1149–1154.

Childs, K.L., Konganti, K., and Buell, C.R. (2012). The Biofuel Feedstock Genomics Resource: a web-based portal and database to enable functional genomics of plant biofuel feedstock species. Database (Oxford) 2012: bar061.

Chou, H.-H., Hayakawa, T., Diaz, S., Krings, M., Indriati, E., Leakey, M., Paabo, S., Satta, Y., Takahata, N., and Varki, A. (2002). Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc. Natl. Acad. Sci. U.S.A. 99: 11736–11741.

Clark, M.B. et al. (2011). The reality of pervasive transcription. PLoS Biol. 9: e1000625; discussion e1001102.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Conner, J.K. (2002). Genetic mechanisms of floral trait correlations in a natural population. Nature 420: 407–410.

Conner, J.K., Sahli, H.F., and Karoly, K. (2009). Tests of adaptation: functional studies of pollen removal and estimates of natural selection on anther position in wild radish. Ann. Bot. 103: 1547–1556.

Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27: 55–71.

Cronn, R.C., Small, R.L., and Wendel, J.F. (1999). Duplicated genes evolve independently after polyploid formation in cotton. Proc. Natl. Acad. Sci. U.S.A. 96: 14406–14411.

Curtis, M.D. and Grossniklaus, U. (2003). A gateway cloning vector set for high-throughput functional analysis of genes *in planta*. Plant Physiol 133: 462–469.

Darwin, C. (1859). On the origin of the species by means of natural selection: Or, The preservation of favoured races in the struggle for life (John Murray).

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. U.S.A 103: 5320–5325.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.-H., Jiang, N., and Robin Buell, C. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J. 71: 492–502.

Davidson, R.M., Hansey, C.N., Gowda, M., Childs, K.L., Lin, H., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M., Jiang, N., and Buell, C.R. (2011). Utility of RNA sequencing for analysis of maize reproductive transcriptomes. The Plant Genome 4: 191–203.

Ding, Y., Zhou, Q., and Wang, W. (2012). Origins of new genes and evolution of their novel functions. Annual Review of Ecology, Evolution, and Systematics 43: 345–363.

Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic 8: 407–423.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol 4: e1000176.

Djebali, S. et al. (2012). Landscape of transcription in human cells. Nature 489: 101–108.

Domazet-Loso, T. and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. Genome Res 13: 2213–2219.

Doyle, J.P. et al. (2008). Application of a translational profiling approach for the comparative analysis of CNS cell types. Cell 135: 749–762.

Dujon, B. (1996). The yeast genome project: what did we learn? Trends Genet. 12: 263–270.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat. Cell Biol 10: 1106–1113.

Eddy, S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4.

Elhaik, E. and Tatarinova, T. (2012). GC3 biology in eukaryotes and prokaryotes. arXiv:1203.3929.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. 12: 861–874.

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

Fares, M.A., Byrne, K.P., and Wolfe, K.H. (2006). Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. Mol. Biol. Evol. 23: 245–253.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Filichkin, S. a, Priest, H.D., Givan, S. a, Shen, R., Bryant, D.W., Fox, S.E., Wong, W.-K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome research 20: 45–58.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Friedman, W.E. (2009). The meaning of Darwin's "abominable mystery." Am. J. Bot. 96: 5–21.

Frith, M.C. et al. (2006). Pseudo-messenger RNA: phantoms of the transcriptome. PLoS Genet 2: e23.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res. 17: 669–681.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11: 725–736.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–1186.

Gregory, B.D., O'Malley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H., and Ecker, J.R. (2008). A link between RNA metabolism and silencing affecting Arabidopsis development. Dev. Cell 14: 854–866.

Guttman, M. et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.

Hahn, M.W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. J. Hered. 100: 605–617.

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., and Shiu, S.-H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics 26: 399–400.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.-H., and Shiu, S.-H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. Genome Res 17: 632–640.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.-H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993–1003.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A., and Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 16: 1252–1261.

He, G. et al. (2010). Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell 22: 17–33.

He, X. and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157–1164.

Heo, J.B. and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79.

Hiller, M., Findeiss, S., Lein, S., Marz, M., Nickel, C., Rose, D., Schulz, C., Backofen, R., Prohaska, S.J., Reuter, G., and Stadler, P.F. (2009). Conserved introns reveal novel transcripts in *Drosophila melanogaster*. Genome Res 19: 1289–1300.

Hu, T.T. et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43: 476–481.

Hurst, L.D. (2009). Evolutionary genomics and the reach of selection. J. Biol. 8: 12.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223.

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463: 763–768.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature 436: 793–800.

Jacob, F. (1977). Evolution and tinkering. Science 196: 1161–1166.

Jiang, W.-K., Liu, Y.-L., Xia, E.-H., and Gao, L.-Z. (2013). Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. Plant Physiol. 161: 1844–1861.

Jiao, Y. et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97–100.

Jiao, Y. and Meyerowitz, E.M. (2010). Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. Mol. Syst. Biol 6: 419.

Joachims, T. (1999). Making Large-Scale Support Vector Machine Learning Practical (MIT Press, Cambridge, MA).

Johnston, J.S., Pepper, A.E., Hall, A.E., Chen, Z.J., Hodnett, G., Drabek, J., Lopez, R., and Price, H.J. (2005). Evolution of genome size in Brassicaceae. Ann. Bot. 95: 229–235.

Jordan, I.K., Mariño-Ramírez, L., and Koonin, E.V. (2005). Evolutionary significance of gene expression divergence. Gene 345: 119–126.

Kapranov, P. et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.

Kawaguchi, R. and Bailey-Serres, J. (2002). Regulation of translational initiation in plants. Curr. Opin. Plant Biol 5: 460–465.

Kellogg, E.A. (2001). Evolutionary History of the Grasses. Plant Physiol 125: 1198–1205.

Kellogg, E.A. and Buell, C.R. (2009). Splendor in the grasses. Plant Physiol. 149: 1–3.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Pääbo, S. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309: 1850–1854.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. 50: 347–363.

Kim, E.-D. and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci. 17: 16–21.

Kimura, M. (1983). The neutral theory of molecular evolution (Cambridge University Press: Cambridge [Cambridgeshire]; New York).

Koch, M.A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol. Biol. Evol. 17: 1483–1498.

Lagercrantz, U. and Lydiate, D.J. (1996). Comparative Genome Mapping in Brassica. Genetics 144: 1903–1910.

Lander, E.S. et al. (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., and Rhee, S.Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nat. Biotechnol. 28: 149–156.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–1858.

Li, L. et al. (2007a). Global identification and characterization of transcriptionally active regions in the rice genome. PLoS ONE 2: e294.

Li, L. et al. (2007b). Global identification and characterization of transcriptionally active regions in the rice genome. PloS one 2: e294.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178–2189.

Li, P. et al. (2010). The developmental dynamics of the maize leaf transcriptome. Nat. Genet. 42: 1060–1067.

Li, W.-H. (1984). Evolution of duplicate genes and pseudogenes. In Evolution of genes and proteins. Edited by M. Nei and R. K. Koehn. (Sinauer Associates: Sunderland, MA), pp. 98–99.

Li, W.-H. (1997). Molecular Evolution (Sinauer Associates).

Li, W.H., Gojobori, T., and Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. Nature 292: 237–239.

Liao, B.-Y. and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol 23: 530–540.

Lim, K.-B. et al. (2007). Characterization of the centromere and peri-centromere retrotransposons in Brassica rapa and their distribution in related Brassica species. Plant J. 49: 173–183.

Lin, H., Moghe, G., Ouyang, S., Iezzoni, A., Shiu, S.-H., Gu, X., and Buell, C.R. (2010). Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana*. BMC Evol. Biol. 10: 41.

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

Liu, J., Gough, J., and Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet. 2: e29.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., and Pallen, M.J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30: 434–439.

Lou, P., Wu, J., Cheng, F., Cressman, L.G., Wang, X., and McClung, C.R. (2012). Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa*. Plant Cell 24: 2415–2426.

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

Margulies, M. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18: 1509–1517.

Matsui, A. et al. (2008). Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. Plant Cell Physiol 49: 1135–1149.

Matsushita, S.C., Tyagi, A.P., Thornton, G.M., Pires, J.C., and Madlung, A. (2012). Allopolyploidization lays the foundation for evolution of distinct populations: evidence from analysis of synthetic Arabidopsis allohexaploids. Genetics 191: 535–547.

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. PLoS Genet 5.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. Proc Natl Acad Sci U S A 36: 344–355.

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

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24: 2818–2824.

Moghe, G.D., Lehti-Shiu, M.D., Seddon, A.E., Yin, S., Chen, Y., Juntawong, P., Brandizzi, F., Bailey-Serres, J., and Shiu, S.-H. (2013). Characteristics and significance of intergenic polyadenylated RNA transcription in Arabidopsis. Plant Physiol. 161: 210–224.

Movahedi, S., Van de Peer, Y., and Vandepoele, K. (2011). Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol. 156: 1316–1330.

Muller, H.J. (1935). The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica 17: 237–252.

Mun, J.-H. et al. (2009). Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol. 10: R111.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet 39: 1507–1511.

Nekrutenko, A., Makova, K.D., and Li, W.-H. (2002). The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res 12: 198–202.

Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R., and Fluhr, R. (2004). Intron retention is a major phenomenon in alternative splicing in Arabidopsis. The Plant journal : for cell and molecular biology 39: 877–85.

Nuzhdin, S.V., Wayne, M.L., Harmon, K.L., and McIntyre, L.M. (2004). Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol. Biol. Evol 21: 1308–1317.

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

Ohta, T. (1992). The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23: 263–286.

Okazaki, Y. et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327: 92–94.

Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22: 1–5.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.

Paterson, A.H. et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.

Pearson, H. (2006). Genetics: What is a gene? Nature 441: 398–401.

Pennisi, E. (2003). Human genome. A low number wins the GeneSweep Pool. Science 300: 1484.

Pesole, G. (2008). What is a gene? An updated operational definition. Gene 417: 1–4.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., and Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 16: 1262–1269.

Ponting, C.P. and Belgard, T.G. (2010). Transcribed dark matter: meaning or myth? Hum Mol Genet 19: R162–R168.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell 136: 629–641.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–65.

Ramsey, J. and Schemske, D.W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics 29: 467–501.

Ruan, Y. et al. (2007). Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res 17: 828–838.

Rutschmann F. (2005). Bayesian molecular dating using PAML/multidivtime. A step-by-step manual.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. 37: 501–506.

Schnable, J.C., Springer, N.M., and 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, P.S. et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Sémon, M. and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. 17: 505–512.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. Genome Research 19: 1117–1123.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. 96: 336–348.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8: 64.

Sparkes, I.A., Runions, J., Kearns, A., and Hawes, C. (2006). Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. Nat Protoc 1: 2019–2025.

Stolc, V. et al. (2005). Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. Proc. Natl. Acad. Sci. U.S.A 102: 4453–4458.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64: 479–498.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol 14: 103–105.

Subramanian, S. and Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168: 373–381.

Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G., Wang, X., Freeling, M., and Pires, J.C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190: 1563–1574.

Thomas, B.C., Pedersen, B., and 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.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.

Tian, E., Jiang, Y., Chen, L., Zou, J., Liu, F., and Meng, J. (2010). Synthesis of a Brassica trigenomic allohexaploid (*B. carinata* × *B. rapa*) *de novo* and its stability in subsequent generations. Theor. Appl. Genet. 121: 1431–1440.

Tirosh, I. and Barkai, N. (2008). Evolution of gene sequence and gene expression are not correlated in yeast. Trends Genet 24: 109–113.

Town, C.D. et al. (2006). Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18: 1348–1359.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol 28: 511–515.

Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21: 2791–2793.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K.-S. (2004). Mouse transcriptome: neutral evolution of “non-coding” complementary DNAs. Nature 431.

Wang, X. et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43: 1035–1039.

Ward, L.D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science.

Warwick, S.I. and Francis, A. (2005). The biology of Canadian weeds. 132. *Raphanus raphanistrum*. L. Canadian Journal of Plant Science 85: 709–733.

Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A., and Stadler, P.F. (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat. Biotechnol. 23: 1383–1390.

Weigel, D. and Mott, R. (2009). The 1001 Genomes Project for *Arabidopsis thaliana*. Genome Biology 10: 107.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., and Rieseberg, L.H. (2009). The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A 106: 13875–13879.

Wright, S.I., Yau, C.B.K., Looseley, M., and Meyers, B.C. (2004). Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol. Biol. Evol 21: 1719–1726.

Wu, T.D. and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.

Xu, A.G. et al. (2010). Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. PLoS Comput. Biol. 6: e1000843.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature 457: 1033–1037.

Yamada, K. et al. (2003a). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302: 842–846.

Yamada, K. et al. (2003b). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science (New York, N.Y.) 302: 842–6.

Yang, L., Takuno, S., Waters, E.R., and Gaut, B.S. (2011). Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation. Mol. Biol. Evol 28: 1193–1203.

Yang, T.-J. et al. (2006). Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. Plant Cell 18: 1339–1347.

Yang, Y.W., Lai, K.N., Tai, P.Y., and Li, W.H. (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, Y.-W., Tai, P.-Y., Chen, Y., and Li, W.-H. (2002). A study of the phylogeny of Brassica rapa, B. nigra, Raphanus sativus, and their related genera using noncoding regions of chloroplast DNA. Molecular Phylogenetics and Evolution 23: 268–275.

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

Zanetti, M.E., Chang, I.-F., Gong, F., Galbraith, D.W., and Bailey-Serres, J. (2005). Immunopurification of polyribosomal complexes of Arabidopsis for global analysis of gene expression. Plant Physiol 138: 624–635.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. Genome Res. 22: 151–162.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39: 61–69.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R., and Shiu, S.-H. (2009). Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol 151: 3–15.