TRANSPOSABLE ELEMENT DIVERSITY, DIVERGENCE, AND CONTRIBUTION TO GENOME EVOLUTION IN PLANTS

Dhanushya Ramachandran

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
Ramachandran, Dhanushya, "TRANSPOSABLE ELEMENT DIVERSITY, DIVERGENCE, AND CONTRIBUTION TO GENOME EVOLUTION IN PLANTS" (2018). Graduate Theses, Dissertations, and Problem Reports. 7236.
https://researchrepository.wvu.edu/etd/7236

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
TRANSPORTABLE ELEMENT DIVERSITY, DIVERGENCE, AND CONTRIBUTION TO
GENOME EVOLUTION IN PLANTS

Dhanushya Ramachandran

Dissertation submitted to the Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Biology

Jennifer Hawkins, Ph.D., Chair
Stephen DiFazio, Ph.D.
Jennifer Gallagher, Ph.D.
Dana Huebert Lima, Ph.D.
Vagner Benedito, Ph.D.

Department of Biology

Morgantown, West Virginia
2018

Keywords: Transposable elements, genome evolution, plants, polyploidy, copy number, next-generation sequencing

Copyright 2018 Dhanushya Ramachandran
ABSTRACT
Transposable element diversity, divergence, and contribution to genome evolution in plants

Dhanushya Ramachandran

Transposable elements (TEs) are mobile genetic elements that comprise a large portion of eukaryotic genomes. With the capacity to self-replicate and reinsert into the host organism’s genome without necessarily providing immediate benefit to the host, TEs have historically been characterized as parasitic, selfish, or junk sequences; however, recent findings suggest that TEs and their associated activity may be powerful drivers of organismal development and adaptation. TE movement creates large insertions and deletions, and such indels in or near genic sequences can affect the gene’s function in myriad ways. The goals of my dissertation work involve describing quantitative and qualitative variation in TE composition at both recent and distant evolutionary timescales. I designed my research objectives to evaluate TE diversity and to compare the evolutionary pathways that induce TE mobility in related but independently domesticated and economically important plants such as sorghum, maize and cotton. Given the abundance, activity, and repetitive nature of mobile elements in these particular plant lineages, it has been challenging to characterize and analyze the TE component from short-read sequence data. In my first study, I developed a method that overcomes this difficulty. Using this method, I determined TE composition and copy number diversity among Sorghum accessions and detected even small differences in TE copy number within and between genomes that diverged as little as 2 million years ago. Therefore, frequent TE proliferation over short evolutionary timescales suggests a significant contribution of TEs in driving intraspecific genome divergence. Following this study, I exploited a larger clade within the grass family to describe interspecific TE diversity among related members that have shared a recent polyploidization event. This phylogenetically informed experimental design allowed detailed insight into TE dynamics in association with genome merger and doubling in both wild species and their economically important cultivated relatives. Additional analyses revealed increased abundance and biased insertion of copia elements near genes involved in various biological processes in the cultivated taxa that had experienced recent genome doubling. The lack of similar insertions near genes in the diploid S. bicolor demonstrates the impact of allopolyploidization and TE activity on genome evolution and domestication. To determine if these observations related to genome doubling and domestication are specific to grasses, I evaluated TE dynamics in five cotton (Gossypium)
allopolyploids, two of which are economically important cultivars. By including both A and D diploid parental genomes, my analyses reveal recent insertions in *G. hirsutum* (AD) that are absent in both diploid progenitors, indicating TE insertions in gene regions post allopolyploidization, similar to findings in maize. Overall, my findings suggest that TE dynamics in species with a history of genome doubling display similar patterns, particularly with respect to gene-proximal TE insertions. Duplicate copies of genes in polyploids may provide stability by creating a particularly well-buffered environment from the deleterious consequences of transposition. The enrichment of TE insertions in gene-rich regions specifically in polyploid domesticates suggests that insertional mutagenesis near functionally important genes may provide evolutionary novelty upon which selection can act during the domestication process.
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my doctoral advisor Dr. Jennifer Hawkins. I met her for the first time in 2010 during the graduate recruitment weekend where we discussed her on-going research topics. One of the interesting topics was about transposable elements and I did not know much about it except the name “jumping genes”. Despite my limited knowledge about transposable elements, I was quite interested to learn more about this after an engaging discussion with Dr. Hawkins where I eventually ended up in my PhD journey under her guidance. I should definitely mention here that I learned the value of thought provoking science from her. The enthusiasm and excitement she has for her research is something I have always admired. Her rich knowledge, friendly personality and fair decisions at right times makes her one of the best. Her invaluable suggestions at many occasions guided me to choose the correct decisions. I would like to convey my sincere thanks for her constant support as a mentor at WVU and more importantly, I am always grateful to her for having confidence in me to achieve this stage.

I am also thankful to my graduate committee members - Dr. Stephen DiFazio, Dr. Jennifer Gallagher, Dr. Dana Huebert Lima, and Dr. Vagner Benedito for generously offering their time and valuable feedback throughout my graduate studies. I am particularly appreciative to Dr. Stephen DiFazio for his questions, ideas, and suggestions on the computational part of my research.

My sincere thanks to my lab-mates for all the fun and support. I would like to thank Mike for his support in form of intellectual discussions and aiding in understanding basic aspects of genomics and computational biology, especially during my early years of graduate school. Many thanks to Jazz for continual support and for being a wonderful friend throughout these years. Thanks to Ashley and Melissa for making the workplace lively and friendly. I also want to acknowledge Raj for sharing knowledge. Every one of them had a significant contribution in completing my research work.

Many thanks to Mickey, Judy, and Chuck for their constant support in administrative related work. Always I owe my thanks to other graduate students in the department, both past and present, who helped me out at many occasions. Special mention goes to Apoorva and
Michael, I deeply cherish those enjoyable lunches, weekend coffee-breaks and late-night dinners that we spent together during our stressful times. Thank you for your invaluable friendship and endless discussion about all kinds of topics.

Finally, thanks must go to my family. My Mom and Dad for love and inspiration throughout my life; my sister for her humor, encouragement and concern for me. And most of all I owe my heartfelt thanks to a special person, my husband Arun, whose unconditional love, patience, and continuous support of my academic endeavors over the past several years enabled me to complete this dissertation.
# TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... ii

ACKNOWLEDGEMENTS ...................................................................................................... iv

TABLE OF CONTENTS ......................................................................................................... vi

LIST OF TABLES ................................................................................................................ viii

LIST OF FIGURES .............................................................................................................. ix

CHAPTER 1: OVERVIEW AND OBJECTIVES ........................................................................ 1

References ........................................................................................................................... 3

CHAPTER 2: INTRODUCTION .............................................................................................. 4

The Discovery of TEs ......................................................................................................... 4

Major TE classes and their genetic structure .................................................................... 4

The C-value paradox and TE contribution to genome size disparity ............................. 5

Epigenetic regulation of TE activity .................................................................................. 7

Stress induced TE activation .............................................................................................. 8

Functional consequences of TE insertions/reactivation .................................................. 10

Study Systems .................................................................................................................... 11

References ........................................................................................................................... 14

CHAPTER 3: METHODS FOR ACCURATE QUANTIFICATION OF LTR-RETROTRANSPOSON COPY NUMBER USING SHORT-READ SEQUENCE DATA: A CASE STUDY IN SORGHUM ........................................................................... 25

Abstract .............................................................................................................................. 25

Introduction .......................................................................................................................... 26

Materials and methods ....................................................................................................... 28

Results .................................................................................................................................. 32

Discussion ........................................................................................................................... 37

Acknowledgements ............................................................................................................ 43

References ........................................................................................................................... 44

Supplemental – Online resources ..................................................................................... 57

CHAPTER 4: EVOLUTIONARY DYNAMICS OF TRANSPOSABLE ELEMENTS FOLLOWING A SHARED POLYPLOIDIZATION EVENT IN THE TRIBE, ANDROPOGONEAE ......................................................................................... 58

Abstract .............................................................................................................................. 58

Introduction ........................................................................................................................... 59
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials and methods</td>
<td>........................................................................</td>
<td>60</td>
</tr>
<tr>
<td>Results</td>
<td>........................................................................</td>
<td>64</td>
</tr>
<tr>
<td>Discussion</td>
<td>........................................................................</td>
<td>69</td>
</tr>
<tr>
<td>References</td>
<td>........................................................................</td>
<td>73</td>
</tr>
<tr>
<td>Supplementals</td>
<td>........................................................................</td>
<td>88</td>
</tr>
<tr>
<td><strong>CHAPTER 5:</strong> COMPARATIVE ANALYSIS OF TRANSPOSABLE ELEMENTS IN GOSSYPIUM DIPLOIDS AND POLYPLOIDS PROVIDES INSIGHT INTO POLYPLOIDY AND CROP DOMESTICATION</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td>........................................................................</td>
<td>89</td>
</tr>
<tr>
<td>Introduction</td>
<td>........................................................................</td>
<td>90</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>........................................................................</td>
<td>91</td>
</tr>
<tr>
<td>Results</td>
<td>........................................................................</td>
<td>93</td>
</tr>
<tr>
<td>Discussion</td>
<td>........................................................................</td>
<td>96</td>
</tr>
<tr>
<td>References</td>
<td>........................................................................</td>
<td>100</td>
</tr>
<tr>
<td>Supplementals</td>
<td>........................................................................</td>
<td>109</td>
</tr>
<tr>
<td><strong>CHAPTER 6:</strong> GENERAL CONCLUSIONS</td>
<td>........................................................................</td>
<td>112</td>
</tr>
<tr>
<td>References</td>
<td>........................................................................</td>
<td>115</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Flow cytometry results for *Sorghum bicolor* and *S. propinquum*. 49

Copy number estimates for the total number of LTRs from short read datasets using both reference-based and *de novo* approaches. 51

Global repeat composition (in Mb) of species with respect to genome size. 78

Global repeat composition (in Mb) of species with respect to genome size. 103

Proximity of retrotransposons near genes. Total number of Illumina paired reads of which one mapped to an LTR-gypsy and the other onto a *G. hirsutum* reference gene model. 103
LIST OF FIGURES

Major classes of transposable elements and examples of insertional mutagenesis. 23

Transposable element content with respect to genome size in grasses. 24

The LTR retrotransposon families of *S. bicolor* BTx623. 52

LTR-retrotransposon dating in *S. bicolor* BTx623. 53

LTR copy number estimates for the ten largest families calculated using a reference-based approach. 54

Copy number estimates for contigs containing LTRs for the ten largest LTR-retrotransposon families via *de novo* analysis. 55

Cladogram indicating the evolutionary relationship of selected grass species, indicating polyploidization and species divergence. 81

Bar graph showing distribution of 248 largest clusters with respect to various repeat families. 82

Comparative graph-based clustering of repetitive analysis in eight species. 83

Activity of retroelements in all eight species pre- and post-divergences of the *Zea-Tripsacum* clade (yellow line) for 15 retrotransposon clusters (shaded gray). 84

Number of paired end reads mapped to one copia element and to nearby gene in *Z. mays*. 85

Evolutionary framework for diploid and allotetraploid cotton. 104

Estimation of total repeat abundance (Mb) in each *Gossypium* genome and repeat composition in diploids and allotetraploids of cotton. 105

UpSet plots showing the interactions of shared repeat clusters among *Gossypium* genotype. 106

GO categories representing genes that were mapped near LTR-gypsy and copia retrotransposons in *G. hirsutum*. Blast hits for *G. hirsutum* genes that were mapped near gypsy and copia contigs in *G. hirsutum* [AD] and its postulated diploid ancestors. 107
CHAPTER 1: OVERVIEW AND OBJECTIVES

Transposable elements (TEs) are mobile genetic elements that can move from one genomic position to another, and as the resultant insertions are quite large, mobilization can lead to mutations of high impact. For example, a TE insertion into the 5’ region of a gene can function as a new regulatory sequence and lead to changes in gene expression, or it can inactivate the gene by disrupting the existing cis-regulatory controls or altering chromatin conformation. Similarly, TE insertions into introns can be incorporated as new exons or lead to premature/truncated transcripts. Although TEs are predominantly silenced, protecting the genome from rampant insertional mutagenesis, the host suppression system can be circumvented in some situations, particularly those situations that present environmental challenges. Indeed, it has been suggested that species or populations that are prone to strong diversifying selection would benefit from such TE-driven genome variability (Naito et al. 2009, Fernandez et al. 2010, Tenaillon et al. 2010, Linquist et al. 2013, Oliver et al. 2013, Vitte et al. 2014, Ong-Abdullah et al. 2015, Lu et al. 2017). In other words, transposition may create genomic diversity with the potential to promote plant adaptation and survival.

The purpose of my dissertation work is to describe quantitative and qualitative variation in TE composition at both recent and distant evolutionary timescales. I have performed three specific aims to evaluate TE diversity in the related but independently domesticated and economically important grasses, maize and sorghum. Further, I extended these analyses into a non-grass system to investigate TE dynamics in response to genome doubling and domestication.

Specific Aim 1 – Determine TE composition and copy number diversity among Sorghum accessions to evaluate intraspecific TE-associated genome evolution over short evolutionary time scales. Though numerous studies have described interspecific TE-associated variation, relatively little is known about TE variation among members of a single species. In this study, I evaluated TE composition and copy number diversity within several representatives of Sorghum bicolor and compared those results to that of Sorghum propinquum, a close wild relative, using short-read sequence data.
Specific Aim 2 – Describe interspecific TE diversity among related members of the Andropogoneae tribe that have undergone a shared recent polyploidization event. Here, I evaluated TE diversity in *Zea mays*, *Tripsacum dactyloides* and *Urelytrum digitatum*, and compare these results with that from *Sorghum*. *Z. mays* and *T. dactyloides* are paleopolyploids that arose from a single genome doubling event involving *U. digitatum* approximately 5-10 million years ago (mya) and have since undergone rediploidization. Comparisons of TE evolutionary dynamics in various *Zea-Tripsacum* species along with closely related diploid species *Urelytrum* and *Sorghum* revealed existing variation in repeat content between pre- and post-polyploid species included in the study.

Specific Aim 3 – Describe interspecific TE diversity in the cotton genus, *Gossypium*, to evaluate TE-associated genome evolution in domesticated polyploids compared to their diploid progenitors. To investigate whether TE dynamics and polyploid-domesticate phenomenon is a common theme for all species, I performed interspecific TE diversity analyses similar to my specific aim 2 in a non-grass system, cotton (*Gossypium*). To characterize both common patterns and lineage-specific differences in TE composition and evolution in response to genome doubling, I have investigated *Gossypium* allopolyploids (AD), which includes leading domesticated cash crops and island-endemics, originating from a single interspecific hybridization event between two divergent diploid (A and D) species.
References


CHAPTER 2: INTRODUCTION

The Discovery of TEs

Historically, genes were thought of as stable entities whose linear-arrangement was highly conserved across diverse organisms; however, Barbara McClintock challenged this view in the late 1940s. McClintock was primarily interested in understanding the mechanisms of chromosome breakage and fusion in maize, and through her studies she observed a frequent breakage event at a specific locus on chromosome 9 during plant development. She named the locus Dissociation (Ds), and further, she discovered that this locus could change its position on the chromosome. She also discovered an unlinked factor called Activator (Ac) that seemingly initiated the break at the Ds locus on chromosome 9 (McClintock 1948, 1951). This work led to the discovery of the first transposable elements, the Ac/Ds system. McClintock referred to these genetic factors as “controlling elements” due to their ability to alter gene function when inserted near or within genic regions; however, the concept of transposition was considered complex at that time, and was largely dismissed due to strong adherence to the stable genome hypothesis. It took two decades for McClintock to receive the prestigious Nobel Prize for her discovery of transposable elements.

Major TE classes and their genetic structure

Since McClintock’s discovery, many different types of TEs have been discovered and described, and they are now classified into two major groups based on their mode of transposition, e.g., whether their transposition intermediate is RNA (Class I – retrotransposons) or DNA (Class II – DNA transposons) (Figure 1). Both classes contain autonomous and non-autonomous elements. Autonomous elements contain open reading frames (ORFs) that encode the proteins necessary for transposition. Non-autonomous elements lack these ORFs, and therefore require proteins encoded by their autonomous counterpart to move throughout the genome. Class II elements, which transpose via a cut-and-paste mechanism, contain a transposase gene that is flanked by two terminal inverted repeats (TIRs). Examples include Ac/Ds and Spm/dspm (En/I) in maize. The inverted terminal repeats are unique to each DNA transposon family and range in size from 11 bp to a few hundred base pairs. Proteins encoded by the autonomous elements of one DNA transposon family recognize the corresponding non-autonomous element by their shared
termini, and thereby assist in its transposition. For example, Ac and Ds elements share 11 bp TIRs and therefore the Ac-encoded transposase can interact with the Ds TIRs to facilitate its movement. The Ac-encoded transposase cannot bind the TIRs of Spm/dspm elements or any other TE family for which it does not share TIR sequence identity.

Class I retrotransposons are RNA elements that are particularly abundant in eukaryotes, especially in plants. Retrotransposons are divided into Long Terminal Repeat (LTR) and non-LTR elements. LTR-retrotransposons vary in size from several hundred base pairs to ~20 kb and contain unique, directly-oriented LTRs at their ends that range in size from ~100 bp to 5 kb. An intact element contains gag, an ORF that encodes a polyprotein responsible for maturation and packaging of retrotransposon RNA, and pol, an ORF that encodes protease (PR), reverse transcriptase (RT), RNase H (RH) and integrase (IN) which aid in synthesis, processing and subsequent integration of new retrotransposons at random locations in the genome (Feschotte 2002). LTR-retrotransposons are classified into Gypsy (RT-RH-IN) and Copia (IN-RT-RH) superfamilies based on the arrangement of RT and IN in the pol ORF (Xiong and Eickbush, 1990). Though several non-autonomous LTR-retrotransposon families have been identified in plants, their autonomous partners have yet to be discovered in most cases (Jin and Bennetzen, 1989; Hu et al. 1995; Lander et al. 2001; Witte et al. 2001; Jiang et al. 2002; Kalendar et al. 2004; Kejnovsky et al. 2006). Non-LTR retrotransposons such as Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) are also found in flowering plant genomes but mostly remain epigenetically silenced and therefore inactive. Although LINEs encode the proteins for their own transposition, SINEs are non-autonomous and therefore depend on the machinery encoded by other retrotransposons.

The C-value paradox and TE contribution to genome size disparity

One of the largest genomes known belongs to the single celled Amoeba dubia (670 billion bp) whose genome is ~200 times larger than that of most multi-cellular organisms. Such lack of correlation between genome size and organismal complexity has been historically termed the C-value paradox (Thomas, 1971). Plant genomes are extremely diverse and exhibit a wide range of C-values (Pellicer et al. 2010, Fedoroff,
2012). It is now known that the two major mechanisms that contribute to genome expansion in plants are genome doubling (polyploidy) and transposable element accumulation (Grover & Wendel 2010, Kejnovsky et al. 2012, Leitch & Leitch 2012). LTR-retrotransposons are predominant in most plant genomes, comprising 50-80% of the maize genome (~2500 Mb) and more than 90% of the bread wheat genome (SanMiguel 1998, Paux 2008, McCarthy 2002). Although TE proliferation occurs frequently and in large detectable bursts, accumulation is counteracted by DNA removal mechanisms such as non-homologous end joining (NHEJ) and unequal homologous recombination. The former excises parts of TEs by generating small deletions, and the latter often occurs between two LTRs of an intact retroelement, removing the internal portion and resulting in a solo LTR. In plants with smaller genomes, DNA removal occurs more frequently via illegitimate recombination (NHEJ) than through unequal homologous recombination (Hawkins 2009, Ma J 2004, Devos KM 2002). It is also thought that smaller genomes purge excess DNA more efficiently relative to larger genomes due to extensive epigenetic silencing (Ibarra-Lacette et al. 2013, Nystedt et al. 2013, Zhong X et al. 2012, Tian Z et al. 2009). Such efficiency could have evolved to decrease the likelihood of a TE insertion near a gene, as such an event is statistically more likely in plants with smaller genomes.

Rapid rates of TE amplification and decay can lead to variable TE composition among closely related plant lineages. For instance, although similar numbers of TE families occupy both the rice and maize genomes, transpositional bursts of a few Long Terminal Repeat (LTR) retrotransposon families have inflated the maize genome to six times that of rice (Baucom 2009, Baucom2008). Variation is also common among more closely related taxa, such as those that belong to the same subfamily (SanMiguel et al. 1998), and significant copy number differences are found even among members of a single genus. Interspecific comparisons performed in Gossypium and Oryza revealed that proliferation of a small subset of TE families were responsible for the observed genome size variation among species (Hawkins et al. 2006, Piegu et al. 2006). TE copy number variation is not always due to the activity of just a few families, however. Proliferation of several different TE families is responsible for the variation in TE content between Zea and Arabidopsis (Tenaillon et al. 2011, Hu et al. 2011, Hollister et al. 2011). In either case, TE accumulation and deletion occurs over very short evolutionary time-scales,
particularly for organisms with shorter life-histories, often contributing to lineage-specific profiles that are therefore phylogenetically uninformative (Hawkins et al. 2008).

**Epigenetic regulation of TE activity**

Although TEs occupy more than 50% of many eukaryotic genomes, the majority are silenced via epigenetic mechanisms. The most well studied epigenetic marks include DNA methylation, several types of histone modification and small interfering RNA (siRNA) mediated silencing. DNA methylation involves the addition of methyl groups at cytosine residues (in eukaryotes) and occurs in all sequence contexts (CG, CHG and CHH) in plants. Subsequent to McClintock’s discovery that sequential breakage and joining of chromosomes during maize development resulted in activation of TEs (class II – Ac and Spm), research demonstrated that the reversible status of these elements was associated with DNA methylation (McClintock B. 1958, McClintock B. 1965, Fedoroff N, et al. 1995, Pan YB and Peterson PA 1988). Later studies in *Arabidopsis* showed that several classes of TEs become hypomethylated and reactivated in lines that are mutant for methyltransferase enzymes (Miura et al. 2001, Singer et al. 2001, Lippman et al. 2003 & 2004, Tsukahara et al. 2009). Recently, a striking example for epigenetic derepression of a TE associated with a deleterious phenotype was found in oil palm (Ong-Abdullah et al. 2015). Hypomethylation of a *Karma* TE insertion within an intron of the gene MANTLED provided an alternative splice site and a premature termination signal, resulting in deformed oil palm fruits. Recent high-throughput silencing studies demonstrate that TEs are enriched for DNA methylation in all three sequence contexts (Lister et al. 2008, Zhang et al. 2006). In addition to DNA methylation, transposon activity is often controlled by histone modifications (Bernatavichute et al. 2008). Various modifications such as methylation, deacetylation, and biotinylation on histone amino-terminal tails impact TE transcription. In plants, both H3K9me2 and H3K27me1 contribute to transcriptional silencing of some TEs in *Arabidopsis* and rice (Ebbs et al. 2005, Jacob et al. 2009 & 2010, Ding et al. 2007, Qin et al. 2010, Mirouze et al. 2009, Zhang et al. 2003). In *Arabidopsis*, mutations in the histone deacetylase gene results in transcriptional activation of several TE families (Lippman et al. 2003). Clearly, the presence or absence of these histone modifications plays an important role in TE silencing.
In addition to these pre-transcriptional mechanisms, TE activity is controlled post-transcriptionally via RNA silencing. Here, the host silencing system identifies RNA molecules that were produced by transposons. TE transcripts are converted into double-stranded RNAs (dsRNAs), which are further processed into small RNA (siRNA) molecules that range from 21 to 35 nucleotides in length. Higher plants evolved specific DNA-dependent RNA polymerases such as RNA polymerase IV and V to produce the initial RNA transcripts for RNA silencing and siRNA induced methylation, respectively (Ream et al. 2009). Therefore, being dynamic and potentially deleterious components of the genome, evolutionary forces have produced various epigenetic ways to facilitate TE silencing.

**Stress induced TE activation**

Given the diversity in TE content among closely related taxa, it seems apparent that transposition occurs despite the presence of epigenetic silencing mechanisms. Indeed, TEs can become activated under stressful conditions, including both genomic and environmental stresses (Grandbastien M-A et al 1989, Pouteau S et al 1991). Environmental stresses include both biotic (microbial infection, wounding etc.,) and abiotic (dehydration, UV radiation) factors that can activate silenced TEs (Wessler 1996; Kalendar et al. 2000; Grandbastien et al. 2005; Ramallo et al. 2008). New TE insertions into coding regions are often deleterious; nevertheless, TE repression mechanisms are often circumvented by biotic and abiotic stress factors (Mhiri et al. 1997; Grandbastien et al. 1998; Takeda 1998; Pecinka 2010; Tittel-Elmer et al. 2010; Fujino et al. 2011; Cavrak et al. 2014; Makarevitch et al. 2015; Finatto et al. 2015). Ac/Ds transposons in maize (Steward et al. 2000), Tam 3 in *Antirrhinum majus* (Hashida et al. 2006), and mPing DNA transposons in rice (Naito et al. 2006) are examples of TEs that are activated in response to cold stress. A specific strain of rice, EG4 (cultivar Gimbozu), contains more than 1,000 active mPing elements when subjected to cold or salt stress (Naito et al. 2006). In *Arabidopsis*, a copia LTR-retrotransposon, ONSEN, is activated under heat stress (Ito et al. 2016). It has since been discovered that ONSEN acquired a heat responsive element, resulting in transcription and production of full-length extrachromosomal DNA copies under higher temperatures (Cavrak et al. 2014). In natural populations, stress may increase TE amplification, as has been shown in wild
barley. Plants grown in drier regions of “evolution canyon” in Israel showed three-fold variation in copy number of a specific retrotransposon, BARE-1, compared to plants grown in other regions of the canyon (Kalendar et al. 2000). This suggests that a stress-sensitive regulatory sequence in the BARE-1 promoter could have activated the retrotransposon in these plants compared to unstressed plants. A recent report identified a mutational event associated with a TE insertion that gave rise to industrial melanism in the English peppered moth (Van’t Hoff et al. 2016). This study found an intronic insertion of a TE that enhances expression of the cortex gene, and this over-expression underlies the adaptive coloration in these moths that occurred during the industrial revolution.

In addition to these environmental stresses, TEs can become activated by genomic stresses (Baack et al. 2005, Noor and Chang 2006). Hybridization and polyploidization, examples of genomic shock in response to the reunion of two divergent genomes in a single nucleus, frequently release TEs from their silenced state (McClintock 1984, Ungerer, 2006, Shan et al 2005, Madlung et al 2005). Although the precise mechanism(s) that induces TE mobility in hybrids and polyploids is unclear, it is obvious that these phenomena lead to global changes in DNA methylation, resulting in transcriptionally active TEs (Liu et al. 2004). A striking example is that of three hybrid sunflower species that independently arose from the same two parental species, and in which the hybrid genomes are 50% larger (~1,130 Mb additional DNA, mostly TEs) than either of the parental genomes (Ungerer et al. 2006). Similarly, studies in polyploid Spartina, wheat, and other plant species have reported transcriptional activation of TEs upon polyploidization due to reduced cytosine methylation (Parisod et al. 2009, Kashkush et al. 2002, Slotkin and Martinussen 2007, Parisod et al. 2009, Madlung et al. 2002, Salmon et al. 2005, Xu et al. 2009, DeFraia and Slotkin 2014). It is speculated that such TE reactivation in response to hybridization and polyploidy could be due to incompatible suppression machinery between the two donor genomes, or that unknown mechanisms are in place that reduce genomic methylation under general stress conditions (Lisch D, 2009). At any rate, it seems likely that epigenetic regulation plays a major role in TE reactivation during hybridization and polyploidy.
Functional consequences of TE insertions/reactivation

TE mobilization can create myriad genetic changes including the creation of alternative expression forms by providing genes with promoters or enhancers, causing gene movement to novel chromosomal locations, and through TE domestication. Regulatory mutations in particular are likely to arise via TE insertions near protein-coding genes due to the presence of regulatory modules within the TE sequence itself (Bennetzen 2014). An example of this is the insertion of an intracisternal A particle (IAP) element in the mouse agouti locus that resulted in mice with altered coat color, obesity and diabetes (Morgan et al. 1999). A similar effect has also been shown for TE insertions in plants belonging to TE families such as Mutator, Helitron, CACTA, and Harbinger (Jiang et al. 2004, Paterson et al. 2009, Vogel et al. 2010, Jin and Bennetzen 1994, Kashkush and Khasdan 2007, Kashkush et al. 2003, Butelli et al. 2012). A genome-wide survey of TE-altered gene expression revealed ~2,000 transposon insertional polymorphisms between two rice cultivars (japonica Nipponbare and indica 93-11), of which 10% of these polymorphisms were located in expressed genic regions and induced a series of genetic differences between two rice subspecies, suggesting that these insertions are an important source of intraspecific genetic variation (Huang et al 2008).

TEs can also move protein coding genes from one location to another within a genome, as has been discovered via synteny studies. For example, several studies have shown that non-collinear genes are often associated with TEs due to the acquisition of these gene fragments for use as “filler sequences” during the double strand break (DSB) repair process (Wicker et al. 2010, Baucom et al. 2009, Jiang et al. 2004, Morgante et al. 2005). TE sequences can also be "domesticated" as exons of new chimeric genes by fusing with nearby coding sequences. Daysleeper, FAR1, and FYH3 transcription factors are some of the genes involved in Arabidopsis development, which were originally derived from the transposase gene of class II elements (Bundock and Hooykaas 2005, Hudson et al. 2003, Lin RC, et al. 2007). Therefore, although TE activity might affect genomic integrity, TE driven genome variability may lead to adaptation upon which selection can act (Martin et al. 2009, Naito et al. 2009, Lockton and Gaut 2010, Fernandez et al. 2010, Tenaillon et al. 2010, Linquist et al. 2013, Oliver et al.2013, Vitte et al. 2014).
Study Systems

For the projects in this dissertation, I have used study systems from the plant families Poaceae and Malvaceae. Below is a general overview of each family. Specific information on the species studied and their relevance to the research objective(s) is provided in the respective chapters.

Grasses

The economic and ecological importance of grasses has paved the way for many whole genome-sequencing efforts, and to date has resulted in completed genomes of important grass species such as *Oryza sativa* (rice) (Matsumoto et al. 2005, Yu et al. 2002), *Sorghum bicolor* (sorghum) (Paterson et al. 2009), *Zea mays* (maize) (Schnable et al. 2009), *Brachypodium distachyon* (purple false broom) (Vogel et al. 2010), *Setaria italica* (foxtail millet) (Bennetzen et al. 2012, Zhang et al. 2012), and *Hordeum vulgare* (barley) (Mascher et al. 2017). These sequenced taxa represent three subfamilies of *Poaceae* and have enabled some of the most comprehensive comparative genomic analyses possible, providing extensive knowledge of the organization and evolution of grass genomes.

Grasses evolved from a common ancestor with a base chromosome number of five (~90 mya) that then underwent a series of whole genome duplications, segmental duplications, chromosome fusions and translocations to produce a 12-chromosome intermediate common ancestor around 70 mya (Salse et al. 2008). Even though most current day grass genomes are functionally diploid, all grasses are ancient polyploids that have gone through a tremendous amount of chromosome fusions and rearrangements resulting in chromosome number reductions. Hence, it is fascinating to study grass genomes from an evolutionary perspective, because of the variation found in ploidy level, genome size, chromosome number, and repetitive DNA content. For example, *Triticum aestivum* (Bread wheat) is a hexaploid (2n=42) with a genome size of ~17 Gb that is >80% repetitive, *Zea mays* (Maize) is a diploid (2n=20) with a genome size of ~2.5 Gb that is 85% repetitive, and diploid rice (2n=24) has a much smaller genome of 0.4 Gb in which 35% is repetitive DNA (Figure 2). In the most comprehensive comparative studies
to date, it has been shown that gene content and order is highly conserved between grasses, even after 50 million years of independent divergence (Dubcovsky et al. 2001, Paterson et al. 2004; Bowers et al. 2003 & 2005); however, with the exception of a very small amount of conserved non-coding sequences, most of the intergenic space is relatively distinct, even between recently diverged species such as sorghum and maize. From these studies it is clear that a considerable amount of genome variation in grasses can be ascribed to repeat variability.

For Chapters 2 and 3, I have chosen two independently domesticated grasses, maize (*Zea mays*) and sorghum (*Sorghum bicolor*) in addition to their close wild relatives as study systems. Both genera belong to the tribe Andropogoneae and differ in morphology, genome size, TE content, and ploidy level. In addition to the difference in their TE proportion and content, the distribution of TEs within each genome varies. For example, sorghum shows a strong separation in the distribution of genes and repeats compared to the maize genome (Paterson et al. 2009, Schnable et al. 2009). Such variation in lineage-specific TE content provides an excellent phylogenetic framework for determining the impact of TEs on genome structure and function in plants.

**Gossypium**

The cotton genus, *Gossypium* L (Malvaceae) comprises about 50 species and is distributed worldwide with several primary centers of diversity in the arid/semi-arid tropics and subtropics (Fryxell 1979, Wendel 2009). Species-rich regions include northwestern Australia, central and southern Mexico, the Horn of Africa and the southern Arabian Peninsula. The genus is extraordinarily diverse and differentiated cytogenetically into eight genome groups (A through G, and K) that differ in DNA content and chromosome size but not in chromosome number (Endrizzi, Turcotte and Kohel 1985; Stewart 1995). The genus contains three-fold variations in genome size: from 800 Mb (1C) in the D-genome to ~2,500 Mb in the K genome. There are three major lineages of diploid (2n=26) species corresponding to three continents: Australia (C, G, and K genomes), the Americas (D genome), and Africa/Arabia (A, B, E and F genomes). Apart from the diploids, this genus contains five recent tetraploids (2n=52) that emerged from a single interspecific hybridization between the A and D genomes approximately 1-2
million years ago. These five tetraploids (AD) are: *G. hirsutum*, *G. tomentosum*, *G. barbendense*, *G. darwinii*, and *G. mustelinum*. *G. hirsutum* and *G. barbendense* are domesticated crops with *G. hirsutum* dominating the world’s textile industry for its superior yield, whereas *G. barbendense* is known for its higher quality. Apart from these two-tetraploid species, there are two diploid A genome species that are also cultivated for textile: *G. arboreum* and *G. herbaceum*. All four-cultivated species were independently domesticated for their specialized fibers and have their own unique history of domestication, diversification and utilization. Given the wide-range of genome size variation reflecting the copy numbers of repeat DNA sequences and a recent polyploidization event that may have contributed to the improvement in cotton fiber yield and quality, I have used *Gossypium* as a model system in Chapter 4 to understand the evolutionary dynamics of repeat components in host genome evolution.
References


Hashida S-N. 2006. The Temperature-Dependent Change in Methylation of the Antirrhinum Transposon Tam3 Is Controlled by the Activity of Its Transposase. The Plant Cell Online. 18:104–118.


Figure 1: Major classes of transposable elements and examples of insertional mutagenesis. The top panel shows the structure of class I Retrotransposons and class II DNA transposons. The bottom panel demonstrates prominent types of mutations caused by TE insertions within or near gene regions.
Figure 2: Transposable element content with respect to genome size in grasses. The TE fraction varies significantly among grass genomes and is positively correlated with genome size.
CHAPTER 3: METHODS FOR ACCURATE QUANTIFICATION OF LTR-
RETROTRANSPOSON COPY NUMBER USING SHORT-READ SEQUENCE
DATA: A CASE STUDY IN SORGHUM

A paper published in the journal *Molecular Genetics and Genomics*
Dhanushya Ramachandran, Jennifer S. Hawkins

Abstract

Transposable elements (TEs) are ubiquitous in eukaryotic genomes and their mobility impacts genome structure and function in myriad ways. Because of their abundance, activity, and repetitive nature, the characterization and analysis of TEs remains challenging, particularly from short-read sequencing projects. To overcome this difficulty, we have developed a method that estimates TE copy number from short-read sequences. To test the accuracy of our method, we first performed an *in silico* analysis of the reference *Sorghum bicolor* genome, using both reference-based and de novo approaches. The resulting TE copy number estimates were strikingly similar to the annotated numbers. We then tested our method on real short read data by estimating TE copy numbers in several accessions of *S. bicolor* and its close relative *S. propinquum*. Both methods effectively identify and rank similar TE families from highest to lowest abundance. We found that de novo characterization was effective at capturing qualitative variation, but underestimated the abundance of some TE families, specifically families of more ancient origin. In addition, interspecific reference-based mapping of *S. propinquum* reads to the *S. bicolor* database failed to fully describe TE content in *S. propinquum*, indicative of recent TE activity leading to changes in the respective repetitive landscapes over very short evolutionary timescales. We conclude that reference-based analyses are best suited for within-species comparisons, while *de novo* approaches are more reliable for evolutionarily distant comparisons.

Reprinted by permission from Springer Nature: Molecular Genetics and Genomics, 291(5), 1871-1883
Introduction

In plant genomes, the transposable elements (TE) community is both quantitatively and qualitatively dynamic. TEs occupy a variable proportion of the genome, and many studies have shown that their differential accumulation and deletion strongly correlates with genome size (Bennetzen 2000; Bennetzen 2002; Tenaillon et al. 2010; Michael 2014). In some cases, rapid rates of amplification and decay can lead to variable TE composition among closely related plant lineages. For instance, although similar numbers of TE families occupy both the rice and maize genomes, transpositional bursts of a few Long Terminal Repeat (LTR) retrotransposon families have inflated the maize genome to six times that of rice (Baucom et al. 2009; Baucom et al. 2009). Variation is also common among more closely related taxa, such as those that belong to the same subfamily (SanMiguel and Bennetzen. 1998), and even among members of a single genus. Interspecific comparisons performed in Gossypium and Oryza revealed that proliferation of a small subset of TE families were responsible for the observed genome size variation among species (Hawkins et al. 2006; Piegu et al. 2006). TE copy number variation is not always due to the activity of just a few families, however, as has been demonstrated by the proliferation of several different TE families in Zea and Arabidopsis (Hollister et al. 2011; Hu et al. 2011; Tenaillon et al. 2011). Given that ongoing and punctuated TE activity has been widely documented in plants, it is important to understand the contribution of TEs to lineage-specific novelty, as insertional polymorphisms may ultimately contribute to species diversification.

Although TEs are predominantly silenced, protecting the genome from rampant insertional mutagenesis, the host suppression system can be circumvented in some situations. TE activation has been reported following both hybridization and polyploidization (Kashkush et al. 2002; Beaulieu et al. 2009; Parisod et al. 2009; Xu et al. 2009). An example is that of hybrid Sunflower species, in which the hybrid genomes are 50% larger (~1.130 Mb additional DNA) than either of the parental genomes (Ungerer et al. 2006). This increase was attributed to recent proliferation (0.5 - 1 mya) of Ty3/Gypsy elements (Ungerer et al. 2009). In addition, TEs are often activated in response to various biotic and abiotic stresses such as infection, temperature, wounding and salinity (Mhiri et al. 1997; Grandbastien et al. 1998; Takeda 1998; Pecinka 2010; Tittel-Elmer et al. 2010;
Fujino et al. 2011; Cavrak et al. 2014; Makarevitch et al. 2015; Finatto et al. 2015). Such TE activity can generate numerous insertional polymorphisms, each with the potential for functional consequences when inserted into genes or gene regions (Kashkush et al. 2003; Kashkush and Khasdan 2007; Chu et al. 2011; Ito et al. 2011; Studer et al. 2011; Butelli et al. 2012). Although TE activation might negatively affect genome integrity, it has been suggested that species or populations that are prone to strong diversifying selection would benefit from TE driven genome variability (Huang et al. 2008; Martin et al. 2009; Fernandez et al. 2010; Lockton and Gaut 2010).

To date, analyses of TE copy number that could inform studies of activity and accumulation have been hindered by the inability to analyze the repetitive fraction from short-read sequencing projects. This difficulty is due to the inability to accurately assemble short sequence reads that belong to repeats, as these reads often create assembly gaps, incorrectly collapse onto a single chromosomal position, and/or map to multiple locations in the genome, resulting in misassembled arrangements. Most modern assemblers attempt to resolve these issues by employing alignment strategies that either discard multiply mapping reads, report all possible mapping locations, or randomly map reads to the position of best alignment; however, the random placement or all-together removal of multiply mapping reads clearly prevents detailed analyses of repeat regions. Additionally, although there are assemblers that are efficient enough to report all possible mapping locations for repetitive sequences, they work best for high-coverage datasets (>20x) or with longer sequence read lengths (Phillippy et al. 2008, Treangen and Salzberg et al. 2012). These limitations pose a serious challenge to the characterization of repeat content and to the determination of repeat copy numbers from NGS datasets.

Here, we demonstrate a method to evaluate TE composition and accurately estimate copy number using Illumina short-read sequence data. We first tested the accuracy of our method by performing an in silico analysis of the reference BTx623 Sorghum bicolor genome, which resulted in copy number estimates that are strikingly mathematically similar to the annotated TE copy numbers. Following this simulated analysis, we tested our methods on real short read datasets by estimating copy numbers in several accessions of S. bicolor and its close wild relative, S. propinquum. Copy number
estimations were performed using both reference-based and de novo methods for comparison. Both methods rank the families in similar order from highest to lowest abundance; however, the estimated copy numbers via de novo analysis differ from that of the reference-based approach, and these differences correlate with the relative insertional timing of individual TE families. Specifically, we find that de novo approaches are more effective in estimating copy numbers for young TE families but tend to underestimate the abundance of older families. In addition, interspecific reference-based mapping failed to fully describe TE content in *S. propinquum* due to inefficient mapping of reads to the *S. bicolor* database. We conclude that reference-based methods are best for estimating within species variation whereas de novo approaches are more reliable for evolutionarily distant comparisons.

**Materials and methods**

*In silico development of method for copy number estimation*

As the availability of a high-quality reference genome was required to develop our approach, we focused on the genus *Sorghum*. The *Sorghum bicolor* reference genome (BTx623, v1.0 Paterson et al. 2009) was downloaded from Phytozome v9.0, and full-length Long Terminal Repeat (LTR)-retrotransposons were identified with LTRharvest (Ellinghaus et al. 2008) using the default settings, except for the following: a motif for 5’ and 3’ LTRs as each LTR should begin and end with TG and CA nucleotides, minimum and maximum length LTRs of 100 – 5,000 bp, and seed length set to 60 bp. Sequences that were incorrectly identified as LTR-retrotransposon (false positives) were removed from the output by performing a nucleotide BLAST against the Plant Genome and Systems Biology (PGSB, formerly MIPS) Poaceae repeat database (Nussbaumer et al. 2013) with an e-value cutoff of 1e-10 and sequence identity of 80%. Sequences that did not match to grass-specific LTR-retrotransposons were removed. An all-by-all BLAST (e-value cutoff 1e-10 and at least 80% sequence similarity) was performed with the 5’LTRs and the result was clustered into families using RepMiner (J. Estill, code available at http://repminer.sourceforge.net, Baucom et al. 2009), and resulting clusters were visualized as a network using the imaging program Cytoscape 3.0.
(Shannon et al. 2003). In addition, the percent sequence divergence between the 5’ and 3’ LTR of each element was extracted from the LTRharvest output and used to determine the insertional timing of each element in the BTx623 reference genome using a grass-specific substitution rate of 1.3 x 10^-8 per site per year (Ma and Bennetzen, 2004).

RepeatMasker (employing default settings) was used to identify additional LTR sequences that are not present as full-length sequences (solo LTRs) using the LTRharvest output sequences as a database (Smit, Hubley and Green RepeatMasker Open-4.0. 2013-2015). All of the full-length elements identified by LTRharvest were first masked using the feature maskfasta from Bedtools 2.17.0 package (Quinlan and Hall 2010), so as not to count these LTRs twice. The RepeatMasker output was filtered for hits that were less than 5% divergent from the database sequences and were at least 150 bp in length. LTR sequences from both the LTRharvest and Repeatmasker analyses were combined into a single dataset for further analysis. In addition, exemplars that represent the entire population of sequences were selected from the extracted 5’ LTRs using affinity propagation clustering, which were then used for de novo analysis (see below) (Frey and Dueck Science 2007; Bodenhofer et al. 2011).

To devise an accurate method to estimate TE copy number from the short read data, we tested the equation from Hawkins et al. (2006) on an in silico dataset generated from the BTx623 reference using the short read simulator, DWGSIM v.0.1.11. The program was run on ‘illumina’ mode to generate 7.5 million reads of 100 bp in length. We used only single-end reads for our analysis in order to treat each read as a mathematically independent sample. The simulated reads were mapped back to the reference genome sequence using Bowtie2 under default settings (Langmead and Salzberg 2012). With the known chromosomal positions for each identified element, the total number of reads that are strictly and uniquely aligned (i.e., entire 100 bp read) within the first and last nucleotide position of each identified 5’ LTR was extracted from the BAM alignment file using the intersectBed and coverageBed tools of Bedtools 2.17.0. The copy number of each element \( n \) was estimated using the following equation (Hawkins et al. 2006):

\[
independent
dependent
Where $X_{obs}$ is the total number of reads aligned to a given element (5′ LTR), $N$ is the total number of reads used during mapping, $lt$ is the length of target sequence (5′ LTR), $m$ is the overlap required to count a match to the target region (in this case, the entire length, or 100 bp), $e$ is the length of the sequence read, and $G$ is genome size. Simply put, this equation estimates how frequently a sequence of a particular length (the LTR) must be present in a genome of a given size, if a proportion of random samples (reads that match/total random samples) from that genome match the particular LTR. The greater the proportion of reads that map to a particular LTR of a given size, therefore, the higher the copy number estimate for that LTR in the genome. To evaluate possible sampling effects, we repeated the in silico subsampling and statistical estimates for 100 independently simulated datasets. Confidence intervals (95%) were calculated for all copy number estimates as described in Hawkins et al (2006).

We also tested our method using a de novo assembly approach, implemented in RepeatExplorer, for the same simulated short read sequence data used in the reference-based approach described above (Novak et al. 2013). RepeatExplorer employs a graph-based clustering method by quantifying the similarities between reads. The program begins by filtering reads that pass a specific threshold (> 90% sequence similarity over 55% of the read length). Using these similarities, the program constructs a graph and creates clusters from frequently connected reads that represent individual repetitive families. Reads within each cluster are then assembled into contigs using CAP3 with an overlap length cutoff of 50 bp (for 100 bp reads) and sequence identity of at least 80% (Novak et al. 2010). To classify clusters into specific LTR-retrotransposon families, BLAST was performed between the RepeatExplorer contigs from the largest clusters and the 5′ LTR exemplars from the reference genome. After classifying the clusters into families, copy numbers were estimated from the contigs present in each cluster employing the same probability equation used in the reference-based quantification. Copy numbers were estimated for the ten largest TE families for each accession.
**Sorghum accessions and genome size determination**

We included five *Sorghum bicolor* (B35, SC56, RTx430, Shanqui red, Tx7000) accession, in addition to four *S. propinquum* accessions to facilitate interspecific comparisons. *S. propinquum* accession PI653737 was obtained from the USDA Agricultural Research Service Plant Genetic Resources Conservation Unit (Griffin, GA) (henceforth referred to as *S. propinquum_USDA*), and an unnamed *S. propinquum* accession (henceforth referred to as *S. propinquum_BR*) was provided courtesy of Dr. William Rooney, Texas A&M University, College Station, TX. Sequences for two additional accessions, *S. propinquum* 369-1 and *S. propinquum* 369-2, were downloaded from the short read archive (Mace et al. 2013). Plants were grown in the WVU Department of Biology greenhouse under normal conditions. Leaves were flash frozen in liquid nitrogen and stored at -80°C. Nuclear DNA content was determined via flow cytometry using chicken erythrocyte nuclei (CEN) as an internal standard with the nuclear DNA content of 2.5 picograms per 2C, performed in triplicate, at the Flow Cytometry Core Lab, Benaroya Research Institute at Virginia Mason (Seattle, WA) (Table 1).

**DNA extraction and Illumina sequencing**

Frozen leaf tissue (~15 g) for all *S. bicolor* accessions, *S. propinquum_USDA*, and *S. propinquum_BR* was ground to a fine powder using liquid nitrogen and suspended in sucrose extraction buffer (SEB), in which 1/20th volume of 10% Triton X-100 solution was added to lyse chloroplasts and mitochondria. The resulting homogenate was double filtered to remove other cellular debris and nuclei were isolated using centrifugation. DNA was extracted from isolated nuclei using the Promega Wizard Genomic DNA Purification Kit (Madison, WI) following the manufacturer’s instructions. For all accessions except *S. propinquum_USDA*, Illumina libraries were constructed and sequenced at the Georgia Genomics Facility (University of Georgia, Athens, GA) using the Illumina HiSeq 2000 (2 x 100 bp, ~500 bp insert size). *S. propinquum_USDA* was sequenced using the Illumina MiSeq (2 × 150 bp, ~500 bp insert size) at the West Virginia University Genomics Core Facility, Morgantown, WV. *S. propinquum*
accessions 369-1 and 369-2 were sequenced by Mace et al. (2013) using the Illumina HiSeq 2000 (2 x 90 bp with 500 bp insert size).

**Copy number estimates for newly sequenced *Sorghum* accessions**

Short read sequence data for the five *S. bicolor* and four *S. propinquum* genomes were trimmed for quality (-q 28) using sickle v1.33 (Joshi, Fass, 2011, available at https://github.com/najoshi/sickle). Reads that contained any N’s and/or that were under 95 bp (85 bp for *S. propinquum* accessions 369-1 and 369-2) were removed using a custom Perl script (available upon request). To estimate the copy number of LTR-retrotransposon families in our newly sequenced accessions, and in addition to the two published *S. propinquum* accessions, we subsampled 7.5 million reads from each filtered sequenced library. The reads were uniquely mapped to the 5’ LTRs mined from the BTx623 reference genome using Bowtie2, allowing one mismatch within the entire read (Langmead and Salzberg, 2012). The total number of reads that strictly aligned to each 5’ LTR was extracted from the BAM output file and the copy number of each element was estimated using the equation describe above. The copy number of each element was also estimated using the *de novo* approach described above.

A two-sample t-test was performed to compare the estimates obtained from both approaches for each retrotransposon family. This test was also performed for each retrotransposon family between *S. bicolor* and *S. propinquum* to determine statistically significant interspecific difference among TE families.

**Results**

**Identification and characterization of LTR-retrotransposons in the *Sorghum* BTx623 reference genome**

UsingLTRharvest, 12,530 LTR-retrotransposons were mined from the *Sorghum bicolor* reference genome (BTx623, v1.0 Paterson et al. 2009). These elements were filtered for false positives using the PGSB Poaceae database as a reference, and 313 false positives (mostly ribosomal repeats) were removed from the dataset. The remaining
12,217 intact LTR-retrotransposons were then grouped into ~210 families based on 5’ LTR sequence similarity (BLASTN, e-value 1e-10) and visualized as a network in Cytoscape v.3.0 (Figure 1). As observed in most grasses, Gypsy-like elements were the most abundant TE sequences in the Sorghum genome (~9,000), followed by ~2,200 unclassified and ~1,100 Copia-like elements. The first and largest cluster in the LTR-retrotransposon network (Figure 1) consists of 7,801 5’ LTR sequences belonging to the Gypsy superfamily. These sequences were further divided into various families such as Onap, Retrosor6, Leviathan, Tekay-like elements, and RLX-CRM. The smallest clusters in the network contained only two sequences, mostly of unknown classification. By using the 5’ LTR sequences as a reference database, we used RepeatMasker to identify an additional 8,240 (presumably solo) LTRs from the reference genome. The ratio of solo-LTR to intact elements is estimated at 0.67:1, similar to previously published findings (Baidouri and Panaud, 2013). Overall, a total of 32,674 LTRs (including both LTRs from the intact elements and solo LTRs) were identified in the Sorghum reference genome assembly.

After grouping the elements into families, we estimated the insertional timing of each element in the genome based on the sequence divergence of the 5’ and 3’ LTRs of each individual full-length retrotransposon (Figure 2A). The estimated insertion times ranged from 0 to 5.8 mya. The average insertion age is ~1.6 mya. As shown in Figure 2B, the majority of LTRs in the “Retrosor-6” and “2-Unknown” clusters share a high degree of sequence similarity (100 - 97.41 %) and were inserted into the genome within the last 1 mya. In addition, more than 80% of the LTRs in the “Retrosor-1”, “4-Gypsy” and “5-Unknown” clusters inserted less than 1 mya (Online Resource 3). In contrast, the LTRs in the Onap cluster share less sequence similarity (92.2 to 85 %) and the estimated insertion age of these elements falls within the range of 3 to 5.8 mya, indicating that these insertions are older in origin, and predate the S. bicolor – S. propinquum divergence, which occurred approximately 1-2 mya (Paterson 2008; Figure 2B & Online Resource 3).
In silico development of methods for copy number estimation

We performed an in silico analysis of the simulated short read dataset generated from the reference genome to develop a framework for accurate copy number estimation. After sequence simulation using DWGSIM, the reads were mapped back to the Sorghum reference genome and the total number of reads that strictly aligned at each of the previously identified LTRs was recorded. From the total read counts for each LTR, we estimated the copy number via a probability statistic that takes into account the total number of reads, number of reads that map to a target, length of the target sequence, and genome size (Hawkins et al. 2006). We identified a total of 32,674 LTR sequences in the Sorghum bicolor reference genome. The estimated copy number (LTRs = 32,912 ± 167) via our statistic is strikingly similar to the actual annotated numbers, indicating that the equation is remarkably accurate at estimating the copy numbers of repetitive sequences from the short-read sequence data using this approach (Online Resource 1).

Further, we performed reference-based in silico copy number estimation for the largest families found in the BTx623 reference genome. Among all families, Onap, a Gypsy LTR-retrotransposon was most abundant (6,744 ± 64) followed by Retrosor-6 and 2_Uknown, estimated at 5,588 ± 48 and 4,551 ± 51, respectively. As with total LTR number, these family-level estimates correlate with the annotated numbers identified via LTR-harvest and RepeatMasker, demonstrating that the equation works well, even at more refined levels. For example, we identified 5,852 full-length and 303 solo Onap elements from LTR-harvest and RepeatMasker (total = 6,155; estimate = 6,744 ± 64). Similarly, estimates correlate with annotated copy numbers for the 5,567 Retrosor-6 (estimate = 5,588 ± 48) and 4,543 2_Uknown (4,551 ± 51) elements. To evaluate possible sampling effects, we repeated this analysis for 100 independently subsampled in silico datasets. The average estimated copy number (LTRs = 33,834 ± 169) from the resampling analysis is similar to the actual annotated number (32,674) as well as the initial estimated copy number (32,912 ± 167) indicating the robustness of the equation in capturing the TE landscape from short read datasets.

After verifying data repeatability using our statistic, we performed an additional analysis to determine its usefulness when employing de novo methods for
characterization of the repetitive fraction of the genome. Copy numbers for each LTR-retrotransposon family were therefore estimated from the consensus sequences of the largest RepeatExplorer clusters using the same statistical equation. Our de novo analysis resulted in an estimated 69,558 ± 197 total LTRs in the reference genome, almost double the total number of LTRs identified by LTRharvest and RepeatMasker (32,674). Among all families, Onap, Retrosor-6 and 2-Unknown are estimated to be the most abundant families, with Retrosor-6 estimated at 12,717 ± 69 followed by 2_Unknown (9,799 ± 68) and Onap (8,636 ± 51). With the exception of Onap, the de novo estimates from the simulated dataset are in agreement with most of the reference-based and all of the de novo estimates for the real short-read sequence data (see below).

Reference-based copy number estimation in Sorghum

We estimated the copy numbers of various LTR-retrotransposon families from nine Sorghum accessions (five S. bicolor and four S. propinquum) using the reference-based approach (Online Resource 2). The estimated total copy number of LTRs is ~73,000 in S. bicolor and ~50,000 in S. propinquum accessions (Table 2). The same ten families contribute the highest number of copies to the total TE fraction in all of the nine Sorghum accessions (Figure 3). Onap, the most abundant retrotransposon in all accessions, varies significantly in copy number among species (Online resource 3). Specifically, Onap copy number is similar among all accessions of S. bicolor (average 25,692 ± 100) and S. propinquum_USDA (27,908 ± 108), but varies approximately three fold compared to S. propinquum_BR (9,673 ± 61), S. propinquum 369-1 (9,055 ± 59) and S. propinquum 369-2 (11,086 ± 65). Indeed, we found that the estimated copy numbers for many families in S. propinquum_USDA are more similar to that of the S. bicolor accessions than the other S. propinquum genomes (Figure 3). For example, 4-Gypsy in S. propinquum_BR (103 ± 8 copies), S. propinquum 369-1 (85 ± 7 copies) and S. propinquum 369-2 (100 ± 7 copies) is composed of twelve-fold fewer copies compared to S. propinquum_USDA (~1,357 ± 27 copies), and approximately seven-fold fewer copies than the five S. bicolor genomes (on average 775 ± 20 copies). To determine whether these observed differences are statistically significant, we performed a two-sample t-test for each retrotransposon family between S. bicolor and S. propinquum (Online resource
The copy numbers for five families (Retrosor-6, RLX-CRM, Giepum, 11_Unclassified and 12_Unclassified) were significantly different (P < 0.05) between *S. bicolor* and *S. propinquum*; however, when we included *S. propinquum USDA* with the *S. bicolor* accessions, 14 families (Online resource 3) showed statistically significant differences in copy number.

**Variation in copy number using a de novo approach**

To identify genome-specific repetitive sequences that were not identified by the reference-based approach, we performed a *de novo* analysis using RepeatExplorer (Novak et al. 2013). As this method is not restricted to LTR-retrotransposons, we also estimated the proportions of a broader range of types of repetitive DNA. The largest cluster in all accessions consisted of satellite repeats. The number of reads in the satellite cluster was variable within and between genomes of *S. bicolor* and *S. propinquum*: B35, RTx430, Shanqui red, *S. propinquum_BR*, *S. propinquum* 369-1 and *S. propinquum* 369-2 contained ~178,000 satellite-associated reads of 137-274 bp that occupy ~22 Mb of the genome, whereas SC56, Tx7000, and *S. propinquum_USDA* contained ~94,000 satellite-associated reads of 21-68 bp which occupy ~2 Mb of the genome. Except for the first one or two largest clusters (satellite and/or DNA transposons), all *Sorghum* genomes contained a greater number of Gypsy-like LTR-retrotransposon clusters than Copia-like elements (Online Resource 4).

The total LTR-retrotransposon copy number estimates for *S. bicolor* via *de novo* analysis are similar to estimates from the reference-based method (Table 2). For *S. propinquum*, *de novo* methods result in significantly higher copy number estimates (~75,900 ± 260), with the exception of the estimate for *S. propinquum USDA* (68,949 ± 270). Nevertheless, both methods indicate that Onap and Retrosor-6 are the two largest LTR-retrotransposon families. For *de novo* estimates, Retrosor-6 is estimated at a higher copy number than Onap in B35 (11,999 ± 83), SC56 (11,615 ± 83) and RTx430 (11,335 ± 86), whereas Onap is estimated at higher copy number than Retrosor-6 in Tx7000 (14,910 ± 81), S. red (12,366 ± 68) and *S. propinquum USDA* (11,874 ± 65) (Figure 4). Results for the other prevalent families, such as 2-unknown, Leviathan, and RLX-CRM, while comparable to copy number estimates from the reference-based approach, are
significantly different in some cases (Figures 3 and 4). Interestingly, 2 Unknown is estimated at higher copy number than either Onap or Retrosor-6 in *S. propinquum_BR* (11,947 ± 80), *S. propinquum_369-1* (12,240 ± 83) and *S. propinquum_369-2* (13,869 ± 92), in concordance with the estimates via the reference-based method (Figures 3 and 4).

We performed a two-sample t-test for each retrotransposon family to determine statistically significant differences in copy number between *S. bicolor* and *S. propinquum*. There were no significant differences in copy number between *S. bicolor* and *S. propinquum*; however, when we include *S. propinquum_USDA* as one of the *S. bicolor* accession, the copy numbers for seven families (2 Unknown, Keama, RLX-CRM, Tekay, 5th Unknown, Retrosor-1 and 12th Unknown) were significantly different between the two species (Online Resource 3). The results for five of these families (Keama, RLX-CRM, 5th Unknown, Retrosor-1 and 12th Unknown) correlate with that from the reference-based t-test comparisons.

Discussion

The LTR-retrotransposon landscape in the *S. bicolor BTx623* reference genome

We identified 12,217 intact LTR-retrotransposons from the *Sorghum bicolor* reference genome and classified these sequences into ~210 different families based on sequence similarity of the 5’ LTRs (Figure 1). With the exception of the largest cluster in the network, all other clusters were composed of sequences from distinct families. The largest cluster contains 7,801 5’ LTR sequences that belong to several families including Onap, Retrosor-6, Leviathan, Tekay, RLX-CRM, and Kaema. Sequence similarity among these families indicates a deep evolutionary connection between the LTR-retrotransposons in the cluster, which can be explained by consecutive sequence evolution of TE families (Khan 2006, Cordaux and Batzer 2010). For instance, Onap retrotransposons share less sequence similarity between their 5’ and 3’ LTRs relative to the other families in the first cluster, suggesting this family is the oldest (3 to 5.8 mya) and could therefore be the progenitor of the younger related sequences, such as those belonging to Retrosor-6 and Leviathan (Figure 2A and B).
Interestingly, the LTR-retrotransposon families of greatest abundance in *S. bicolor* differ significantly from those in *Zea mays*. Huck, the most abundant family in maize and a few other grasses, is found in very low copy number (2 copies) in *S. bicolor* (Peterson et al. 2002). Even at an e-value cutoff of 1e-05, we could not identify additional copies of this element. The same is true for Ji and Opie, which are abundant in maize, but found at low frequency (Ji - 62 copies, Opie – 0 copies) in *Sorghum*. In contrast, Onap, Retrosor-6, and Leviathan are highly repetitive in *Sorghum*, but found in very low copy number in maize, indicating activation of different TE families over very short evolutionary timescales (Estep et al. 2013). In other words, although *Sorghum* and maize diverged only 12 mya (Swigoňová et al. 2004), each species has undergone independent activation of lineage-specific TEs.

Though several studies report recent LTR-retrotransposon bursts in related grasses (Piegu et al. 2006, Bennetzen et al. 2012, Senerchia et al. 2013), we did not detect a large amount of recent activity in the *Sorghum* reference genome. Indeed, very few intact LTR-retrotransposons share 100% sequence similarity between 5’ to 3’ LTRs (Figure 2A). Retrosor-1 is one of the few families that contained a higher proportion of sequences with at least 99% similarity (70 out of 99) between 5’ and 3’ LTRs of the same element, indicative of relatively recent amplification and insertion (< 1 mya). This family consists of very few copies, having little effect on genome size. Nevertheless, most of these recent insertions are located within 5-10 kb of protein coding genes (data not shown), and therefore carry the potential to induce possible functional consequences on neighboring genes (i.e, loss-of-function or altered gene expression), such as has been observed for tb1, ZmCCT, and ZmRAP2.7 in maize (Salvi et al. 2007; Studer et al. 2011; Yang et al. 2013).

One possible explanation for low levels of detected recent transposon activity could be artifactual in nature. The *Sorghum* reference genome assembly, consisting of ~730 Mb, is ~100 Mb less than the flow cytometry measurements for the *S. bicolor* genomes included in this study (Table 2), suggesting a significant portion of the genome sequence is missing. Since recently transposed elements share high sequence similarity, reads belonging to these elements would collapse in the assembly and appear as a single
or small number of repeats, rather than a number of dispersed repeats. This would also explain why the estimated copy numbers from the Illumina data are significantly higher than that from the in silico analyses (see below).

**Methods to estimate the repetitive fraction using short read sequences**

We aimed to develop a method to accurately estimate repetitive sequence copy number from short-read sequence data, and to use this method to detect inter- and intraspecific variation in TE abundance for five accessions of *S. bicolor* and four accessions of *S. propinquum*. To this end, we performed an in silico analysis to test the accuracy of our statistical equation (Hawkins et al. 2006) using both reference-based and de novo approaches. We annotated a total of 32,674 LTRs from the reference genome, and the estimated copy number for both the total LTRs (32,912 ± 167) and for LTRs from individual families (Onap = 6,744 ± 64; retrosor6 = 5,588 ± 48) from our reference-based in silico analysis was strikingly similar to the annotated number.

The estimated total number of LTRs from the de novo in silico analysis, however, was much higher. This discrepancy can easily be explained by the fact that, for the referenced-based in silico analysis, we focused specifically on the number of reads that mapped to precisely defined genomic locations, namely the 32,674 bioinformatically identified LTRs. Reads that map to unidentified LTRs would be excluded from the 32,674 regions of interest, drastically reducing the mathematically estimated copy number. In addition, our mathematical equation required the entire read to map within the first and last nucleotides of an LTR, which would rarely occur by chance for shorter LTRs. These same sequence reads would, however, be included in the in silico de novo assembly, providing a clear explanation for the increased copy number via that method. Indeed, the de novo in silico estimates correlate strongly with most of the reference-based and all of the de novo estimates for the real short-read sequence data. Importantly, the referenced-based in silico analysis was not designed to determine the actual number of LTRs in the reference genome, but rather to verify the accuracy of our statistical equation.
Repeat diversity is most accurately characterized from short-read data via a combination of reference-based and de novo approaches

Comparative characterization of TE diversity and abundance in the newly sequenced accessions using both reference-based and de novo approaches suggests that the former is primarily suitable when estimating within-species variation while the latter is more reliable for more evolutionarily distant comparisons. We initially expected reference-based mapping to efficiently describe TE content in *S. propinquum*, given that *S. bicolor* and *S. propinquum* diverged as little as 1-2 million years ago (Paterson 2008), but this was not the case. Although the reference-based and de novo estimates for the total number of LTRs are in strong agreement for *S. bicolor* (and *S. propinquum_USDA*), suggesting that either approach will accurately characterize within-species diversity, the reference-based estimates for total LTRs in *S. propinquum* were considerably low (Table 2). At the individual family level, we note that the reference-based estimate for the most ancient and abundant TE family (Onap- Figure 2B) in *S. propinquum* is unexpectedly low, given its genome size and in comparison to the estimated copy numbers in the other genomes (Figure 3). Onap, which has accumulated near gene-poor regions and is composed of much older sequences based on molecular clock dating (Figure 2B), has likely been retained due to limited selection pressure and recombination suppression in this part of the genome. As most Onap insertions predate the *S. bicolor* - *S. propinquum* divergence and have therefore accumulated a large number of lineage-specific mutations, these sequences were more easily identified in *S. bicolor* accessions using reference-based methods in comparison to that for *S. propinquum*. In addition, we suspect that de novo assembly is ineffective at accurately estimating copy numbers from short read data for older TE families due to difficulties with assembling reads that contain a larger number of polymorphisms. We conclude that, although *S. bicolor* and *S. propinquum* diverged recently, the substitution rates for LTR-retrotransposons in *Sorghum* are sufficiently high to prevent interspecific comparisons of repetitive content via shared sequence similarity alone (reference-based mapping), particularly for older sequences. Therefore, caution should be used when performing interspecific reference-based TE annotation, even among closely related species, as the results will likely underestimate
the actual number of copies in the genome, especially for sequences of more ancient origin, or those that are undergoing accelerated rates of diversification.

**Comparative analyses reveal detectable repeat variation over short evolutionary timescales**

Our comparative analyses using copy number estimates from the Illumina data reveal small but detectable variation in LTR-retrotransposon content and copy number among accessions of *S. bicolor*, and larger variation between *S. propinquum_USDA* and the other three *S. propinquum* accessions. This result was expected, as the *S. propinquum* accessions include the individuals with both the smallest (833 Mb) and largest (902 Mb) genomes. We anticipated that this size disparity would be associated with recent transpositional activity in *S. propinquum_USDA*; however, we could not detect significantly elevated copy numbers for any specific element in the *S. propinquum_USDA* genome. We also estimated the copy numbers for satellite repeats, but again could not identify large differences that would explain the genome size disparity. There are two possible explanations for this observation: 1) Genome size variation among the *S. propinquum* accessions is due to the accumulation of a small number of TE copies from a large number of families in *S. propinquum_USDA*, and would therefore be undetectable in our analysis, or 2) the excess nuclear content in *S. propinquum_USDA* is composed of older decaying TE sequences that can no longer be identified at 80% sequence similarity, and/or that have been more effectively removed from the other *Sorghum* genomes. From comparisons between our two approaches, we suspect the most likely explanation is the latter. The LTR copy numbers for the largest retrotransposon family, Onap, are much higher for *S. propinquum_USDA* than for other *S. propinquum* genomes using the reference-based method (Figure 3), but drops significantly using the de novo approach (Figure 4) suggesting that there are a greater number of older sequences in the *S. propinquum_USDA* genome.

Alternatively, we observed that the repeat profile for *S. propinquum_USDA* is surprisingly more similar to that of the *S. bicolor* genomes than *S. propinquum*, and unlike the results for the *S. propinquum* accessions, reference-based methods of analysis appear to effectively characterize the *S. propinquum_USDA* content. This could be
explained by introgression of *S. bicolor* chromatin into this particular accession, as has been suggested by Tang et al. (2013) based on microsatellite genotyping. For example, our results indicate significant interspecific copy number differences for many prominent families, but only when the USDA accession is included as one of the *S. bicolor* species (Online resource 3). In addition, the estimates for two specific families in *S. propinquum_USDA* show a unique pattern of increase compared to *S. bicolor* and the other *S. propinquum* genomes. The copy number estimates for 4-Gypsy and Keama are significantly higher in *S. propinquum* USDA (1,357 ± 27 and 5,208 ± 84 copies, respectively) compared to *S. bicolor* and other *S. propinquum* genomes (Figure 3 and Online resource 3). Further, in a recent phylogenetic analysis of *Sorghum*, this accession of *S. propinquum* contained *S. bicolor*-like alleles at all studied loci and resolved firmly within the *S. bicolor* clade (Hawkins et al. 2015). In contrast, phylogenetic analysis employing the assembled chloroplast genomes resolves *S. propinquum_USDA* outside of the *S. bicolor* clade (Govindarajulu and Hawkins, unpublished data). Therefore, it seems likely that *S. propinquum_USDA* has resulted from hybridization with and introgression of *S. bicolor* DNA. It is widely recognized that interspecific hybridization can cause reactivation of cryptic TEs, commonly referred to as the “genomic shock hypotheses” (McClintock 1984). Therefore, it is possible that hybridization-induced transposition has contributed to the larger genome size of this *S. propinquum* accession.

**Conclusions**

Our study provides a powerful method for accurate copy number estimation using short-read sequence data, and demonstrates it's use in describing LTR-retrotransposon diversity both between species that diverged as little as 2 million years ago, and among individuals of a single species. This approach detected even small differences in TE copy number within and between genomes over extremely short evolutionary timescales, suggesting that TE proliferation may be more frequent than expected, and therefore may contribute significantly to the genetic diversity that is driving intraspecific genome divergence. Further, we detected cryptic evidence of introgressive hybridization in one of the four *S. propinquum* accessions, as TE content of one of the *S. propinquum* accession is more similar to the *S. bicolor* genomes. Future work should focus on
determining the functional consequences of TE insertions to determine the extent to which TEs serve as drivers of divergence and speciation in plants.

**Acknowledgements**
The authors wish to thank James C. Estill and Michael Carlise for assistance with bioinformatic analyses, and Alex Harris for providing laboratory assistance. We acknowledge the use of the Super Computing System Spruce Knob at West Virginia University, which is funded by the National Science Foundation EPSCoR Research Infrastructure Improvement Cooperative Agreement 1003907, the state of West Virginia (WVEPSCoR via the Higher Education Policy Commission) and West Virginia University. This work was funded by the Department of Biology, West Virginia University and the WVU Advanced Energy Initiative.
References


El Baidouri M, Panaud O (2013) Comparative genomic paleontology across plant


Sanmiguel P, Bennetzen JL (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. Ann Bot 82:37–44.


Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize


Table 1: Flow cytometry results for *Sorghum bicolor* and *S. propinquum*. The DNA content of each sample is calculated based on the values of G0+G1 peak means [(Sample/Standard chicken erythrocyte nuclei (CEN)) x 2.5)]. Haploid genome size is derived by converting the mean picogram (pg) weight into Mbp [(DNA content in pg x 978 Mb)/2) where 1pg = 978 Mb]. The average and standard error for each accession is shown in bold font.

<table>
<thead>
<tr>
<th>Sample Accessions</th>
<th>Sample G0+G1 mean</th>
<th>Standard G0+G1 mean CEN</th>
<th>DNA content 2C (pg)</th>
<th>SE</th>
<th>Genome Size (Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B35</td>
<td>257.44</td>
<td>365.08</td>
<td>1.76</td>
<td></td>
<td>867</td>
</tr>
<tr>
<td></td>
<td>275.81</td>
<td>389.48</td>
<td>1.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>298.40</td>
<td>422.31</td>
<td>1.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>1.77</strong></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>RTx430</td>
<td>253.58</td>
<td>353.86</td>
<td>1.79</td>
<td></td>
<td>877</td>
</tr>
<tr>
<td></td>
<td>277.63</td>
<td>384.57</td>
<td>1.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>296.05</td>
<td>416.38</td>
<td>1.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>1.79</strong></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>SC56</td>
<td>256.07</td>
<td>359.31</td>
<td>1.78</td>
<td></td>
<td>876</td>
</tr>
<tr>
<td></td>
<td>277.48</td>
<td>386.37</td>
<td>1.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>298.78</td>
<td>420.85</td>
<td>1.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>1.78</strong></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Shanqui red</td>
<td>247.08</td>
<td>356.64</td>
<td>1.73</td>
<td></td>
<td>843</td>
</tr>
<tr>
<td></td>
<td>265.49</td>
<td>387.17</td>
<td>1.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>288.50</td>
<td>417.91</td>
<td>1.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>1.72</strong></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Tx7000</td>
<td>247.25</td>
<td>354.24</td>
<td>1.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>--------</td>
<td>----</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>263.95</td>
<td>382.40</td>
<td>1.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>284.63</td>
<td>410.67</td>
<td>1.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S. propinquum USDA</strong></td>
<td>232.78</td>
<td>315.21</td>
<td>1.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>251.62</td>
<td>340.97</td>
<td>1.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>269.29</td>
<td>368.56</td>
<td>1.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S. propinquum BR</strong></td>
<td>240.94</td>
<td>354.33</td>
<td>1.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>261.97</td>
<td>384.12</td>
<td>1.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>287.22</td>
<td>420.09</td>
<td>1.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>1.70</strong></td>
<td><strong>0.002</strong></td>
<td><strong>833</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Copy number estimates for the total number of LTRs from short read datasets using both reference-based and de novo approaches. Asterisk (*) indicates unknown genome size, where the *S. propinquum BR* genome size was used for estimating copy numbers in *S. propinquum* 369-1 and *S. propinquum* 369-2.

<table>
<thead>
<tr>
<th>Accessions</th>
<th>Genome size (Mb)</th>
<th>LTR (Reference)</th>
<th>LTR (de novo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTx623 <em>(In silico)</em></td>
<td>730</td>
<td>31409 ± 158</td>
<td>69558 ± 196</td>
</tr>
<tr>
<td>B35</td>
<td>867</td>
<td>72000 ± 237</td>
<td>76741 ± 282</td>
</tr>
<tr>
<td>RTx430</td>
<td>877</td>
<td>74561 ± 189</td>
<td>72369 ± 274</td>
</tr>
<tr>
<td>SC56</td>
<td>876</td>
<td>71780 ± 236</td>
<td>79954 ± 291</td>
</tr>
<tr>
<td>Shanqui red</td>
<td>843</td>
<td>74665 ± 185</td>
<td>73007 ± 269</td>
</tr>
<tr>
<td>Tx7000</td>
<td>848</td>
<td>74052 ± 239</td>
<td>74376 ± 262</td>
</tr>
<tr>
<td><em>S. propinquum USDA</em></td>
<td>902</td>
<td>77836 ± 251</td>
<td>68949 ± 270</td>
</tr>
<tr>
<td><em>S. propinquum BR</em></td>
<td>833</td>
<td>54025 ± 221</td>
<td>77232 ± 276</td>
</tr>
<tr>
<td><em>S. propinquum 369-1</em></td>
<td>*</td>
<td>43131 ± 186</td>
<td>72986 ± 251</td>
</tr>
<tr>
<td><em>S. propinquum 369-2</em></td>
<td>*</td>
<td>54106 ± 213</td>
<td>77779 ± 261</td>
</tr>
</tbody>
</table>
Figure 1: The LTR retrotransposon families of *S. bicolor* BTx623. Each node represents a single 5’ LTR while the edges/lines connecting the nodes indicate the sequence similarity between LTRs. Nodes are clustered into individual families according to a community structure interpretation. The clusters that are identified as known families from the grass database are uniquely colored, whereas the clusters with unknown family association are pale pink. The unknown clusters were named based on their superfamily designations (gypsy/copia/unknown).
Figure 2: LTR-retrotransposon dating in *S. bicolor* BTx623. A. Each node (representing a single full-length LTR-retrotransposon insertion in the BTx623 genome) is colored based on the amount of sequence divergence among the 5’ and 3’ LTR, which was further used to determine time since insertion. Cool colors (blue) represent older insertions with the most divergent LTRs at 85% similarity and an insertion estimate of approximately 5.8 million years ago (mya), while warm colors (red) represent recent insertions with LTRs of up to 100% sequence similarity. B. Graph representing estimated insertional timing for the three largest LTR-retrotransposon families as determined by the percent sequence similarity of 5’ and 3’ LTR of each element. The majority of the Retrosor-6 and 2-Unknown elements were inserted in the genome more recently (<1 mya) compared to elements belong to Onap.
Figure 3: LTR copy number estimates for the ten largest families calculated using a reference-based approach. Asterisks indicate two different clusters of Retrosor-6, which were deeply connected but formed two distinct clusters (Figure 1 – Retrosor-6-a –pale green and Retrosor-6-b - red). Because these clusters may represent the autonomous and non-autonomous sequences, we estimated copy numbers separately for each distinct cluster.
**Figure 4:** Copy number estimates for contigs containing LTRs for the ten largest LTR-retrotransposon families via de novo analysis.
Supplemental – Online resources

Please use the link below to access supplemental figures and excel files

CHAPTER 4: EVOLUTIONARY DYNAMICS OF TRANSPOSABLE ELEMENTS FOLLOWING A SHARED POLYPLOIDIZATION EVENT IN THE TRIBE, ANDROPOGONEAE

Dhanushya Ramachandran¹, Michael McCain², Elizabeth Kellogg², Jennifer S. Hawkins¹

A paper prepared for submission to the journal *Genome Biology and Evolution*

Abstract

Both polyploidization and transposable element activity are known to be major drivers of plant genome evolution. Here, we exploit the *Zea-Tripsacum* clade to investigate the contribution of TE activation and accumulation on genomic divergence after a recent shared polyploidization event. Comparisons of TE evolutionary dynamics in various *Zea-Tripsacum* species, along with closely related diploid species *Urelytrum* and *Sorghum*, revealed existing variation in repeat content between all genomes included in the study. The repeat composition of *Urelytrum* is more similar to *Zea* and *Tripsacum* compared to *Sorghum*, irrespective of similarity in genome size with the latter. The similarity in the proportion of *copia* retrotransposons and satellite DNA in the *Zea*, *Tripsacum*, and *Urelytrum* genomes suggests amplification of these elements after the maize-sorghum split but before the allopolyploidization event leading to the *Zea-Tripsacum* lineage. Although the genomes of all species studied were abundant with LTR-retrotransposons, we observed an expansion of the *copia* superfamily exclusively in *Z. mays* (maize) and *T. dactyloides*. Additional analyses of the genomic distribution of *copia* elements in maize provided evidence of biased insertion proximal to genes involved in various biological processes including plant development and defense. The lack of *copia* insertions near the orthologous genes in cultivated *S. bicolor* suggests that duplicate gene copies may offer new neutral sites for TEs to insert, thereby providing an avenue for subfunctionalization via TE insertional mutagenesis. Therefore, TE amplification and polyploidization may complement one another in shaping genetic architecture during maize domestication.
**Introduction**

Transposable element (TE) activation and accumulation generates significant genetic variation that can confer a range of effects on genome structure and function. As TEs carry ‘ready-to-use’ cis-elements, their insertions can impact gene regulation on a genome-wide scale by providing assorted regulatory elements to the adjacent genes. The new regulatory elements offered by inserted TEs can amplify and/or redistribute transcription factor binding sites, therefore creating new regulatory networks or even participate in re-wiring of pre-existing networks (Hènaff et al. 2014, Lavialle et al. 2013, Krupovic et al. 2014, Huang et al. 2016, Carmona et al. 2016, Joly-Lopez et al. 2016). Several empirical studies have demonstrated TE-induced phenotypic changes associated with domestication and/or diversification of cultivated plants, including rice, maize, wheat, soybean, melon, palm etc. (Naito et al 2009, Fernandez et al 2010, Studer et al. 2011, Uchiyama et al. 2013, Sanseverino et al. 2015, Ong-Abdullah et al. 2015, Lu et al. 2017). Indeed, TE-related polymorphisms are largely responsible for phenotypic variation in many agronomically important crops, demonstrating their importance in creating the genetic variability that contributes to plant genome evolution.

Hybridization, polyploidy, and stress are considered the primary triggers of transposable element movement (Steward et al. 2000, Kalendar et al. 2000, Madlung et al. 2005, Ungerer et al 2006, Ito et al. 2011, Cavrak et al 2014, Bardil et al. 2015, Guo et al. 2017). Flowering plants are known to tolerate hybridization and polyploidy, both of which have promoted species diversification (Payseur and Rieseberg, 2016, Soltis et al. 2016, Goulet et al. 2017). These phenomena result in TE mobilization leading to local mutations and genome size changes (Liu and Wendel 2000; Josefsson et al. 2006; Ungerer et al. 2006; Kawakami et al. 2010; Parisod et al. 2010; Piednoël et al. 2013). Furthermore, such bursts of TE activity result in insertional polymorphisms, often with deleterious effects on genome function; however, these effects could be nullified or shielded via gene duplication in polyploid genomes. Although the precise mechanism(s) that induce TE mobility in hybrids and polyploids is unclear, it is speculated that such TE reactivation in response to genomic stresses could be due to incompatible suppression machinery between the two donor genomes, or that unknown mechanisms are in place that reduce genomic methylation under general stress conditions (Ha et al. 2009, Yaakov...

Previous studies of polyploidy in Zea have revealed evidence for a whole genome duplication (WGD) event at or shortly after the origin of grasses, followed by another, more recent, WGD in the Zea history that promoted the origin of the Zea-Tripsacum clade. Being emerged from a common ancestral allotetraploid (n=20), both the Zea (n=10) and Tripsacum (n=18) genomes differentially responded to the rediploidization process (Swignova et al. 2004, Schnable et al. 2009, Schnable & Freeling, 2011). In addition to these chromosomal rearrangements, there is also evidence for retrotransposon invasion post divergence in both Zea and Tripsacum (Gaut et al. 2000). Hence, being divergent descendants of a common allopolyploid ancestor, the Zea-Tripsacum clade is a good model system to understand various evolutionary processes including the contribution of TEs to polyploidy, rediploidization, and species diversification.

Here, we describe TE activation and contribution to genome diversity in the Zea-Tripsacum clade that has undergone a recent shared polyploidization event. We included a close diploid progenitor, Urelytrum digitatum, which provides an opportunity to explore TE-associated evolutionary events induced by hybridization and genome doubling. By using clustering analysis, we have characterized the repetitive landscape in six Zea-Tripsacum species (post allopolyploidization) compared to the diploid sister taxa Urelytrum and Sorghum (pre allopolyploidization). Our findings suggest post-divergence and recent activity of TEs in Zea and Tripsacum with an expansion of copia elements in cultivated lineages compared to wild relatives. Biased insertions in euchromatic regions in Z. mays but not in S. bicolor suggests allopolyploidy induced retrotransposition in Z. mays. Also, with more insertions near developmental and defense genes and as TEs carry their own cis-elements, these elements may have influenced the evolution of the maize genome during domestication.

**Materials and methods**

**Plant material sources and Illumina sequencing of DNA**

The following eight panicoid grasses were used in this study: Zea mays, Z. diploperennis, Z. luxurians, Tripsacum dactyloides, T. laxum, T. australe, Urelytrum
digitatum and Sorghum bicolor. Short-read sequence data for Zea mays (SRS291653), Zea luxurians (SRR088692), Tripsacum dactyloides (SRS302460), and Sorghum bicolor (SRS1323776) were downloaded from the NCBI short read archive (Chia et al. 2012, Tenaillion et al. 2011, Ramachandran et al. 2016). Genome sequences of Zea diploperennis (XXXXXX), Tripsacum laxum (MIA34792), Tripsacum austral (MIA34499) and Urelytrum digitatum (SM3109) were obtained from Dr. Elizabeth Kellogg, Donald Danforth Plant Science Center, St. Louis, Missouri. See Supplementary Table 1 for more information on genome sequencing.

Identification of TE families

Sequences were quality trimmed using Trimmomatic v0.33 (Bolger et al. 2014) using a sliding window of 4:25 and minimum length of 50 bp. Graph-based clustering of quality-trimmed reads was performed with RepeatExplorer, a pipeline designed to identify repeats from NGS reads (Novak et al. 2013). RepeatExplorer employs a clustering algorithm that quantifies similarities between all sequence reads and produces a graph that consists of nodes (sequence reads) and edges (connecting overlapping reads). Nodes are frequently connected to one another if they pass a threshold of 90% similarity over at least 55% of the sequence length, representing individual repetitive families.

Three million reads (approximately 0.2x to 0.5x genome coverage) were subsampled from each dataset and processed to the format required by RepeatExplorer. Species-specific clustering analysis provides information regarding repeat quantities by reporting the number of reads per cluster, which can then be used to estimate the genome space occupied by each particular repeat, i.e., (total length of each cluster (in Mb) x genome size (in Mb)) / total length of all clusters (in Mb) (Kelly et al. 2015, Ramachandran et al. 2016). Subsequently, all of the processed reads from all species were concatenated into one combined dataset, and the RepeatExplorer clustering was repeated in order to facilitate comparative analysis. All clusters were annotated using the Viridiplantae RepeatMasker library and categorized into repeat families. A plot representing interactions between repeat clusters among species was created using UpSetR (Lex et al. 2014).
Quantitative analysis of TE activity using molecular clock analysis

To estimate the timing of TE activity in each lineage, species-specific LTR sequences were extracted from each LTR-retroelement cluster. These species-specific reads were assembled using the Geneious de novo assembler to obtain a consensus sequence (Kearse et al. 2012). A grass-specific database was then used to extract LTRs from each consensus contigs (blastn, e-value 1e-10, 85% identity). The best match for each species was chosen and the corresponding hit region was extracted using BEDTools v2.17.0 (Quinlan and Hall 2010).

To calculate LTR divergence (a rough measurement to estimate the age of a specific retrotransposon family) the reads that were used for de novo assembly were mapped to the consensus LTR sequence using the Geneious reference genome assembler. The percent identity of each read mapped to its respective LTR consensus sequence was derived from the reference alignment. Using a grass specific transposable element substitution rate of $1.3 \times 10^{-8}$ per site per year (Ma and Bennetzen, 2004), we estimated the activity of each major TE family in each species.

Genomic distribution of copia retroelements

To test whether the copia elements that have expanded in select species demonstrate an insertional bias, Illumina paired-end reads from Z. mays were mapped to a library consisting of Z. mays copia clusters assembled by RepeatExplorer and to a filtered gene set containing the protein-coding genes from the Z. mays reference genome. Reference mapping of paired-end reads to the library was carried out using BWA version 0.7.12 (Li and Durbin, 2009) with the following parameters: aln -t 4 -l 12 -n 4 -k 2 -o 3 -e 3 -M 2 -O 6 -E 3 (Mascagni et al, 2015). The results were used to generate a “sam” file via the BWA “sample” module, and then converted to a “bam” file using SAMtools (Li et al, 2009). A copia element was considered proximal to a gene if one of the paired-end read mapped to a copia element and the other to a gene. Genes proximal to copia elements were further analyzed for their presence in gene-dense or gene-poor regions by determining the number of TEs present within various distances (1 kb, 5 kb, and 10 kb).
both upstream and downstream of genes using BEDTools v2.17.0 (Quinlan and Hall 2010).

**Phylogenetic analysis of retroelement families**

To assess the evolutionary relationships of the shared gypsy and copia families, the reverse transcriptase (RT) and integrase (INT) amino acid domains were used for phylogenetic analysis. RepeatExplorer clusters were filtered for LTR-gypsy and copia elements with RT and INT domain blastx hits. RT reads were extracted from each cluster using the blastx output file and placed in separate genome-specific files. The reads were assembled for each cluster using the Geneious de novo assembler (Kearse et al. 2012). The resulting contigs were then confirmed to contain reverse transcriptase domains using blastx against the Cores-RT database (Llorens et al. 2011). RT sequences were then combined into a final query file for further analysis. The same analysis was performed for INT reads using Cores-INT database (Llorens et al. 2011).

Rpsblastn (e-value = 1e-10) was performed for the sequence dataset against the Conserved Domain Database (Marchler-Bauer et al. 2015) to identify and extract conserved regions. The best hits for each sequence were extracted, and the filtered blast output was converted to three-column bed format with matching coordinates for each hit. BEDTools v2.170 (Quinlan and Hall 2010) was used to extract the conserved regions (~540 bp for the RT domain and ~340 bp for the INT domain). The correct open reading frame from each sequence was identified using ORFfinder. All amino acid sequences were globally aligned with MUSCLE v3.8.31 (Edgar 2004). Alignments were manually inspected and adjusted in Bioedit v7.3.5 (Hall, 1999). The optimal model of amino acid substitution for each alignment was estimated using Prot-test v3.4.5 (Abascal et al. 2005). In all cases except RT-copia, the best model selected was LG+G (Le and Gascuel. 2008). Blosum62+G was chosen as the optimal model for RT-copia (Henikoff and Henikoff. 1992). Likelihood analyses with 1,000 bootstrap replicates were performed in RAxML v.8 (Stamatakis et al. 2008) using the best model for each alignment. Bayesian analysis of alignments was performed in MrBayes v3.2.6 using rates=gamma and respective substitution model (Ronquist and Huesenbeck. 2003). Two independent MCMC runs of
10 million generations were performed, sampling each run every 1,000 generations. All trees were visualized using FigTree v1.4.0.

**Results**

**Repeat composition in the genomes of Zea, Tripsacum, Urelytrum, and Sorghum**

To evaluate the repeat content with respect to genome size, we performed a separate clustering analysis for each species. Individual clustering allows the maximum number of reads to assemble in each cluster, which increases the accuracy of the repeat estimates. We estimated the quantities of each repeat family in the genome using the following equation: (total length of each cluster (in Mb) x genome size (in Mb)) / total length of all clusters (in Mb) (Macas et al. 2015). The estimated repeat compositions are shown in **Table 1**.

As expected, LTR-retrotransposons are the most abundant repeat in all eight genomes. Although all Zea species used in this study are diploid and contain the same number of chromosomes, the genome size of *Z. luxurians* (~4,479 Mb) is nearly double the size of other two Zea species (~2,600 Mb). From the clustering analysis, *copia* elements were found to contribute approximately 710 Mb, 930 Mb, and 1,110 Mb to the *Z. diploperennis*, *Z. mays* and *Z. luxurians* genomes, respectively. *Gypsy* elements account for ~1,240 Mb and 1,420 Mb of the *Z. mays* and *Z. diploperennis* genomes, respectively, whereas ~2,390 Mb of *Z. luxurians* genome is comprised of *gypsy* elements (**Table 1**). The greater repeat abundance in *Z. luxurians* correlates with its larger genome size. The *Tripsacum* species contain genomes of similar size (~3,200 Mb) and chromosome number (2n=36), in which *T. laxum* contains the smallest genome (2,974 Mb). *Copia* elements occupied ~740 and 780 Mb in *T. australis* and *T. laxum* genomes, in contrast to 1,050 Mb in the *T. dactyloides* genome. Approximately 1,760 Mb and 1,825 Mb of the genome is composed of *gypsy* elements in *T. laxum* and *T. australis*, respectively, whereas the *T. dactyloides* genome contains 1,230 Mb of *gypsy* elements. *Gypsy* elements contributed to more of the genome space (~53-59%) compared to *copia* (22-28%) in all Zea-Tripsacum species except *Z. mays* and *T. dactyloides*, where both *gypsy* and *copia* were equally distributed (**Figure 1B**). *Urelytrum* and *Sorghum* contain ~167 Mb and 92 Mb of *copia*, and 438 Mb and 536 Mb of *gypsy* elements, respectively.
DNA transposons were found to contribute only 2-6% to the Zea and 2-3% to the Tripsacum genomes, in contrast to 10-11% in Urelytrum and Sorghum. Other groups of repeat elements such as satellite repeats made up a significant fraction of the genome in several species. Approximately 755 Mb of the Z. luxurians and 510 Mb of the T. dactyloides genomes were occupied by satellite repeats. Although Urelytrum and Sorghum contain genomes of similar size, the former is composed of only 1.76 Mb of satellite DNA whereas the latter contained ~100 Mb of satellite DNA in its genome.

The most abundant repeat families and their contribution to genome size

From the individual repeat clustering analysis, we identified 24 copia and 30 gypsy families. Among the 24 copia families, Ji was the most abundant family in both the Z. mays (444 Mb) and Z. diploperennis (363 Mb) genomes, whereas Opie was the most abundant in Z. luxurians (535 Mb) and in all of the Tripsacum genomes (Table 1). Dijap was estimated at 146-240 Mb in the three Tripsacum genomes, but contributed very little to the genome size of Zea.

Among the gypsy families, Cinful-Zeon, Prem1, Flip, Gyma, Huck and Xilon-Diguus were abundant in both the Zea and Tripsacum genomes. The Cinful-Zeon family ranges from 224 - 583 Mb among the three Zea genomes with the greatest abundance in the larger Zea genome; however, this family contributes only ~70 Mb to the Tripsacum genomes. This is also true for the Xilon-Diguus family, with estimates ranging from 125 - 226 Mb in Zea and ~42 Mb in the Tripsacum genomes. The Huck family is estimated at 246 in Z. diploperennis, 321 Mb in Z. luxurians, 152 in T. laxum and 276 Mb in T. australe; however, Huck occupies only ~15 Mb of the Z. mays genome and ~1.4 Mb of the T. dactyloides genome. Similarly, elements such as Doke, Puck, Lata and CRMI were more abundant in the wild species relative to the domesticated species.

There were 13 gypsy families that were specific to Urelytrum and/or Sorghum. Athila and Leviathan elements (~19 – 66 Mb) were identified in both Urelytrum and Sorghum. Apart from these two families, the remaining 11 gypsy families were predominantly present in Sorghum, but present in low copy number in Urelytrum, or absent altogether. For example, Retrosor6 is estimated at ~180 Mb in the Sorghum genome but is completely
absent in all other species; however, there are a large number of unclassified gypsy elements in the Urelytrum genome (See Table 1). Although we used a grass specific database to annotate the elements, the majority of this repeat content could not be annotated, suggesting the presence of species-specific repeats and retroelements.

**Insertional biases in Z. mays**

The copia superfamily was found to be more abundant in Z. mays and T. dactyloides compared to the other species included in the study, suggesting recent proliferation of some copia families in both genomes. Investigating the genomic distribution of this expansion, we discovered that the frequency of Z. mays copia reads mapping to stress-associated genes (~34%) was higher compared to other genes (on average ~6%). Copia elements mapped in close proximity to genes involved in plant defense, leaf morphogenesis, photoreceptors, homeobox proteins, signal transduction, and transcription (Figure 5). Further analysis revealed that these genes were surrounded with approximately four to five genes within 5kb windows both upstream and downstream.

**Comparative analysis of Zea, Tripsacum, Urelytrum and Sorghum**

We performed comparative repeat analysis by simultaneously clustering reads from all eight species. This approach facilitated the identification of repeat families that are shared between multiple species, and allowed us to determine their fate during Andropogoneae evolution, especially during the divergence of Zea and Tripsacum. This analysis resulted in four major cluster configurations, for which examples are shown in Figure 3A-D. Figure 3A shows an example of a cluster (2: Prem1, LTR-gypsy) in which the repeat family is common to all species. In this example, reads from both Zea and Tripsacum are tightly clustered, and reads from Sorghum and Urelytrum are peripherally connected, as would be expected based on their evolutionary relationships. Cluster 6 (Opie, LTR-copia) is an example of a lineage-specific repeat family, where sequences are shared between Zea and Tripsacum but absent in Urelytrum and Sorghum (Figure 3B). In Cluster 21 (Flip, LTR-gypsy), the graph indicates three separate groups (Z. mays and T. dactyloides [top], Z. diploperennis and Z. luxurians [right], T. australe and T. laxum
[left]) in which Z. mays and T. dactyloides are more similar to one another than either is to their sister species (Figure 3C). Finally, cluster 64 (Angela, LTR-copia) is an example of a tightly knitted graph in a linear arrangement shared between all eight species, demonstrating the conserved nature of ancient Angela elements across all included taxa (Figure 3D).

From a total of two million reads from eight genomes, 248 significant clusters were formed of various sizes and repeat families. On average, ~81% of the reads from each species clustered with LTR-retrotransposons (127 LTR-gypsy and 48 LTR-copia clusters, Figure 2A). Among the 175 LTR-RT clusters (or families) identified, 85 families were present exclusively in the Zea-Tripsacum clade. For all species except Z. mays and T. dactyloides, the proportion of reads from LTR-gypsy families (53%) was higher compared to LTR-copia families (28%), whereas gypsy and copia were equally abundant in Z. mays and T. dactyloides. Compared to the other genomes, Sorghum contained the smallest proportion of reads from copia families.

Among the 127 gypsy clusters, four clusters were shared among all eight species, two clusters were common to Zea, Tripsacum and Urelytrum (but absent in Sorghum), 34 clusters were exclusive to Zea and Tripsacum species, and 15 clusters were found only in Urelytrum and Sorghum. In addition, we observed lineage-specific gypsy families: 10 in Zea, 17 in Tripsacum, 21 in Urelytrum, and 14 in Sorghum (Figure S1. A). Of the 48 copia clusters, only two were common to all species, ten were common to Zea, Tripsacum and Urelytrum, 19 clusters were exclusive to Zea and Tripsacum, and 3 were exclusive to Urelytrum and Sorghum. Compared to gypsy super-families, there were fewer species-specific copia families (Figure S1. B).

**Evolutionary relationships and timing of transposition events**

To assess the timing of major transposition events that occurred pre- and post-divergence of the Zea-Tripsacum clade, we constructed maximum likelihood trees using INT and RT (data not shown) of both gypsy and copia elements. Of the 127 shared gypsy clusters, 15 (total of 82 sequences) shared sufficient sequence identity within the integrase domain to allow amino acid sequence alignment. Major repeat families such as Cinful-xeon, Prem1, Flip, Gyma, Xilon-Diguus, and Huck were among these 15 clusters.
With a few exceptions, most clades formed as expected in regard to species relationships, such as all *Zea* and *Tripsacum* species clustered together with *Urelytrum* and *Sorghum* being more distantly related (data not shown). The gypsy families *Flip* and *Gyma* clustered together. The *Sorghum* and *Urelytrum* sequences from the *Flip* family clustered with *Zea* sequences of *Gyma*, whereas the *Zea* and *Tripsacum* sequences of *Flip* clustered with *Tripsacum* and *Sorghum* of *Gyma*. Several families such as *huck*, *puck*, and *grande* were clustered together with high support values, suggesting a recent origin of these families. Clusters such as CL24 (unclassified), *uwum* (CL82) and *guhis* (CL132) also clustered with high sequence similarity.

We employed comparative sequence analyses of LTRs from 15 prominent clusters to estimate the temporal activity of retroelements both pre- and post-divergence of the *Zea-Tripsacum* clade (Figure 4). The clusters chosen for this analysis are comprised of the following repeat families: *Prem1*, *Flip*, *Cinfu-Xeon*, *Gyma*, *Ji*, *Opie*, *Dijap*, *Retrosor-6*, and several prominent unclassified elements. In Figure 4A, the peak activity of each element per species per cluster is plotted against a TE-specific grass molecular clock (11 mya to present). The approximate timing of the *Zea-Tripsacum* divergence is highlighted in yellow (5-6 mya). *Zea* and *Tripsacum* have experienced post divergence lineage-specific activity for most repeat families. For example, *Ji*, *Opie*, and *Dijap* (CL7, CL12, CL15, CL42, and CL51) were active between 0-3 mya for all species in which they are present. The *Opie* element represented in CL7 is shared between *Zea*, *Tripsacum* and *Urelytrum* and has been active within the last ~1-3 mya (Figure 4A & 4B) indicating that amplification of *Opie* occurred in all three lineages after species divergence. In contrast, the amplification of CL2 (*prem1*) occurred recently only in *Z. luxurians* (2-3 mya) compared to all other species. Although *T. dactyloides* and *T. laxum* experience increased activity of *Prem1* around the time of divergence, the activity of this element in *Z. mays*, *Z. diploperennis*, *T. australae*, and *Sorghum* dates as an older amplification event. Similarly, the activity of CL5 (*gyma*) in *Z. diploperennis* is recent but lost in *Z. luxurians*. Several families were shared only between *Sorghum* and the wild relatives of *Zea* (*Z. diploperennis*, *Z. luxurians*) and *Tripsacum* (*T. laxum*, *T. australae*). Despite their presence, the activity of these families varies between species. For example, the activity of CL11 in *Z. luxurians* is recent (0-1 mya) but is dated as an old insertion in
the other species (Figure 4A and 4C). In contrast, elements in CL19 display post-divergence activity in all species. For example, Z. diploperennis and S. bicolor experienced CL19 activity around 1-2 mya, whereas in Z. luxurians and T. dactyloides, CL19 elements were active around 2-3 mya (Figure 4A).

**Discussion**

The present study evaluates TE dynamics in divergent descendants (Zea-Tripsacum) of a common allopolyplloid ancestor within a phylogenetic framework that is rooted with two diploid relatives (Urelytrum and Sorghum). The comparative analysis of repeat elements from Zea, Tripsacum, Urelytrum, and Sorghum provides insight into the contribution of retrotransposons to genome evolution post a shared polyploidization event. Inclusion of additional Zea and Tripsacum species provided an opportunity to assess the genomic variability in repeat content between wild and cultivated genotypes.

As expected, LTR-retrotransposons account for the majority of the repeat composition in the genomes of all species included in this study. Individual clustering analyses indicate that a diversity of LTR-retrotransposons contribute to genome size variation in this taxonomic group. Based on our comparative and molecular clock analyses, the majority of retrotransposon families are common to the Zea – Tripsacum clade in comparison to their diploid relatives, suggesting an occurrence of retrotransposon invasion after allopolyplloidization but before the split between the two species (Figure 4A). Previous studies have hypothesized an occurrence of retroelement bursts just before the divergence of Zea and Tripsacum based on maize retroelement activity (Gaut et al. 2000, Estep et al. 2013). The results for the Tripsacum species included in the current analysis supports this hypothesis, revealing a high number of shared retrotransposon families between the two species. For example, Ji and Opie of the copia superfamily have been especially active (0-2 mya, Figure 4A, >300 Mb, Table 1) in both Zea and Tripsacum; however, these families contribute little (~35 Mb) to genome composition in Urelytrum and are absent in Sorghum. The presence and hyper activity of these families in the Zea-Tripsacum-Urelytrum clade but not in Sorghum suggests amplification after the maize-sorghum split but before the allopolyplloidization event leading to the Zea-Tripsacum lineage. Similarly, five gypsy families (Cinfu-Xeon,
Prem1, Flip, Gyma, and Huck) are abundant in the Zea-Tripsacum clade but present in low copy numbers in the other lineages. Molecular clock analysis reveals recent activity (1-4 mya) for these families in Zea and Tripsacum, suggesting amplification post divergence from other taxa (Figure 4A). Conversely, families such as Athila and Leviathan have accumulated in the Urelytrum and Sorghum genomes, but are absent in Zea and Tripsacum, suggesting independent activation of LTR-retrotransposon families in different lineages over short evolutionary time scales.

Additionally, both Zea and Tripsacum contained genus-specific families, indicating variation in retroelement amplification in each species post divergence. Compared to Zea, Tripsacum contained more unique gypsy and copia families. There were nine gypsy families that were common to all Tripsacum genomes and seven families that were shared only between T. laxum and T. australine (Figure 2B & Figure 4A). The larger number of unique and recently active retroelements (~1-4 mya) in the Tripsacum lineage indicates the independent expansion of these families post divergence of the two genera. Overall, the abundance and recent activity of these genus-specific LTR-retrotransposons shared only between Zea and Tripsacum, suggests that the activation of these families might be an outcome of shared polyploidization as proposed by the genomic shock hypotheses (McClintock 1984, Comai et al. 2003).

Surprisingly, clustering analyses suggest that Z. mays and T. dactyloides share greater similarity in TE composition than either do to the other members of their respective genera. Also, a greater number of reads from both genomes are derived from the copia superfamily, suggesting independent expansion of copia clades in both lineages, reminiscent of results from comparisons in Asian rice varieties (Li et al. 2017). For a few shared copia clusters, the peak activity level of Z. mays overlaps with T. dactyloides, suggesting both species experienced copia activity during a similar time period (Figure 4A). Considering the independent evolution of both species post divergence, and the role of artificial selection in maize domestication, the similarity in composition and activity of copia elements in Z. mays and T. dactyloides indicates that natural and artificial selection have acted in a similar way in both lineages.

Additionally, considering the intensity of retroelement accumulation in maize as reported in other studies, it is likely that these copia elements have been active during
maize domestication. Studies demonstrating TE involvement in plant domestication predominantly show that the insertions were proximal to functional genes that were important to plant function and/or development. Well-known examples include *Hopscotch* involvement in apical dominance in maize (Studar et al, 2011), *Gret1* in berry color variation in *Vitis vinifera* (Cadle-Davidson et al. 2008), and LTR-mediated control of the blood orange phenotype (Butelli et al. 2012). Because the clustering analysis revealed recent *copia* expansion in cultivated lineages, we explored the frequency of *copia* insertions near genes in *Z. mays*. In *Z. mays*, a large number of reads from *copia* elements mapped in close proximity to several functionally relevant genes. Indeed, the majority of *copia*-associated genes are involved in plant defense, homeobox proteins responsible for shoot apical meristem and leaf morphogenesis, cytokinin response, signal transduction, and transcription (*Figure 5*). Additionally, although gene density in *Z. mays* is approximately one gene per 3.2 kb (Fu et al. 2001), the *copia* insertions identified in this study were surrounded by approximately 4-5 genes upstream and downstream in a 5kb interval, suggesting biased insertion in more gene-rich regions. Such close proximity provides the potential for TEs to affect the function of neighboring genes, as seen in other plants (Makarevitch et al. 2015, Cao et al. 2016, Pietzenuk et al. 2016). For example, the tobacco *Tnt1* and the rice *To17 copia* elements were found near stress-related genes, and the expression of these elements is linked with the biological responses of the plant to the external stresses (Grandbastein et al. 1997, Miyao et al. 2003, Le et al. 2007).

Though we report *copia* insertions in gene-rich regions of the *Z. mays* genome, we could not confirm the same for *T. dactyloides* due to the lack of a high-quality genome assembly. To test whether this pattern is common in other domesticates, however, we performed the same analysis for *S. bicolor*. In *Z. mays*, there were 32 *copia* elements inserted within a 1kb interval of a stress gene (GRMZM2G047919) whereas in *S. bicolor*, there were only two *copia* elements found near a stress related gene (SORBI_3009G188300). It is possible that the disruptive nature of insertional mutagenesis was buffered in *Z. mays* by the presence of duplicate genes in the allopolyploid.
Conclusion

In this study, we provide insight into interspecific TE diversity and its contribution to genome evolution in related members of the Andropogoneae tribe that have undergone a shared polyploidization event. By including multiple accessions of two divergent species (Zea and Tripsacum) originating from a common allopolyploidy ancestor, in addition to close diploid relatives (Urelytrum and Sorghum), we described LTR-retrotransposon diversity with respect to the hybridization and genome doubling process. Though the genome size of Urelytrum is similar to that of Sorghum, the repeat composition of Urelytrum is more like that of Zea and Tripsacum. Similarities in the proportion of the copia superfamily and satellite DNA in the Zea-Tripsacum-Urelytrum clade suggests that Urelytrum or a close relative may have played a role in the evolution of the Zea-Tripsacum lineages. Our clustering analysis revealed an expansion of the copia superfamily exclusively in Z. mays and T. dactyloides, suggesting participation of new copia insertions during the domestication process. Further analyses provided evidence for insertion in euchromatic gene-rich regions, specifically near genes involved in plant development and defense. Though we could not perform the same for T. dactyloides and the other wild relatives, the presence of copia insertions proximal to genes in Z. mays but not S. bicolor suggests allopolyploidy induced retrotransposition in Z. mays. As a majority of these insertions are near genes involved in plant development and defense, the cis-regulatory effects of gene-proximal TEs may have influenced the evolution of plant architecture and host defense mechanisms during maize domestication.
References


Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon


COPIA elements. Genome Biol. 17.


Table 1: Global repeat composition (in Mb) of species with respect to genome size. Genome size (in Mb) for each species is given below each species name. Estimated repeat content (in Mb) for each repeat family is listed below using individual repeat clustering analysis. Bold text represents the most abundant families in each genome.

<table>
<thead>
<tr>
<th>LTR/Copia</th>
<th>Z. mays (2600 Mb)</th>
<th>Z. dip (2591 Mb)</th>
<th>Z. lux (4479 Mb)</th>
<th>T. dact (3200 Mb)</th>
<th>T. aus (3229 Mb)</th>
<th>T. lax (2974 Mb)</th>
<th>Urelytrum (711 Mb)</th>
<th>Sorghum (848 Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ji</td>
<td><strong>444.02</strong></td>
<td><strong>362.72</strong></td>
<td><strong>519.58</strong></td>
<td><strong>313.39</strong></td>
<td><strong>102.17</strong></td>
<td>89.23</td>
<td>34.41</td>
<td>0.00</td>
</tr>
<tr>
<td>opie</td>
<td>327.96</td>
<td>224.42</td>
<td><strong>534.88</strong></td>
<td><strong>444.17</strong></td>
<td><strong>286.28</strong></td>
<td><strong>356.65</strong></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>giepum</td>
<td>45.85</td>
<td>26.29</td>
<td>22.95</td>
<td>84.73</td>
<td>20.09</td>
<td>48.38</td>
<td>7.13</td>
<td>13.74</td>
</tr>
<tr>
<td>machiavelli</td>
<td>12.07</td>
<td>3.73</td>
<td>3.95</td>
<td>3.17</td>
<td>0.00</td>
<td>6.61</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>wiwa</td>
<td>10.28</td>
<td>3.33</td>
<td>8.16</td>
<td>9.01</td>
<td>13.90</td>
<td>23.07</td>
<td>5.92</td>
<td>0.00</td>
</tr>
<tr>
<td>uloh</td>
<td>7.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>gudyeg</td>
<td>7.03</td>
<td>0.00</td>
<td>2.79</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>eninu</td>
<td>2.44</td>
<td>2.35</td>
<td>0.00</td>
<td>5.15</td>
<td>6.86</td>
<td>9.71</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dijap</td>
<td>2.27</td>
<td>2.63</td>
<td>0.00</td>
<td>145.58</td>
<td>240.43</td>
<td>159.12</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>stonor</td>
<td>1.22</td>
<td>0.00</td>
<td>0.00</td>
<td>6.06</td>
<td>2.71</td>
<td>4.62</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>anar</td>
<td>1.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>raider</td>
<td>0.00</td>
<td>3.73</td>
<td>0.00</td>
<td>0.00</td>
<td>11.01</td>
<td>3.71</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>fourf</td>
<td>0.00</td>
<td>6.67</td>
<td>11.39</td>
<td>1.48</td>
<td>0.97</td>
<td>1.37</td>
<td>0.00</td>
<td>3.44</td>
</tr>
<tr>
<td>ebel</td>
<td>0.00</td>
<td>9.42</td>
<td>0.00</td>
<td>0.00</td>
<td>43.14</td>
<td>17.72</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>maximus</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>47.48</td>
<td>15.94</td>
<td>0.00</td>
</tr>
<tr>
<td>gudyeg</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>3.10</td>
<td>0.00</td>
</tr>
<tr>
<td>gina</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.79</td>
<td>0.00</td>
</tr>
<tr>
<td>Angela</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.58</td>
<td>1.66</td>
</tr>
<tr>
<td>osr14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.48</td>
<td>0.00</td>
</tr>
<tr>
<td>Ale</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.30</td>
<td>0.48</td>
</tr>
<tr>
<td>ruda</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.00</td>
</tr>
<tr>
<td>Os4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>7.25</td>
</tr>
<tr>
<td>nuhan</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>6.49</td>
</tr>
<tr>
<td>ekasi</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.28</td>
</tr>
<tr>
<td>Unclass.Copia</td>
<td>70.06</td>
<td>67.44</td>
<td>4.20</td>
<td>35.85</td>
<td>13.57</td>
<td>62.07</td>
<td><strong>63.53</strong></td>
<td><strong>40.61</strong></td>
</tr>
<tr>
<td>Total</td>
<td>931.50</td>
<td>712.72</td>
<td>1107.90</td>
<td>1048.58</td>
<td>741.14</td>
<td>782.26</td>
<td>166.85</td>
<td>91.90</td>
</tr>
<tr>
<td></td>
<td>LTR/Gypsy</td>
<td>Gypsy cinful-zeon</td>
<td>premI</td>
<td>flip</td>
<td>gyma</td>
<td>xilon-diguus</td>
<td>teKay</td>
<td>uwum</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>--------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>289.37</td>
<td>223.63</td>
<td>583.02</td>
<td>58.70</td>
<td>80.87</td>
<td>65.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>220.81</td>
<td>70.62</td>
<td>308.55</td>
<td>170.72</td>
<td>51.08</td>
<td>104.28</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>194.03</td>
<td>150.07</td>
<td>351.10</td>
<td>279.59</td>
<td>323.29</td>
<td>166.96</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>146.91</td>
<td>169.35</td>
<td>183.39</td>
<td>213.95</td>
<td>167.51</td>
<td>223.19</td>
<td>4.68</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>142.19</td>
<td>124.76</td>
<td>225.79</td>
<td>45.64</td>
<td>35.02</td>
<td>45.12</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>52.17</td>
<td>23.23</td>
<td>23.41</td>
<td>83.83</td>
<td>14.26</td>
<td>63.59</td>
<td>0.00</td>
<td>32.73</td>
</tr>
<tr>
<td></td>
<td>43.27</td>
<td>15.26</td>
<td>18.71</td>
<td>80.17</td>
<td>88.45</td>
<td>64.29</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>36.39</td>
<td>80.04</td>
<td>136.62</td>
<td>0.00</td>
<td>86.64</td>
<td>44.80</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>27.21</td>
<td>17.46</td>
<td>13.73</td>
<td>17.69</td>
<td>18.29</td>
<td>37.73</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>20.93</td>
<td>47.00</td>
<td>50.30</td>
<td>30.26</td>
<td>55.60</td>
<td>119.78</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>14.69</td>
<td>246.19</td>
<td>321.64</td>
<td>1.42</td>
<td>275.99</td>
<td>152.18</td>
<td>0.00</td>
<td>5.37</td>
</tr>
<tr>
<td></td>
<td>7.55</td>
<td>74.35</td>
<td>78.71</td>
<td>0.00</td>
<td>39.71</td>
<td>46.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>4.53</td>
<td>4.12</td>
<td>9.87</td>
<td>1.41</td>
<td>0.94</td>
<td>1.59</td>
<td>0.00</td>
<td>9.71</td>
</tr>
<tr>
<td></td>
<td>3.99</td>
<td>79.25</td>
<td>23.28</td>
<td>0.00</td>
<td>96.93</td>
<td>90.92</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>17.19</td>
<td>0.00</td>
<td>44.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>31.73</td>
<td>0.00</td>
<td>2.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.61</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>56.64</td>
<td>29.53</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>31.73</td>
<td>0.00</td>
<td>2.38</td>
<td>0.00</td>
<td>0.00</td>
<td>15.04</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>18.95</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>20.66</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>11.10</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.15</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>75.87</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>181.60</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>40.13</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.23</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.95</td>
</tr>
<tr>
<td>Unclass.Gypsy</td>
<td>37.05</td>
<td>93.91</td>
<td>29.59</td>
<td>226.53</td>
<td>488.30</td>
<td>492.92</td>
<td>296.70</td>
<td>77.75</td>
</tr>
<tr>
<td>Total</td>
<td>1241.09</td>
<td>1419.24</td>
<td>2389.45</td>
<td>1227.10</td>
<td>1825.87</td>
<td>1762.61</td>
<td>437.76</td>
<td>536.43</td>
</tr>
<tr>
<td></td>
<td>Unclass.</td>
<td>0.00</td>
<td>87.57</td>
<td>159.20</td>
<td>126.08</td>
<td>102.82</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Retroelements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINE</td>
<td>2.72</td>
<td>0.00</td>
<td>1.00</td>
<td>2.91</td>
<td>0.00</td>
<td>0.00</td>
<td>1.22</td>
<td>0.94</td>
</tr>
<tr>
<td>SINE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.60</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>DNA transposons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA.CMC.EnSpm</td>
<td>80.22</td>
<td>61.68</td>
<td>77.51</td>
<td>85.00</td>
<td>88.81</td>
<td>107.27</td>
<td>36.06</td>
<td>48.78</td>
</tr>
<tr>
<td>DNA.MULE.MuDR</td>
<td>25.64</td>
<td>4.59</td>
<td>3.20</td>
<td>1.17</td>
<td>15.25</td>
<td>12.62</td>
<td>27.78</td>
<td>6.49</td>
</tr>
<tr>
<td>DNA.hAT.Ac</td>
<td>18.51</td>
<td>0.94</td>
<td>0.00</td>
<td>10.56</td>
<td>1.64</td>
<td>1.59</td>
<td>0.74</td>
<td>2.18</td>
</tr>
<tr>
<td>Helitron</td>
<td>17.78</td>
<td>9.44</td>
<td>6.75</td>
<td>2.28</td>
<td>0.00</td>
<td>0.00</td>
<td>0.35</td>
<td>1.78</td>
</tr>
<tr>
<td>DNA.PIF.Harbinger</td>
<td>14.96</td>
<td>0.00</td>
<td>1.29</td>
<td>4.51</td>
<td>0.00</td>
<td>3.62</td>
<td>2.47</td>
<td>29.57</td>
</tr>
<tr>
<td>DNA.TcMar.Stowaway</td>
<td>2.40</td>
<td>0.00</td>
<td>0.00</td>
<td>3.04</td>
<td>0.00</td>
<td>0.00</td>
<td>3.60</td>
<td>6.25</td>
</tr>
<tr>
<td>Total</td>
<td>159.90</td>
<td>76.64</td>
<td>88.75</td>
<td>116.75</td>
<td>105.70</td>
<td>125.09</td>
<td>71.00</td>
<td>95.05</td>
</tr>
<tr>
<td>Other Repeats</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satellite</td>
<td>65.84</td>
<td>272.15</td>
<td>754.91</td>
<td>509.75</td>
<td>239.95</td>
<td>379.73</td>
<td>1.76</td>
<td>95.85</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>63.14</td>
<td>46.10</td>
<td>0.00</td>
<td>52.94</td>
<td>107.11</td>
<td>13.80</td>
<td>31.42</td>
<td>12.58</td>
</tr>
<tr>
<td>rRNA</td>
<td>5.88</td>
<td>61.01</td>
<td>21.74</td>
<td>0.00</td>
<td>34.84</td>
<td>51.58</td>
<td>1.70</td>
<td>14.41</td>
</tr>
<tr>
<td>Low complexity</td>
<td>13.86</td>
<td>6.61</td>
<td>5.21</td>
<td>15.97</td>
<td>0.00</td>
<td>0.00</td>
<td>1.71</td>
<td>0.00</td>
</tr>
<tr>
<td>Unknown</td>
<td>20.72</td>
<td>3.65</td>
<td>20.38</td>
<td>43.23</td>
<td>0.00</td>
<td>7.41</td>
<td>1.19</td>
<td>4.03</td>
</tr>
<tr>
<td>Genome proportion</td>
<td>75%</td>
<td>74%</td>
<td>66%</td>
<td>66%</td>
<td>61%</td>
<td>60%</td>
<td>68%</td>
<td>61%</td>
</tr>
</tbody>
</table>
Figure 1. A. The evolutionary relationships of selected grass species, indicating polyploidization and species divergence. B. Proportional repeat composition. Genome size in Mb shown for each species in the y-axis. An expansion of copia families is observed in both cultivated Z. mays and T. dactyloides compared to related sister species. Sorghum displays a predominance of gypsy elements with a low level accumulation of copia families compared to Zea-Tripsacum-Urelytrum genomes.
Figure 2. A. Bar graph showing the distribution of 248 largest clusters with respect to various repeat families. B. UpSet plot showing the interactions of shared repeat clusters among eight species. Each species is represented in one row with filled and empty cells. Each column represents the intersection between each species. From left to right, elements shared in all eight species to elements unique to each species is shown. Filled cells indicate that the element is shared with other species. The bars above each intersecting row represent the intersection size.
Figure 3. Comparative graph-based clustering. Graphs of individual repeat clusters that are shared between species demonstrating existing sequence variants within species. Highlighted dots represent sequences from individual species and lines connecting the dots represent sequence similarity. Each species is represented with a unique color: Red (Z. mays), purple (Z. diploperennis), pink (Z. luxurians), green (T. dactyloides), orange (T. laxum), and grey (T. austral.). A. Cluster 2 shows shared LTR-gypsy elements in all genomes, in which sequences of Zea and Tripsacum are tightly connected with each other and sequences from Urelytrum and Sorghum are peripherally connected, concordant with their evolutionary relationships. B. Cluster 6 (Opie, LTR-copia) is an example of a lineage-specific repeat family, where sequences are shared between Zea and Tripsacum; however, there is a clear separation in clustering of both lineages. C. Cluster 21 (Flip, LTR-gypsy) shows three separate groups in which the cultivated genomes Z. mays and T. dactyloides are more similar to one another than either is to their sister species. D. Cluster 64 (Angela, LTR-copia) is an example of a tightly knitted graph in a linear arrangement shared between all eight species, demonstrating the conserved nature of ancient Angela elements across all included taxa.
Figure 4. A. Activity of retroelements pre-and post-divergences of the Zea-Tripsacum clade (yellow line) for 15 prominent retrotransposon clusters (shaded gray). Concentric circles indicate time scale per million years from 11 mya (center) to present (outer circle). For each cluster, the corresponding repeat family and shared species information is given below each cluster name. Each data point represents the peak activity of that element. B & C display retrotransposon activity of CL7 & CL11 based on percent identity of shared LTR sequences (bottom axis) and the corresponding grass molecular clock (mya) along the top axis. CL11 is absent in domesticated Zea and Tripsacum but present in the wild relatives.
**Figure 5:** Paired end read mapping to *copia* elements and nearby gene in *Z. mays*. The majority of the TEs proximal to genes are involved in plant development and defense, such as terpene synthase, beta galactosidase, profilin-3, and blue-light receptor phototropin.
Figure S1. UpSet plots showing the interactions of shared gypsy (A) and copia (B) clusters. Each species is represented in one row with filled and empty cells. Each column represents the intersection between each species. Clusters are displayed from most common (on the left) to least common (on the right). Cells are either filled or empty indicating whether the element is shared with other species. The bars above each intersecting row represent the intersection size.
CHAPTER 5: COMPARATIVE ANALYSIS OF TRANSPOSABLE ELEMENTS IN GOSSYPIUM DIPLOIDS AND POLYPLOIDS PROVIDES INSIGHT INTO POLYPLOIDY AND CROP DOMESTICATION

Dhanushya Ramachandran, Jennifer S. Hawkins

Abstract

Transposable element (TE) proliferation along with polyploidization play a major role in shaping plant genome architecture and evolution. To characterize both common patterns and lineage-specific differences in TE composition and evolution in response to genome doubling, we have investigated Gossypium allopolyploids (AD) that originated from a single interspecific hybridization between two divergent diploid (A and D) taxa. Using clustering analysis, we have estimated global abundances of various repetitive composition both within and among diploid and polyploid Gossypium relatives. Comparative analyses of repeats revealed tetraploids being ~62% repetitive, exhibiting an additive pattern of the diploid progenitor’s overall repeat composition (A-genome: ~77%; D-genome:55%). Also, we observed particularly high similarity in repeat composition between allopolyploids (AD) and diploid A-genomes relative to the D-genome diploid. Although we find nearly all TE lineages shared between AD and A-genomes, the number of repeat families shared between polyploids and G. arboreum is higher compared to G. herbaceum. Upon analyzing the proximity of recent polyploid-specific TEs near gene regions, it appears that many are near genes involved in various steps of biosynthesis of cotton fiber and mechanisms of pathogen defense. Additionally, we observed contrasting patterns of subgenome-specific insertions near genes between gypsy and copia retroelements. Genes found near young copia insertions are primarily located in the dominant and highly expressed A-subgenome, whereas gypsy insertions are predominantly located in the D-subgenome. Finally, the presence of recent, intact insertions (≤ 1 Mya) near genes in polyploids but absent in the parental diploids suggests that these insertions originated post genome merger and doubling. This study provides a starting point for future studies aimed at understanding the roles of TEs on Gossypium allopolyploid genome evolution and cotton fiber cell development.
Introduction

Originally described as “junk” or “selfish” DNA, transposable elements (TEs) are now recognized as factors that create genetic novelty involved in host adaptation and genome evolution. Although transposition of resident TEs is largely controlled by the host genome to prevent significant mutagenic effects, circumstances that trigger quiescent TEs are often described, some with the ability to alter the structure and expression of neighboring genes (Steward et al. 2000, Kalendar et al. 2000, Madlung et al. 2005, Ungerer et al. 2006, Ito et al. 2011, Cavrak et al. 2014, Bardill et al. 2015, Guo et al. 2017). In addition to these local affects on genes, insertions can impact the genome at a global scale. For instance, recombination between two TEs can result in deletions of the interleaving genome sequence, or create chromosomal rearrangements, as observed in maize and Arabidopsis (Weil and Wessler, 2003, Hughes et al. 2003). TE mediated chromosomal rearrangements may be an important mechanism contributing to reproductive isolation, species diversification in plants, and crop domestication.

Although TEs are largely dormant in many genomes, possible triggers for transpositional bursts include environmental and genomic stresses such as whole genome duplication (WGD) events and interspecific hybridization (Parisod et al. 2009, Petit et al. 2010, Piednoël et al., 2013, Senerchia et al. 2014, Ågren et al. 2016). Polyploidization plays an important role in shaping genomes structure and function in many plants and animals. By introducing an additional complement set of chromosomes, polyploid genomes go through one of the most dramatic mutational event that profoundly impacts entire cellular architecture, meiosis and mitosis pairing, gene redundancy, gene regulatory changes, evolution rate etc. (Otto 2007, Soltis et al. 2015). Despite posing a major challenge to genome stability, gain of a complete set of chromosomes provides genomic variation for polyploid lineages that is not available to their diploid progenitors.

Polyploidization is frequently accompanied by an increase in TE content because of transpositional bursts following genome collision. Unlike diploid genomes, the presence of duplicate copies of all essential genes provide a relatively well-buffered genomic environment from the deleterious consequences of transposition; however, although genome merger can induce TE activation, these changes are mostly limited to the first few generations after polyploidy that are subsequently followed by re-establishment of the TE silencing mechanisms (Vicient and Casacuberta 2017). Nevertheless, the consequences of such transpositional bursts
can extend to many generations. Therefore, even in the absence of new transposition events, higher abundances of resident TEs can lead to recombination and induce gene losses/mutations and genome restructuring (Soltis et al. 2015). As polyploidy is a common theme in plants and often it involves fusion of two diverged genomes, it is of interest to explore the consequences of genomic merger and doubling on TE activation in polyploids.

Here, we used the cotton genus, *Gossypium*, to describe TE dynamics in response to interspecific hybridization and whole genome duplication. The well-established evolutionary framework for the allopolyploid formation makes *Gossypium* as an ideal system to investigate emergent consequences of polyploidy on TE activation. *Gossypium* includes ~45 diploid (2n=2x=26) species divided into eight genome groups (A-G, K) and five polyploids (AD). A-genome diploids, native to Africa, and Mexican D-genome diploids differ two-fold in genome size and diverged from one other ~5-10 million years ago (Mya). Allotetraploids (AD) arose in the New World from interspecific hybridization between an A- and a D-genome species, which occurred 1-2 Mya (Wendel et al. 2009). The nascent allopolyploid diverged into five species (Figure 1). *G. mustelinum* is restricted to a small region in NE Brazil; Each of the other two lineages are represented by two species, one of which is cultivated and other is an island-endemic: *G. hirsutum* (domesticated) and *G. tomentosum* (Hawaii); *G. barbadense* (domesticated) and *G. darwinii* (Galapagos island). The New World allotetraploids *G. hirsutum* (AD)1 and *G. barbadense* (AD)2 and the Old-World diploids *G. herbaceum* (A1) and *G. arboreum* (A2) were independently domesticated and become a major oilseed for fiber production (Wendel et al. 2009). By using resequencing data, clustering analysis, and bioinformatic approaches, we assessed the impact of polyploidization on the dynamics of transposable elements in polyploids compared to their diploid progenitors.

Materials and methods

Plant material sources and Illumina sequencing

The following eight *Gossypium* genomes were used in this study: *Gossypium herbaceum* (A), *Gossypium arboreum* (A), *Gossypium raimondii* (D), *Gossypium hirsutum* (AD), *Gossypium tomentosum* (AD), *Gossypium barbadense* (AD), *Gossypium darwinii* (AD), and *Gossypium mustelinum* (AD). Short-read sequence data of *Gossypium herbaceum* (SRR2039552), *Gossypium arboreum* (SRR959522), *Gossypium raimondii* (SRR067329), *Gossypium hirsutum*
(SRR768357), *Gossypium tomentosum* (SRR1975565), *Gossypium barbendense* (SRR1975557), *Gossypium darwinii* (SRR1975567), and *Gossypium mustelinum* (SRR769542) were downloaded from the NCBI short read archive. All libraries were sequenced using Illumina HiSeq (100 bp paired end reads with ~500 bp insert size) as mentioned in the SRA details online.

**Qualitative identification of repeat families**

Sequences were quality trimmed using Trimmomatic v0.33 (Bolger et al. 2014) with the following settings: sliding window 4:25, minimum length 50 bp. Graph-based clustering of quality-trimmed reads was performed using RepeatExplorer, a pipeline designed to identify repeats from NGS reads (Novak et al. 2013). RepeatExplorer employs a clustering method that quantifies similarities between all sequence reads and produces a graph that consists of nodes (sequence reads) and edges (connecting overlapping reads). Nodes are frequently connected to one another if they pass a threshold of 90% similarity over at least 55% of the sequence length, representing individual repetitive families.

One to five million reads (0.2x genome coverage) were sub-sampled from each dataset and processed to the format required by RepeatExplorer. Individual clustering analysis provides species-specific information regarding repeat quantities by reporting the number of reads per cluster, which can then be used to estimate the genome space occupied by each repeat, i.e., \( \frac{\text{total length of each cluster (in Mb) x genome size (in Mb)}}{\text{total length of all clusters (in Mb)}} \) (Kelly et al. 2015, Ramachandran et al. 2016). After the individual analysis described above, the processed reads were concatenated into one combined dataset for comparative clustering analysis. Reads were prefixed with species names (-c 8 -f 5). All resulting clusters were annotated based on repeat families using a *Gossypium* specific repeat database retrieved from the Plant Genome and Systems Biology (PGSB) repeat database and *G. raimondii* transposable elements database (http://www.grtedb.org/). Clusters that were common to all species in addition to clusters found only in a subset of species were identified.

**Insertional bias near protein coding genes**

To test whether the retroelements are biased towards insertion in gene-rich regions, Illumina paired-end reads of all eight species were mapped to a library consisting of *gypsy/copia/TXX* contigs assembled by RepeatExplorer and reference protein coding genes
(Ramachandran et al. in preparation). With the availability of gene annotations for *G. arboreum* (A), *G. raimondii* (D), *G. hirsutum* (AD), and *G. barbendense* (AD), paired end reads from each genome were mapped to a library consisting of their respective RepeatExplorer (RE) contigs and closest extant relative reference gene sequences. For example, *G. tomentosum* reads were mapped to RE contigs assembled from *G. tomentosum* reads and *G. hirsutum* reference protein coding gene models; similarly, *G. darwinii* reads were mapped to RE contigs assembled from *G. darwinii* reads and *G. barbendense* protein coding gene models.

Reference mapping of paired-end reads to the library was carried out using BWA version 0.7.12 (Li and Durbin, 2009) with the following parameters: aln -t 4 -l 12 -n 4 -k 2 -o 3 -e 3 -M 2 -O 6 -E 3 (Mascagni et al, 2015). The resulting paired-end mappings were used to generate a sam file via BWA “sample” module, and then converted to a bam file using SAMtools (Li et al, 2009). A gypsy/copia/TXX element was considered proximal to a gene if one of the paired-end reads mapped to a gypsy/copia/TXX element and the other to a gene. Genes proximal to repetitive elements were further analyzed for their presence in gene-dense or gene-poor regions using BEDTools v2.17.0 (Quinlan and Hall 2010). In addition, for genes with recent retrotransposon (based on 5’LTR and 3’LTR similarity of a full-length element) insertions in the AD genomes, we examined the presence/absence of similar insertions near homologs in their diploid ancestor genomes in order to determine if the insertion occurred pre- or post-polyploid formation. We performed this analysis only for species with available genome assemblies such as *G. raimondii* (D), *G. arboreum* (A), and *G. hirsutum* (AD).

Results

**Clustering of Next-Generation Sequences from Eight Species of *Gossypium***

To characterize the repeat content in diploid and polyploid genomes of *Gossypium*, we randomly subsampled short read sequences from each genome and performed comparative repeat analysis by simultaneously clustering reads from all eight species. From a total of 30 million reads, ~10 million reads clustered to produce ~600,000 clusters ranging from a minimum of 2 reads to over 21,000 reads per cluster. We focused the remainder of our analyses on clusters that contained at least 1000 reads (total of 296 clusters) and repeat families from 6,895,304 reads (Figure 2A). Both diploid D (*G. raimondii*) and A genomes (*G. arboreum*, *G. herbaceum*) contained ~55% and ~77% of repetitive DNA, respectively. Tetraploid cotton genomes (AD) are
comprised of ~62% repetitive DNA. Not surprisingly, LTR-retrotransposons represent the major fraction of all genomes, particularly LTR-gypsy elements (Figure 2B).

Depending on the species, gypsy elements comprise between 297 and 1,030 Mb of the genome (Table 1). On average, 67% of reads from each species contributed to 132 gypsy clusters, in which 47 clusters were common between all eight genomes, 68 clusters were shared only between the two diploid A genomes and polyploids, 14 clusters were shared by G. raimondii (D) and polyploids, and 3 clusters were common between G. raimondii (D), G. arboreum (A) and all AD genomes, but absent in G. herbaceum (A) (Figure 3B). As expected, copia families were less abundant when compared to gypsy retroelements, comprising between 84 and 192 Mb of the genomes (Table 1). The overall copia abundance remains similar (~88 Mb) among all diploids irrespective of the variation in genome size, in contrast to the gypsy superfamily. Also, 44 of 56 copia clusters were shared between all eight genomes and 50% of the remaining clusters were shared between the A and AD genomes (Figure S1). Other than retroelements, unknown mobile elements (TXX) are the next largest category comprising between 142 and 182 Mb in all genomes, with the exception of G. raimondii (9 Mb). Among the 34 clusters in this category, 27 clusters were shared between the two A diploids and all polyploids (Figure S1). Other repeat classes such as DNA transposons, satellite, simple repeats etc., follow a similar pattern where most clusters in each category were common between the A and AD genomes compared to D and AD genomes (Table 1). When repeat families were ranked by their prevalence in each genome, the first four groups, gypsy, copia, TXX, and DNA-TE, were ranked similar in all AD genomes; however, the fifth most abundant group was simple sequence repeats in G. hirustum (28 Mb) and G. mustelinum (24 Mb) but not in the other three tetraploids (Table 1). In diploids, following the gypsy superfamily, the TXX content (~173 Mb) was higher than that of copia families (~90 Mb) in the A-genomes, thereby comprising the second most abundant group. Last, the SSR abundance (Mb) in these genomes was similar to G. barbendense, G. darwinii, and G. tomentosum.

Gene proximity near functional genes

Upon determining the genomic distribution of retroelements, we discovered that paired end reads corresponding to repeat families frequently mapped near protein coding genes. Analyses were performed separately for each repeat category. By mapping to genotype-specific
retrotransposon consensus sequence and reference genes, we observed an increase in the total number of paired end reads mapped to a gene and repeat sequence in *G. tomentosum* and *G. darwinii* compared to domesticated allopolyploids (Table 2). This observation was common to both gypsy and copia elements. Although the total number of reads mapped near genes was higher in the A-genome diploids (similar to island-endemic polyploids), reads from the diploid A-genomes mapped to a greater number of genes (~10,500) compared to that of the AD genome species (~7,500).

The protein coding genes proximal to gypsy contigs were similar in evolutionarily closely related genomes. For example, reads from *G. hirsutum*, *G. tomentosum*, and *G. mustelinum* mapped near genes such as UDP-Glycosyl transferase superfamily protein, protein kinase superfamily protein, long-chain acyl-CoA synthetase 7, leucine-rich repeat transmembrane protein kinase, iron-sulfur cluster binding, tetratricopeptides repeat (TPR) like superfamily protein, and multidrug-resistance associated proteins. The *Arabidopsis* gene annotation report was used to retrieve the biological processes of the observed protein coding genes. The majority of the genes are involved in lipid, fatty acid, and carbohydrate metabolisms, and response to abscisic acid, ethylene, ozone, salt, and osmotic stress; however, reads from *G. barbendense* and *G. darwinii* mapped near genes that were responsible for DNA/RNA binding and repeat related proteins. In diploids, the genes that mapped proximal to gypsy contigs were mostly involved in photosynthetic electron transport systems. By examining the presence/absence of intact LTR-retrotransposons near these genes, we observed insertions dating 0.19-0.77 Ma present in *G. hirsutum* genome, whereas similar insertions were absent near orthologs in both diploid ancestors (Figure S2). Also, *G. hirsutum* genes that mapped near newly inserted gypsy elements were predominantly distributed in the D-subgenome (Figure 4B).

For copia elements, *G. hirsutum* reads mapped near genes such as photosynthetic electron transfer B, GDSL-like lipase, glycosyl hydrolase, NB-ARC and TIR-NBS-LRR domain containing disease resistance proteins. The majority of copia reads from *G. barbendense* and *G. darwinii* mapped near cysteine-rich receptor like protein kinase (CRK8), which is involved in biotic defense response to bacterium (Figure 3B). Similar to gypsy insertions, recent copia insertions (0.19-1.15 mya) were present only in *G. hirsutum* (AD). In contrast to gypsy insertions, the majority of *G. hirsutum* genes near copia elements are more frequently located in the A-subgenome (Figure 4D).
Discussion

Here, we used next-generation sequencing data to analyze the global repeat composition within and among eight cotton (Gossypium) species, which includes five allotetraploids formed 1-2 Mya from intergenomic hybridization between the A- and D-genome diploids (Figure 1). Comparative analysis of repeats revealed tetraploids are ~62% repetitive, exhibiting an additive pattern of the diploid progenitor’s overall repeat composition (A-genome: ~77%; D-genome:55%). As expected from genomic analyses in other plant species, LTR-gypsy retrotransposons account for the majority of diploid and polyploid cotton genomes (Figure 2). Being the smallest genome included in the study, G. raimondii (880 Mb/1C) had the smallest abundance (297 Mb) of gypsy elements and G. barbendense had the greatest abundance (1,030 Mb, Table 1). Among the largest repeat clusters identified in the comparative analysis, 132 clusters represented LTR-gypsy elements, in which 50% of the clusters were exclusively shared only between polyploids (AD) and the A-genome species (Figure 3A and 3B). In addition, among the two A-genome species included in the analysis, it appears G. arboreum shares more similarity with tetraploids compared to G. herbaceum. In contrast to our results, previous studies suggested that G. herbaceum is the closest extant ancestor of the original A-genome donor (Brown and Menzel 1950, Gertsel 1953, Menzel and Brown, 1954); however, as both extant A-genome species are phylogenetically sister to each other and hence equidistant from the A subgenome of allopolyploid cotton, the closest descendant of the ancestral diploid maternal parent is still under debate (Endrizzi, Turcotte & Kohel, 1985; Wendel & Cronn, 2003).

Overall, copia retrotransposon abundance is similar in all diploids (~88 Mb) in contrast to gypsy retrotransposons, irrespective of the disparity in their genome sizes. Hawkins et al., (2006) reported differential accumulation of copia-like sequences between the D-genome, G. raimondii and A-genome diploids; however, the results presented here reveal a predominance of gypsy-like sequences in both the A and D genome diploids and as reported by Renny-Byfield et al. 2016 (Table 1). Their analysis was based on cloned, whole genome shotgun sequences that were matched to the NCBI database with relatively low repeat content, whereas the use of Illumina short read sequences and plant specific repeat database in this study allows for more accurate annotation. Like gypsy elements, more copia clusters were shared between the A and AD genomes. A sizeable fraction of each genome except G. raimondii was attributable to unknown
mobile elements (TXX). Interestingly, these unknown sequences are specific to the *Gossypium* genus as we did not see any similarity with repeats present in other closely related taxa including *Arabidopsis*. In an attempt to identify any coding potential of these elements, we found only a few sequences matching to the *gag* internal coding domain of large LTR-*gypsy* retrotransposon families in soybean and pea genomes such as *Cyclops, Calypso, Peabody, Diaspora,* and *Ogre* at less stringent settings. In addition, these elements of unknown origin are the second most abundant repeat class in A-genome diploids following *gypsy* superfamily. The abundance of these elements in one of the parental genome as well as in all polyploids suggests that the expansion would have occurred in the progenitor genome and been retained after allopolyploid formation.

As polyploidization is frequently accompanied by an increase in TE content, and as the effects of retrotransposon mobility on plant phenotype is related to their insertion near genes, we evaluated the proximity of LTR-RTs to genes in all *Gossypium* genotypes. In polyploids, reads from cultivated genomes mapped near 8,000 genes on average with more reads mapped to genes in *G. hirsutum* compared to *G. barbadense*. In diploids, reads from both A-genomes mapped to 10,700 genes on average with an increase in the total number of reads mapped to retrotransposon and genes compared to polyploids. Overall, there are elevated numbers of gene-proximal insertions in three out of four domesticates, regardless of ploidy, and these insertions may have contributed to domestication, similar to our previous observations in maize (Ramachandran et al., in preparation).

By categorizing genes proximal to retrotransposons based on GO category, we found that the majority of the genes in polyploids were involved in lipid, fatty acid, and carbohydrate metabolism for both *gypsy* and *copia* super families (*Figure 4A and 4C*). As fiber-associated factors are involved in various steps in carbohydrate, lipids, fatty acid biosynthesis and elongation, it is interesting to note that retrotransposons are proximal to these genes (Gou et al. 2007, Liu et al. 2012, Yoo and Wendel, 2014, Fu et al. 2015, Hu et al. 2016, Parekh et al. 2018). Also, we found *copia* retrotransposons proximal to stress-related genes such as oxidative stress induced proteins and disease resistance proteins associated with Leucine-rich repeat (LRR) domains (*Figure 4C*), similar to our previous findings in maize (Ramachandran et al., in preparation). The leucine-rich repeats (LRR)-containing domain is evolutionarily conserved in
many proteins associated with innate immunity in plants and animals (McHale et al. 2006, Mendy et al. 2017, Wang et al. 2018, Huang et al. 2018). In plants, it has long been hypothesized that TEs play a major role in reconstruction of the genome in response to various stress factors including pathogen infection. As cotton is more susceptible to soil-borne fungal infections such as *Verticillium* wilt, *copia* insertions near LRR domains could contribute to resistance gene activation and diversification (Zhang et al. 2017).

On identifying the genome-wide location of genes proximal to retrotransposons, we found that genes near *gypsy* insertions were located predominantly in chromosomes of the D-subgenome in *G. hirsutum* (AD) genome (Figure 4B). While this biased distribution of new, polyploid-specific *gypsy* insertions near genes in the D-subgenome is clearly displayed, the underlying reason is unclear. Additionally, genes near *copia* insertions are more frequent in the A-subgenome compared to D-subgenome (Figure 4D). Li et al. (2015) described a large segmental replacement from the D-subgenome to the A-subgenome in *G. hirsutum*, causing a greater number of genes to be expressed in the current A-subgenome compared to its ancestral diploid A progenitor. This large replacement might cause substantial difference in TE activities between the two subgenomes compared to its progenitors. Further, the presence of young LTR-retrotransposon insertions (0.19 - 1.15 Mya) near genes in *G. hirsutum*, but absent in both diploid progenitors indicate post polyploidization insertions that were specific to AD genomes (Figure S2 and S3). Taken together, the results presented here demonstrate greater similarity in TE composition between diploid A genomes and allopolyploids, with gene proximal insertions biased towards the highly expressed A-subgenome compared to the D-subgenome in polyploids. Therefore, being inserted close to genes in the dominant genome, TEs may potentiate the genome in generating evolutionary novelty during domestication followed by polyploidy.

Conclusions

The present study provides insight into the evolution of TEs in cultivated *Gossypium* polyploids compared to their diploid progenitors. Comparative clustering analysis and other bioinformatic approaches revealed greater similarity in repeat composition between polyploids and diploid A-genomes compared to the D-genome, suggesting that the majority of TEs in polyploids may have been inherited from A-genome species. Additionally, gene-proximal TE insertions in cultivated cotton polyploids are predominantly near genes involved in various steps
of biosynthesis of cotton fiber and pathogen defense mechanisms; however, genes near copia-specific insertions were almost exclusively found in the highly expressed dominant A-subgenome, suggesting a higher probability of influencing host gene expression or function. As even a low number of insertions close to genes can have dramatic effects on plant form and function, preferential insertion of copia elements near functionally important genes in cultivated polyploids highlights their potential regulatory power during the evolution of the cotton polyploid genome and domestication.
References


Fu W et al. 2015. Acyl-CoA N-acyltransferase influences fertility by regulating lipid metabolism and jasmonic acid biogenesis in cotton. Scientific Reports. 5:11790.


Steward N, Kusano T, Sano H. 2000. Expression of ZmMET1, a gene encoding a DNA methyltransferase from maize, is associated not only with DNA replication in actively proliferating cells, but also with altered DNA methylation status in cold-stressed quiescent cells. Nucleic Acids Res. 28:3250–3259.


**Table 1.** Global repeat composition (in Mb) of species with respect to genome size. Genome size (in Mb) for each species is given below each species name. Estimated repeat content (in Mb) for each repeat family is listed below using separate repeat clustering analysis.

<table>
<thead>
<tr>
<th>Repeat_class</th>
<th>G. arb (1,682)</th>
<th>G. her (1,672)</th>
<th>G. hir (2,347)</th>
<th>G. bar (2,455)</th>
<th>G. tom (2,386)</th>
<th>G. dar (2,367)</th>
<th>G. mus (2,377)</th>
<th>G. rai (880)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR.gypsy</td>
<td>921.09</td>
<td>826.45</td>
<td>917.72</td>
<td>1029.75</td>
<td>1016.70</td>
<td>1012.86</td>
<td>889.23</td>
<td>296.96</td>
</tr>
<tr>
<td>LTR.copia</td>
<td>84.28</td>
<td>94.45</td>
<td>173.50</td>
<td>191.75</td>
<td>188.58</td>
<td>184.58</td>
<td>162.03</td>
<td>87.92</td>
</tr>
<tr>
<td>TXX</td>
<td>181.92</td>
<td>164.92</td>
<td>143.00</td>
<td>158.07</td>
<td>149.18</td>
<td>142.29</td>
<td>142.47</td>
<td>9.03</td>
</tr>
<tr>
<td>DNA-TE</td>
<td>49.01</td>
<td>55.57</td>
<td>104.12</td>
<td>90.64</td>
<td>82.47</td>
<td>85.28</td>
<td>93.82</td>
<td>36.17</td>
</tr>
<tr>
<td>Simple_repeat</td>
<td>6.07</td>
<td>8.62</td>
<td>27.95</td>
<td>9.14</td>
<td>10.01</td>
<td>8.47</td>
<td>23.20</td>
<td>7.37</td>
</tr>
<tr>
<td>rRNA</td>
<td>31.54</td>
<td>18.97</td>
<td>16.95</td>
<td>24.98</td>
<td>28.97</td>
<td>22.19</td>
<td>14.65</td>
<td>27.96</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.00</td>
<td>4.41</td>
<td>12.31</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>11.92</td>
<td>0.00</td>
</tr>
<tr>
<td>LINE</td>
<td>2.37</td>
<td>2.45</td>
<td>6.42</td>
<td>5.80</td>
<td>4.90</td>
<td>4.74</td>
<td>5.78</td>
<td>2.14</td>
</tr>
<tr>
<td>Low_complexity</td>
<td>5.02</td>
<td>5.56</td>
<td>5.17</td>
<td>6.15</td>
<td>7.28</td>
<td>6.88</td>
<td>5.04</td>
<td>2.44</td>
</tr>
<tr>
<td>Satellite</td>
<td>3.23</td>
<td>1.02</td>
<td>4.26</td>
<td>5.66</td>
<td>4.44</td>
<td>2.43</td>
<td>4.00</td>
<td>3.62</td>
</tr>
<tr>
<td>LTR.Caulim.</td>
<td>0.67</td>
<td>1.49</td>
<td>4.08</td>
<td>4.07</td>
<td>3.96</td>
<td>4.25</td>
<td>3.94</td>
<td>1.38</td>
</tr>
<tr>
<td>RC.Helitron</td>
<td>1.24</td>
<td>1.45</td>
<td>3.14</td>
<td>3.14</td>
<td>3.04</td>
<td>3.38</td>
<td>2.70</td>
<td>1.32</td>
</tr>
</tbody>
</table>

**Total Mb** 1294.15 1196.50 1432.76 1544.06 1514.69 1491.76 1372.97 485.34

**Table 2.** Proximity of retrotransposons near genes. Total number of Illumina paired reads of which one mapped to an LTR-gypsy and the other onto a G.hirsutum reference gene model. Asterisk indicates domesticated genomes.

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Genome</th>
<th>Total no. of mapped reads</th>
<th>Discordant PE reads</th>
<th>Discordant: one read to a LTR-RT and other to a gene</th>
<th>Reads mapped to total # of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. hirsutum</td>
<td>AD*</td>
<td>7,657,141</td>
<td>1,439,046</td>
<td>103,819</td>
<td>8,824</td>
</tr>
<tr>
<td>G. tomentosum</td>
<td>AD</td>
<td>7,377,297</td>
<td>2,634,882</td>
<td>198,281</td>
<td>7,459</td>
</tr>
<tr>
<td>G. barenense</td>
<td>AD*</td>
<td>5,851,661</td>
<td>1,187,050</td>
<td>75,205</td>
<td>6,363</td>
</tr>
<tr>
<td>G. darwinii</td>
<td>AD</td>
<td>7,539,153</td>
<td>2,322,207</td>
<td>220,750</td>
<td>7,817</td>
</tr>
<tr>
<td>G. mustelinum</td>
<td>AD</td>
<td>7,625,334</td>
<td>834,992</td>
<td>83,210</td>
<td>8,795</td>
</tr>
<tr>
<td>G. arboreum</td>
<td>A-genome*</td>
<td>8,799,322</td>
<td>2,775,576</td>
<td>237,432</td>
<td>10,342</td>
</tr>
<tr>
<td>G. herbaceum</td>
<td>A-genome*</td>
<td>8,055,060</td>
<td>3,263,926</td>
<td>370,358</td>
<td>11,133</td>
</tr>
<tr>
<td>G. raimondii</td>
<td>D-genome</td>
<td>1,970,593</td>
<td>217,547</td>
<td>22,165</td>
<td>3,044</td>
</tr>
</tbody>
</table>
**Figure 1.** Evolutionary framework for diploid and allotetraploid cotton. The ancestral A and D genome diploids diverged from a common ancestor ~5-12 million years ago (Mya), prior to interspecific hybridization ~1-2 Mya. The nascent allopolyploid (AD) further diverged into five species that radiated into three lineages. Blue colored star indicates independently domesticated species.
Figure 2. A. Estimation of total repeat abundance (Mb) in each *Gossypium* genome. B. Bar graph showing distribution of 296 clusters with respect to various repeat families. C. Repeat composition in diploids and allotetraploids of cotton.
Figure 3. A. UpSet plot showing the interactions of shared repeat clusters among *Gossypium* genotypes. Each species is represented in one row with filled and empty cells. Each column represents the intersection between each species. From left to right, elements shared in all eight species to elements unique to each species. Cells are either filled or empty representing whether the element is shared with other species or not. The bars above each intersecting row represent the intersection size. Plot created using UpSetR (Lex et al. 2014). B. UpSet plot showing interactions exclusively for LTR-Gypsy clusters. More common clusters between A diploid genome and AD tetraploids.
Figure 4. A. GO categories representing genes that were mapped near LTR-gypsy retrotransposons in *G. hirsutum*. B. Blast hits for *G. hirsutum* genes that were mapped near gypsy contigs in *G. hirsutum [AD]* and its postulated diploid ancestors (*G. raimondii [D]* & *G. arboreum [A]*). More hits (Yellow triangles on chromosomes) on Dt subgenome compared to At subgenome in *G. hirsutum*. C. GO categories representing genes that were mapped near LTR-copia retrotransposons in *G. hirsutum*. D. Blast hits for *G. hirsutum* genes that were mapped near Copia contigs in *G. hirsutum [AD]* and its postulated ancestors (*G. raimondii [D]* & *G. arboreum [A]*).
Supplementals

Figure S1. A. A plot showing the interactions of shared and unique LTR-copia retrotransposon clusters among *Gossypium* genotypes. Majority of clusters are shared between all eight species. B. Similar plot showing interactions of an unknown mobile element (TXX) among all genotypes. 27 out of 34 TXX clusters were common only between polyploids and A genome diploids.
CHAPTER 6: GENERAL CONCLUSIONS

Since McClintock's discovery TEs have been viewed as highly mutagenic elements as well as a prolific source for evolutionary innovation. As reviewed in Chapter 2, these self-replicating genetic elements are found in virtually all organisms and contribute nearly half of the genomic content, particularly in plants. Given the abundance and inherent mutagenic potential of these elements, their impact and significance are entirely dependent on the scale and perspective at which one focuses. Although the majority of TEs are selected against, there are a few elements that have conferred benefits to host adaptation and survival over the course of evolution. Despite multi-omics technologies and tools that have transformed the understanding of current molecular biology phenomena, it is still a difficult task to analyze and extract information from the non-genic (TEs) component of the genome. Having mentioned the regulatory roles of TEs elsewhere in Chapter 2 and realizing the complexity of an organism arises from the complexity in gene regulation, it is critical to analyze and decipher the potential of these non-genic sequences in this post-genomics era. In this dissertation, I have focused on the evolutionary dynamics of TEs and their contribution to host genome evolution. Beginning with our novel method and bioinformatics pipeline, I quantified intra- and interspecific TE diversity in various genera in the Andropogoneae tribe (Chapters 3 and 4). In addition, I have performed similar analyses of genome wide TE dynamics in an economically important dicot, Gossypium (Chapter 5).

In Chapter 3, I described repeat composition in Sorghum accessions and evaluated intraspecific TE-associated genome evolution over a short evolutionary time scale. As their repetitive nature and abundance makes characterization difficult, particularly from short-read sequencing projects, I developed a method that effectively estimates TE copy number from short-read sequences (Ramachandran et al. 2016). By testing the method on an in silico data set generated from the reference S. bicolor genome, I found that the estimated TE copy numbers were strikingly similar to the annotated numbers, proving the accuracy of the method. With the developed method in hand, and using both reference and de novo approaches, I further described intraspecific TE diversity among various accessions of cultivated (S. bicolor) and its close wild relative S. propinquum from real short read data. Results from comparative analyses revealed small but detectable variation in LTR-retrotransposon content and copy number among S. bicolor accessions and larger variation between the S. propinquum USDA accession and the other three S. propinquum accessions. In addition, I observed that the repeat profile of S.
propinquum USDA is more similar to that of the S. bicolor genomes, supporting the proposal that this germplasm is the product of introgressive hybridization of S. bicolor chromatin into this particular accession (Tang et al. 2013).

In Chapter 4, I extended my analysis to a larger group in the tribe Andropogoneae, to evaluate interspecific TE diversity at greater evolutionary timescales. As genome doubling and transposable element proliferation greatly influence plant genome evolution, and because genome merger is thought to induce TE activity, I included Zea, Tripsacum, Urelytrum, and Sorghum as my study systems, where Zea and Tripsacum emerged from a common allopolyploid ancestor 5-10 mya. By including the diploid sister taxa, Urelytrum and Sorghum, I was able to characterize the repetitive landscape and describe TE-associated dynamics pre- and post-polyploidization. The similarity in the proportion of copia super family and satellite DNA in Zea-Tripsacum-Urelytrum genomes suggests that Urelytrum or a close relative may have played a role in the evolution of the Zea-Tripsacum lineage. Surprisingly, we observed significant expansion of the copia superfamily exclusively in the cultivated accessions, Z. mays and T. dactyloides. Further analyses provided evidence of the presence of recent copia insertions near genes involved in plant development and defense mechanisms in Z. mays. As duplicate gene copies provide an avenue for subfunctionalization through TE insertional mutagenesis, the presence of TEs in genic regions indicate their possible involvement in maize genome evolution and domestication.

In Chapter 5, I set out to determine if gene-proximal TE insertions in cultivated polyploids is a phenomenon specific to the Zea-Tripsacum lineage by performing similar analyses in the cotton genus, Gossypium (Malvaceae). Containing five allopolyploid species, two of which are major cultivated cash crops, that emerged from a single allopolyploidization event just 1-2 mya, cotton serves as an ideal study system to investigate emergent consequences of polyploidy and TE activation on plant genome evolution. I performed clustering analysis and other bioinformatic approaches on the closest extant relatives of the allotetraploid parents along with the current five AD genome species. I observed higher similarity in repeat composition between allopolyploids and the diploid A-genomes compared to that of the D-genome. Upon analyzing the proximity of TEs near genes, I found that TE insertions were proximal to protein coding genes involved in various steps of biosynthesis of cotton fiber and pathogen defense
mechanisms. Altogether, we report differential TE dynamics in *Gossypium* polyploids following polyploidization, with a possible role of gene-proximal *copia* insertions in the evolution of domesticated polyploids.

In summary, this dissertation work has contributed a method to estimate TE copy number from short read sequences for genomes with or without reference transposable element annotation. Also, this work has provided insights on the genomic distribution and evolutionary dynamics of gene-proximal TEs in cultivated polyploids that may have played a role in shaping genome architecture and led to the emergence of genetic innovations in plant lineages.
References
