GENE PAIR REARRANGEMENT DURING THE EVOLUTION OF FLOWERING PLANTS

by

LIANG FENG

(Under the Direction of Jeffrey L. Bennetzen)

ABSTRACT

Flowering plants (angiosperms), comprising ~250,000 species, vary tremendously at the levels of chromosome number and nuclear genome size. At the moment, there is no agreed-upon format for quantitation of the degree of conservation of genic content and colinearity across species. The nature, origins and biases of the numerous small genic rearrangements that differentiate genomes have not been comprehensively investigated, especially inconsideration of possible lineage-specificities in the quality or quantity of rearrangements. The research goals of this project are to investigate the retention or loss of gene pair linkage in various plant species, to quantify the frequencies of various types of genome rearrangements, to understand the lineage-specificity of the genomic instability, and to gain insights into the mechanisms responsible for genic rearrangements.

The great complexity of most angiosperm genomes leads to significant challenges in precise genome comparison, so I will pursue a process of local and global genome comparison through investigating pairs of adjacent genes, and a sampling approach to manually inspect the retention and rearrangement of gene pair linkage. Use of this approach indicates that relative gene pair orientation is random for most plant genes in

most flowering plant genomes, with the dramatic exception of genes that are very tightly linked, where convergent genes are highly over-represented. Careful manual inspection suggests that ~59% of adjacent gene pairs are conserved in rice compared to sorghum. Less than 3% of gene pairs in this comparison are disrupted by gene loss or gene creation. Gene deletions and insertions are observed to be the most common disruptor of gene pairs, relative to other genome rearrangement types, such as inversion and translocation, but most genome rearrangements appear to be the results of multiple events. The gene pair comparison approach has also been extended to a number of plant genomes, including foxtail millet, Brachypodium, Arabidopsis, and Medicago, and suggests that more than 50% of adjacent gene pairs are conserved in every grass pair investigated. The recently sequenced banana and date palm genomes are the first two sequenced monocot genomes outside the grass family, which serve as outgroups to determine the lineage-specificity of the genomic rearrangements that were observed.

Mutation is one of the most important genetic processes, which generate genetic variation between individuals within a species. However, it is not fully clear in plants whether different rates or types of mutation are found in different parts of the genome. By comprehensively investigating mutations that differentiate pairs of LTRs on rice chromosomes 3 and 4, we found that point mutations in chromosome 3 are more abundant near the centromeres, while the transition to transversion ratio (averaging 2.9) does not exhibit any genome location bias. The overall number of these small mutations is significantly correlated with LTR retrotransposon age, but there is no correlation between the transition to transversion ratio and the age of LTR retrotransposons.

This work represents the first to quantify genomic instability during the evolution of flowering plants by combining both high throughput characterization and manual inspection. It advances our understanding of the mechanistic basis of genomic instability in flowering plants. The investigation of rates and natures of genome rearrangement across lineages allows us to identify the evolutionary origins of changes in genome instability, and may provide insights into the mechanisms of the adaptation to various environments for certain species.

INDEX WORDS: flowering plant, genome evolution, gene pair, genome rearrangement, LTR retrotransposon, mutation, Poaceae, synteny

GENE PAIR LINKAGE AND REARRANGEMENT DURING THE EVOLUTION OF FLOWERING PLANTS

by

LIANG FENG

B.S., Liaoning Normal University, P.R. China, 2004

M.S., Beijing Normal University, P.R. China, 2007

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

GENE PAIR LINKAGE AND REARRANGEMENT DURING THE EVOLUTION OF FLOWERING PLANTS

by

LIANG FENG

Major Professor:	Jeffrey L. Bennetzen
Committee:	David W. Hall Richard B. Meagher Chung-Jui Tsai

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

DEDICATION

I dedicate this dissertation to my parents Jinggang Feng and Di Li, without whom I would not accomplish my goals, and to my husband Xinyu Liu, without whom I would not be finished.

ACKNOWLEDGEMENTS

As the Chinese saying goes, “once a mentor, a parent forever”. I would first like to thank Jeff Bennetzen who served as my mentor for my graduate study at the University of Georgia. Jeff is the most knowledgeable and intelligent scientist I have worked with, and I greatly appreciate all that I have learned in his lab. Jeff has always been very supportive of my career. I am extremely grateful for his guidance and constructive criticism on my research, and the freedom and support that he has given me to pursue my own interests. Six years of graduate study in Athens have been one of the most memorable experiences in my life, and I will truly miss being a member of his research team.

My advisory committee also deserves great thanks. I thank CJ Tsai for being so accessible for advice and help, and David Hall for his critical comments on my research projects and advice on my course work. In particular, Richard Meagher and I work on the same floor, and we talk with each other frequently. I truly appreciate his invaluable suggestions on scientific questions, proposal writing, and his encouragement and support on my career development.

I also want to thank some other faculty members of the Institute of Bioinformatics (IOB) and the Department of Genetics, including Russell Malmberg and Liming Cai for taking good care of me during my first rotation, Katrien Devos for providing constructive ideas in our collaboration, Kelly Dyer and Regina Baucom for their helpful comments and revisions on my grant proposal, and Jeremy DeBarry for his advice on my

comprehensive exams and Dissertation Completion Award application. In addition, I want to thank the administrative staff members of IOB for their assistance throughout my studies, and the Graduate School of the University of Georgia for providing the assistantship and fellowship for three years.

My dissertation research also benefited from the stimulating conversations and discussions with fellow members of the Bennetzen lab, including Regina Baucom, Srinivasa Chaluvadi, Jeremy DeBarry, Taoran Dong, Matt Estep, Jamie Estill, Ansuya Jogi, Soo-Jin Kwon, Fang Lu, Jennifer Hawkins, Ryan Percifield, Justin Vaughn, Hao Wang, Lixing Yang, and Qihui Zhu. I particularly want to thank Srinivasa Chaluvadi, Hao Wang, and Qihui Zhu for their valuable suggestions and help on my research projects.

Above all, a special thanks is extended to my parents and my husband, Xinyu, for their endless love, encouragement, and patience.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Patterns in Plant Genome Evolution

The flowering plants (angiosperms), consisting of ~250,000 species, are considerably diverse in their growth habit and nuclear genome structure (Bennetzen 2000). The grasses (*Poaceae*), a monophyletic family with more than 9000 species (Clark et al. 1995) including the world's major cereals, have served as a model family for comparative genetics and genomics in plants (Magallon and Castillo 2009).

The current species of grasses are all derived from a common ancestor that lived about 50-80 million years ago (Mya), but grass genomes have diverged tremendously at the levels of chromosome number, nuclear genome size, and the frequency of genic rearrangement (Bennett and Leitch 1995; Vitte and Bennetzen 2006; Bennetzen 2007). It has been demonstrated that this variation is the outcome of a set of highly active processes, including polyploidy, transposable element (TE) amplification, and small genomic rearrangements, such as deletion, inversion, transposition/translocation and duplication (Bennetzen 2007). Ancient and recent polyploidy and segmental duplication followed by increasing gene numbers and subsequent removal of some post-polyploid gene copies, are believed to be standard phenomena in the history of all flowering plants. (Kashkush et al. 2002; Ilic et al. 2003; Lai et al. 2004). Studies in *Oryza* (Piegu et al. 2006*)* and *Gossypium* (Hawkins et al. 2006) have shown that transposable element amplification, especially amplification of LTR retrotransposons, has been the most important factor responsible for the great variation in plant genome size. Comparison of the genomes of the two domesticated rice subspecies, *japonica* and *indica*, also indicated that theses two genomes had grown more than 2% in size over the past few hundred thousand years because of LTR retrotransposon amplification (Ma and Bennetzen 2004). The abundance of LTR retrotransposons is positively correlated with the overall genome size across flowering plants. For instance, there are > 15% LTR retrotransposons in the ~140Mb Arabidopsis genome, while LTR retrotransposons of the common grass genomes, such as barley or wheat, are > 70% (Liu and Bennetzen 2008; Bennetzen 2009). DNA removal by unequal homologous recombination or illegitimate recombination is an equally important factor in the determination of plant genome sizes (Devos et al. 2002; Bennetzen et al. 2005). The presence of partially deleted LTR retrotransposons and solo LTRs in all studied plant genomes are a reflection of these processes. Truncated elements are thought to be derived from illegitimate (nonhomologous) recombination, but the mechanism of illegitimate recombination is not absolutely clear. Double-strand breaks repair is one likely mechanism to be responsible for most DNA removal. (Kirik et al. 2000; Bennetzen 2009). Solo LTRs are generated exclusively by unequal homologous recombination, which serves as another significant factor for DNA removal (Bennetzen and Kellogg 1997; Shirasu et al. 2000; Ma et al. 2004). In all plants investigated, centromeres show frequent and extensive DNA rearrangement due to the high rates of unequal homologous recombination in centromeric regions, even though meiotic chromosomal exchange is suppressed in most or all pericentromeric regions (Ma and Jackson 2006; Ma et al. 2007). It has been demonstrated that > 190 Mb of LTR retrotransposons from the rice genome have been deleted by the processes of unequal homologous recombination and illegitimate recombination within the past four million years (Myr), leading to a current rice genome of ~400 Mb that is ~20% detectable LTR retrotransposons or fragments thereof (Ma et al. 2004; Tian et al. 2009). Many, perhaps all, of these processes are quite variable in levels of activity in different plant lineages, even when comparing closely related species, and are thus responsible for the great range of different genome sizes and structures observed in flowering plants (Bennetzen and Chen 2008).

However, the rapid and dramatic changes in genome composition and structure have led to little change in genic content, genetic function, and gene order. For instance, more than 80% of Arabidopsis genes appear to have excellent homologs in rice (Feng et al. 2002; Goff et al. 2002). More than 90% of genes are expected to be shared by any two grass species, even though the copy numbers and expression patterns of these shared genes might have diverged to some extent (Bennetzen 2007). Most genes retain similar or identical function despite their very different chromosomal environments in different plant species (Ilic et al. 2003). Moreover, early studies by comparative genetic mapping revealed the extensive conservation of gene content and gross gene order across various grass species (Moore et al. 1995) that was later confirmed and extended by DNA sequence comparisons in orthologous regions (Chen et al. 1997; Tikhonov et al. 1999). The observed colinearity of genes within compared segments of different plant species demonstrated that most gene loci have been retained since their divergence from a common ancestor. In contrast, numerous chromosomal rearrangements, such as telomeric fusions, nested insertions, inversions, and translocations, were observed by detailed comparative genetic and physical mapping among more distantly related species, such as rice, sorghum and maize (Bennetzen and Ramakrishna 2002; Wang and Bennetzen 2012). Most of the detected rearrangements are less than a cM in size, indicating that thousands of smallish rearrangements are much more common than the large rearrangements that are detected when comparing genetic maps (Feuillet and Keller 2002; Bennetzen and Ma 2003; Bowers et al. 2003).

Over-estimation of Genic Content in Plant Genomes

Precise genome comparison in plants is particularly challenging because of the complexity of their genomes. A large number of gene fragments, pseudogenes, and low-copy-number TEs commonly lead to a dramatic over-estimation of gene content (Bennetzen et al. 2004). The copy numbers of most TE families are only a few per genome, even as low as one copy per genome (Bennetzen 1996; Baucom et al. 2009; Yang and Bennetzen 2009). Low-copy-number TEs commonly encode genes that are annotated as hypothetical nuclear genes due to a failure in repeat masking. For instance, mis-identifying TEs as genes resulted in an approximate two-fold over-estimation of gene number in the original rice sequence releases (Bennetzen and Ma 2003; Bennetzen et al. 2004). Misannotation of gene fragments as genes also results in inaccurate gene number prediction. For example, the gene number of maize is decreased by ~20,000 by using a manual inspection approach, with genes only considered real if the gene sequences are more than 70% of the length of the protein-encoding gene in other species, and if they have homologous to a gene in rice or other distant relative of maize (Liu et al. 2007). With a similar manual inspection, the actual number of protein-coding genes is predicted to be ~17% lower than the published result in the *Sorghum bicolor* genome sequence paper (Bennetzen el al. unpub. res.; Paterson et al. 2009). *Helitrons*, LTR retrotransposons and *Mutator* elements have been shown to acquire transpose and express gene fragments sometimes (Jin and Bennetzen 1994; Gandolfo et al. 1998; Jiang et al. 2004). Furthermore, many truncated or inactivated genes left over from the ancient polyploidization events are still annotated as genes (Ilic et al. 2003). A tiny in-frame deletion or nonsynonymous substitution can turn a gene into a pseudogene, but no sequence-based annotation will be sufficient to annotate such sequences as pseudogenes (Bennetzen 2009).

Although the process continues to improve, especially in speed and to a smaller degree in accuracy, genome annotation and genome comparisons tend to be imprecise and full of errors. One of the very important reasons for this inaccuracy is that gene identification continues to be an imperfect process. Low-copy-number TEs and gene fragments are very abundant in plant genomes, and it is never certain whether apparent pseudogenes might actually have functions (Bennetzen 2007). A necessary step towards assessing annotation accuracy is to randomly extract a subset from the entire annotated gene models and then to annotate these sequences with careful manual inspection (Bennetzen et al. 2004). Future studies on full genome sequences in plants should require a careful and largely manual determination of the accuracy of the high throughput annotation that was performed. Such a requirement will correct this persistent misunderstanding of the nature of gene content and gene novelty, and will shed light on the precise genome comparison in plants to uncover the mechanisms of evolutionary processes.

Studying Pairs of Adjacent Genes

"A pair of adjacent genes is a natural unit of gene co-localization" (Rogozin et al. 2004). There are three types of gene pairs in terms of the directions of their transcription: convergent (tail-to-tail, →←), divergent (held-to-head, ←→), and unidirectional (tail-to-head, →→/←←). There are a number of reasons to study the dynamics of genome rearrangements from the perspective of adjacent gene pairs. First, most detected rearrangements are small, often including only one or two genes (Feuillet and Keller 2002; Bennetzen and Ma 2003). In this case, adjacency may be the aspect of genome structure most influenced by inversion, translocation, or DNA deletion. Second, there are some known cases of functional interaction between adjacent genes in eukaryotes. The coexpression of some linked genes is well-documented in various organisms, such as human (Fukuoka et al. 2004; Semon and Duret 2006), *Drosophila melanogaster* (Boutanaev et al. 2002; Kalmykova et al. 2005), *Arabidopsis* (Williams and Bowles 2004), and *Saccharomyces cerevisiae* (Fukuoka et al. 2004). Often, coordinate expression is due to a bidirectional promoter (Hansen et al. 2003), promoter cross talk (Hampf and Gossen 2007), or transcriptional interference (Callen et al. 2004; Shearwin et al. 2005). Third, the orientation of closely linked (intergenic distance < 1 kb) genes is not random. At this distance, there are more divergent gene pairs than expected in a number of vertebrate genomes (Franck et al. 2008), while convergent genes were found to be more frequent in flowering plants, such as rice, Arabidopsis, and poplar (Krom and Ramakrishna 2008). The usual explanation for the origin of closely linked gene pairs is that, once created by chance, the linkage should be evolutionarily maintained by purifying selection based on a degree of shared regulation (Hurst et al. 2002; Singer et al.

2005). Fourth, and most important for our purposes, the adjacent gene pair can serve as a perfect and unambiguous unit to characterize genome rearrangement. The study of rearrangements between adjacent gene pairs will allow us to determine the absolute nature and frequency of different types of genome rearrangement that affect gene composition. By generating the novel terminology of gene pair conservation, we can quantify the frequency of rearrangements, and determine precisely what percentages are caused by deletion, inversion, translocation or other process, for any investigated pair of species. For all of these reasons, I believe that the study of adjacent gene pairs will become a valuable tool to be extensively utilized for comparing genomes and tracing genome evolution.

We currently lack terminology to precisely describe the degree of conservation of genic content and colinearity between any two plant species. We do not yet know whether there are general or lineage-specific patterns in the numerous small genic rearrangements. For instance, it may be that gene order rearrangement is primarily by the accumulation of small deletions, by gene inversion through unequal homologous recombination, or by translocation through recombination between nonhomologous chromosomes. We do not even know whether high conservation at one scale (e.g., the genetic map) in any way correlates with high conservation at other scales (e.g., microcolinearity). Now that the number of plant species targeted for extensive DNA sequence analyses has expanded tremendously, it is time for comprehensive studies to address all of these questions by an effective process of local and global genome comparison.

Objectives and Overview of Dissertation Chapters

At the moment, there is no agreed-upon standard for the quantitation across species of the degree or the nature of change in gene content or colinearity. Possible patterns in the rates or nature of the numerous small genic rearrangements have not been sought. My dissertation research aims to investigate the retention or loss of gene pair linkage in various plant species, to quantify the frequencies of various types of genome rearrangements, to understand the lineage-specificity of the genomic instability, and to gain insights into the mechanisms responsible for genic rearrangements. I also expect to generate a novel terminology and perspective, namely gene pair conservation, to precisely describe the absolute nature of genomic instability during angiosperms evolution. This research represents the first effort to carefully quantify genomic instability during the evolution of flowering plants.

In Chapter 2, the research goals are to identify the distribution of convergent, divergent, and unidirectional gene pairs in rice and sorghum, to precisely investigate the retention and rearrangement of gene pair linkage by considering their genomic context, and to quantify the frequencies of different types of genome rearrangement since the divergence of specific species pairs. The complexity of most angiosperm genomes leads to significant challenges in precise genome comparison, so I have pursued a process of local and global genome comparison through investigating pairs of adjacent genes, and a sampling approach to manually inspect the retention and rearrangement of gene pair linkage. The results of these studies indicate that relative gene pair orientation is random for most plant genes in most flowering plant genomes. Careful manual inspection suggests that ~59% of adjacent gene pairs are conserved in rice compared to sorghum.

Less than 3% of gene pairs in this comparison are disrupted by gene loss from the entire genome or apparent gene creation, verifying our hypothesis of rare genic content change in plant genomes. Gene insertions or deletions are observed to be the most common disruptors of gene pairs, relative to other rearrangement types, such as inversion and translocation.

In the studies described in Chapter 3, the research goals were to apply the approach of manual genome comparison through adjacent gene pairs to multiple plant species, including Brachypodium, foxtail millet, Arabidopsis, and Medicago, to precisely characterize genome rearrangement in several species pairs, to see if different rearrangement frequencies exhibit any lineage-specificity or genome structure correlations by using monocots outside the grass family as the outgroups for grass comparisons, and to gain insights into the mechanisms responsible for genic rearrangements. There are a number of classes of events that are responsible for the observed gene order rearrangements, such as inversion, translocation, and deletion. By a precise survey of a set of randomly selected gene pairs in each plant pair investigated, I find that more than 50% of adjacent gene pairs are conserved in each pairwise grass comparison, while only 7% of gene pairs are retained in the two distantly related dicots, Arabidopsis and Medicago. It also appears that the rice lineage is more unstable than the sorghum lineage by using the date palm genome as the outgroup.

In the experiments described in Chapter 4, the research goal was to understand the mechanisms of genomic instability from the perspective of small-scale mutations, including point mutations and small indels. Mutation is one of the most important genetic processes generating genetic variation. In this chapter, two undergraduate students, Melanie Buser and Zack Farmer, worked with me to comprehensively identify LTR mutations on rice chromosomes 3 and 4, and I analyzed their distribution across the chromosomes. The results demonstrate that point mutations in chromosome 3 are more abundant near the centromeres, while the transition to transversion ratio (averaging 2.9) does not exhibit any genome location bias. The overall number of these small mutations is significantly correlated with LTR retrotransposon age, but there is no correlation between the transition to transversion ratio and the age of LTR retrotransposons.

Significance of This Work

This research will contribute to the interdisciplinary study of evolutionary mechanisms. This study is the first to quantify genomic instability during the evolution of flowering plants by combining both high throughput characterization and manual inspection. Studying genome instability via precise annotation of adjacent gene pair conservation provides a valuable new technique that can be extensively utilized in the field of comparative genomics and molecular evolution. The novel terminology and characterization of gene pair retention between closely related plant species will provide a uniquely definitive tool for the genome annotation community to characterize genomic features with high standards of accuracy and a comparable vocabulary. Following the rates and natures of genome rearrangement across lineages may identify the evolutionary origins of changes in genome instability, and this could suggest why certain species have or have not been able to adapt to changed environments. Understanding mutant generation is an essential, and (in plants) highly understudied, component of understanding adaptation. As climates change rapidly, it will be useful to know what types of gene changes give rise to the creation of mutants with higher adaptation potential. Also important, foxtail millet and sorghum are close relatives of several important bioenergy crops, such as switchgrass (*Panicum virgatum*), napiergrass (*Pennisetum purpureum*), *Miscanthus x giganteus*, and sugarcane (*Saccharum* species) that are all polyploids with rather large genomes and challenging genetics. Our research on monocot genome evolution will shed light on the evolutionary patterns that influence these biofuel crops, and will enhance the research potential on their functional genomics for improved biomass production.

CHAPTER 2

RETENTION AND LOSS OF GENE PAIR LINKAGE DURING THE DESCENT OF RICE AND SORGHUM GENOMES FROM A COMMON ANCESTRAL GENOME[1]

[1] Feng, L. and J.L. Bennetzen. To be submitted to *Genome Research.*

Abstract

The great complexity of most higher eukaryotic genomes leads to significant challenges in precise genome comparison. At the moment, there is no agreed-upon standard for quantitation across species of the degree or the nature of change in gene content or colinearity, and lineage specificity in these traits has barely been investigated. Using gene pair linkage as the assessment criterion, the genomic instabilities of rice and sorghum lineages were investigated, using the *Musa* genome as the outgroup. The results indicate that relative gene pair orientation is random for most plant genes in most flowering plant genomes. However, with genes < 1 kb apart, both rice and sorghum exhibit a great excess (~2-fold) of convergent gene pairs, and these are particularly unlikely to exhibit rearrangement. Using a random sampling process and manual annotation, it was found that ~59% of the adjacent gene pairs have been conserved after > 50 million years of divergence between the rice and sorghum genomes. Analysis of the median/intergenic distances between conserved and rearranged gene pairs in rice and sorghum did not indicate any distance bias, except with genes that were < 1 kb apart. For genes separated by > 1 kb, it appears that the "rearrangement space" between genes is much more constant than the physical distance. Gene movement to other chromosomes is observed to be the most common disruptor of gene pairs, with single gene insertions/deletions within the pairs serving as the second most common event. Inversions that usually involved only a single gene were found in ~2% of the cases examined. Some gene pair rearrangements appear to be the result of multiple events.

Introduction

The great variation in angiosperm genome size, organization and complexity (Bennett and Leitch 1995; Vitte and Bennetzen 2006) is the outcome of a set of highly active processes, including polyploidy, transposable element (TE) amplification, and small genomic rearrangements such as deletions and duplications (Bennetzen 2007). There is evidence that all of these processes are variable in levels of activity in different plant lineages, even when comparing closely related species. Hence, inherited differences in the levels of these activities are expected to be responsible for the great range of different genome sizes and structures observed (Bennetzen and Chen 2008). In dramatic contrast, the highly consistent genome sizes, chromosome numbers and gene contents of gymnosperms over the last 300 million years may be an outcome of consistently low and largely unchanging activity levels of all of these instability mechanisms (Pavy et al. 2012).

Despite rapid and dramatic changes in genome composition and structure, there has been much less change in genic content or gene order. For instance, more than 80% of genes in Arabidopsis appear to have excellent homologs in rice (Feng et al. 2002; Goff et al. 2002), and more than 90% of gene families are shared by any two grass species, even though the copy numbers and expression patterns of the shared genes may have diverged significantly (Bennetzen 2007). Moreover, comparative genetic mapping has revealed extensive conservation of gene content and gross gene order across various grass species (Moore et al. 1995), a result that was later confirmed and extended by DNA sequence comparisons in orthologous regions (Chen et al. 1997; Tikhonov et al. 1999). Within this general colinearity, numerous chromosomal rearrangements (e.g., inversions and translocations) were observed by detailed comparative genetic and physical mapping among distantly related grasses, such as rice and maize (Bennetzen and Ramakrishna 2002). Most of the detected rearrangements are less than a cM in size, indicating that smallish rearrangements are orders of magnitude more common than the large rearrangements that are detected when comparing genetic maps (Feuillet and Keller 2002; Bennetzen and Ma 2003; Bowers et al. 2003).

Genome comparisons currently lack terminology to precisely describe the degree of conservation of genic content and colinearity between any two genomes. Possible subtle patterns and natures of the genic rearrangements within lineages are also unknown. For instance, it may be that gene order rearrangement proceeds by different primary mechanisms, with different primary outcomes, in different taxa. In order to facilitate such analyses, this manuscript describes a gene pair approach to develop a terminology and objective quantitation for genome rearrangement.

There are a number of reasons to study the dynamics of genome rearrangements from the perspective of adjacent gene pairs. The most important reason for our purposes is that the constancy of adjacent gene pairs is an unambiguous and unbiased criterion for measuring genome stability. In plant genomes, most of the detected rearrangements are small, often including only one or two genes (Feuillet and Keller 2002; Bennetzen and Ma 2003). In this case, adjacency may be the aspect of genome structure most influenced by inversion, translocation, or DNA deletion. Hence, characterization of gene pair rearrangement will allow absolute quantitation of the nature and frequency of the different types of genome rearrangement that affect gene content or order.

Previous studies of gene pair conservation relied completely on high throughput characterization, and suggested a very low percentage of gene pair conservation (25.4%) between rice and sorghum (Krom and Ramakrishna 2008; Liu and Han 2009; Krom and Ramakrishna 2010). However, in such analyses, every gene mis-annotation would be counted as a rearrangement, always leading to a higher level of detected instability than is actually present. When genomes routinely contain 15-50% of inaccurately annotated genes (Bennetzen et al. 2004), this can dramatically influence the gene pair analysis outcome. This study uses careful manual inspection to precisely characterize genome rearrangement in rice and sorghum, using the non-grass monocot, banana (*Musa* spp.), as the outgroup. This analysis provides the first detailed assessment of the rates, types, extent and lineages of all types of local gene rearrangement in plants.

Materials and Methods

Non-paralogous Gene Pair Identification

Genome sequence and annotation data for rice (*Oryza sativa ssp. japonica*) Build5 and sorghum (*Sorghum bicolor*) version 1.0 were downloaded from the Rice Annotation Project Database (RAP-DB, http://rapdb.dna.affrc.go.jp/download/index.html) and the DOE Joint Genome Institute database (JGI, http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html). Genes that were annotated as non-coding genes (e.g., RNA genes), pseudogenes, unanchored genes, hypothetical genes without EST evidence, and known transposon-related genes were excluded from the study. If genes were annotated as exhibiting alternative splicing variants, the first annotated variant was selected for further analyses.

A pair of adjacent genes was defined as any two adjacent real genes with no unacceptable sequence gaps between them. A non-paralogous gene pair in this study refers specifically to a pair of adjacent non-duplicated genes. An unacceptable gap was defined as more than 50 consecutive N's in the genome assembly. To identify and remove adjacent pairs of duplicated genes, a homology comparison between the two members of each gene pair was performed by using the BLAST2Seq program (Tatusova and Madden 1999). A cutoff e-value of 1E-10 was used to exclude duplicated gene pairs. Gene pairs containing overlapping genes were also excluded.

For each genome, the numbers of convergent (→←), divergent (←→), and unidirectional (→→/←←) gene pairs were determined, together with the intergenic distance, by means of Perl scripts. Pearson's Chi-Square test was conducted to test whether gene pair orientation is random (i.e., convergent, divergent, and unidirectional genes pairs in a ratio of 1:1:2) by using R.

Manual Inspection of Gene Pair Conservation and Rearrangement

A sampling approach was used to provide the raw material for manual inspection of the retention of gene pair linkage. Four hundred non-paralogous gene pairs were selected from the whole genome dataset; 100 convergent, 100 divergent, and 200 unidirectional gene pairs, of which half were initially selected from rice and half were initially selected from sorghum for each type of gene pair. The selected gene pairs were required to be composed of genes with clear predicted functions, rather than hypothetical or transposon-related genes. No tandem gene families were included in this analysis in order to avoid the confusion associated with orthology::paralogy determination when doing comparisons between genomes.

Gene pair retention and rearrangement were carefully inspected for the randomly selected dataset. In the example shown in Figure 2.1A, a convergent gene pair in rice consists of two adjacent genes (gene 1 and gene 2). Starting with gene 1, the protein sequence of gene 1 was aligned with the full set of protein sequences of sorghum using BLASTP to identify its ortholog (gene 1'), which is defined as the best BLASTP hit with e-value less than 1E-10. Then, the 100 kb of genomic sequence downstream of gene 1' was analyzed for a homologue of gene 2. If the adjacent downstream gene of gene 1' (gene 1'D) is an ortholog of gene 2, and they are in the same orientation, with no intervening gene, then gene pair 1-2 and gene pair 1'-1'D were scored as a conserved gene pair. If there was no annotated ortholog of gene 2 found in this region, an alignment was performed between gene 2 and the 100 kb sequence of genomic DNA with TBLASTN to determine whether there is an ortholog of gene 2 which has been missed by inaccurate annotation or is a partial gene deletion. Because only 100 kb was being analyzed, an e-value of 1E-3 was considered significant. If gene 2' is only found far away from gene 1', with an intervening gene or genes, then the region containing gene pair 1-2 was scored as having undergone rearrangement. If no BLAST hit was found anywhere in the sorghum genome for gene 2, it became a candidate for a gene loss or gene gain. The size 100 kb was chosen because the great majority of nested transposon blocks in angiosperms are smaller than this size, so this scan should find the correct position for the ortholog of gene 2 if it is not separated by a multi-gene segment. Because only 100 kb was investigated, even weak expect values would be significant across such a small data set, thus possibly detecting tiny legacies of partially deleted genes, when present.

To confirm a real gene loss or gene gain event, the apparently novel gene was screened for homology in the NCBI non-redundant (nr) database, NCBI dbEST of non-human non-mouse cDNA sequences, and a plant repeat database. The plant repeat database was constructed with sequences downloaded from Repbase (http://www.girinst.org/server/RepBase/index.php), MIPS PlantsDB (ftp://ftpmips.helmholtz-muenchen.de/redat/), and the TIGR Plant Repeat Databases (http://plantrepeats.plantbiology.msu.edu/downloads.html). For example, a rice gene was scored as a loss/gain relative to sorghum if it met two criteria. First, there was no BLAST hit with e-value lower than 1E-5 when searching against the known repetitive sequence or TE databases. Second, there was also no BLAST hit with e-value lower than 1E-5 found in either annotated proteins or genomic DNA of sorghum. Another factor investigated was whether this potential rice-specific gene is present in at least one other plant species. If the candidate gene was found in multiple species, but was not found in the sequenced sorghum genome (Paterson et al. 2009), then it was scored as a sorghum gene loss.

The same analysis was then performed by starting with gene 2 instead of gene 1, and so on for the analysis of the 400 randomly-selected gene pairs for the two plant species. Mis-annotation cases were also identified with this manual inspection. Some gene models had incorrect gene boundaries or their annotation way by translation in an incorrect reading frame. The most common mis-annotation was for the same sequence to be called two adjacent and unidirectional genes in one species but called a single gene in the other species. Any of the initial gene pairs containing such a mis-annotation candidate was removed and replaced by another randomly selected gene pair. Pearson's Chi-Square test was used to test for differences in conservation and rearrangement frequencies among convergent, divergent, and unidirectional gene pairs, and to test for differences between the numbers of conserved and rearranged gene pairs.

Gene Pair Distance Analysis

The distance between two adjacent genes was measured using both the median distance between gene “midpoints” and intergenic distance in base pairs. The “midpoint” of the gene was defined as the bp located midway between the 5’ and 3’ coding sequence (CDS) ends. The intergenic distance was defined as the number of bp between annotated 5’ and 3’ ends of the CDS of adjacent genes. The Kolmogorov-Smirnov test was used to test normality. The parametric t-test or nonparametric Mann-Whitney U-test was used to test the significance of differences for median/intergenic distance between conserved and rearranged gene pairs. The nonparametric Spearman's Rank Correlation and parametric Pearson's Correlation for log-transformed data were performed to test the significance of the correlation coefficient of the median/intergenic distances in conserved gene pairs. A contingency table was constructed to record and analyze the relation of median/intergenic distance with different ranges (i.e., 0-5 kb, 5-10 kb, 10-15 kb, > 15 kb for median distance; 0-3 kb, 3-6 kb, 6-9 kb, > 9 kb for intergenic distance) between rice and sorghum. Given the observation of median/intergenic distance in one species, the conditional probability of median/intergenic distance in the other species observed in the same distance range was also calculated.

Lineage Specificity of Genome Rearrangements

The genome sequence and annotation data for banana (*Musa acuminata ssp. malaccensis*) version 1 were downloaded from the Banana Genome Hub centralises databases (http://banana-genome.cirad.fr/) (D'Hont et al. 2012). If a gene was annotated as containing alternative splicing variants, the first annotated variant was selected for further analyses. To determine the lineage specificity of genome rearrangement between rice and sorghum, the unconserved gene pairs in the rice and sorghum genome comparison were used to search orthologs in the banana genome. Probable orthologs were defined as the best BLASTP hit with e-value less than 1E-10. In addition, a random sampling of gene pairs in banana was used to provide a larger data set. Four hundred gene pairs (excluding tandem duplicates) were selected from banana; 100 convergent, 100 divergent, and 200 unidirectional gene pairs. As with the grass genes pairs, only pairs that did not include non-coding genes, pseudogenes, unanchored genes, hypothetical genes, or TE-related genes were selected. The selected gene pairs were also required to contain only genes with clear predicted function. With this manual inspection, the numbers of banana gene pairs conserved in both rice and sorghum, rearranged in both rice and sorghum, conserved in rice but rearranged in sorghum, and conserved in sorghum but rearranged in rice, were identified.

<u>Results</u>

Gene Pair Identification, Classification, and Organization

The IRGSP Build5 pseudomolecules of the rice genome contains 34,780 annotated genes and the JGI v1.0 assembly of sorghum genome contains 34,496 annotated genes. After excluding RNA genes, pseudogenes, unanchored genes, hypothetical genes, and known transposon-related genes from the gene sets, a total of 33,195 and 27,444 genes were used for further analyses in rice and sorghum, respectively. By discarding gene pairs containing sequencing gaps or overlapping genes, a total of 30,903 rice and 22,857 sorghum gene pairs were identified. The numbers and percentages of each type of gene pairs are shown in Table 2.1. The fractions of convergent, divergent, and unidirectional gene pairs in rice are 24.4%, 23.8%, and 51.8%, respectively, while the respective frequencies in sorghum are 23.7%, 22.0%, and 54.3%. If gene orientation were random, the ratio of convergent, divergent, and unidirectional gene pairs should be 25% : 25% : 50%. Hence, this analysis indicates that the percentage of unidirectional gene pairs is slightly higher than the expectation.

A non-paralogous gene pair in this study refers to a pair of adjacent non-duplicated genes. Because tandemly duplicated genes can make it difficult to identify orthologues rather than paralogues, and because rearrangements of these genes are both frequent and primarily caused by the well-characterized process of unequal homologous recombination, we decided to exclude tandem gene duplications from further analysis in this study (Table 2.2). The percentages of tandem-duplicated gene pairs in direct (unidirectional) orientation (12.3% in rice; 16.3% in sorghum) are significantly higher than for convergent or divergent tandem gene pairs (2.4% and 2.6% in rice; 4.3% and 4.3% in sorghum). After removing tandemly duplicated gene pairs, the frequencies of the three types of gene pairs are 25.8% convergent, 25.1% divergent, 49.1% unidirectional in rice, and 25.5% convergent, 23.6% divergent, 50.9% unidirectional in sorghum (Table 2.1). Pearson's Chi-Square test indicated that these values were not random ($P < 0.05$)

because of an excess of convergent gene pairs. However, the orientation of non-paralogous gene pairs is closer to random (*P* = 1.3E-6) than that of all gene pairs (*P* = 0.0), indicating that gene duplication is one of the most important factors that influences gene pair orientation.

If the intergenic distance of adjacent gene pairs is taken into account, a different picture emerged. The orientation of closely-linked (intergenic distance < 1 kb) gene pairs was found to be highly biased toward convergent gene pairs, 46.3% in rice and 46.8% in sorghum (Table 2.3). This observation is consistent with previous results in *Arabidopsis* and *Populus* (Krom and Ramakrishna 2008), but contrasts dramatically with analysis in vertebrate genomes, where there are more divergent closely linked gene pairs than expected (Adachi and Lieber 2002; Li et al. 2006; Franck et al. 2008).

Gene Pair Retention and Rearrangement during Grass Genome Evolution

Precise genome comparison among plants is particularly challenging because of the complexity of plant genomes. Over-estimation of gene content, recent polyploidization, chromosomal duplication, and single gene duplication will affect the accuracy of automated high throughput characterization and lead to incorrect conclusions. A sampling approach was used to provide material for detailed manual inspection of gene pair linkage. A randomly extracted data set consisting of 100 convergent, 100 divergent, and 200 unidirectional gene pairs, in which each gene had a clear predicted function, was carefully annotated for the precise degree and nature of rearrangement when comparing rice and sorghum genomes. Figure 2 provides a few examples of the main outcome categories.

Figure 2.2A represents gene pair retention, in which a convergent gene pair in rice possesses homologs in sorghum in the same orientation and with no intervening genes. In some cases, conserved gene pairs may be internal to a large inversion event, but this does not alter gene pair linkage in any way, so these events were annotated as gene pair retention (Figure 2.2B). In Figure 2.2C, gene n, annotated as a hypothetical gene, is inserted between two rice genes. Gene n was then compared against the plant repeat database, the nr database, and dbEST. In this case, no significant homology was found in any of these databases, so the candidate gene was considered likely to be an annotation error, and the chosen gene pair in sorghum was classified as a conserved gene pair. In two cases, as shown in Figure 2.2D, the orthologues of gene 1 and gene 2 are separated by two other genes on a separate chromosome in sorghum, but the second best BLAST hits of these two genes are adjacent to each other in the same orientation on another sorghum chromosome. This class of gene pair was also considered a conserved gene pair. With the manual inspection for each chosen gene pair, a total of 64 convergent, 61 divergent, and 111 unidirectional gene pairs were found to be conserved in rice relative to sorghum (Table 2.4).

To determine the frequencies and the scopes of gene rearrangements, each rearranged gene pair was carefully inspected. Figure 2.3 shows the basic patterns of gene rearrangement observed. Figure 2.3A depicts the most frequent type of gene pair disruption (96 cases), where one of the members of the gene pair in one species was found on another chromosome in the compared species. In Figure 2.3B, a case is given where it is unclear whether it is a gene insertion in one species or a gene deletion in the other spaces. This was the second most-frequent type of rearrangement (41 cases). To determine whether the insertion/deletion event is real, the intervening gene (gene n) would be tested for homology in the repeat database, the nr database, and dbEST. This rearrangement would be confirmed if these two gene candidates have no significant BLAST hit in the repeat database but have at least one significant hit in the nr database and/or dbEST. In Figure 2.3C, the orthologs of a divergent gene pair 1-2 are adjacent to each other in sorghum, but the relative orientation of the two genes is different. This gene pair is considered a rearranged gene pair with a single gene inversion, an event observed in 7 out of the 400 gene pairs investigated. Figure 2.3D shows the apparent outcome of an inversion involving a single segment carrying 32 genes plus a single gene insertion/deletion.

Gene loss/gain is a special case of genic rearrangement. A total of 6 gene pairs were observed that appeared to contain genes that were not found anywhere in the other species genome, including 2 divergent cases and 4 unidirectional cases. An example is given in Figure 2.4A, where gene 1 in rice was annotated as “similar to autophagy-related protein 8D”. There are cDNA hits in dbEST for this gene in rice, Brachypodium, and Triticum, but no significant hit was found when the BLAST searches were against sorghum protein sequences, sorghum genomic DNA, the repeat database, or the nr database. This gene is thus classified as a gene loss rearrangement in sorghum. Some sequence legacies of deleted genes in the genome were also found, as shown in Figure 2.4B. Gene 2 in sorghum was annotated as “similar to ethylene-responsive small GTP-binding protein”, which exists in many other plant species. It has cDNA evidence in dbEST for sorghum, maize, Brachypodium, Setaria and Panicum, and has a significant TBLASTN hit (e-value: 2E-83) in the intergenic region of rice chromosome 6. This fragment in rice is a pseudogene associated with a partial gene deletion. Less than 3% of randomly selected gene pairs exhibited potential gene loss/gain loci in this study.

With the precise genome comparison between rice and sorghum, 64% of convergent, 61% of divergent, and 56% of unidirectional gene pairs were found to be conserved between the two grasses (Table 2.4). Application of Pearson's Chi-Square test demonstrated that there is no significant difference in the conservation frequencies among convergent, divergent, and unidirectional gene pairs (P = 0.429), and that the frequencies of conserved gene pairs are significantly higher than that of rearranged gene pairs (P = 0.002). The overall ratio of gene pair conservation to rearrangement is 59:41. For the gene pairs less than 1 kb apart, 3 out of 17 of convergent pairs were rearranged, while 8 out of 25 divergent or unidirectional pairs were rearranged (Table 2.5). Although this suggests almost 2-fold greater stability of the convergent gene pairs, this data set was too small to indicate statistical significance for this difference. Overall, however, the ratio of stable to rearranged copies for those genes < 1 kb apart was 74:26, which is significantly less rearrangement than seen for the more distantly linked gene pairs.

Frequencies of Different Types of Rearrangements

The most frequent rearrangements are exemplified by the presence of linked gene pairs on one chromosome for the selected pair, but with those genes on different chromosomes in the other species (96 events). Although we can use the comparable genetic maps to clearly determine whether one or both of the paired genes are on orthologous chromosomes, this has not yet been done. Of the simpler events, single gene insertions or deletions were the most abundant, accounting for 41 confirmed rearrangement events. The next most frequent events were inversions, with 7 cases of single gene inversion (Figure 2.3C) and 1 case involving a segment with 32 genes (Figure 2.3D). No cases of translocation-related gene pair rearrangement were observed. A few of the gene rearrangements appear to be the result of multiple events like the one depicted in Figure 2.3D, where a single gene insertion is found adjacent to the inversion of the segment that contained 32 genes.

Misannotation of Plant Genomes

Since the genome sequences of Arabidopsis and rice have been completed at high standards of accuracy, plant genome sequencing efforts have been expanding, but without such detailed (and expensive efforts) at completeness. Moreover, the near-total reliance on high throughput characterization for many of these genomes was accompanied by minimal effort to assess the accuracy of gene annotation. With the very careful manual inspection of a randomly chosen subset of gene pairs, a number of mis-annotations were identified in the relatively well-annotated sorghum and rice genomes. A common mis-annotation type is presented in Figure 2.5A. A pair of unidirectional genes in rice aligned perfectly to two different parts of one single gene in sorghum. In this case, the intergenic distance of the gene pair is 7357 bp in rice, and it appears that the “one” sorghum gene is the mis-annotation, because EST evidence and further annotation in Setaria, Brachypodium, and *Musa* indicate two genes. A total of 5 cases of this type of mis-annotation were observed. Two cases of mis-annotation caused by inappropriate reading frame translation were also observed. For instance, in Figure 2.5B, gene 2 of a unidirectional gene pair in sorghum has no protein orthologs when searching against predicted proteins in the rice genome, but it has a very significant hit (e-value: 4E-48) in rice genomic DNA. This genomic region overlaps with an adjacent upstream gene of gene 1' (gene 1'U). Careful inspection indicated that the product of gene 2 and the annotated protein of gene 1'U are different primarily because they were translated in different frames for the same region of genomic DNA. Protein 2 has many homologs in various species, while protein 1'U has no homolog in any other species, indicating that 1'U is the more likely annotation error. This facile observation of examples of mis-annotation by comparative analysis indicates that this approach could be used to help determine the accuracy of any high throughput annotation.

Median and Intergenic Distance Conservation

Distances between the two adjacent genes that constituted each gene pair were measured using the median distance and intergenic distance in base pairs. To identify whether the distances between conserved and rearranged gene pairs are significantly different, both the median distance and intergenic distance were compared between the 236 conserved and 164 rearranged gene pairs identified in this study. No significant difference was found for either median distances observed between conserved and rearranged gene pairs (Mann-Whitney test, $P > 0.05$; Table 2.6) or for intergenic distances ($P > 0.05$, data not shown).

However, when the distances between the 236 conserved gene pairs were compared between sorghum and rice, a positive correlation was observed for both median and intergenic distances (Figure 2.6). Hence, if no genic rearrangements are observed, the distance between genes has a tendency toward conservation over a period of ~50 million years. The most common median distance of conserved gene pairs is 5-10 kb in both rice and sorghum, followed by 0-5 kb. A contingency table to record the median distances in different ranges (i.e., 0-5 kb, 5-10 kb, 10-15 kb, > 15 kb) was constructed between rice and sorghum, and presented the distance frequency distribution and their mutual association across species (Table 2.7). The numbers on the diagonal represent the median distances with the same size range in rice and sorghum. For instance, there are 40 conserved gene pairs, of which the median distances in rice and sorghum are in the same range of 0-5 kb. Similarly, 61, 9, and 18 conserved gene pairs have conserved median distance ranges of 5-10 kb, 10-15 kb, and > 15kb, respectively. Given the observation of median distance in one species, the observed frequencies of median distance in the other species observed in the same range was calculated based on the observation in the contingency table (Table 2.8). For instance, given the median distance of 0-5 kb for a gene pair in rice, the likelihood of its conserved gene pair in sorghum having the same distance range of separation by chance is 0.284, yet the observed frequency is 0.678. The nonparametric Spearman's Rank Correlation indicates a highly significant correlation for median distance of conserved gene pairs when comparing rice and sorghum (Spearman’s correlation coefficient, 0.646; P = 1E-6), and Pearson's Correlation for log-transformed data yielded similar results (Pearson’s correlation coefficient, 0.629; P = 2.2E-27). Similar trends were also observed in the analyses of intergenic distance (data not shown).

Lineage Specificity of Genomic Instability

Comparative genetic maps in grasses indicate that different lineages exhibit different degrees of genomic instability. In order to identify the lineages of the rearrangements that differentiate sorghum and rice genomes, an outgroup is needed. Date palm provided the first sequenced monocot genome that is not from a grass (Al-Dous et al. 2011), but the depth of sequence coverage and thus overall completeness of the genome assembly is lower than for another sequenced monocot, *Musa* (D'Hont et al. 2012). Manual analysis of 164 rearranged gene pairs, including the 6 loss/gain cases, form the rice/sorghum comparison, uncovered 8 cases where the *Musa* and rice arrangements were the same, 9 cases where the sorghum and *Musa* arrangements were the same, and 129 cases where the *Musa* arrangement did not have the same composition as either rice or sorghum. The other 18 gene pairs could not be identified in *Musa* due to annotation or assembly issues.

Given the high rate of local rearrangement that differentiates *Musa* from the two grasses, a random sampling approach selecting 400 non-paralogous gene pairs from banana was used to increase the data set. These randomly chosen pairs were composed of 100 convergent, 100 divergent, and 200 unidirectional gene pairs, with our stringent criteria that each gene not be an annotation artifact. Manual inspection found that 343 banana gene pairs were rearranged relative to both rice and sorghum, while only 22 were conserved relative to both rice and sorghum (Table 2.9). Of the other 35 *Musa* gene pairs, 13 had the same arrangement as in rice but not in sorghum, 12 had the same structure as in sorghum but were different from rice, 2 of them are present in rice but absent in sorghum, and 8 of them are present in sorghum but absent in rice. In this context, “absence” is defined as absence of at least one gene of the pair. In all the gene pairs investigated, only the loss of one gene in the pair was observed. Hence, both analyses indicate comparable frequencies of genome rearrangement in rice and sorghum lineages.

Discussion

Studying Non-Paralogous Adjacent Gene Pairs

One of the advantages of comparing rice and sorghum genomes, rather than those from other plant species, is that these two grass genomes have not undergone polyploidization since their divergence from a common ancestor ~50 Mya. In contrast, a maize ancestor underwent a polyploidization some time in the last few million years (Swigonova et al. 2004), making it more complicated to quantitate stability if one needs to factor in two maize orthologous pairs for every single orthologous pair in rice or sorghum. In addition, polyploidization leads to a higher acceptable rate of gene loss because every locus is initially duplicated, and also exhibits properties of "fractionation" (Freeling 2008) where one homoeologous segment tends to lose genes more rapidly than another. Although these will be interesting factors to assess in future studies, we felt that this first demonstration of the precise quantitation of rearrangement of plant genes should investigate as simple a case as possible. In this study, focus was placed on non-paralogous gene pairs, because tandem gene duplication also confuses orthologue identification. Moreover, the rearrangement of tandemly duplicated genes is a well-studied process involving unequal homologous recombination that does not desperately need additional analysis in our study.

Orientation and Intergenic Distance Conservation in Rice and Sorghum

By excluding tandem gene duplications, gap-containing gene pairs, hypothetical genes, transposon-related genes, and overlapping genes, characterization of gene pair organization can be quite robust. Adjacent paralogous genes were found to be common in both rice (17.3%) and sorghum (24.9%), of which the great majority are directly ("unidirectionally") repeated. About 5% of tandem gene duplicates were found to be in inverted orientation, a structure that can arise during replication if the DNA polymerase switches to the complementary strand before switching back to the original strand (Bi and Liu 1996; Kato et al. 2000) or by unequal recombination between flanking repeats (e.g., TEs) in opposite orientations. The overall paralogous gene copy number results agree with previous studies that indicated many more (~1.9X) duplicated genes in sorghum than in rice (Paterson et al. 2009).

After removing paralogous and very tightly linked gene pairs, the relative frequencies of convergent, divergent and unidirectional gene pairs were found to be statistically identical to a random choice of orientations, which would yield a respective 1:1:2 ratio. A highly dramatic exception to this randomness was observed with gene pairs that exhibited intergenic distances of < 1 kb. At this range of separation, the frequency of convergent genes is ~2X higher than random. This observation is quite different from investigations in mammalian genomes, in which closely linked gene pairs are more often divergently transcribed than would be expected by chance. (Adachi and Lieber 2002; Li et al. 2006; Franck et al. 2008). For instance, the ratios of convergent: divergent: unidirectional closely linked gene pairs (intergenic distance < 600bp) are 28%: 49%: 23% in human, 30%: 53%: 17% in chimpanzee, and 28%: 50%: 22% in mouse (Franck et al. 2008). The non-random distribution of closely linked gene pairs may result from regulatory interactions between adjacent genes, perhaps caused by bidirectional promoters (Hansen et al. 2003), promoter cross talk (Hampf and Gossen 2007), or transcriptional interference (Callen et al. 2004; Shearwin et al. 2005). The establishment of gene pair orientation could be initially random, but those random beneficial mutations that led to incidental overlap in adjacent gene regulation would then make it difficult for genes to be separated subsequently. If this accidental overlap in regulation required a certain geometry (e.g., relative arrangement or distance between regulatory modules), then this could also explain the observation we made that adjacent genes are often about the same distance apart in sorghum and rice. It is interesting, though, that the ~2X bias towards tightly linked divergent gene pairs in mammals is very similar to the nearly 2X bias we see toward convergent gene pairs in these two grasses.

An excess of convergent genes among tightly linked pairs suggests that these genes might be antagonisticly regulated by transcriptional interference or by the creation of dsRNA through overlap of the 3’ ends of the mRNAs. Further analysis will be needed to identify whether such antagonism exists, both by characterization of the mRNAs encoded by these genes and by analysis of their representation in siRNA libraries. Interestingly, the conservation of convergent gene pairs < 1 kb apart was particularly high (~82%). Of course, genes that are tightly linked have less space for an intervening rearrangement event, but the equally tightly linked divergent and unidirectional gene pairs exhibited a conservation frequency of ~68%. Higher conservation suggests a stronger selection against rearrangement of tightly-linked convergent pairs, a selection difference that was not seen for gene pairs of any orientation that were less tightly linked, but larger data sets will be needed to test the significance of the < 1 kb results.

Comprehensive Description of Gene Pair Retention and Rearrangement

Currently, we have no controlled vocabulary to describe, much less quantitate, the level or nature of genome rearrangement. From the perspective of gene pair conservation, one can precisely determine the frequency of overall gene pair rearrangement and of the different types of rearrangement; namely insertion, deletion, inversion, and translocation. Our results indicated that insertions or deletions of single genes outnumber any other class of simple genic rearrangement. Because the outgroup used in these studies (*Musa*) was so heavily rearranged relative to these two grasses, it was rarely possible for us to differentiate insertions from deletions. A more closely related outgroup, or a more robust analysis across a broader sampling of the grass phylogenetic tree, should resolve this issue.

Although inversions represented a fairly numerous class of rearrangement, they were not particularly diverse in their nature. All but one involved only a single gene. This small usual size could be a mechanistic outcome, related either to a common short length of template switching (Bi and Liu 1996; Lin et al. 2001) or an indication of a greater likelihood of unequal recombination between gene-flanking inverted repeats if those repeats are close together. The paucity of discovered translocations (none were seen) was not a surprise, because of their very negative effects on fertility in rearrangement heterozygotes and on fitness in individuals with segmental duplication/deletion outcomes. Comparative genomic studies in all eukaryotes show a low frequency of translocations, and we would only classify such a rearrangement if the break point were between the two genes of a gene pair. Hence, if two species differ by a single translocation, and they contain about 25,000 adjacent gene pairs, then there is only one gene pair in that 25,000 that should be scored as a rearrangement caused by translocation.

Previous studies of gene pair conservation in plants have relied completely on high throughput characterization, and indicated only 25.4% gene pair conservation between rice and sorghum (Krom and Ramakrishna 2008; Liu and Han 2009; Krom and Ramakrishna 2010), compared to our conservation level of 59%. Because so many candidate genes are actually annotation artifacts (commonly TE-vectored gene fragments) (Bennetzen et al. 2004), and these errors always lead to an artifactual "gene rearrangement" conclusion, we believe that current annotation accuracy requires a very careful manual analysis to provide an accurate measure of gene pair conservation. Segmental duplication, like polyploidy, can also confuse both gene pair analysis and removes some selective constraints to rearrangement, and we expect this to be a factor on rice chromosomes 11 and 12 that contain large segmental duplications dating 5-7 Mya (Choisne et al. 2005; Wang et al. 2007). All forms of gene duplication, including polyploidy and segmental duplication, lead to a loss of selection for retention of both gene copies (Thomas et al. 2006; Doyle et al. 2008) and also impact the accuracy of genome comparisons, especially those that rely exclusively on high throughput characterizations that are replete with paralogy/orthology errors. The random sampling approach employed in this study allowed the choice of genes that had very low artifactual potential and very clear orthology relationships. The only limitation to this approach is the size of the data set that can be assayed with this labor intense analysis. By using random resampling, we are able to show that the 95% confidence interval for the measured frequency of gene pair conservation is 59% +/- 4.8% by resampling 1000 times.

One rare category of gene pair rearrangement was comprised of the 3% of gene pairs that were attributable to apparent gene gain or gene loss from the entire genome. This rarity is consistent with models stating that gene loss/gain in closely-related plant species is much rarer than indicated by many full genome comparisons, because these comparisons rely primarily on high throughput annotation (Bennetzen 2007). The exceptional diversity of TEs (especially the very numerous low copy number TE families) and pseudogenes in higher plant genomes will consistently lead to over-estimation of gene content, and a large number of these misannotated sequences are thus characterized as “novel genes” (Bennetzen et al. 2004). In a small study of two very-well-sequenced and annotated genomes, we found 9 misannotation cases even after all hypothetical genes were removed. We suggest that future publications on full genome sequences in plants should require a manual characterization of the accuracy of the high throughput annotation that was performed, and a comparative gene pair analysis could greatly assist this task.

For the few apparent gene gain or gene losses that we identified, several of the presumed losses left behind a legacy of the previous gene with a reasonable BLASTN expect value, but either no gene model or a gene model with a BLASTP homology below the cutoff value to identify significant homology in a full-genome scan. When only investigating a stretch of 100 kb, however, these homologies were sufficient to show that the “lost gene” was actually still there, but in a degraded or rapidly evolving form. Further studies with larger data sets and a broader selection of species will be needed to determine whether this is simply pseudogene formation or whether some of these genes are diversifying into neofunctional states. It should also be noted, however, that a “gene loss" or "gene gain" is an outcome of the comparison of two individuals, not of a comprehensive comparison of two species. It is known that the germplasms of many (perhaps all) organisms can be polymorphic for a certain percentage of gene presence/absences (Flint-Garcia et al. 2009; Cao et al. 2011). Hence, some of the apparent gene losses in rice, for instance, might be genes that are found in other rice varieties, while some of the gene gains that differentiate one species from the other might not do so if a broader spectrum of germplasm was investigated.

Evolutionary Conservation of Median/Intergenic Distance

Intergenic distance has been identified as a strong determinant of gene pair conservation in a number of eukaryotes (Basu et al. 2008; Liu and Han 2009; Davila Lopez et al. 2010), making the intrinsically obvious suggestion that genes that are farther apart should have higher chances of rearrangement. However, analysis of the median/intergenic distances between conserved and rearranged gene pairs in rice and sorghum did not indicate any distance bias, except with genes that were < 1 kb apart. One possible explanation for this result would be that the "rearrangement" is very different from the physical space between genes. When a pair of adjacent genes is very far apart, most of the DNA is usually comprised of TEs or fragments of ancient TEs (Bennetzen 2007; Bennetzen 2009). This TE-derived DNA is likely to be epigenetically silenced, primarily as heterochromatin. Perhaps this heterochromatin is itself resistant to acting as sites for rearrangement, meaning that the actual rearrangement space between distantly linked gene pairs might be as small as the gene itself and the short stretches of flanking euchromatin. One rearrangement process, homologous recombination, seems to be largely limited to euchromatin (Dooner and He 2008), and maybe this is also true for other rearrangement mechanisms that might be responsible for the local genomic instability that we have observed. It will be particularly interesting to see if there is a gene pair distance effect in larger genomes like maize, where intergenic distances show a very broad range (< 1 kb to megabases) and average > 50 kb (Schnable et al. 2009), but may still have very similar sizes of flanking euchromatin blocks regardless of the physical distance.

The observed positive correlation of median/intergenic distance when comparing conserved gene pairs in rice and sorghum suggests an important functional role for gene spacing. It is not clear whether this unexpected result is caused by requirements for chromatin folding (Kalmykova et al. 2005), how regulatory modules are spaced (Hansen et al. 2003; Hampf and Gossen 2007) or some other factor.

Lineage Specificity of Genome Rearrangements

By comparing genome arrangement patterns in several plant species of known phylogenetic relatedness, one should be able to determine the lineage and approximate timing of any rearrangement. Because of the low level of conservation between the chosen outgroup in this study, *Musa*, and the grasses, very few rearrangements could be attributed in either their nature (e.g., are particular indels insertions or deletions) or their lineage of origin. The diversification of the grasses from a common ancestor has been estimated at 50–70 Mya, while the divergence time of Poales and Zingiberales from a common ancestor has been estimated at 109–123 Mya (Magallon and Castillo 2009; Vogel et al. 2010). It is interesting that the ~2-fold longer time since divergence for *Musa* versus the grasses compared to sorghum versus rice is accompanied by a much higher level of rearrangement. With ~40% rearrangement seen in ~60 million years in these two grasses, then the same rate of change would lead to ~65% rearrangement for ~120 million years of divergence. However, the *Musa* versus rice and *Musa* versus sorghum degrees of rearrangement of gene pairs are ~91% and ~90%, respectively, suggesting a much higher rate of instability in the *Musa* lineage or a higher rate of instability in the two grass lineages before they diverged, or both.

Comparing the two grass genomes, with this sample size of 400 gene pairs, it is clear that these two lineages have exhibited little or no accumulated difference in their degree of gene pair instability over the last 50 million years. However, there is not enough data in this study to determine whether different types of rearrangements are more frequent in one lineage than in another. Future studies will need to apply these approaches to more organisms to provide both larger data sets and a better set of phylogenetically-appropriate outgroups. Development of this approach and the vocabulary to describe rearrangement types and lineages should make future analyses of larger data sets easier, so that more robust numbers can be generated to determine the natures and rates of genomic rearrangements that differentiate the genomes of any and all families of organisms.

Table 2.1. Numbers and percentages of adjacent and non-paralogous adjacent gene pairs in rice and sorghum

	All gene pairs				Non-paralogous gene pairs			
Species	Total	Convergent	Divergent	Unidirectional	Total	Convergent	Divergent	Unidirectional
Oryza	30,903	7549 (24.4%)	7360 (23.8%)	15,994 (51.8%)	28,567	7368 (25.8%)	7171 (25.1%)	14,028 (49.1%)
Sorghum	22,857	5427 (23.7%)	5029 (22.0%)	12,401 (54.3%)	20,384	5193 (25.5%)	4813 (23.6%)	10,378 (50.9%)

Table 2.2. Numbers and percentages of tandemly duplicated gene pairs in rice and sorghum

Species	Categories	Total gene pairs	Duplicated gene pairs
Oryza	convergent	7549	181 (2.4%)
	divergent	7360	189 (2.6%)
	unidirectional	15,994	1966 (12.3%)
Sorghum	convergent	5427	234 (4.3%)
	divergent	5029	216 (4.3%)
	unidirectional	12,401	2023 (16.3%)

Table 2.3. Numbers and percentages of non-paralogous gene pairs separated by < 1 kb

Species	Total	Convergent	Divergent	Unidirectional
Oryza	2919	1352 (46.3%)	597 (20.5%)	970 (33.2%)
Sorghum	2396	1122 (46.8%)	396 (16.5%)	878 (36.7%)

Table 2.4. Detailed results for 400 randomly selected gene pairs compared between rice and sorghum

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	64	61	111
Rearranged gene pairs	36	37	85
Gene pairs with gene loss/gain	0	2	4
Total	100	100	200
Misannotation	1	0	11
Unknown	3	8	4

Table 2.5. Results for randomly selected gene pairs < 1 kb apart in rice and sorghum

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	14	7	10
Rearranged gene pairs	3	2	5
Gene pairs with gene loss/gain	0	1	0
Total	17	10	15

Table 2.6. Comparison of median distances between conserved and rearranged gene pairs

Category	Group	N	Mean	Median	Std. Dev.	*P**
Convergent	Conserved pairs	64	7500	5420	5480	0.585
	Rearranged pairs	36	8920	5480	8530	
Divergent	Conserved pairs	61	12,600	8150	114,00	0.435
	Rearranged pairs	39	167,00	9720	200,00	
Unidirectional	Conserved pairs	111	12,200	7250	19,100	0.394
	Rearranged pairs	89	11,400	8260	11,000	

* $P < 0.05$ is considered a significant difference between the means of the two groups by Mann–Whitney tests.

Table 2.7. Contingency table for median distances

Os Sb	0-5 kb	5-10 kb	10-15 kb	> 15 kb
0-5 kb	**40**	23	2	2
5-10 kb	14	**61**	13	11
10-15 kb	1	9	**9**	13
> 15 kb	4	6	10	**18**

Table 2.8. Frequencies for the given range of median distance

Distance range	0-5 kb	5-10 kb	10-15 kb	> 15 kb
Given Os range	0.678	0.616	0.265	0.409
Given Sb range	0.597	0.616	0.281	0.474

Table 2.9. Lineages of genome rearrangement in the comparison between rice and sorghum gene pairs using banana as the outgroup

Categories	Convergent	Divergent	Unidirectional	Total
Conserved in both	6	6	10	22
Rearranged in both	79	83	159	321
Absent in both*	7	2	13	22
Conserved in rice only	2	4	7	13
Conserved in sorghum only	4	2	6	12
Absent in sorghum only*	1	1	0	2
Absent in rice only*	1	2	5	8
Total gene pairs studied	100	100	200	400
Misannotation	0	0	13	13
Unknown	0	2	1	3

* “Absent” means that at least one of the genes in Musa gene pair was lost in rice or sorghum.

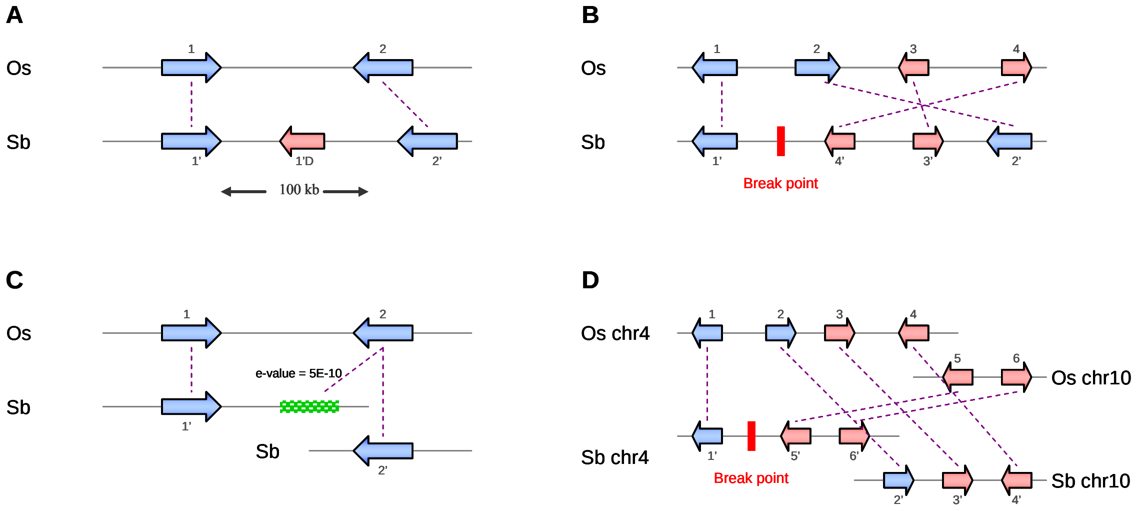

Figure 2.1. Possible types of genome rearrangement that might be detected. Arrows indicate genes and their predicted direction of transcription, and alphanumeric characters above or below arrows indicate the particular gene involved. Dotted lines connect orthologous genes. Genes 1 and 2 are a pair of adjacent genes randomly selected from the rice or sorghum genomes, while gene 1' and 2' are the orthologs of gene 1 and 2 in the other genome, respectively. These four genes are depicted in blue, while all other genes are shown in red. Green lines or arrows represent gene fragments or misannotated genes. Os = *Oryza sativa* and Sb = *Sorghum bicolor*. (A) Manual inspection of gene pair conservation and rearrangement. Gene 1'D denotes the adjacent downstream gene of gene 1'. (B) (C) (D) Depictions of different types of genome rearrangement, comprised of an inversion involving genes 2, 3, and 4 (B), a deletion in the sorghum genome (C), and a translocation of genes 2, 3, and 4 (D).

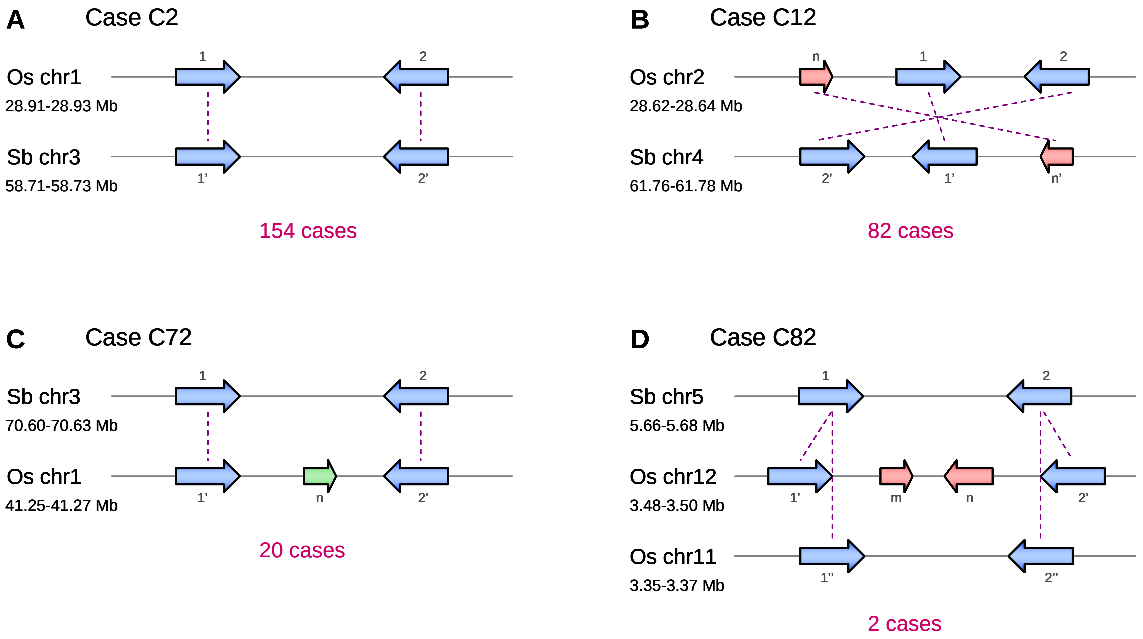

Figure 2.2. Schematic view, not to scale, of actual cases of gene pair conservation that were observed. (A) Basic pattern of gene pair conservation. (B) Conserved gene pair involved in segmental inversion event. (C) Gene pair conservation after excluding misannotation. (D) Gene pair conserved with paralogues rather than orthologues.

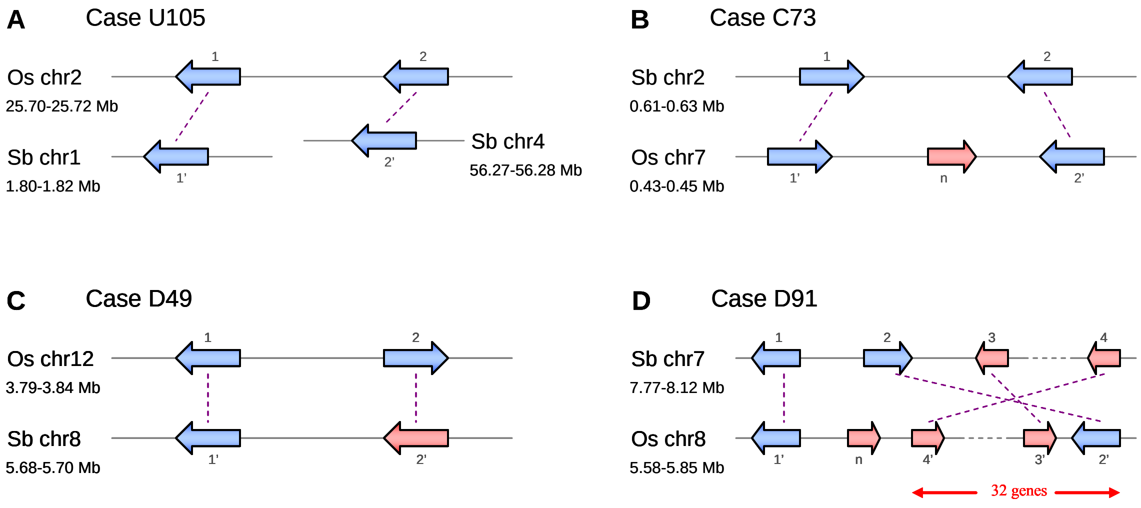

Figure 2.3. Actual cases of gene pair rearrangement observed in the rice versus sorghum comparison. (A) Basic pattern of gene pair rearrangement, in which two genes in rice are separate to different chromosomes in sorghum. (B) Gene pair rearrangement by gene insertion/deletion. (C) Gene pair rearrangement by single gene inversion. (D) Gene pair rearrangement by at least two events, including one gene insertion (n) and a segmental inversion involving 32 genes.

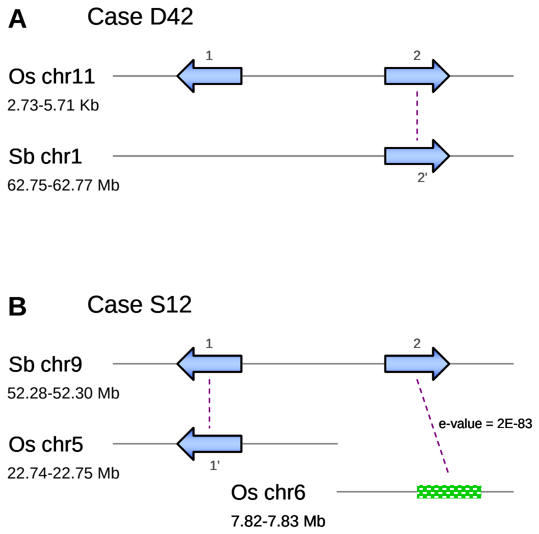

Figure 2.4. Schematic view detected cases of apparent gene gain and gene loss from either the rice or sorghum genomes. (A) Gene 1 apparently created in the lineage leading to rice. (B) Possible loss of gene 2 from the rice genome, but vestiges of that gene remain with an excellent e-value, but without any detected open reading frame or gene model arising from the annotation of this genome.

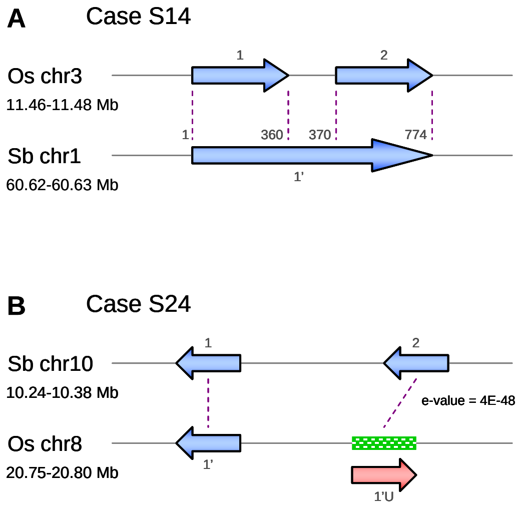

Figure 2.5. Various rare types of mis-annotation detected in the rice and sorghum genomes. The common types of genome mis-annotation, such as the characterization of TE genes as regular genes, were removed from this analysis in the first step of gene pair choice, and thus were not found in the eventual manual comparisons. (A) A single gene annotation in rice compared to a two annotation of the same gene(s) in sorghum. The numbers near the dotted line indicate the start and end of protein alignment against gene 1' in sorghum. (B) Misannotation of protein 1'U, which was translated with an incorrect reading frame of gene 2. Gene 1'U denotes the adjacent upstream gene of gene 1'. The e-value shown here is for BLASTN. Hence, a single gene inversion, without manual annotation, would have been characterized as a gene insertion or deletion.

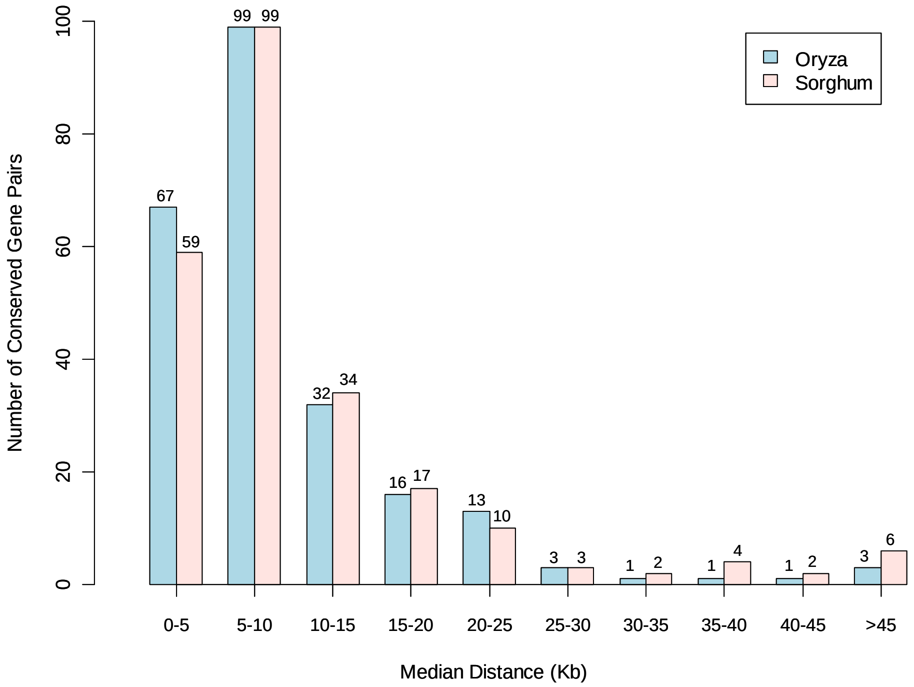

Figure 2.6. Distribution of median distance for conserved gene pairs in rice and sorghum. The x axis indicates distance ranges in kilobase pairs, while the y axis indicates the number of gene pairs in each distance range. The blue and pink boxes denote the median distance distribution in rice and sorghum, respectively.

CHAPTER 3

THE NATURES, FREQUENCIES AND LINEAGE-SPECIFICITIES OF GENOME REARRANGEMENTS ACROSS MULTIPLE PLANT SPECIES[1]

[1] Feng, L. and J.L. Bennetzen. To be submitted to *PLOS Genetics.*

Abstract

In plants, comparative genome analysis has lacked precision and reproducibility due to a lack of robust annotation techniques and a controlled vocabulary. Gene pair analysis has recently been proposed as a solution to both of these problems. A study of gene pair stability and rearrangement was conducted using the genome sequences of two monocots (Setaria and Brachypodium) and two dicots (Arabidopsis and Medicago). As previously shown in rice and sorghum, non-paralogous gene pairs that are > 1 kb apart show no bias toward a convergent, divergent or unidirectional orientation, but genes < 1 kb are greatly biased toward a convergent orientation. The results indicate that more than 50% of adjacent gene pairs have been preserved in structure over the last ~50 million years in the two grasses, while only ~7% of gene pairs have retained their structure since the divergence of Arabidopsis and Medicago ~92 million years ago. The frequency of gene pair retention in closely linked convergent gene pairs was found to be significantly higher than that of divergent and unidirectional gene pairs in sorghum and Setaria. The phylogenetic placement of the multiple analyzed species often allowed determination of the lineage in which particularly rearrangements occurred and their precise nature in the grasses. Taken together, all of these analyses indicated that sorghum is the most stable among the grass lineages investigated. The two dicot genomes, Medicago and Arabidopsis, have exhibited a much higher relative instability per unit time than any of these grasses, but we lacked enough phylogenetic sampling to determine the relative contributions of each lineage. The results also demonstrated that movement of one gene of the pair to another chromosome in the other compared species is the most frequent type of rearrangement in all species comparisons.

Introduction

The primary processes responsible for the rapid evolution of flowering plant genome structure are polyploidy, transposable element amplification, chromosome breakage, and DNA removal by unequal homologous recombination and illegitimate recombination. These are not mutually exclusive mechanisms, as shown for instance by the ability of TEs to occasionally break chromosomes (McClintock 1947) and the frequent DNA loss through illegitimate recombination associated with double strand break repair (Kirik et al. 2000). The differences in frequencies, specificities, and amplitudes of these rapid and dynamic genome rearrangement processes have been responsible for the great variety in plant genome structure. (Bennetzen and Chen 2008). The comparative circular map of the grasses, also known as the crop circle, has provided insight into the major rearrangements and the timing of large rearrangement events within the grass species after the divergence from a common ancestor more than 50 million years ago (Mya) (Moore et al. 1995; Devos 2005). The comparative map indicates that large genomic rearrangements are relatively rare and are unevenly distributed across lineages. Although there are many rearrangements in local gene order that differentiate closely related species, gene content is much less variable (Bennetzen 2007).

Orthologous sequence comparisons in different time frames have the potential to reveal the rates and mechanisms that are responsible for the disruption of colinearity. For instance, a sequence comparison of the *adh1*-orthologous regions for maize and sorghum, two panicoid grasses that shared a common ancestor ~12 Mya, indicated that nested LTR retrotransposon insertion is the major reason for the more than three-fold size difference between the two genomes. (SanMiguel et al. 1996; Tikhonov et al. 1999). In contrast, in the comparison of the *adh1*-orthologous regions between sorghum and sugarcane, strong homology of both non-coding regions and colinear genes were observed, indicating that these two lineages have been very stable over the shorter time frame (~8 Myr) since their divergence. In comparisons of distantly related species, lower degrees of colinearity across orthologous regions were both expected and observed. For instance, the divergence date between banana and rice lineages is ~110 Mya, but more than 50% of the banana genes were found to be non-colinear in the comparison to rice (Lescot et al. 2008). This greater degree of rearrangement is not exclusively a function of time, because some lineages like the one leading to banana (Chapter 2 of the dissertation) appear to have been unusually unstable. With even more distant comparison to the eudicots, only rare genic colinearity was observed at either whole genome scales or local genome scales (Devos et al. 1999; Liu et al. 2001).

We currently lack a vocabulary to precisely describe the degree of conservation of genic content and colinearity between any two species. Possible subtle patterns and natures of the genic rearrangements within lineages are also unknown. At the level of the genetic map, it is well known that some lineages, such as pearl millet and maize, are more unstable than other lineages, such as rice and foxtail millet (Devos et al. 2000; Ramakrishna et al. 2002; Bennetzen and Ma 2003; Ilic et al. 2003). It is also known that types of genome rearrangement vary greatly in their frequency, as exemplified by the great rarity of translocations compared to the high frequencies of tandem duplication or gene deletion. In order to understand the reasons for these differences, and to quantify the retention and rearrangement of gene pair linkage, as well as the frequencies of different types of rearrangements, a random sampling approach was developed to manually inspect the retention and rearrangement of gene pair linkage between rice and sorghum (see Chapter 2). This analysis allowed the first definitive determination of the frequencies of different types of genome rearrangement that affect gene content or order. Here, this approach is applied to additional flowering plants: Setaria, Brachypodium, Arabidopsis, and Medicago. The results indicate high levels of gene pair conversation in the grasses examined, but much less in the two investigated dicots. For the grasses, the frequency of gene pair retention was found to be significantly higher in closely linked genes, and was especially high in the convergent gene pairs separated by < 1 kb. Taken together, these results indicate genomes with different rates of rearrangement, and with significant specificity in the types of rearrangement that can accumulate over time.

Materials and Methods

Experimental System

In this chapter, the gene pair comparison approach has been extended to more plant species, including foxtail millet (*Setaria italica*), Brachypodium (*Brachypodium distachyon*), Arabidopsis (*Arabidopsis thaliana*), and Medicago (*Medicago truncatula*). Date palm (*Phoenix dactylifera*) was used as an outgroup, along with banana (*Musa acuminata ssp. malaccensis*), to help determine the lineage specificity of genome rearrangements, including those previously seen between rice (*Oryza sativa ssp. japonica*) and sorghum (*Sorghum bicolor (L.) Moench*). Genome sequences and annotation data for rice build 5 were downloaded from the Rice Annotation Project Database (RAP-DB, http://rapdb.dna.affrc.go.jp/download/index.html), for sorghum v1.0 were downloaded from DOE Joint Genome Institute (JGI, http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html), for foxtail millet were downloaded from Phytozome 8.0 (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Sitalica/), for Brachypodium version 1.0 were downloaded from ftp://brachypodium.org/, for banana version 1 were downloaded from the Banana Genome Hub centralized databases (http://banana-genome.cirad.fr/), for date palm version 3 were downloaded from (http://qatar-weill.cornell.edu/research/datepalm/index.html), for Arabidopsis TAIR 8 were downloaded from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/), and for Medicago Mt2.0 were downloaded from http://www.medicago.org/ (Table 3.1).

Non-paralogous Gene Pair Identification

Gene pair selection was performed as described in Chapter 2, with the following modifications. The annotation of the Brachypodium genome is somewhat for incomplete than for many other genomes. The 25,532 annotated genes have been divided into 6 different classes based on the confidence of the annotation by the International Brachypodium Initiative. Classes 0, 1, and 2 are defined as predicted genes that have a low level of certainty as real genes. Classes 3 and 4 are defined as true gene loci with an incomplete model or minor problems with the predicted gene structure, while class 5 contains only loci with complete gene models. In this study, only the 18,540 genes in classes 3, 4 and 5 of Brachypodium were used for further analyses. The current genome assembly of date palm is composed of 57,277 scaffolds, and no attempt was made to create large pseudomolecules. In addition, the number of sequencing gaps is high, providing 76,584 gaps across the scaffolds. Hence, all adjacent genes on the same scaffold with no gaps between them are defined as the only usable gene pairs (for our purpose) in date palm. For each investigated genome, the number of convergent (→←), divergent (←→), and unidirectional (→→/←←) gene pairs was determined, together with the intergenic distance, by means of Perl scripts. Pearson's Chi-Square test was conducted to test whether gene pair orientation is random (i.e., convergent, divergent, and unidirectional genes pairs in a ratio of 1:1:2) by using R.

Manual Inspection of Gene Pair Retention and Genome Rearrangement Types

Rice, sorghum, and Setaria were compared to each other to identify the retention and rearrangement of gene pair linkage. Rice was also compared to Brachypodium, but no Brachypodium:Sorghum or Brachypodium:Setaria comparison was attempted. Arabidopsis and Medicago gene pairs were also randomly selected and compared by the same approach. Four hundred non-paralogous gene pairs were selected from the whole genome dataset of rice-sorghum, rice-Setaria, sorghum-Setaria, rice-Brachypodium, and Arabidopsis-Medicago comparisons, composed of 100 convergent, 100 divergent, and 200 unidirectional gene pairs. Half of the gene pairs were initially selected from one genome and half were selected from the other species for each type of gene pairs. When the same species was used for different pairwise comparisons among rice, sorghum and Setaria, the same set of 200 gene pairs was selected from that species. For instance, rice was compared to one species in this study (Setaria) and to one species in another study (sorghum, see Chapter 2), using the same 200 initial rice gene pairs. The Brachypodium-rice comparison used a different 200 rice gene pairs. For these previously used gene pairs and the new ones selected (from rice, Brachypodium, Setaria, Arabidopsis and Medicago), the stringency rules used to confirm that both genes in the gene pair were real genes and not annotation artifacts were the same criteria as described in Chapter 2. The identification process for the frequencies and types of gene pair rearrangement are also provided in Chapter 2.

Lineage Specificity of Genome Rearrangements

The use of date palm as an outgroup was similar to the approach using banana as an outgroup to understand the lineage specificity of genome rearrangements between rice and sorghum described in Chapter 2, In addition, 400 non-paralogous gene pairs with clear predicted function were selected from date palm: 100 convergent, 100 divergent, and 200 unidirectional gene pairs. As before, genes annotated as non-coding genes, pseudogenes, unanchored genes, hypothetical genes, and transposon-related genes were excluded. The numbers of date palm gene pairs conserved in both rice and sorghum, rearranged in both rice and sorghum, conserved in rice but rearranged in sorghum, and conserved in sorghum but rearranged in rice, were identified to investigate relative rates of rearrangement across these lineages. As we used the same 200 gene pairs chosen from rice for both rice-sorghum and rice-Setaria comparisons, the numbers of rice gene pairs conserved in both sorghum and Setaria, rearranged in both sorghum and Setaria, conserved in sorghum but rearranged in Setaria, and conserved in Setaria but rearranged in rice, were also scored to identify the lineages of specific rearrangements.

The Relationship between Gene Size and Genome Rearrangement

Inserted genes were characterized by pairwise comparisons among rice, sorghum, and Setaria, and by considering the phylogeny of the three organisms (Figure 3.1). For instance, a sorghum gene was considered an inserted gene if the gene pair is conserved in rice and Setaria, but is interrupted by the gene inserted between the same gene pair in sorghum. Inserted gene size was measured using the coding sequences (CDS) length in base pairs. The Kolmogorov-Smirnov test was used to test normality. The Student's t-test was used to test the significance of the differences in gene size between inserted genes and all genes in each organism.

Results

Gene Pair Organization across Multiple Plant Species

There are ~35,000 annotated genes in the current releases of rice (IRGSP Build 5), sorghum (JGI 1.0), and Setaria (Phytozome 8.0), while the number of annotated genes for Brachypodium and date palm are 25,532 and 28,889, respectively. The Arabidopsis genome sequence still lacks full coverage of centromeric regions, but the parts that have been sequenced were finished at a very high level of accuracy in both assembly and annotation. The sequenced part of the Arabidopsis genome has 32,423 annotated genes (Table 3.1). After excluding RNA genes, pseudogenes, unanchored genes, hypothetical genes, known TE-related genes from the gene set, and discarding gap-containing gene pairs and overlapping genes, a total of 30,903, 22,857, 27,802, 13,285, 4341, 21,346, and 26,941 gene pairs were identified in rice, sorghum, Setaria, Brachypodium, date palm, Medicago and Arabidopsis, respectively. The numbers and percentages of each type of adjacent gene pairs are shown in Table 3.2. The number of gene pairs in Brachypodium and date palm are low because only genes in class 3, 4, and 5 in the Brachypodium genome annotation and only genes on the same (usually short) date palm scaffold were used to identify gene pairs. The percentages of convergent and divergent gene pairs in most species are slightly lower than 25%, while the proportion of unidirectional gene pairs is slightly higher than 50%. If the gene orientation is random, the ratio of convergent, divergent, and unidirectional gene pairs should be 25% : 25% : 50%.

A non-paralogous gene pair in this study refers to a pair of adjacent non-duplicated genes. Because tandemly duplicated genes can confuse the discrimination of orthologues from paralogues, and because rearrangements of these genes are both frequent and primarily caused by the well-characterized process of unequal homologous recombination, tandem gene duplications were excluded from this study. After removing the duplicated gene pairs by using the BLAST2Seq program, the percentages of the three types of gene pairs in all organisms, except date palm and Medicago, are very close to 25% convergent, 25% divergent, and 50% unidirectional gene pairs. In date palm, the frequency of divergent gene pairs (16.5%) is substantially lower than expected (25%). Medicago, in contrast, is deficient in both divergent and convergent gene pairs, even after correction for the frequent unidirectional tandem gene families. Using Pearson's Chi-Square test to investigate whether gene pair orientation is random for both the adjacent gene pairs and non-paralogous adjacent gene pairs in flowering plant genomes led to the observation that the ratio of the three possible orientation does not fit a 25% : 25% : 50% model ($P < 0.05$), but the orientation of non-paralogous adjacent gene pairs is closer to random than that of all adjacent gene pairs. Hence, gene pair duplication is a major factor in determining non-random properties in angiosperm genome orientation.

The frequencies of convergent closely linked gene pairs (intergenic distance < 1 kb) are significantly higher (> 35%) in most species, but the frequencies of divergent gene pairs are less than 21% (Table 3.3). This result is consistent with my previous study in rice and sorghum and previous investigations in *Arabidopsis* and *Populus* (Krom and Ramakrishna 2008), but is in contrast to the similar studies in mammalian genomes, where an excess of divergently transcribed genes have been observed in tightly linked gene pairs (Adachi and Lieber 2002; Li et al. 2006; Franck et al. 2008).

Gene pair Retention and Rearrangement during the Evolutionary Processes

Previous manual inspection for a randomly selected subset of adjacent gene pairs in rice and sorghum indicated that 59% of gene pairs have been conserved in content and organization (see Chapter 2). This frequency of gene pair retention is significantly higher than the 25.4% predicted by automatic high-throughput characterization (Krom and Ramakrishna 2008; Liu and Han 2009; Krom and Ramakrishna 2010). As precise genome comparison is challenging because of polyploidization, chromosomal duplication and high levels of gene mis-annotation, the same sampling and manual inspection approaches were used on Setaria, Brachypodium, Arabidopsis and Medicago, none of which have not undergone a polyploidization for more than 25 million years.

Precise genome comparison between rice and Setaria indicated that 57% of convergent, 46% of divergent, and 47.5% of unidirectional gene pairs are conserved between the two grasses, and only 8 out of the 400 randomly selected gene pairs contain potential gene loss/gain loci (Table 3.4). To test whether the frequencies of gene pair conservation and rearrangement show significant differences among each type of gene pairs, Pearson's Chi-Square test was performed and indicated that there is no significant difference of conservation frequencies among convergent, divergent, and unidirectional gene pairs (*P* = 0.173). For this pair of species, the ratio of gene pair conservation to rearrangement is ~1 : 1. The gene pair comparison between sorghum and Setaria suggest that 76% of convergent, 63% of divergent, and 57% of unidirectional gene pairs are conserved between the two grasses, and only 3 out of the 400 randomly selected gene pairs contain potential gene loss/gain loci (Table 3.5). The Chi-Square test suggests that the ratio of gene pair retention to rearrangement in convergent gene pairs are significantly higher than that of divergent and unidirectional gene pairs (*P* = 0.021). About 76% of convergent gene pairs are conserved in sorghum and Setaria, while ~59% of divergent and unidirectional gene pairs are retained.

The frequencies of the retention/rearrangement of closely linked gene pairs among rice, sorghum, and Setaria are different. For instance, in the rice to sorghum comparison, there is an approximately 2-fold difference of the conservation frequencies between convergent compared to divergent plus unidirectional gene pairs, but this difference is not statistically different because of the small sample size (*P* = 0.496). However, the overall frequencies of conserved closely linked gene pairs (74%) are significantly higher than that of rearranged gene pairs (26%, Table 2.5), and significantly higher than conversation of all gene pairs in rice versus sorghum (59%, Chapter 2). In rice and Setaria comparison, the retention frequency of divergent closely linked gene pair (27%) is significantly lower than that of the convergent and unidirectional gene pairs

(67%, Table 3.6). For the closely related sorghum and Setaria comparison, there is significant difference of the retention and rearrangement frequencies among three types of gene pairs (*P* = 0.004). 96% of convergent, 78% of divergent, and 55% of unidirectional closely linked gene pairs are conserved (Table 3.7).

The comparison between rice and Brachypodium indicates that 58% of convergent, 51% of divergent, and 56% of unidirectional gene pairs have been conserved since the divergence of rice and Brachypodium lineages (Table 3.8). Statistically, there is no significant difference of conservation frequencies among convergent, divergent, and unidirectional gene pairs (*P* = 0.865), and the overall ratio of gene pair conservation to rearrangement is 55:45. Comparing the three grasses to rice, the percent retained gene pair composition and structure is 59%, 55%, and 50% for sorghum, Brachypodium and Setaria. Brachypodium is a closer relative to rice than either panicoid grass, but the sorghum and Setaria lineages diverged from the rice lineage at exactly the same time. Hence, sorghum appears to have been a more stable genome lineage than Setaria in the time since these two lineages diverged.

The Medicago genome was sequenced by a BAC-by-BAC approach that did not use Arabidopsis as an assembly guide and did not lead to a great number of questionable assemblies in either contig or scaffold order. Genomic comparison between Arabidopsis and Medicago indicated that only 5% of convergent, 4% of divergent, and 8% of unidirectional gene pairs are conserved, while 21% of convergent, 25% of divergent, and 26% of unidirectional gene pairs contain apparent gene loss/gain loci (Table 3.9). The estimated divergence time between Arabidopsis and Medicago is ~92 Mya (Gandolfo et al. 1998; Grant et al. 2000). This is less than 2 fold longer than the divergence time (~50 million years) for rice and sorghum, two species that show ~59% gene pair conservation. At this rate of gene pair disruption, 92 million years would yield a predicted ~40% conservation, rather than the ~7% that was observed.

Lineage Specificity of Genomic Instability

Comparison to a lineage outside the Poales can help determine the lineage of specific genome rearrangements observed in the grass pairwise comparisons. In Chapter 2, I did not observe the lineage specificity of rearrangement frequencies in rice and sorghum by using the non-grass monocot, banana. In this chapter, another non-grass monocot, date palm, is employed as an outgroup for this same purpose. There are 57,277 scaffolds and 76,584 sequencing gaps in the date palm genome assembly, so the number of identifiable adjacent gene pairs in date palm is relatively few. To minimize the influence of genome assembly, 400 non-paralogous adjacent gene pairs were randomly selected from date palm, consisting of 100 convergent, 100 divergent, and 200 unidirectional gene pairs, in which each gene must have a clear predicted function. Manual inspection indicated that 329 gene pairs were rearranged in both rice and sorghum compared, while 35 were conserved in both grasses (Table 3.10). A total of 31 out of 400 date palm gene pairs have different structure and/or composition in rice and sorghum, of which 8 of them are conserved in rice but rearranged in sorghum, 16 of them are conserved in sorghum but rearranged in rice, 2 of them are present in rice but absent in sorghum, and 5 of them are present in sorghum but absent in rice. From this analysis, gene pairs conserved in sorghum but rearranged in rice are two-fold more frequent than gene pairs conserved in rice but rearranged in sorghum. This suggests that the rice lineage has been more unstable than the sorghum lineage.

Almost one quarter (96) of the gene pairs selected in date palm appeared to be mis-annotation products, so these mis-annotation cases were replaced with other randomly selected gene pairs from date palm. Most (79/96) of these mis-annotations from date palm appeared to be the same sequence to be called a single gene in both rice and sorghum but called two adjacent unidirectional genes in date palm.

Because the same 200 gene pairs were chosen from rice for both rice-sorghum and rice-Setaria comparisons, the lineages associated with these gene pairs in sorghum and Setaria were also characterized (Table 3.11) by using rice as the outgroup for Setaria-sorghum comparisons. A total of 72 out of 200 rice gene pairs were rearranged in both sorghum and Setaria, while 85 of them are conserved in both sorghum and Setaria. The remaining 43 rice gene pairs are comprised of 18 that are conserved in sorghum but rearranged in Setaria, 17 that are conserved in Setaria but rearranged in sorghum, and 2 that have both genes present in Setaria but one absent in sorghum. These results indicate very similar rates of genome instability in sorghum and Setaria lineages since their descent from a common ancestor.

The Relationship between Gene Size and Genome Rearrangements

One might expect that insertions of genes into a region, by whatever mechanism, could be affected by the size of the gene that is inserted. To test this possibility, gene insertion events that disrupted gene pairs in rice, sorghum, and Setaria genomes were identified for comparison to average gene sizes in these same species. By considering the phylogeny of the three species, 13, 12, and 16 inserted genes were identified in rice, sorghum and Setaria, respectively (Table 3.12). The sizes of the inserted genes exhibited no significant difference from the average gene sizes in these species (Table 3.13).

Discussion

The first manual curation of randomly selected gene pairs was the rice-sorghum comparison described in Chapter 2. This Chapter extends this approach to foxtail millet, Brachypodium, Arabidopsis and Medicago. There are several reasons for choosing these species for investigation. First, the great advantage of rice, sorghum, foxtail millet, and Brachypodium is that they have not undergone any polyploidization after the divergence of grasses from a common ancestor, ~50 Mya. Second, foxtail millet diverged from sorghum approximately 28 Mya, and from rice about 50 Mya (Gaut 2002), thus providing two different time frames for characterization of genome rearrangement. Comparison of the genetic map for rice and foxtail millet indicate that their genomes are highly collinear, and thus have been very stable over the past 50 million years at this level of comparison (Devos et al. 1998). Hence, genome alignments were facile, allowing specific rearrangements to be precisely characterized. Third, *Brachypodium distachyon* was the first member of the Pooideae subfamily of grasses to have its genome sequenced (Vogel et al. 2010). Genome comparison of Brachypodium, rice, sorghum, and foxtail millet allow us to trace the evolutionary history of the Poaceae family across a broad diversity of grasses. Finally, Arabidopsis is widely used as the model organism for studying all plants, and Medicago has been adopted as a model plant for studying legumes. The assembly of the Medicago genome sequence was by a BAC-by-BAC approach that did not use Arabidopsis as an assembly guide and did not lead to a great number of questionable assemblies in either contigs (contiguous sequences) or scaffold order. The same cannot be said for any other dicot genome sequence, other than that of Arabidopsis and perhaps lotus (*Lotus japonicus*) (Sato et al. 2008). The gene pair comparison between Arabidopsis and Medicago yields our first insights into dicot genome evolution at this level of investigation.

As in the rice and sorghum genome comparisons (Chapter 2), a focus was placed on non-paralogous gene pairs in this study, because a substantial proportion of adjacent genes are gene duplicates that can confuse orthology determination. After excluding hypothetical genes, transposon-related genes, overlapping genes, and tandemly duplicated genes, the orientations of gene pairs are close to random in most of the plant species investigated. The exceptions to this rule were Medicago and date palm. Although the Medicago genome is relatively far from complete, with sequencing emphasis placed on gene-rich regions (Cannon et al. 2006; Young et al. 2011), it is not clear why this would yield an excess of unidirectional genes. The date palm genome, because of its assembly from a relatively sparse dataset, contains relatively small number of large contigs or scaffolds. It is not clear why this would affect perceived gene order, somehow giving rise to an artifactual shortage of divergent genes. Hence, it is possible that these two plant species have evolved certain types of gene pair interaction that are not as common in the other species that we have analyzed. From such a comparative analysis of any process across a breadth swath of phylogenetically-placed species, an unusual property in one lineage indicates an exciting opportunity to search for an evolutionary switch that has led to the evolution of novel regulatory or functional potential.

For non-paralogous gene pairs with intergenic distances of < 1 kb, the frequencies of convergent gene pairs are significantly higher than predicted by chance in all species investigated, as seen in the previous rice-sorghum comparison (Chapter 2). It is not known why this orientation is represented almost 2-fold more frequently than expected by chance, but a functional interaction, perhaps based on the opportunity for antisense RNA regulation, seems a logical possibility.

Many existing studies of gene pair conservation relied completely on high throughput characterization, and suggested a very low percentage of gene pair conservation between rice and sorghum (Krom and Ramakrishna 2008; Liu and Han 2009; Krom and Ramakrishna 2010). However, the manual inspection for a randomly selected subset of adjacent gene pairs in rice and sorghum suggests that 59% of gene pairs have been conserved after the divergence of rice and sorghum ~50-70 Mya. With the same approach, rice-Setaria, sorghum-Setaria, rice-Brachypodium, and Arabidopsis-Medicago were compared. In general, the frequencies of gene pair retention and rearrangement were found to correlate with phylogenetic relationships. The estimated divergence time of the lineages leading to the two dicots Arabidopsis and Medicago is ~92 Mya, and only 7% of gene pairs were still intact. This rate of rearrangement is much higher than a simple extrapolation would predict in 92 million years for even the fastest grass rearrangement rate that we detected. It seems likely that this higher instability is explainable by a combination of two factors. First, the recent polyploidy (~24 Mya) in the Arabidopsis lineage (Vision et al. 2000; Blanc et al. 2003) would have relieved selective constraints on many genes, so they could be lost or moved even if this affected function. Many of these genes are likely to be lost through a fractionation process (Freeling 2008), and the lost of many of the genes from earlier polyploidies has been very well documented in Arabidopsis (Thomas et al. 2007). Second, this same polyploidy may have given rise to some confusion during the annotation process in whether orthologues were truly being compared. This artifactual issue could be significant, especially when combined with the fact that some genic regions containing the actual orthologues sought had not yet been sequenced in Medicago.

In contrast to the dicot instability observed, > 50% of gene pairs were observed to be conserved in rice-sorghum, rice-Setaria, sorghum-Setaria, and rice-Brachypodium comparisons. Because sorghum and Setaria diverged from a common ancestor only ~28 Mya, the overall conservation level of 63% was not surprising. The fact that quite different levels of conservation was seen for the three different orientations in sorghum and Setaria is surprising (76% for convergent, 63% for divergent, and 57% for unidirectional), because we have not seen this in any other comparison. The higher level of conservation of the convergent gene pairs in sorghum, compared to the divergent and unidirectional pairs, is statistically significant ($P = 0.003$).

When the retention and rearrangement of gene pairs with intergenic distances < 1 kb were investigated for the sorghum-Setaria pairwise analysis, it was observed that the frequency of gene pair retention in convergent gene pairs (96%) was significantly higher than that of divergent (78%) and unidirectional gene pairs (55%). This statistically significant bias was also observed in the other grass comparisons. One expects that the bias in the frequency of convergent pairs at short intergenic distances has a functional (probably regulatory) explanation, so it makes perfect sense that there would be strong selection against breaking up this evolved function by gene pair rearrangement. As with all other classes of mutation, the great majority of gene pair rearrangements will be negative or neutral, so we expect that our analysis uncovers mostly neutral changes, plus the very rare positive changes or negative changes that have not yet been removed by selection.

With appropriate outgroups, it was possible to determine whether several gene gain/loss rearrangements were actually caused by gene insertion or gene deletion. A total of 41 inserted genes were found across rice, Setaria and sorghum. It seems obvious that any process that inserts genes, like transposon vectoring or insertion during double strand break repair (Kirik et al. 2000) would be size limited, and that this might be detected by observing that our 41 genes over-represented small genes. However, no significant difference in gene size was observed between inserted genes and average gene size in any of these species. In fact, even a statistically insignificant trend towards smaller genes was not observed. We know of only mechanism for single gene movement that is not terribly size limited, and that result is unequal homologous recombination.

Comparative genetics map in grasses indicate that different lineages have different levels of genomic instability. For example, foxtail millet has more similar genetic map to that of rice (last shared ancestor ~50 Mya) than it does to pearl millet, despite their divergence from a common ancestor ~20 Mya (Devos et al. 1998; Devos 2005). I did not observe lineage-specific differences in rearrangement frequencies in rice and sorghum by using banana as an outgroup in Chapter 2, but this study indicated that the rice lineage is more unstable than the sorghum lineage when using date palm as the outgroup. We do not know the reason for this inconsistency, but suspect that it may be caused by some non-randomness in the analysis associated with the very small numbers of gene pair retentions (compared to the grasses) seen in both the date palm and banana lineages. Further studies are needed with larger data sets and a more closely related outgroup to the grasses in order to resolve this issue.

This study, manually investigating randomly chosen gene pairs, provides insights into the natures, lineages and frequencies of different patterns of plant genome evolution. We hope that this approach becomes a routine tool that is extensively utilized in the fields of comparative genomics and molecular evolution. Currently, we do not know which of the many possible mechanisms of instability are responsible for the different types of genome rearrangements observed in any case or in any lineage. Future research should be performed to use sequence data from closely linked species, such as the cultivated rice and wild rice (Huang et al. 2012), to investigate recent genome rearrangements. If very recent events can be detected, then any possible legacies of the structures associated with the genome rearrangement may not have yet been obscured by the very rapid processes for DNA removal in angiosperms (Ma et al. 2004). If these legacies are still visible, then we hope to be able to gain mechanistic insights that might also differentiate the evolving properties of different plant lineages.

Table 3.1 Genome properties for the plant species investigated

Species	Number of Chromosomes	Genome Size (Mb)	Genome Assembly	Annotated Genes
Oryza sativa	12	~ 390	IRGSP Build 5	34,780
Sorghum bicolor	10	~750	JGI 1.0	34,496
Setaria italica	9	~515	Phytozome 8.0	35,471
Brachypodium distachyon	5	~300	Version 1.0	25,532
Musa acuminata	11	~520	Version 1	36,542
Phoenix dactylifera	18	~650	Version 3	28,889
Medicago truncatula	8	~500	Mt 2.0	38,844
Arabidopsis thaliana	5	~140	TAIR 8	32,423

Table 3.2. Numbers and percentages of adjacent and non-paralogous adjacent gene pairs in each species

	Adjacent gene pairs				Non-paralogous adjacent gene pairs			
Species	Total	Convergent	Divergent	Unidirectional	Total	Convergent	Divergent	Unidirectional
Oryza	30903	7549 (24.4%)	7360 (23.8%)	15,994 (51.8%)	28567	7368 (25.8%)	7171 (25.1%)	14,028 (49.1%)
Sorghum	22857	5427 (23.7%)	5029 (22.0%)	12,401 (54.3%)	20384	5193 (25.5%)	4813 (23.6%)	10,378 (50.9%)
Setaria	27802	6716 (24.2%)	6142 (22.1%)	14,944 (53.7%)	25160	6508 (25.9%)	5929 (23.6%)	12,723 (50.5%)
Brachypodium	13285	3223 (24.3%)	3124 (23.5%)	6938 (52.2%)	12146	3092 (25.5%)	2977 (24.5%)	6077 (50.0%)
Musa	21177	5235 (24.7%)	4866 (23.0%)	11,076 (52.3%)	20570	5173 (25.1%)	4783 (23.3%)	10,614 (51.6%)
Phoenix	4341	1279 (29.5%)	717 (16.5%)	2345 (54.0%)	4158	1253 (30.1%)	702 (16.9%)	2203 (53.0%)
Medicago	21346	4257 (19.9%)	4145 (19.4%)	12,944 (60.7%)	18868	4076 (21.6%)	3961 (21.0%)	10,831 (57.4%)
Arabidopsis	26941	6412 (23.8%)	6409 (23.8%)	14,120 (52.4%)	24391	6281 (25.8%)	6243 (25.6%)	11,867 (48.6%)

Table 3.3. Numbers and percentages of non-paralogous gene pairs < 1 kb apart that were available to this analysis

Species	Total	Convergent	Divergent	Unidirectional
Oryza	2919	1352 (46.3%)	597 (20.5%)	970 (33.2%)
Sorghum	2396	1122 (46.8%)	396 (16.5%)	878 (36.7%)
Setaria	4657	1765 (37.9%)	837 (18.0%)	2055 (44.1%)
Brachypodium	1887.	931 (49.3%)	299 (15.9%)	657 (34.8%)
Musa	3930	1315 (33.4%)	541 (13.8%)	2074 (52.8%)
Phoenix	641	188 (29.3%)	98 (15.3%)	355 (55.4%)
Medicago	4251	950 (22.4%)	349 (8.2%)	2952 (69.4%)
Arabidopsis	10,419	4156 (39.9%)	1807 (17.3%)	4456 (42.8%)

Table 3.4. Results for 400 randomly selected gene pairs in rice and foxtail millet

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	57	46	95
Rearranged gene pairs	43	50	101
Gene pairs with gene loss/gain	0	4	4
Total	100	100	200
Misannotation	3	0	4
Unknown	6	2	9

Table 3.5. Results for 400 randomly selected gene pairs in sorghum and foxtail millet

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	76	63	114
Rearranged gene pairs	23	37	84
Gene pairs with gene loss/gain	1	0	2
Total	100	100	200
Misannotation	0	0	7
Unknown	10	4	11

Table 3.6. Results for randomly selected gene pairs < 1 kb apart in rice and foxtail millet

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	16	3	10
Rearranged gene pairs	7	6	6
Gene pairs with gene loss/gain	0	2	0
Total	23	11	16

Table 3.7. Results for randomly selected gene pairs < 1 kb apart in sorghum and foxtail millet

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	24	7	12
Rearranged gene pairs	1	2	10
Gene pairs with gene loss/gain	0	0	0
Total	25	9	22

Table 3.8. Results for 400 randomly selected gene pairs in rice and Brachypodium

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	58	51	112
Rearranged gene pairs	40	46	82
Gene pairs with gene loss/gain	2	3	6
Total	100	100	200
Misannotation	0	1	0
Unknown	7	1	4

Table 3.9. Results for 400 randomly selected gene pairs in Arabidopsis and Medicago

Categories	Convergent	Divergent	Unidirectional
Conserved gene pairs	5	4	16
Rearranged gene pairs	74	71	132
Gene pairs with gene loss/gain	21	25	52
Total	100	100	200
Misannotation	7	6	20
Unknown	2	2	4

Table 3.10. Lineage specificity of genome rearrangements between rice and sorghum, using date palm as the outgroup

Categories	Convergent	Divergent	Unidirectional	Total
Conserved in both	11	6	18	35
Rearranged in both	79	88	162	329
Absent in both	0	1	4	5
Conserved in rice only	2	2	4	8
Conserved in sorghum only	8	1	7	16
Absent in sorghum only	0	0	2	2
Absent in rice only	0	2	3	5
Total gene pairs studied	100	100	200	400
Misannotation	6	2	88	96
Unknown	0	1	13	14

* “Absent” means that at least one of the genes in date palm gene pair was lost in rice or sorghum.

Table 3.11. Lineage specificity of genome rearrangements in sorghum versus Setaria using rice as the outgroup

Categories	Convergent	Divergent	Unidirectional	Total
Conserved in both	22	21	42	85
Rearranged in both	17	16	39	72
Absent in both	0	3	3	6
Conserved in sorghum only	6	5	7	18
Conserved in Setaria only	5	5	7	17
Absent in Setaria only	0	0	0	0
Absent in sorghum only	0	0	2	2
Total gene pairs studied	50	50	100	200
Misannotation	3	0	7	10
Unknown	6	6	5	17

* "Absent" means that at least one of the genes in rice gene pair was lost in sorghum or Setaria.

Table 3.12. Inserted genes and their sizes in rice, sorghum, and Setaria

Oryza	Gene Size (bp)	Sorghum	Gene Size (bp)	Setaria	Gene Size (bp)
Os07g0108100	594	Sb03g009020	2680	Si014374m	2507
Os01g0269900	2171	Sb10g000410	15,545	Si010022m	6846
Os04g0606000	1401	Sb05g003860	396	Si015086m	882
Os08g0196700	2069	Sb03g033240	7279	Si016621m	2328
Os09g0507300	2139	Sb01g047440	2349	Si032824m	2833
Os11g0107266	1011	Sb04g026220	754	Si022295m	2890
Os01g0158100	804	Sb05g027730	2125	Si003766m	2394
Os07g0113600	3609	Sb02g042470	2636	Si004415m	699
Os11g0124500	456	Sb06g032440	2336	Si027804m	2033
Os04g0486600	3291	Sb03g032740	5175	Si000276m	2634
Os07g0576100	4161	Sb03g043720	720	Si014612m	426
Os03g0317000	3600	Sb01g044160	669	Si039751m	1034
Os03g0299700	300			Si005194m	624
				Si008008m	5739
				Si032927m	2451
				Si011849m	219

Table 3.13. Size comparisons between inserted genes and average genes in each genome

Categories	*Oryza*		*Sorghum*		*Setaria*	
	All	Inserted	All	Inserted	All	Inserted
Number of genes	33,265	13	34,496	12	35,471	16
Mean of gene size	2140	1970	2620	3560	2220	2280
Median of gene size	1480	2070	1810	2340	1630	2360
Std. Dev.	2290	1340	3160	4280	2150	1830
*P**	0.649		0.463		0.886	

* $P < 0.05$ is considered a significant difference between the means of the two groups by *t* test.

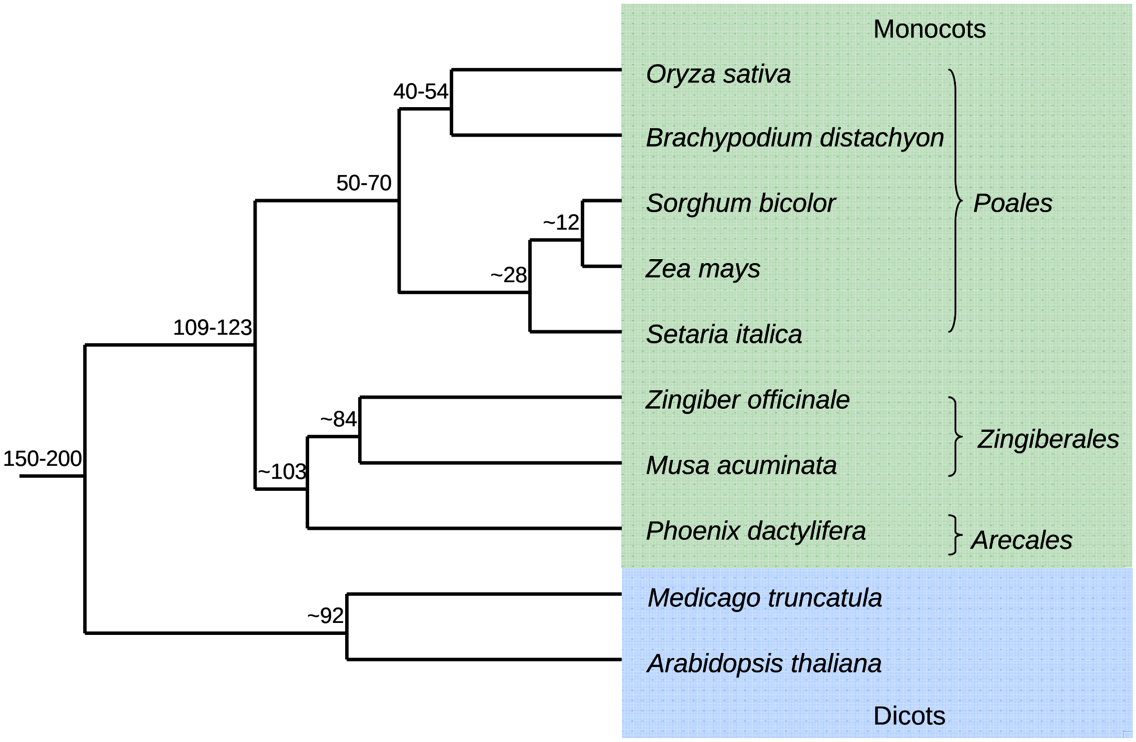

Figure 3.1. The phylogeny of the sequenced monocot genomes. The numbers on the nodes represent the estimated divergence times in millions of years (Gandolfo et al. 1998; Grant et al. 2000; Gaut 2002; Swigonova et al. 2004; Basu et al. 2008; Magallon and Castillo 2009; Paterson et al. 2009; Vogel et al. 2010; D'Hont et al. 2012).

CHAPTER 4

MUTATION RATES AND ACCUMULATION ACROSS DIFFERENT REGIONS OF RICE CHROMOSOMES[1]

[1] Feng, L. and J.L. Bennetzen. To be submitted to *Genetics*.

Abstract

It is well known that nucleotide substitution rates vary across mammalian genomes over many different scales, from adjacent sites to whole chromosomes. In plants, the distribution and dynamics of point mutations and indels have not been extensively investigated. To shed light on the evolutionary mechanisms that determine the distribution and dynamics of these small and very frequent mutations in flowering plants, mutations in the LTRs of LTR retrotransposons were analyzed across rice chromosomes 3 and 4. The results indicate that point mutations in chromosome 3 are more abundant near the centromeres, while the transition to transversion ratio (averaging 2.9) does not exhibit any genome location bias. Indels are unevenly distributed across both chromosomes. The overall number of these small mutations is significantly correlated with LTR retrotransposon age, but there is no correlation between the transition to transversion ratio and the age of LTR retrotransposons. These results suggest that the LTR retrotransposons across the rice genome are all or nearly all fully cytosine methylated, regardless of genomic location, leading to this high transition to transversion ratio. Significant negative correlation between the indel to point mutation ratio and the age of LTR-RTs was observed, suggesting that indels occur more rapidly early after the insertion of an LTR-RT than they do later, while point mutations maintain a more constant rate.

Introduction

Mutation is vital in producing the raw genetic variation on which natural selection and other evolutionary processes can act. Mutations and mutation rates can vary at many different scales, from single nucleotide substitution up to variation between whole chromosomes (Benzer 1961; Hodgkinson and Eyre-Walker 2011). In mammalian genomes, it has been shown that the greatest variation in rates of mutation are seen at the smallest scales, with single nucleotide variations occurring more than tenfold faster at some sites compared to others in the same genome (Hodgkinson and Eyre-Walker 2011). In all bacterial and eukaryotic species that have been studied to date, G and C nucleotides are more mutable than A and T nucleotides (Lynch 2007; Hershberg and Petrov 2010). The analysis of non-coding DNA showed that the identities of the adjacent nucleotides have significant effects on the mutation rate in mammals (Gojobori et al. 1982; Blake et al. 1992; Hwang and Green 2004). For instance, the frequency of transition mutations the cytosines in CG dinucleotides is ~30-fold higher relative to the average rate of mutation in great apes and is ~15-fold higher in other mammals (Zhao and Boerwinkle 2002; Hwang and Green 2004; Keightley et al. 2011). The rate of transversion mutations at CG dinucleotides are also a few-fold higher than other sites (Nachman and Crowell 2000; Hwang and Green 2004), suggesting that the overall mutation rate is elevated by tenfold at CG dinucleotides. Larger scale mutations, including gene insertion/deletion, inversion, duplication, and translocation have been discussed in Chapter 2 and 3.

Small-scale mutations are defined in this Chapter as point mutations in the form of transitions (purine-purine or pyrimidine-pyrimidine interchanges) and transversions (purine-pyrimidine interchanges) or small insertions and deletions (indels). Transitions are more frequent than transversions in coding sequences, because silent/neutral mutations are more commonly transitions (Noe and Kucherov 2004). Hence, such studies of point mutation variation do not actually measure mutation rate, because they are strongly affected by different levels of subsequent selection. When mutations in introns are investigated, the ratio of transitions to transversions tends to be closer to 1:1 in all eukaryotes investigated (Vitte and Bennetzen 2006). In the maize genome, it was observed more than 15 years ago that the highly cytosine methylated LTR retrotransposons have a ratio of transitions to transversions of ~3.1, compared to ~1.5 inside introns (SanMiguel et al. 1998). This difference was attributed to the higher instability of 5-methyl cytosine to transition mutation relative to unmodified cytosine.

To our knowledge, the relative frequencies of only two types of mutation have been investigated in plants. The frequency of rearrangements caused by unequal homologous recombination appears to be much higher in gene-rich areas, presumably caused by the higher rate of all forms of recombination in these regions (Ma and Bennetzen 2006; Ma et al. 2007). Also, transposable element insertions show different rates of targeting different genes for *de novo* mutation, while many accumulate preferentially in different regions of the genome (e.g., gene poor heterochromatin rather than gene-rich euchromatin, or vice versa) (Baucom et al. 2009).

Flowering plants vary considerably in nuclear genome size, and long terminal repeat retrotransposons (LTR retrotransposons, LTR-RTs) are the primary factor responsible for genome expansion in plants (Bennetzen et al. 2005). The abundance of LTR-RTs is positively correlated with the overall genome size across flowering plants. For instance, the ~140 Mb Arabidopsis genome is composed of only ~15-20% LTR-RTs, while the much larger grass genomes, such as maize (~2400 Mb) and barley (~5600 Mb), are > 70% LTR-RTs (Bennetzen 2009). Most of these elements have amplified in the past few million years (SanMiguel et al. 1996; SanMiguel et al. 1998; Wicker et al. 2001). Flowering plants have been shown to rapidly remove this and other DNA without selected host value by small deletions that can remove several hundred Mb of DNA per million years (Devos et al. 2002; Ma et al. 2004). The presence of partially deleted LTR-RTs in all the plant genomes is a reflection of this process. In addition, solo LTRs are generated by unequal intra-element homologous recombination between the two LTRs of an LTR-RT, thus providing a second mechanism for rapid DNA removal. With the combination of these two processes, > 190 Mb of LTR-RTs have been deleted from the rice genome within the past four million years, leading to a current genome of ~400 Mb containing 20-25% detectable LTR-RT elements or fragments (Ma et al. 2004; Tian et al. 2009).

In all plants investigated, the core of the centromere (the kinetochores) show frequent and extensive DNA rearrangements due the high rates of unequal homologous recombination in centromeric regions, even though meiotic chromosomal exchange is suppressed in these regions (Ma and Bennetzen 2006; Ma and Jackson 2006; Ma et al. 2007). In the areas flanking the centromeres, the so-called paracentromeric heterochromatin, this unequal recombination is highly suppressed (Ma and Bennetzen 2006). However, it is not known in plants whether point mutations or small indels occur at different rates across chromosomes. Because the two LTRs of an LTR-RT are usually identical at the time of insertion, investigation of the divergence of these LTRs provides a uniquely powerful tool to assess the relative frequencies of all types of small mutations

(SanMiguel et al. 1998). To shed light on this question, LTR mutations were investigated for their distribution across rice chromosomes 3 and 4. The relationships between LTR mutations and the age of LTR-RTs were also investigated.

Materials and Methods

LTR Retrotransposon Data Resources and Processing

The sequences and genomic features of the LTR retrotransposons in 12 rice chromosomes (IRGSP Build 4.0 pseudomolecules) were provided by Zhixi Tian and Jianxin Ma at Purdue University (Tian et al. 2009). The age of LTR-RTs they provided were determined in a manner described in previous study based on the number of point mutations (Ma et al. 2004). Rice chromosomes 3 and 4 were investigated in this study because the LTR-RTs in these two chromosomes were fully annotated, including those across the completely sequenced centromere of chromosome 4. LTR-RTs that were annotated as solo LTRs, truncated elements, and intact elements without target site duplication (TSD) were excluded from the study. The LTR-RTs with a predicted size greater than 100 kb were also removed, because these might have arisen from an artifactual assembly or because they might give rise to artifactual interpretations of which LTRs were the actual ends of a single LTR-RT.

Manual Inspection of LTR Mutations

The point mutations and indels in LTR-RT pairs across rice chromosomes 3 and 4 were manually inspected by two undergraduate students, Melanie Buser and Zack Farmer. ClustalW (Higgins et al. 1996) and Jalview (Clamp et al. 2004) were used for the sequence alignment of LTRs. If the sequence identity between the LTR pair of an intact element was less than 90%, the element were removed from the study because we did not wish to miss or mis-score mutations that overlapped with or reverted previous mutations, a strong possibility in very old LTR-RTs. The number of transitions (A ↔ G and C ↔ T), transversions (A ↔ C, A ↔ T, C ↔ G, and G ↔ T), and indels between the LTR pairs of the intact elements were documented on rice chromosomes 3 and 4. All the intact elements were analyzed in rice chromosome 3, while only the first ~60% (according to the genomic location) of the intact elements in chromosome 4 were investigated within the time frame of the project.

LTR Mutation Frequencies across Rice Chromosomal Regions

Each chromosome was divided into 2 Mb bins, and the numbers of transitions, transversions, indels, transition to transversion ratio, and indel to point mutation (total of transition and transversion) ratio were calculated in each 2 Mb bin. The distribution of the LTR mutations per 10 kb of LTRs analyzed in each 2 Mb bin across chromosomes is presented in bar plots. The constancy of the transition to transversion ratio and the indel to point mutation ratio in each 2 Mb bin across chromosomes was analyzed by means of logistic regression. The fold difference between transition and transversion, as well as the fold change between indel and point mutation was also calculated. Pearson’s Correlation was used to test the relationship between the number of mutations per 1 kb of compared LTRs and the age of LTR-RT insertion. The relationships between the transition to transversion ratio, as well as the indel to point mutation ratio, and the age of LTR-RTs on rice chromosomes 3 and 4 were also calculated. All statistical analyses were performed using the R statistical package.

Results and Discussion

Manual Discovery and Description of LTR Mutations

A recent comprehensive investigation of LTR retrotransposons in rice suggested a total of 16,013 LTR retrotransposons and fragments in the 12 chromosomes, including 4937 intact elements, 7981 solo LTRs, 2006 truncated elements, and 1089 other elements (Tian et al. 2009), such as the chimeric LTR-RT formed by inter-element unequal recombination, and thus without TSD (Devos et al. 2002). The length of chromosome 3 is estimated to be ~37.3 Mb, and it contains 343 intact elements. The length of chromosome 4 is ~36.1 Mb, and it contains 556 intact elements (Table 4.1). By excluding LTR-RTs with size > 100 kb and elements with LTR pair identity < 90%, a total of 274 intact elements on chromosome 3 and 278 intact elements on chromosome 4 were used for further analyses. The distribution of theses LTR-RTs across chromosomes is shown in Figure 4.1. The centromere of rice chromosome 3 is located at 19.4 Mb of the chromosome, and that of chromosome 4 is located at 9.7 Mb (Ouyang et al. 2007). The LTR-RTs investigated in this study are more abundant near the centromere regions. With manual sequence alignment for a pair of LTRs in each element, a total of 4247 transitions, 1484 transversions, and 483 indels were identified on chromosome 3, and a total of 5009 transitions, 2326 transversions, and 854 indels were characterized across chromosome 4 (Table 4.2).

The number of transitions is significantly higher than that of transversions, consistent with previous observations (SanMiguel et al. 1998). It should be noted that our selection only of elements with > 90% LTR identity, leads to an underestimation of the total number of LTR mutations in the genome, but < 10% of rice LTR-RTs fell into this discarded category.

LTR Mutation Preference across Rice Chromosomes

Transition, transversion and indels numbers per 10 kb of LTRs analyzed, along with the transition/transversion ratio and indel/point mutation ratio, were calculated in 2 Mb bins across each chromosome. Uneven distributions of each type of mutation are observed along the rice chromosomes. Most of the point mutations are located near the centromere region of rice chromosome 3, while this pattern was not observed in chromosome 4 (Figure 4.2). In the 4-6 Mb bin and 10-12 Mb bin, the total lengths of the analyzed LTRs are 3355 bp and 7761 bp respectively, so the numbers of mutations in these two bins were not shown. The number of transitions is significantly higher than that of transversions in all bins, which is consistent with previous observations (SanMiguel et al. 1998). The number of indels showed uneven distribution across both chromosomes, and there is no clear pattern for their genome location preferences (Figure 4.3). On chromosome 3, the most abundant regions for indels are the 6-8 Mb and 28-30 Mb bins, and the 12-14 Mb bin on chromosome 4 has more indels per 10 kb of compared LTRs than any other region analyzed.

The transition/transversion ratio was also calculated in each 2 Mb bin, and tested if the ratios in each bin were constant across chromosomes by means of logistic regression (Figure 4.4). Statistically, the overall transition/transversion ratio is not consistent in each 2 Mb bin (*P* = 0.004 in chromosome 3, *P* = 0.029 in chromosome 4). There is an exceptional region (32-34 Mb bin) with a very high transition/transversion ratio on chromosome 3. If this bin is excluded, the logistic regression showed a constant transition/transversion ratio across chromosome 3 (*P* = 0.053). On an average, transitions are 2.9 fold more abundant than transversions on chromosome 3, while the ratio on chromosome 4 is 2.2 fold. The 32-34 Mb bin on chromosome 3 that has a very small number of transversions, leads to the significantly higher ratio of transition to transversion than calculated for chromosome 4. It is believed that a high transition to transversion ratio is evidence of extensive cytosine 5-methylation, which will increase the C to T transition rate (SanMiguel et al. 1998). The > 2:1 ratio seen in plant LTR retrotransposons suggests that most or all of these elements are in an epigenetically silenced state associated with extensive cytosine 5-methylation (Gruenbaum et al. 1981; Vitte and Bennetzen 2006).

Significant correlations between the number of indels and point mutations per 10 kb of compared LTRs in each 2 Mb bin were observed on chromosome 3 (Pearson's correlation coefficient = 0.679, *P* = 0.001), but there is no such correlation observed on chromosome 4 (Pearson's correlation coefficient = 0.508, *P* = 0.163). The indel/point mutation ratio was also calculated in each 2 Mb bin (Figure 4.5). With logistic regression, the indel/point mutation ratio is not consistent across 2 Mb bins (*P* = 0.0001 in chromosome 3, *P* < 0.0001 in chromosome 4). In the 8-10 Mb bin of chromosome 3 and the 2-4 Mb of chromosome 4, the indel/point mutation ratios are relatively high. The average ratio of indel to point mutation is 0.08 on chromosome 3, while the ratio on chromosome 4 is 0.12. The ratio of indel to point mutations has been shown to be quite variable in plants, but does not correlate with either phylogenetic relatedness or genome size (Vitte and Bennetzen 2006).

LTR Mutations and the Age of LTR Retrotransposons

There is a significant positive correlation observed between the age of LTR-RTs and the number of mutations per 1 kb of compared LTRs, including transition (Pearson's correlation coefficient = 0.889, *P* = 2.2E-16), transversion (Pearson's correlation coefficient = 0.768, *P* = 2.2E-16), and indels (Pearson's correlation coefficient = 0.429, *P* = 1.1E-13) on rice chromosome 3 (Figure 4.6 and 4.7). This is as expected, given that age is calculated from the number of point mutations that differentiate two LTRs (SanMiguel et al. 1998). However, the same trends were not observed on chromosome 4 (*P* > 0.05 for all analyses between transition, transversion, and indels versus the age of LTR-RTs). The relationship between the transition/transversion ratio and the age of LTR-RTs was also investigated by means of Pearson's Correlation. There is no significant correlation observed between the transition/transversion ratio and the age of LTR-RTs in chromosome 3 (Pearson's correlation coefficient = 0.079, *P* = 0.245) and chromosome 4 (Pearson's correlation coefficient = 0.083, *P* = 0.393) (Figure 4.8). These results suggest that the LTR-RTs across the rice genome are fully cytosine methylated, regardless of genomic location, leading to this high transition to transversion ratio. We also analyzed the relationship between the ratio of indel to point mutation and the age of LTR-RTs with the same approach (Figure 4.9). Interestingly, there is a significant negative correlation observed on chromosome 3 (Pearson's correlation coefficient = -0.217, *P* = 0.0005), but there is no significant correlation observed on chromosome 4 (Pearson's correlation coefficient = -0.081, P = 0.368). The negative correlation between indel/point mutation ratio and the age of LTR-RTs suggests that indels occur more rapidly early after the insertion of an LTR-RT than they do later, while point mutations maintain a more constant rate.

The difference of the observations between chromosome 3 and 4 may be due to the incomplete analysis of the LTR mutations on chromosome 4. Only 60% of the LTR-RTs have been analyzed on this chromosome. Future studies will need to apply this analysis to larger data set and more chromosomes in multiple plant species to uncover the relationship between LTR mutations, genomic location, and age of LTR-RTs. This study represents the first detailed assessment of the chromosome-wide mutation rates and their distribution in flowering plants. The results indicate that different regions of the genome do not have dramatically different mutation rates for transitions, transversions or small indels when one investigates the highly methylated (and presumably heterochromatic) segments of these different regions. Future studies are needed to investigate relatively unmethylated regions of the genome that are under relatively low selection pressure (e.g., introns, 3' trailers and 5' leaders) to see the relative frequency of accumulated mutations. Better still would be next generation-sequence-based analysis of the mutations found in first generation progeny, where mutation across the entire genome could be investigated in a time frame where natural selection has had little chance to remove negative mutations.

Acknowledgements

We thank Zhixi Tian and Jianxin Ma for sharing the LTR retrotransposon annotation results for rice chromosomes 3 and 4, Melanie Buser and Zack Farmer for the manual inspection of LTR mutations in rice chromosomes 3 and 4, and Xinyu Liu for suggestions on statistical analyses.

Table 4.1. LTR retrotransposons across rice Chromosomes 3 and 4

Chromosomes	Length	Centromere location	Total number of LTR-RTs	Intact element	Truncated element	Solo LTRs	Others
Chr3	37.3 Mb	19.4 Mb	1217	343	145	632	97
Chr4	36.1 Mb	9.7 Mb	1746	556	228	863	99

Table 4.2 LTR mutations across rice Chromosomes 3 and 4

Chromosomes	Number of LTR-RTs used in this study	Transitions	Transversions	Indels
Chr3	274	4247	1484	483
Chr4	278	5009	2326	854

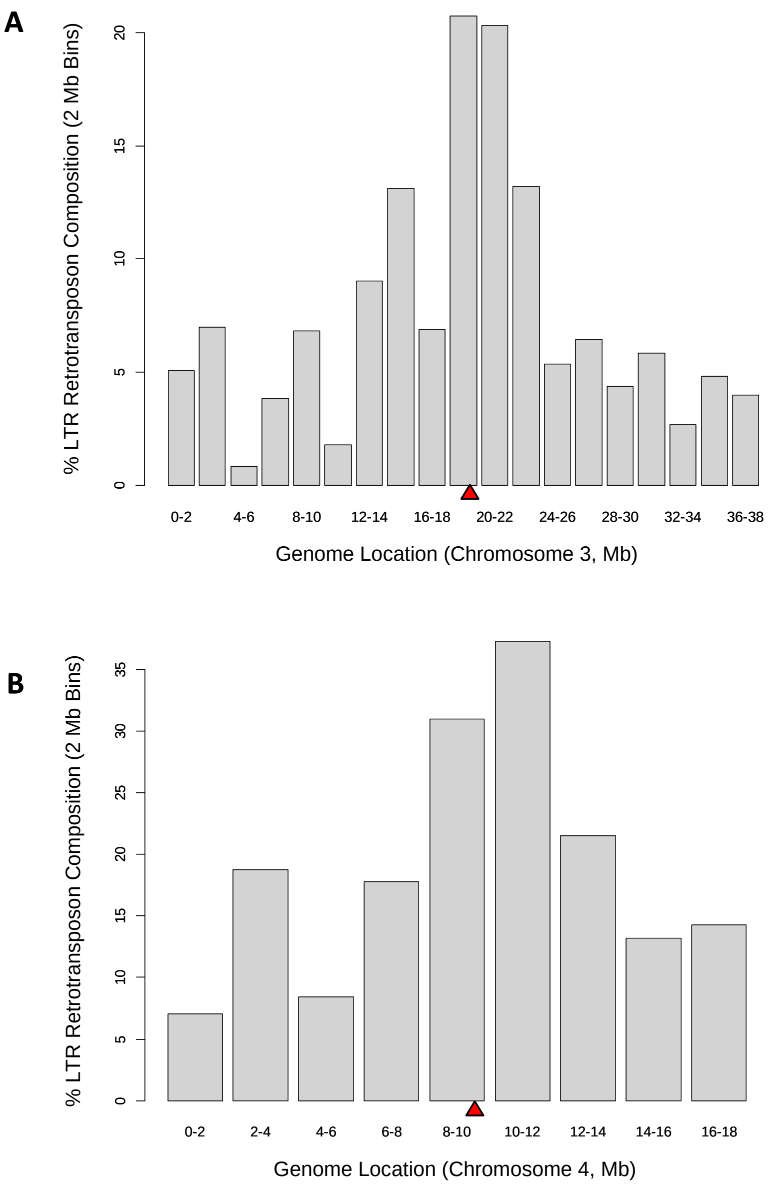

Figure 4.1. LTR retrotransposon abundance across rice Chromosomes 3 (A) and 4 (B). The x axis indicates chromosome locations while the y axis shows % LTR retrotransposon abundance in 2 Mb windows. Red triangles on the x axis indicate approximate centromere positions.

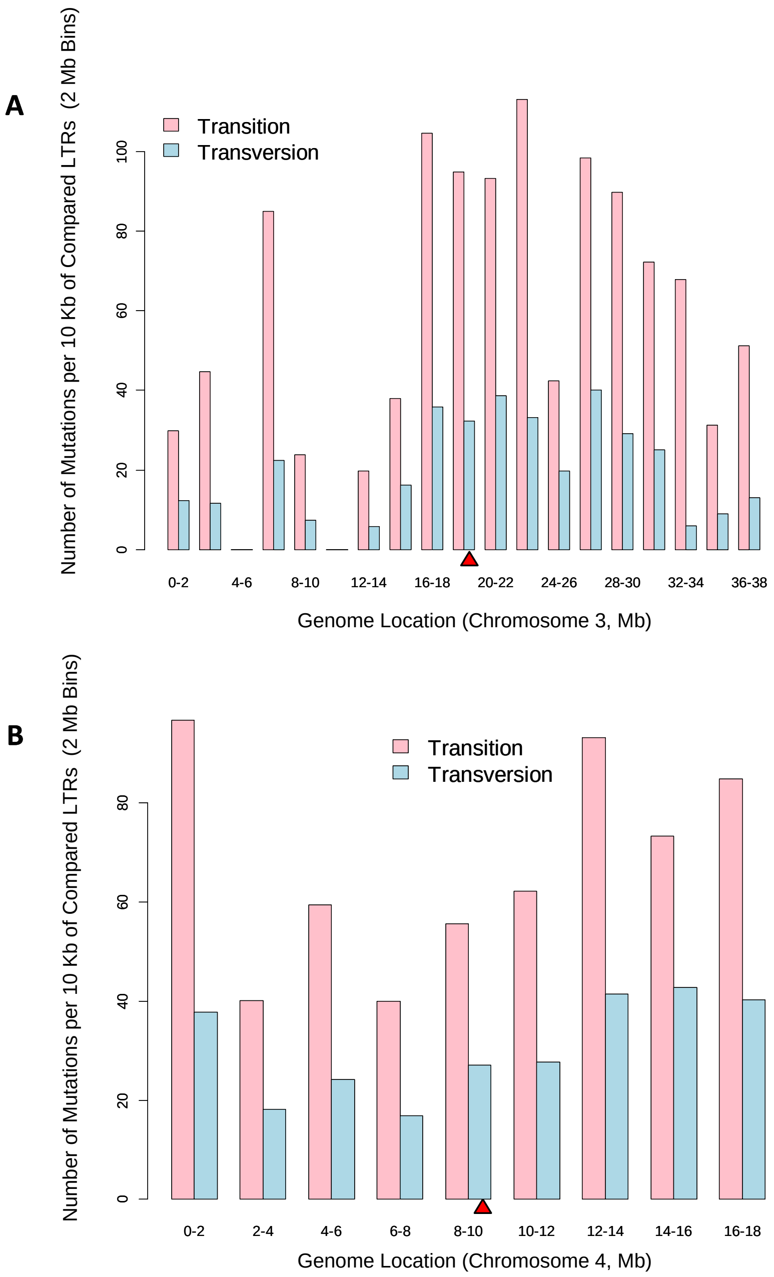

Figure 4.2. Point mutations distribution across rice Chromosomes 3 (A) and 4 (B). The x axis indicates chromosome locations while the y axis shows the number of transitions and transversions per 10 Kb of compared LTRs in 2 Mb windows. Red triangles on the x axis indicate approximate centromere positions.

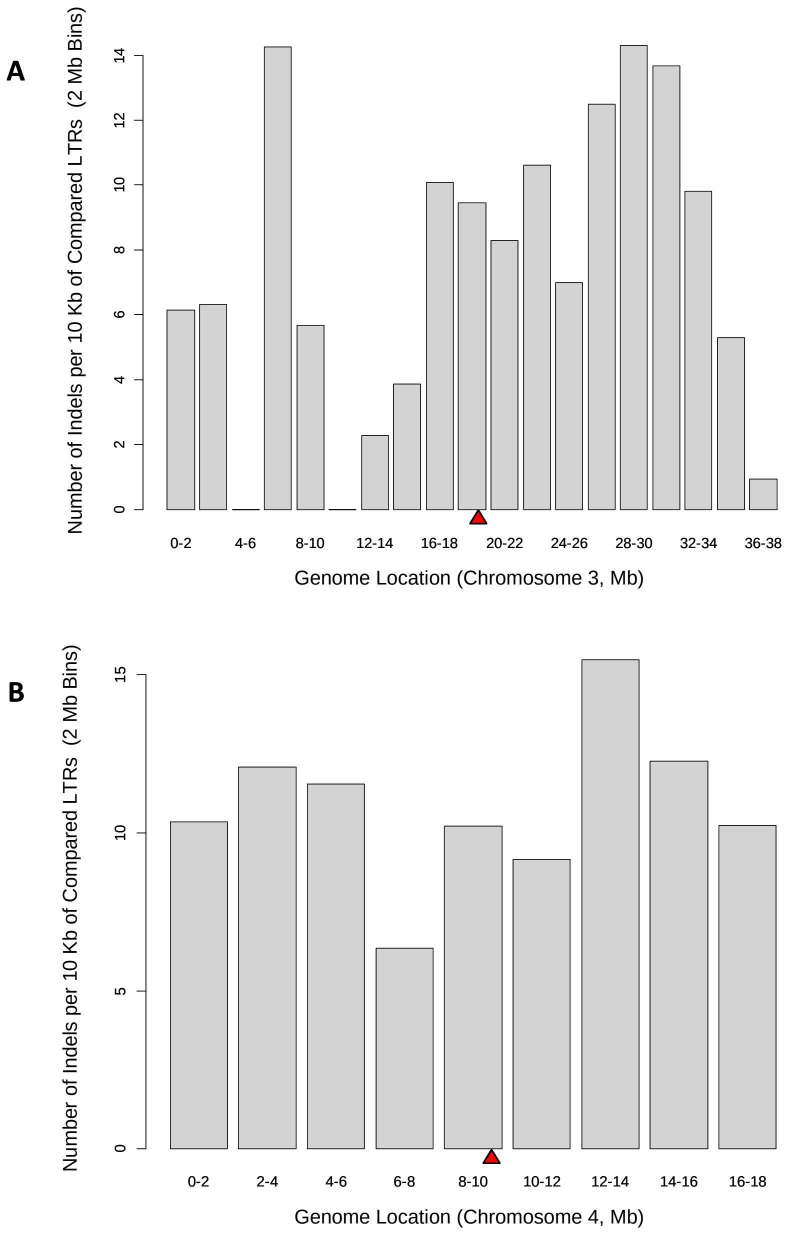

Figure 4.3. Indels distribution across rice Chromosomes 3 (A) and 4 (B). The x axis indicates chromosome locations while the y axis shows the number of indels per 10 Kb of compared LTRs in 2 Mb windows. Red triangles on the x axis indicate approximate centromere positions.

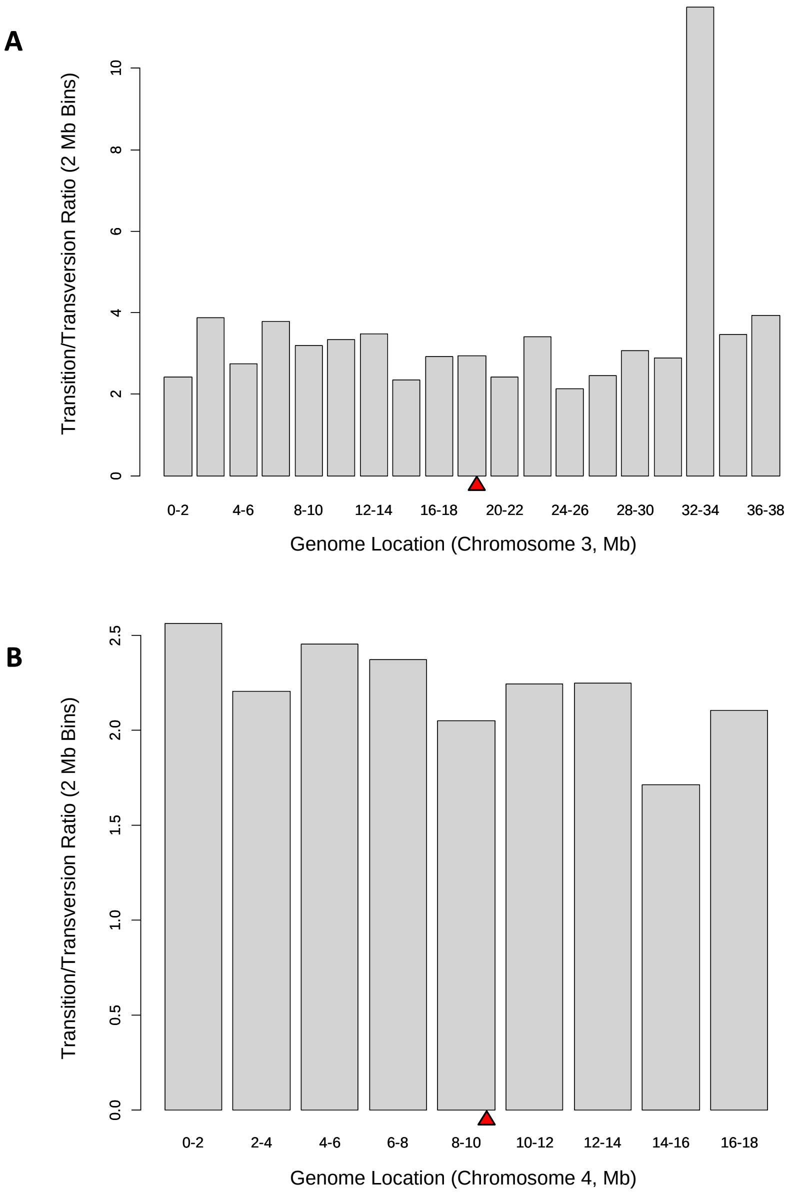

Figure 4.4. Transition/transversion ratio across rice Chromosomes 3 (A) and 4 (B). The x axis indicates chromosome locations while the y axis shows the transition/transversion ratio in 2 Mb windows. Red triangles on the x axis indicate approximate centromere positions.

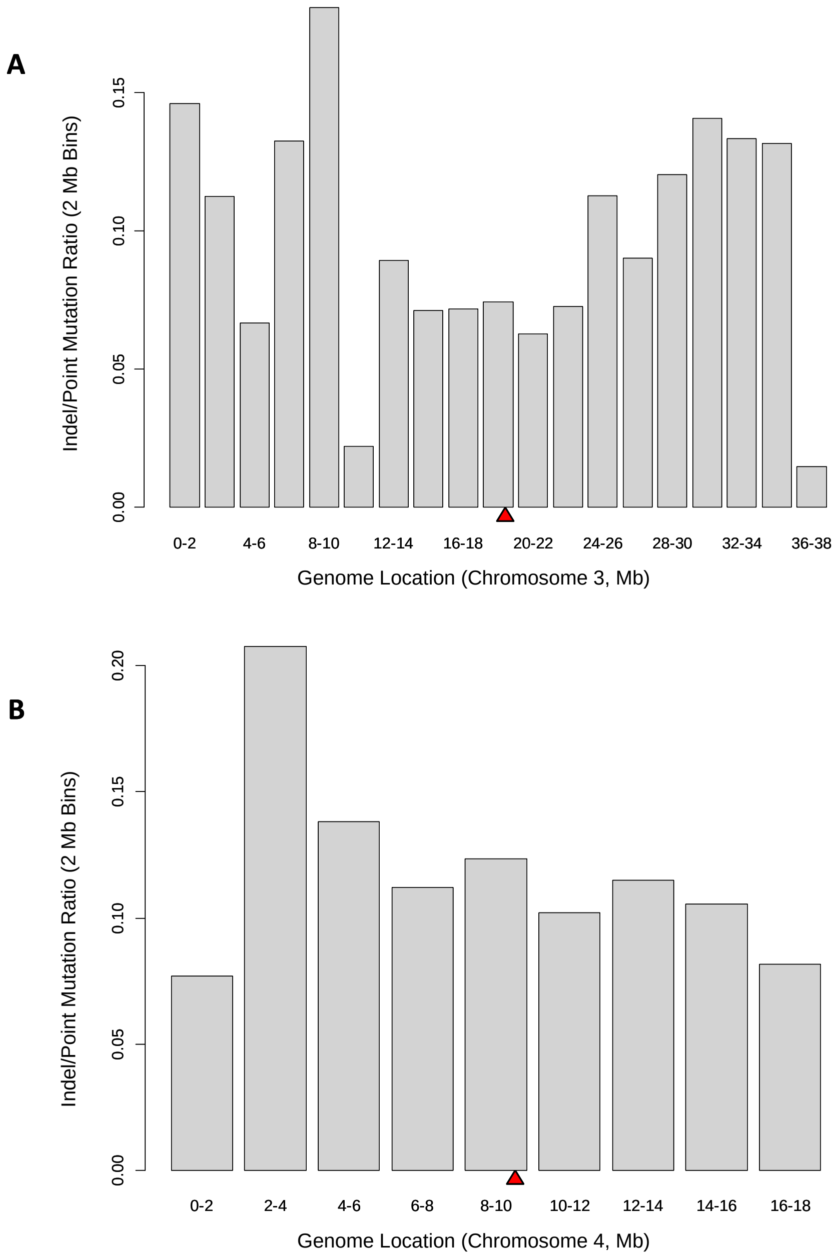

Figure 4.5. Indel/point mutation ratio across rice Chromosomes 3 (A) and 4 (B). The x axis indicates chromosome locations while the y axis shows the indel/point mutation ratio in 2 Mb windows. Red triangles on the x axis indicate approximate centromere positions.

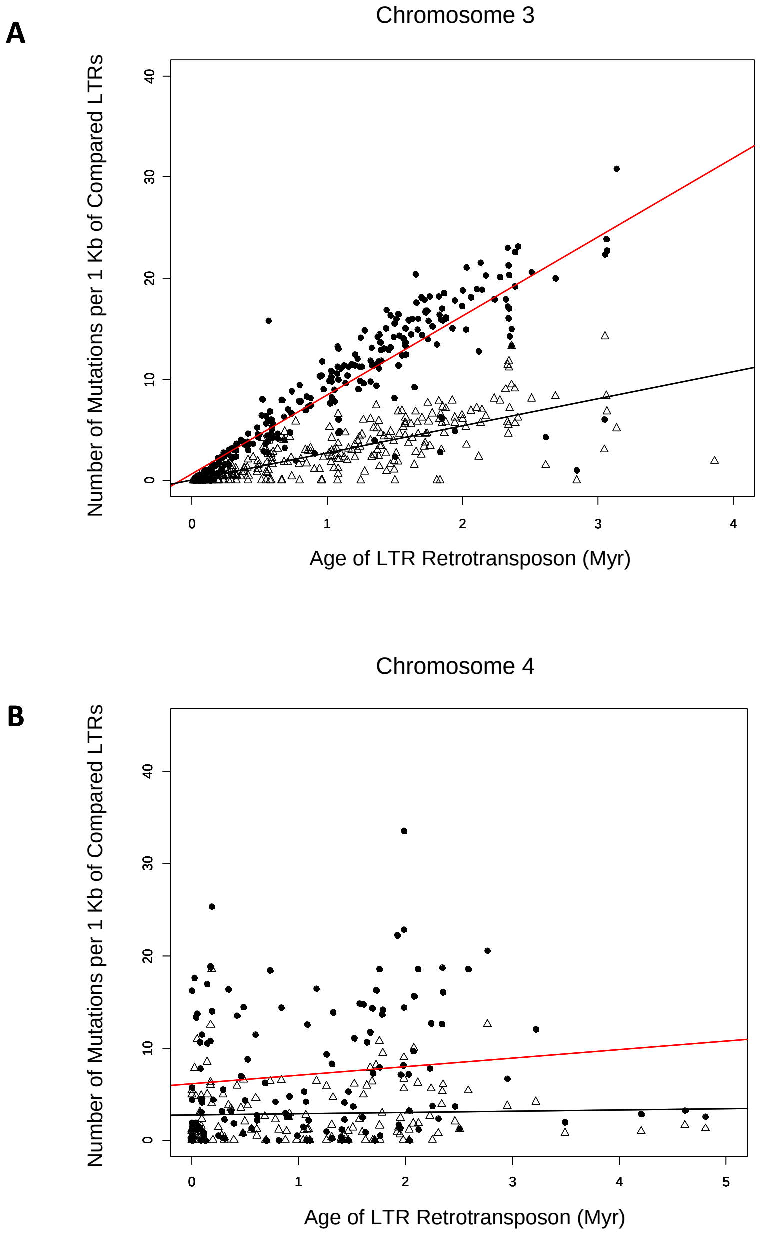

Figure 4.6. The relationship between point mutations and the age of LTR retrotransposon in rice Chromosomes 3 (A) and 4 (B). The x axis indicates the age of LTR retrotransposon in million years while the y axis shows the number of point mutations for each intact element. Dots represent transitions and triangles represent transversions.

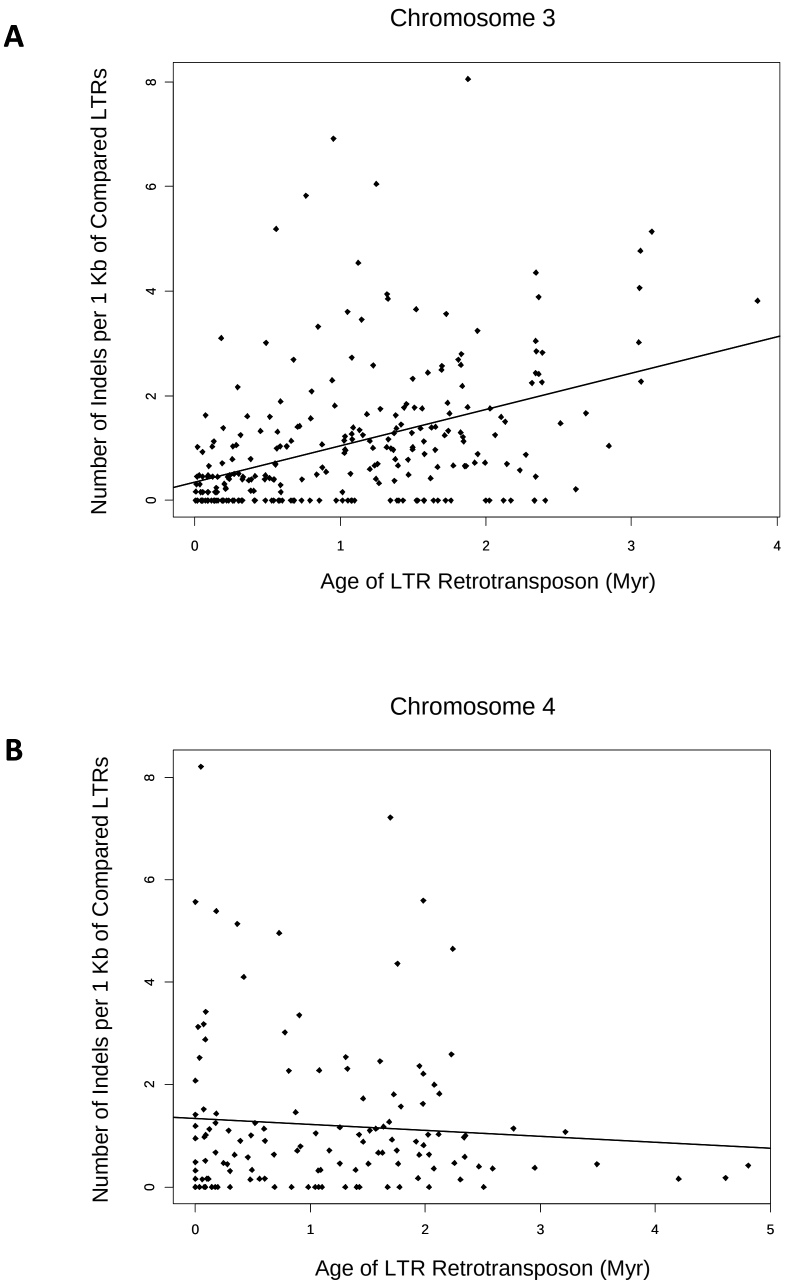

Figure 4.7. The relationship between indels and the age of LTR retrotransposon in rice Chromosomes 3 (A) and 4 (B). The x axis indicates the age of LTR retrotransposon in million years while the y axis shows the number of indels for each intact element.

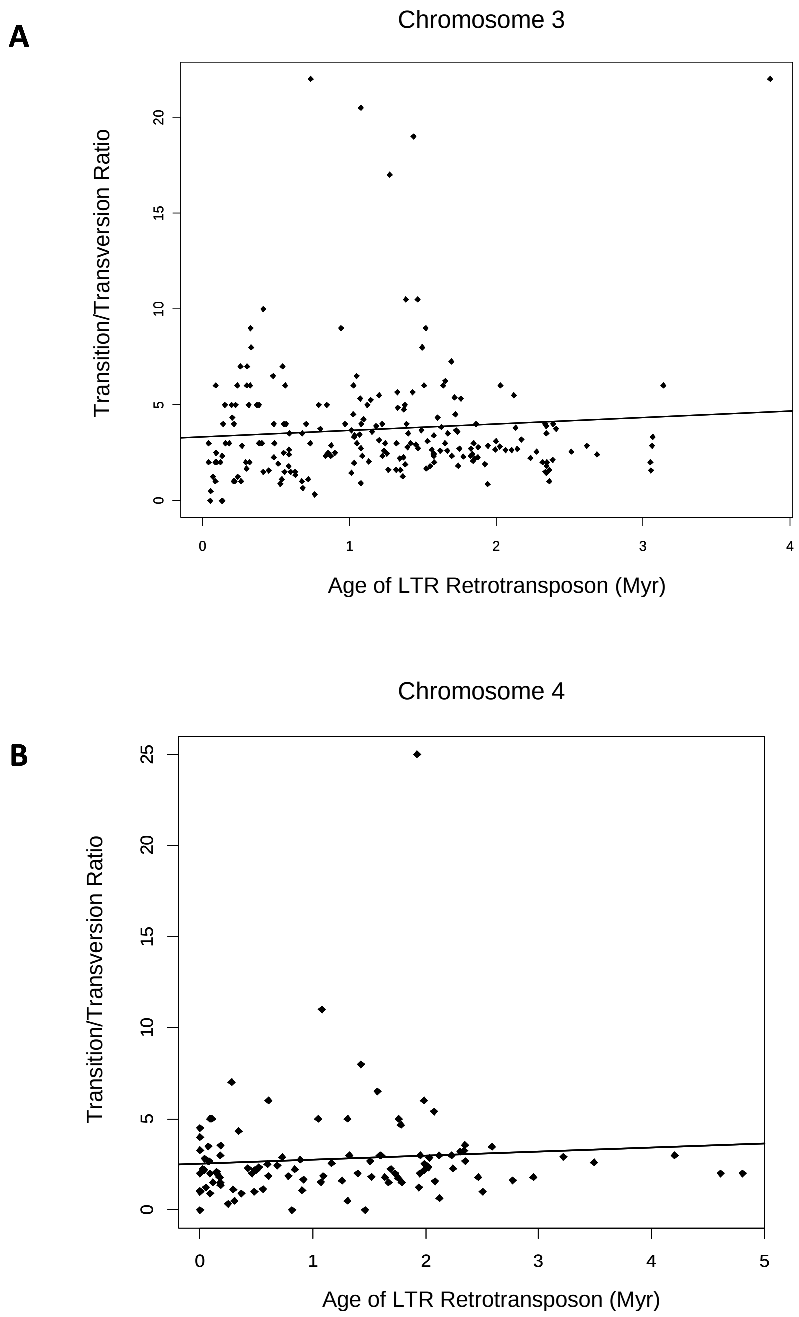

Figure 4.8. The relationship between the transition/transversion ratio and the age of LTR retrotransposon in rice Chromosomes 3 (A) and 4 (B). The x axis indicates the age of LTR retrotransposon in million years while the y axis shows the transition/transversion ratio for each intact element.

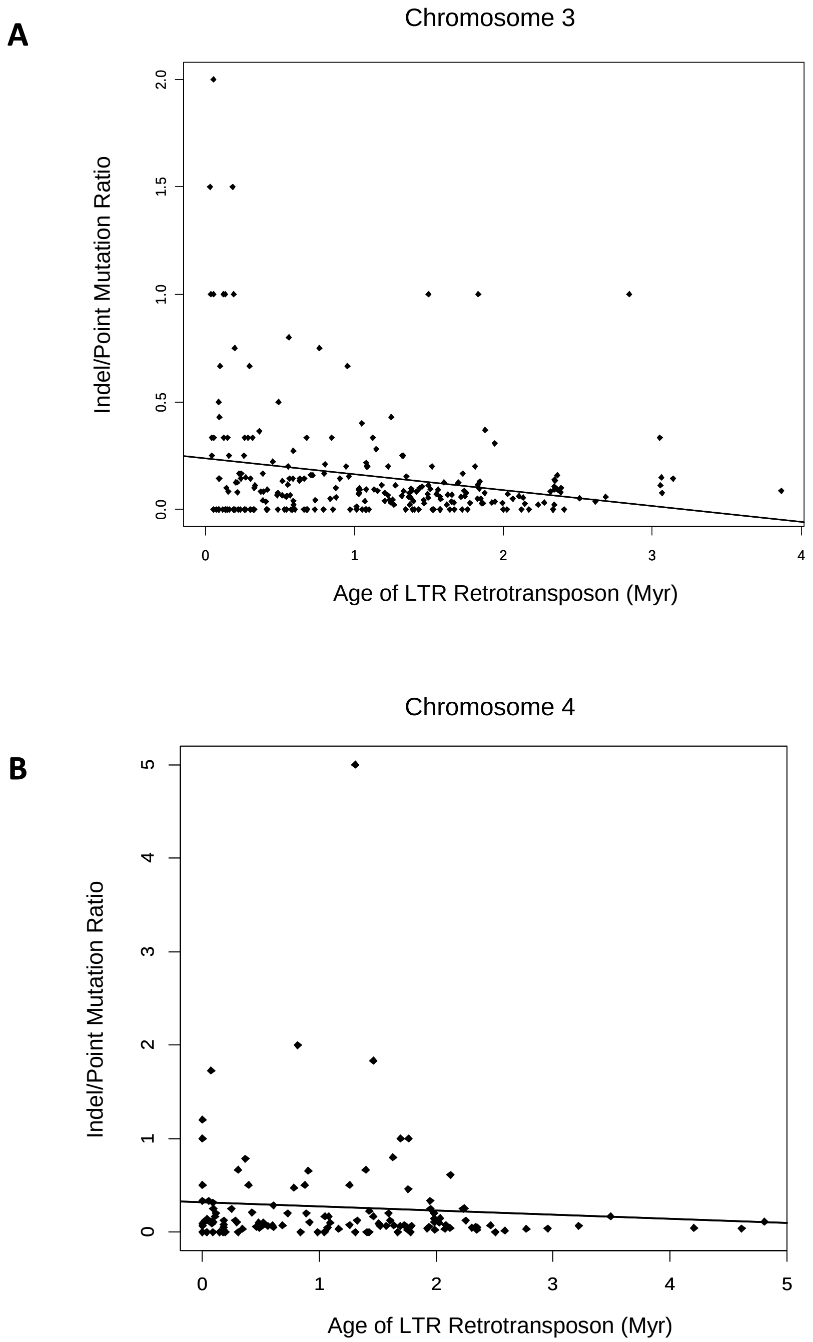

Figure 4.9. The relationship between the indel/point mutation ratio and the age of LTR retrotransposon in rice Chromosomes 3 (A) and 4 (B). The x axis indicates the age of LTR retrotransposon in million years while the y axis shows the indel/point mutation ratio for each intact element.

CHAPTER 5

CONCLUDING REMARKS

The work presented here advances our understanding of the mechanistic basis of genomic instability in flowering plants. The central focus of this work has been the idea of generating a novel terminology and perspective, gene pair conservation, to precisely describe the absolute nature of genomic instability during the evolution of flowering plants. The primary intents of this study were to investigate the retention or loss of gene pair linkage in various plant species, to quantify the frequencies of various types of genome rearrangements, to understand the lineage-specificity of the genomic instability, and to gain insights into the mechanisms responsible for genic rearrangements. To this end, I pursued a process of local and global genome comparison through investigating pairs of adjacent genes, and a sampling approach to manually inspect the retention and rearrangement of gene pair linkage across an array of plant species: rice, sorghum, Setaria, Brachypodium, Arabidopsis, and Medicago. The recently sequenced non-grass monocots, banana and date palm, served as outgroups in this study to determine the lineage-specificity of the genomic rearrangements that were observed. By using this gene pair approach, a novel terminology and objective quantitation for genome rearrangement was developed.

Flowering plants exhibit highly variable genomes. Small rearrangements are abundant in plant genomes, but the degree of variation in these traits is just beginning to be investigated (Vitte and Bennetzen 2006) and the reasons for the quantitative differences in these rearrangements are unknown. The adjacent gene pair can serve as an unambiguous and unbiased unit to characterize genome rearrangement. The study of rearrangements between adjacent gene pairs allowed us to quantify the frequencies of conservation and rearrangement between a number of grass pairs (i.e., rice-sorghum, rice-Setaria, sorghum-Setaria, rice-Brachypodium), and two distant related dicots (i.e. Arabidopsis-Medicago), and determine precisely what percentages are caused by deletion, inversion, translocation or other processes, for any investigated pair of species.

In this study, a focus was placed on non-paralogous gene pairs, because a substantial proportion of adjacent genes are gene duplicates that can confuse orthology determination. After identifying all the adjacent gene pairs in the plant genomes investigated, and removing hypothetical genes, transposon-related genes, paralogous and very tightly linked gene pairs, the relative gene pair orientations are close to random in most of the plant species investigated. A dramatic exception to this randomness was observed in gene pairs with intergenic distance < 1 kb. At this separation range, gene pairs are greatly biased toward a convergent orientation with ~2X higher frequency than expected by chance. It is not known why this excess convergent orientation in tightly linked gene pairs is observed, but a functional interaction for antisense RNA regulation seems a logical possibility. Hence, this study also provides a useful dataset of closely linked genes to investigate certain types of gene pair interaction, and to begin to investigate the selection on these gene pairs as it relates to a possible degree of shared regulation.

One of the major outcomes of this research is the retention and rearrangement frequencies of gene pair linkage across a number of plant species. By precise manual inspection for a randomly selected subset of gene pairs in each grass pair investigated, it suggests that more than 50% of adjacent gene pairs were observed to be preserved in rice-sorghum, rice-Setaria, sorghum-Setaria, and rice-Brachypodium comparisons after their divergence from a common ancestor ~50-70 Mya. This frequency of gene pair retention is substantially higher than the 25.4% predicted by automatic high throughput characterization (Krom and Ramakrishna 2010), in which every gene mis-annotation would be counted as a rearrangement, resulting in an inappropriately high level or predicted instability.

When the retention and rearrangement of gene pairs with intergenic distances < 1 kb were investigated, the gene pair rearrangement was found to be significantly less than seen for the more distantly linked gene pairs. However, for genes separated by various distances > 1 kb, the rearrangement space between genes was much more constant than the physical distance. This suggests that much of the DNA between distantly linked genes (commonly composed of epigenetically silenced transposable elements) is not particularly active in rearrangement. Given that my analysis of point mutations and small indels across the rice genome did not indicate any differences in mutation rates across genomic regions, it seems likely that the major rearrangement processes that do seem to work only in a small rearrangement space are ones that are quite sensitive to the epigenetic status. *De novo* transposon insertion and excision are both strongly affected by the epigenetic status of the DNA involved, as is homologous recombination, so these results further support the idea that these entities are important factors in genome rearrangement.

In contrast to the grass instability observed, only 7% of gene pairs were still intact when compared the two dicots Arabidopsis and Medicago, whose ancestors are though to have diverged ~92 Mya. In general, the frequencies of gene pair retention and rearrangement were found to correlate with phylogenetic relationships. By considering the divergence time of each investigated grass pair and their frequency of gene pair retention, our data suggest that sorghum is the most stable among the grass lineages investigated, and the two dicot genome, Medicago and Arabidopsis, have exhibited a much higher relative instability per unit time than any of the grasses. In future analyses, a larger number of investigated species and breadth of phylogenetic sampling will allow us to determine the relative contributions of each lineage.

Characterization of gene pair rearrangement also allowed us to quantify the absolute nature and frequency of the different types of genome rearrangement that affect gene content and order. Manual inspection for the genomic context and possible patterns of gene pair rearrangement suggest that movement of one gene of the pair to another chromosome in the other compared species is the most frequent type of rearrangement in all species comparisons, with single gene insertions/deletions within the pairs serving as the second most common event. The next most frequent events are inversions that usually involved only a single gene, found in ~2% of the cases examined. No cases of translocation-related gene pair rearrangement were observed, and this was not a surprise because comparative genomic studies in all eukaryotes show a low frequency of translocations. This quantitation is unique among all the existing studies for the genomic instability in plants, and provides insights into the natures of different patterns of plant genome evolution. We hope that this approach becomes a routine tool that is extensively utilized in the field of comparative genomics.

Another important outcome of this dissertation is the determination of lineage specificity of the genomic rearrangements that were observed. By comparing genome arrangement patterns in several plant species of known phylogeny, we should be able to determine the lineage-specificity and timing of the rearrangements. To identify the lineages of the rearrangements that differentiate sorghum and rice genomes, I used banana and date palm, the first two sequenced monocot genomes outside the grass family, as the outgroup. The rice lineage appears to be more unstable than the sorghum lineage by using date palm as the outgroup, while this difference was not seen when using banana as the outgroup. We do not know the reason for this inconsistency, but a possible reason is the non-randomness in the analysis of very small numbers of gene pair retentions seen in date palm and banana. In order to resolve this issue, larger data sets and a more closely related outgroup to the grasses are need in future studies. The gene pair comparisons among rice, sorghum, and Setaria suggest that very similar rates of genomic instability in sorghum and Setaria lineages since their descent from a common ancestor.

With the very careful manual inspection of a randomly chosen subset of gene pairs, several mis-annotations were identified in relatively well-annotated genomes, such as rice and sorghum. However, in date palm, almost one quarter of the gene pairs selected appeared to be mis-annotation products. Most of these mis-annotations from date palm appeared to be a sequence that was called a single gene in both rice and sorghum but called two adjacent unidirectional genes in date palm. The second most common mis-annotation events observed are the mis-annotation of TEs and pseudogenes as novel genes. We suggest that future publications on full genome sequences in plants should require a manual characterization of the accuracy of the high throughput annotation that was performed.

In this dissertation, the distribution and dynamics of small-scale mutations (i.e., point mutations and indels) have also been extensively investigated. The nearly constant distribution of transition to transversion ratio across chromosomes suggests that LTR retrotransposons across the rice genome are fully cytosine methylated, regardless of genomic location, leading to this high transition to transversion ratio. The significant negative correlation observed between indel to point mutation ratio and the age of LTR-RTs indicates that, over time, the relative frequency of indels decreases relative to point mutations. Perhaps early in the history of an LTR-RT, indels occur quite frequently before it has yet been fully silenced by epigenetic processes, while point mutations occur at a relatively constant rate over time. Future analyses with a larger data set on more chromosomes in multiple plant species will provide insights into the evolutionary mechanisms that determine the distribution and dynamics of these small and very frequent mutations in flowering plants.

Currently, we do not know which of the many possible mechanisms of instability, such as unequal homologous recombination and illegitimate recombination, are responsible for the different types of genome rearrangements observed in different lineages. Future research will need to investigate the very recent genome rearrangement from closely linked species, such as the cultivated rice and wild rice (Huang et al. 2012), to uncover the absolute nature of the genomic instability. If very recent events can be detected, then we hope to be able to gain mechanistic insights into the processes that can differentiate the evolving properties of different plant lineages.

In conclusion, these analyses indicate plant genomes with different rates of rearrangement, and with significant specificity in the types of rearrangement that can accumulate over time. This study, manually investigating randomly chosen gene pairs, provides the first and unique detailed assessment of the rates, types, extent and lineages of all types of local gene rearrangement in plants. It advances our understanding of the mechanistic basis of genomic instability in flowering plants. The investigation of rates and natures of genome rearrangement across lineages allows us to identify the evolutionary origins of changes in genome instability, and may shed light on the mechanisms of the adaptation to various enviroments for certain species. Development of this gene pair approach and the vocabulary to describe rearrangement types and lineages will make future analyses of larger data sets easier to generate more robust findings and determine the natures and rates of genomic rearrangements that differentiate the genomes of all families of organisms.

BIBLIOGRAPHY

Adachi N, Lieber MR. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**(7): 807-809.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J et al. 2011. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol* **29**(6): 521-U584.

Basu MK, Rogozin IB, Deusch O, Dagan T, Martin W, Koonin EV. 2008. Evolutionary dynamics of introns in plastid-derived genes in plants: Saturation nearly reached but slow intron gain continues. *Mol Biol Evol* **25**(1): 111-119.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet* **5**(11): e1000732.

Bennett MD, Leitch IJ. 1995. Nuclear-DNA amounts in angiosperms. *Ann Bot* **76**(2): 113-176.

Bennetzen JL. 1996. The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol* **4**(9): 347-353.

Bennetzen JL. 2000. Comparative sequence analysis of plant nuclear genomes:m microcolinearity and its many exceptions. *Plant Cell* **12**(7): 1021-1029.

Bennetzen JL. 2007. Patterns in grass genome evolution. *Curr Opin Plant Biol* **10**(2): 176-181.

Bennetzen JL. 2009. Maize genome structure and evolution. *Maize Handbook Volume II: Genetics and Genomics* **2**.

Bennetzen JL, Chen M. 2008. Grass genomic synteny illuminates plant genome function and evolution. *Rice* **1**: 109-118.

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Curr Opin Plant Biol* **7**(6): 732-736.

Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**(9): 1509-1514.

Bennetzen JL, Ma J. 2003. The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr Opin Plant Biol* **6**(2): 128-133.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot (Lond)* **95**(1): 127-132.

Bennetzen JL, Ramakrishna W. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol Biol* **48**(5-6): 821-827.

Benzer S. 1961. On the topography of the genetic fine structure. *Proc Natl Acad Sci U S A* **47**(3): 403-415.

Bi X, Liu LF. 1996. DNA rearrangement mediated by inverted repeats. *Proc Natl Acad Sci U S A* **93**(2): 819-823.

Blake RD, Hess ST, Nicholsontuell J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol* **34**(3): 189-200.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* **13**(2): 137-144.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**(6916): 666-669.

Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lennington J, Li Z et al. 2003. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* **165**(1): 367-386.

Callen BP, Shearwin KE, Egan JB. 2004. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol Cell* **14**(5): 647-656.

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Scheix T et al. 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc Natl Acad Sci U S A* **103**(47): 18026-18026.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**(10): 956-963.

Chen M, SanMiguel P, de Oliveira AC, Woo SS, Zhang H, Wing RA, Bennetzen JL. 1997. Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proc Natl Acad Sci U S A* **94**(7): 3431-3435.

Choisne N Demange N Orjeda G Samain S D'Hont A Cattolico L Pelletier E Couloux A Segurens B Wincker P et al. 2005. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol* **3**: -.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. *Bioinformatics* **20**(3): 426-427.

Clark LG, Zhang WP, Wendel JF. 1995. A phylogeny of the grass family (Poaceae) based on ndhF - sequence data. *Syst Bot* **20**(4): 436-460.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**(7410): 213-217.

Davila Lopez M, Martinez Guerra JJ, Samuelsson T. 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One* **5**(5): e10654.

Devos KM. 2005. Updating the 'crop circle'. *Curr Opin Plant Biol* **8**(2): 155-162.

Devos KM, Beales J, Nagamura Y, Sasaki T. 1999. Arabidopsis-rice: will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Res* **9**(9): 825-829.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12**(7): 1075-1079.

Devos KM, Pittaway TS, Reynolds A, Gale MD. 2000. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. *Theor Appl Genet* **100**(2): 190-198.

Devos KM, Wang ZM, Beales J, Sasaki Y, Gale MD. 1998. Comparative genetic maps of foxtail millet (*Setaria italica*) and rice (*Oryza sativa*). *Theor Appl Genet* **96**(1): 63-68.

Dooner HK, He L. 2008. Maize genome structure variation: Interplay between retrotransposon polymorphisms and genic recombination. *Plant Cell* **20**(2): 249-258.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annu Rev Genet* **42**: 443-461.

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X et al. 2002. Sequence and analysis of rice chromosome 4. *Nature* **420**(6913): 316-320.

Feuillet C, Keller B. 2002. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Ann Bot (Lond)* **89**(1): 3-10.

Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM. 2009. Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS One* **4**(10).

Franck E, Hulsen T, Huynen MA, de Jong WW, Lubsen NH, Madsen O. 2008. Evolution of closely linked gene pairs in vertebrate genomes. *Mol Biol Evol* **25**(9): 1909-1921.

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn* **4**: 25-40.

Fukuoka Y, Inaoka H, Kohane IS. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* **5**(1): 4.

Gandolfo M, Nixon K, Crepet W. 1998. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). *Am J Bot* **85**(7): 964.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. *New Phytol* **154**(1): 15-28.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**(5565): 92-100.

Gojobori T, Li WH, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol* **18**(5): 360-369.

Grant D, Cregan P, Shoemaker RC. 2000. Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc Natl Acad Sci U S A* **97**(8): 4168-4173.

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A. 1981. Sequence specificity of methylation in higher plant DNA. *Nature* **292**(5826): 860-862.

Hampf M, Gossen M. 2007. Promoter crosstalk effects on gene expression. *J Mol Biol* **365**(4): 911-920.

Hansen JJ, Bross P, Westergaard M, Nielsen MN, Eiberg H, Borglum AD, Mogensen J, Kristiansen K, Bolund L, Gregersen N. 2003. Genomic structure of the human mitochondrial chaperonin genes: *HSP60* and *HSP10* are localised head to head on chromosome 2 separated by a bidirectional promoter. *Hum Genet* **112**(1): 71-77.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* **16**(10): 1252-1261.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PloS Genet* **6**(9).

Higgins DG, Thompson JD, Gibson TJ. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**: 383-402.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**(11): 756-766.

Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**(7421): 497-501.

Hurst LD, Williams EJB, Pal C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet* **18**(12): 604-606.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* **101**(39): 13994-14001.

Ilic K, SanMiguel PJ, Bennetzen JL. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci U S A* **100**(21): 12265-12270.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**(7008): 569-573.

Jin YK, Bennetzen JL. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* **6**(8): 1177-1186.

Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. 2005. Regulated chromatin domain comprising cluster of co-expressed genes in *Drosophila melanogaster*. *Nucleic Acids Res* **33**(5): 1435-1444.

Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**(4): 1651-1659.

Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S. 2000. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res* **7**(6): 323-330.

Keightley PD, Eory L, Halligan DL, Kirkpatrick M. 2011. Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. *Genetics* **187**(4): 1153-1161.

Kirik A, Salomon S, Puchta H. 2000. Species-specific double-strand break repair and genome evolution in plants. *EMBO J* **19**(20): 5562-5566.

Krom N, Ramakrishna W. 2008. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, *Arabidopsis*, and *Populus*. *Plant Physiol* **147**(4): 1763-1773.

Krom N, Ramakrishna W. 2010. Conservation, rearrangement, and deletion of gene pairs during the evolution of four grass genomes. *DNA Res* **17**(6): 343-352.

Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, Llaca V, Tanyolac B, Park YJ, Jeong OY, Bennetzen JL et al. 2004. Gene loss and movement in the maize genome. *Genome Res* **14**(10A): 1924-1931.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CMR, D'Hont A, Garsmeur O et al. 2008. Insights into the *Musa* genome: Syntenic relationships to rice and between *Musa* species. *BMC Genomics* **9**: 58.

Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX. 2006. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* **2**(7): e74.

Lin CT, Lin WH, Lyu YL, Whang-Peng J. 2001. Inverted repeats as genetic elements for promoting DNA inverted duplication: implications in gene amplification. *Nucleic Acids Res* **29**(17): 3529-3538.

Liu H, Sachidanandam R, Stein L. 2001. Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order. *Genome Res* **11**(12): 2020-2026.

Liu R, Bennetzen JL. 2008. Enchilada redux: how complete is your genome sequence? *New Phytol* **179**(2): 249-250.

Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL. 2007. A GeneTrek analysis of the maize genome. *Proc Natl Acad Sci U S A* **104**(28): 11844-11849.

Liu X, Han B. 2009. Evolutionary conservation of neighbouring gene pairs in plants. *Gene* **437**(1-2): 71-79.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Assocs., Inc., Sunderland, MA.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101**(34): 12404-12410.

Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14**(5): 860-869.

Ma J, Wing RA, Bennetzen JL, Jackson SA. 2007. Plant centromere organization: a dynamic structure with conserved functions. *Trends Genet* **23**(3): 134-139.

Ma JX, Bennetzen JL. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci U S A* **103**(2): 383-388.

Ma JX, Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res* **16**(2): 251-259.

Magallon S, Castillo A. 2009. ANGIOSPERM DIVERSIFICATION THROUGH TIME. *Am J Bot* **96**(1): 349-365.

McClintock B. 1947. Cytogenetic studies of maize and *Neurospora*. *Carnegie Inst Washington Year Book* **46**: 146-152.

Moore G, Devos KM, Wang Z, Gale MD. 1995. Grasses, line up and form a circle. *Curr Biol* **5**(7): 737-739.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1): 297-304.

Noe L, Kucherov G. 2004. Improved hit criteria for DNA local alignment. *BMC Bioinformatics* **5:** 149.

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L et al. 2007. The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res* **35**: D883-D887.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**(7229): 551-556.

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. 2012. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol* **10**: 84.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA et al. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* **16**(10): 1262-1269.

Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162**(3): 1389-1400.

Rogozin IB, Makarova KS, Wolf YI, Koonin EV. 2004. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. *Brief Bioinform* **5**(2): 131-149.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**(1): 43-45.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**(5288): 765-768.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K et al. 2008. Genome structure of the legume, *Lotus japonicus*. *DNA Res* **15**(4): 227-239.

Schnable PS Ware D Fulton RS Stein JC Wei F Pasternak S Liang C Zhang J Fulton L Graves TA et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956): 1112-1115.

Semon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* **23**(9): 1715-1723.

Shearwin KE, Callen BP, Egan JB. 2005. Transcriptional interference - a crash course. *Trends Genet* **21**(6): 339-345.

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res* **10**(7): 908-915.

Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* **22**(3): 767-775.

Swigonova Z, Lai JS, Ma JX, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res* **14**(10A): 1916-1923.

Tatusova TA, Madden TL. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174**(2): 247-250.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16**(7): 934-946.

Thomas BC, Rapaka L, Lyons E, Pedersen B, Freeling M. 2007. Arabidopsis intragenomic conserved noncoding sequence. *Proc Natl Acad Sci U S A* **104**(9): 3348-3353.

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res* **19**(12): 2221-2230.

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc Natl Acad Sci U S A* **96**(13): 7409-7414.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. *Science* **290**(5499): 2114-2117.

Vitte C, Bennetzen JL. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc Natl Acad Sci U S A* **103**(47): 17638-17643.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan M. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**(7282): 763-768.

Wang H, Bennetzen JL. 2012. Centromere retention and loss during the descent of maize from a tetraploid ancestor. *Proc Natl Acad Sci U S A* **109**(51): 21004-21009.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**(3): 1753-1763.

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum L.*) reveals multiple mechanisms of genome evolution. *Plant J* **26**(3): 307-316.

Williams EJ, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res* **14**(6): 1060-1067.

Yang L, Bennetzen JL. 2009. Structure-based discovery and description of plant and animal *Helitrons*. *Proc Natl Acad Sci U S A* **106**(31): 12832-12837.

Young ND Debelle F Oldroyd GED Geurts R Cannon SB Udvardi MK Benedito VA Mayer KFX Gouzy J Schoof H et al. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**(7378): 520-524.

Zhao Z, Boerwinkle E. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12**(11): 1679-1686.