Transposable elements and genome size dynamics in Gossypium

Jennifer S. Hawkins
Iowa State University
Transposable elements and genome size dynamics in *Gossypium*

by

Jennifer S. Hawkins

A dissertation submitted to the graduate faculty
In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Botany

Program of Study Committee:
Jonathan F. Wendel, Major Professor
  Lynn G. Clark
  John D. Nason
  Thomas Peterson
  Randy Shoemaker
  Dan Voytas

Iowa State University
Ames, Iowa
2007

Copyright © Jennifer S. Hawkins, 2007. All rights reserved.
# TABLE OF CONTENTS

## LIST OF FIGURES

iv

## LIST OF TABLES

v

## ABSTRACT

vi

## CHAPTER ONE. GENERAL INTRODUCTION

1

- Description of Research Objectives 1
- Dissertation Organization 2
- Literature Cited 3

## CHAPTER TWO. AN INTRODUCTION TO EUKARYOTIC GENOME SIZE VARIATION

4

- Introduction 4
- Mechanisms of Genome Size Change 5
- Transposable Elements and Genome Size Variation 8
  *Gossypium* as a Model System for the Study of Genome Size Evolution 12
- Literature Cited 13

## CHAPTER THREE. DIFFERENTIAL LINEAGE-SPECIFIC AMPLIFICATION OF TRANSPOSABLE ELEMENTS IS RESPONSIBLE FOR GENOME SIZE VARIATION IN GOSSYPIUM

25

- Abstract 25
- Introduction 26
- Methods 29
- Results 34
- Discussion 38
- Concluding Remarks 42
- Acknowledgments 43
- Literature Cited 43

## CHAPTER FOUR. RAPID DNA LOSS AS A COUNTERBALANCE TO RETROTRANSPOSON PROLIFERATION IN PLANT GENOME SIZE EVOLUTION

61

- Abstract 61
- Introduction 62
- Methods 64
- Results 67
- Discussion 70
- Concluding Remarks 72
- Acknowledgments 73
- Literature Cited 73
<table>
<thead>
<tr>
<th>CHAPTER FIVE. PHYLOGENETIC DETERMINATION OF THE PACE OF TRANSPOSABLE ELEMENT PROLIFERATION IN PLANTS: COPIA AND LINE-LIKE ELEMENTS IN GOSSYPIUM</th>
<th>84</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>84</td>
</tr>
<tr>
<td>Introduction</td>
<td>85</td>
</tr>
<tr>
<td>Methods</td>
<td>87</td>
</tr>
<tr>
<td>Results</td>
<td>89</td>
</tr>
<tr>
<td>Discussion</td>
<td>90</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>93</td>
</tr>
<tr>
<td>Literature Cited</td>
<td>94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER SIX. GENERAL CONCLUSIONS</th>
<th>103</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Cited</td>
<td>107</td>
</tr>
</tbody>
</table>

| ACKNOWLEDGMENTS | 108 |
**LIST OF FIGURES**

Evolutionary relationships among diploid members of *Gossypium* 58

Copy number estimates for repetitive sequences in *Gossypium* 59

Neighbor-joining analysis of *Gossypium* gypsy-like *Gorge* 1, 2, and 3 reverse transcriptase sequences 60

Lineage-specific nature and timing of *Gorge* 3 transposition in *Gossypium* 80

Phylogenetic relationships and estimated rates of *Gorge* 3 gain and loss among diploid members of *Gossypium* 81

Neighbor-joining analysis of 724 PCR amplified *Gorge* 3 reverse transcriptase sequences 82

Density distributions of the parametric solution space possible for the model based on empirical sampling error 83

Genome sizes and evolutionary relationships among diploid members of *Gossypium* and a phylogenetic outgroup, *Gossypioides kirkii* 100

Neighbor-joining analysis of PCR amplified RT fragments of copia and LINE retrotransposons from *Gossypium* species and phylogenetic outgroup, *Gossypioides kirkii* 101

Timing of lineage-specific retrotransposon proliferation in *Gossypium* 102
LIST OF TABLES

Library construction and sequencing effort for three species representing different *Gossypium* genomes and one phylogenetic outgroup  

55

Repetitive element copy number and density estimates  

56

Estimated copy number and total number of Mb of *Gorge3* from various time points  

79
ABSTRACT

Although eukaryotic organisms display a wide range in genome size, from as little as 9 Megabases (Mb) in some fungi to over 690,000 Mb in the diatom Navicula pelliculosa, there is only an approximate 20-fold variation in the number of protein coding genes. Additionally, this variation in genome size is not directly correlated with organismal or morphological/physiological complexity. This striking contradiction has historically been termed the "C-value paradox". Research conducted over the past half century has revealed that most genome size variation in plants can be ascribed to the repetitive fraction of the genome, particularly LTR-retrotransposons. Furthermore, recent studies in maize and rice have revealed the tremendously dynamic nature of LTR-retrotransposons, where transposition has occurred recently and rapidly, resulting in a highly dynamic genome. To date, most studies of genome size evolution and transposable element dynamics has been conducted in the grasses, or among a few distantly related model dicots. In the work presented here, the cotton genus, Gossypium, was used to study genome size evolution and transposable element dynamics among closely related, long-lived species, whose members diverged within the last 5-7 million years. Provided is a description of the various types of repetitive sequences present and their relative contributions to overall extant genome size among three Gossypium species and a phylogenetic outgroup, Gossypioides kirkii. Results indicate that one type of gypsy-like LTR-retrotransposon, Gorge3, is responsible for much of the genome size variation in the genus. However, different transposable elements behave differently in different genomes, and estimated copy numbers are not always correlated with genome size. Further detailed analysis of the evolutionary history of Gorge3 suggests that this LTR-retrotransposon underwent lineage-specific proliferation in each clade, and that these proliferation events occurred in an episodic manner at different time points in each of the genomes investigated. Using a novel modeling approach, it is shown that although lineage-specific transposition of Gorge3 has occurred in the taxa having small genomes, the magnitude of DNA removal outweighs DNA gain through transposition, ultimately leading to genome downsizing. Investigation of the transpositional nature and timing of
two other relatively abundant repetitive sequences, *copia*-like LTR-retrotransposons and non-LTR LINEs, indicates that lineage-specific amplification has occurred for each sequence type in each lineage. *Copia*-like LTR-retrotransposons, similar to that found for *Gorge3*, undergo episodic proliferation at different time points in each genome. In contrast, non-LTR LINEs appear to accumulate in a more stochastically regular manner, and phylogenetic analysis suggests that the majority of the *Gossypium* LINE population is dominated by ancient sequences that predate divergence events within the genus. The cumulative results of this dissertation work indicate that transposable element proliferation has occurred recently and rapidly in all genomes investigated and their differential accumulation or retention contributes significantly to variation in genome size in *Gossypium*. 
CHAPTER ONE

General Introduction

Description of Research Objectives

For the past half century is has been widely recognized that there is no correlation between total physical nuclear content and organismal or genetic complexity (Price 1988; Sparrow et al. 1972). This well-documented deficiency in an overall correlation between genome size and morphological or physiological complexity of an organism has historically been termed the "C-value paradox" (Thomas 1971). Originally thought to reflect differences due to polyploidy, it is now evident that the majority of genome size variation is due to differential accumulation of the repetitive fraction of the genome. Since this discovery, the once seemingly contradictory "C-value paradox" has now graduated to the now perplexing "C-value enigma" in order to highlight the fact that genome size variation results from a complex combination of forces (Gregory 2002; Gregory 2004). However, many questions remain to be addressed. As most studies to date have been conducted within model grass systems and a few select but distantly related dicots, our current knowledge with respect to pace, tempo and directionality of genome size change reflects evolutionary trends within primarily short-lived annual taxa. Few comparative analyses have been conducted among closely related species, posing the potential caveat of detecting multiple overlapping mutational events that may be easily misinterpreted.

The purpose of my doctoral research is to contribute to the overall understanding of the pace, tempo and directionality of genome size change by employing the phylogenetically informed, recently diverged, non-grass system, *Gossypium*. In this work, I describe the spectrum and frequency of the various kinds of sequences that are responsible for genome size variation, and determine if this variation is due to an increase/decrease in the number of transposable element families, an increase/decrease in element copy numbers from particular families, or both. The following questions are addressed:
1. What types of repetitive elements in general and transposable elements in particular are present in *Gossypium* and what are their copy numbers?
2. How does this spectrum vary phylogenetically?
3. Is there a correlation between transposable elements copy number and genome size?
4. Do all transposable element families expand in a linear fashion ("one-way ticket to genomic obesity" hypothesis) or do some increase while others decrease, and are some families consistently over-represented/under-represented in larger genomes?
5. Do some species appear to be more efficient in removal of non-genic DNA and is there a correlation between this efficiency and genome size?

**Dissertation Organization**

The body of this dissertation work is organized into five chapters. Chapter two, entitled "An Introduction to Eukaryotic Genome Size Variation", provides an overview of the current literature in the field, with a particular emphasis on our understanding of the effects of transposable elements on genome size evolution. The following three chapters consist of original research results on the effects of transposable elements on genome size evolution in the cotton genus, *Gossypium*. Chapter 3, entitled "Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*" and published in the journal *Genome Research*, describes work in which we sequenced a portion of three *Gossypium* genomes that vary 3-fold in total nuclear DNA content, and described the various repetitive sequences in each genome, with respect to their relative contributions to overall genome size (Hawkins et al. 2006).

Chapter 4, entitled “Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants” and submitted to the *Proceedings of the National Academy of Sciences of the USA*, provides an investigation into the evolutionary history of the *gypsy*-like LTR-retrotransposon, *Gorge3*, and presents a novel modeling approach to describe the various rates of *Gorge3* gain and loss in each *Gossypium* lineage. This work is extended in Chapter 5, entitled "Phylogenetic determination of the pace of transposable element proliferation in plants: *copia* and LINE-like elements in
"Gossypium" prepared for submission to the journal *Genome*, where we evaluate the evolutionary history of *copia*-like LTR-retrotransposons and LINEs among the various *Gossypium* species. A final chapter summarizing the results of this dissertation work follows these original research chapters.

**Literature Cited**


CHAPTER TWO

An Introduction to Eukaryotic Genome Size Variation

Introduction

The genomes of eukaryotic organisms vary approximately 80,000-fold in size, ranging from 9 Mb in some fungi to over 690,000 Mb in the diatom *Navicula pelliculosa* (Cavalier-Smith 1985; Li and Graur 1991). Large ranges in genome size are common in various types of organisms, varying 5800-fold among protozoans, 250-fold among arthropods, and 5000-fold among algae (Gregory 2001; Neafsey and Palumbi 2003). Angiosperm genome sizes range from approximately 38 Mb in *Cardamine amara* (Brassicaceae) to over 120,000 Mb in some members of the Liliaceae (Bennett and Leitch 1995; Bennett and Leitch 1997; Flavell et al. 1974; Leitch et al. 1998). Not only are wide ranges in genome size common across distantly related organisms, but have also been observed in closely related species: genome sizes range approximately 6-fold among members of the genus *Vicia* (Chooi 1971), and 9-fold within the genus *Crepsis*, (Jones and Brown 1976). Interspecific genome size variation has also been observed (Price 1988). Along a 400 meter transect in Evolution Canyon, Mount Carmel, Israel, in which the microclimate differs in both solar irradiation and aridity, genome size was weakly correlated with slope orientation in local populations of wild barley (*Hordeum spontaneum*) (Kalander et al. 2000). Plants sampled from the south facing slope typically had larger genomes than those of the north facing slope.

A fraction of this genome size variation can be ascribed to differences in gene number due to segmental duplication and polyploidization, in addition to limited amounts of gene loss (Bancroft 2001; Bennetzen and Ramakrishna 2002; Blanc et al. 2000; Grant et al. 2000; Ku et al. 2000; Tikhonov et al. 1999; Vision et al. 2000; Wendel 2000). Nevertheless, greater than 90% of plant genes possess close homologs within most other plant species, indicative of highly conserved gene content (Bennetzen 2000a). In fact, there is only a 20-fold variation in the number of protein coding genes among all eukaryotic organisms (Li 1997 and references therein). It is generally agreed that the vast
majority of this genome size variation can be ascribed to the repetitive fraction of the genome (Bennetzen 2000b; Bennetzen 2002; Kidwell 2002).

**Mechanisms of Genome Size Change**

**Genome Size Expansion.**

There are several well-known factors that contribute to genome size increase (Bennetzen and Ramakrishna 2002; Blanc et al. 2000; Ku et al. 2000; Tikhonov et al. 1999; Wendel 2000). One of these factors is polyploidy (Leitch and Bennett 1997; Wendel 2000), in which the entire genomic content of an organism is doubled and both subgenomes coexist within the same nucleus with little to no disruption of the genic balance of that organism. Estimates suggest that were this the only means of genome duplication, as few as 10 rounds of polyploidization would be required to account for the large range in genome size observed between *E. coli* and mammals (Nei 1969).

Particularly relevant is the recent work by Schlueter et al. in which ESTs from eight diverse angiosperms were mined to identify potential gene duplicates (Schleuter et al. 2004). By calculating the synonymous and nonsynonymous distances for the 1392 gene duplicates recovered from the EST libraries, the authors were able to identify several rounds of whole-genome duplication events in each of the lineages evolutionary histories. Additionally, Schlueter et al. were able to identify a shared duplication event among members of the Poaceae. Blanc and Wolfe performed a similar study in which they compared gene duplicates among 14 model plant species (Blanc and Wolfe 2004). The author's results were similar to that of Schlueter et al., in which they found evidence for multiple ancient duplication events in each of the genomes investigated. More recently, Cui et al. (2006) extended this work to lineages outside of previously studied model crop species by performing comparisons of duplicate genes among basal angiosperms (Cui et al. 2006). The authors found evidence of multiple whole-genome duplication events throughout the history of angiosperms, although they were unable to find evidence for a duplication event in pine. This work indicates that polyploidy is widespread, and not necessarily limited to crop species where a predisposition to polyploidy would increase the likelihood of obtaining traits needed for domestication. Presently, it is unclear why
some organisms undergo polyploidy while others do not. Polyploidy has also been observed in animals (Pebusque et al. 1998; Smith et al. 1999; Wolfe and Shields 1997), although this phenomenon appears to be more common in plants and likely contributes to the wider range in genome size observed across plants relative to animals.

Activation and accumulation of transposable elements may also contribute to genome size expansion by allowing amplification and insertion of newly synthesized elements. Evidence from maize indicates that its genome size has doubled over the last 3 million years due to transposable element proliferation alone (SanMiguel and Bennetzen 1998). Additionally, recent studies in rice show massive, lineage-specific amplification of LTR-retrotransposons in a large genome member of the genus, *Oryza australiensis*, suggesting these new transposon insertions have lead to its relatively bloated genome (Piegu et al. 2006). The contribution of transposable element amplification to genome size variation will be discussed further below.

Less often, a large-scale duplication of a part of the genome or fixation of an accessory chromosome may be responsible for flux in genome size. Segmental duplication has been demonstrated in rice, where a 3 Mb duplication has occurred between chromosomes 11 and 12, two chromosomes rich in disease resistance genes (The rice chromosomes 11 and 12 sequencing consortia 2005; Wang et al. 2005). These types of mutations are less likely to have a large impact on variation in genome size (compared to polyploidy and amplification of repetitive DNA) because of their increased potential for disruption of the genic balance of the organism. In most cases, the results of chromosomal duplication are fatal, but in some cases such as duplication of the smaller chromosomes, the progeny are viable yet sterile (Hamerton 1971; Lindsley et al. 1972).

**Genome Size Contraction.**

Although it has been suggested that organisms may have a "one-way ticket to genomic obesity" (Bennetzen and Kellogg 1997), it is possible that differences in genome size are not only the outcome of an organism's tolerance for accrual of non-genic DNA, but the efficiency of an organism in the removal of non-essential DNA (Petrov 2002; Petrov and Hartl 1997; Petrov et al. 2000; Wendel et al. 2002). Many organisms with
smaller genomes are strikingly lacking in non-genic DNA. Examples include birds, *Arabidopsis* and *Drosophila*, all harboring very few transposable elements. Studies of indel patterns within "dead-on-arrival" (DOA) non-LTR retroelements (pseudogenes) in *Drosophila* compared to *Laupala* (cricket), whose genome sizes vary 11-fold, suggest an increased rate of DNA loss in organisms with smaller genomes compared to species of insects with larger genomes (Petrov and Hartl 1997; Petrov et al. 2000). These indel patterns observed in "DOA" non-LTR retrotransposons appear to be congruent with patterns observed in other parts of the genome, including both euchromatic and heterochromatic regions, transposable and non-transposable non-genic DNA, in addition to repetitive and unique sequences (Petrov 2002). The rate of DNA loss in *Drosophila* is 40-fold higher than the rate of DNA loss in *Laupala* (Petrov et al. 1996; Petrov et al. 2000). The authors suggest that an inverse correlation exists between genome size and rate of DNA loss. A study by Kirik et al., (2000) parrots this conclusion. In a comparison of *Arabidopsis* and tobacco, whose genomes vary approximately 20-fold in size (Bennett and Leitch 1997), deletions in *Arabidopsis* were on average one-third larger than those found in tobacco (Kirik et al. 2000). The authors found no insertions associated with deletions in *Arabidopsis*, but conversely, approximately one-half of the repair events in tobacco were associated with an insertion (Salomon and Puchta 1998).

DNA loss may also occur via other mechanisms. In some instances, whole chromosome loss may occur upon hybridization of distantly related species (Laurie and Bennett 1989; Riera-Lizarazu et al. 1996), although this often results in intermediates with low fertility due to aneuploid gametes (Bennetzen 2002). The same instability is observed in the event of unequal recombination between homologous chromosomes. Because of highly unstable intermediates, these processes likely contribute little to genome size evolution. However, unequal intra-strand homologous recombination between two tandem repeats in the same orientation, such as the LTRs of retrotransposable elements, can result in DNA loss of one of the repeats and the intervening DNA sequence (Bennetzen 2002). Indeed, solo LTRs have been observed within various organisms (Chen et al. 1998; SanMiguel et al. 1996; Shepherd et al. 1984; Vicient et al. 1999). Intra-strand homologous recombination between the LTRs of a
single retrotransposon can lead to attenuation of genome expansion but would not reverse the expansion process due to the remaining LTR. However, SanMiguel et al. have shown that large blocks of nested LTR retrotransposons exist within the maize genome (SanMiguel et al. 1996). Should intrastrand recombination take place between the LTRs of adjacent retrotransposons or of entire retrotransposon blocks, a net loss of DNA would be expected (Bennetzen 2002).

Transposable Elements and Genome Size Variation
Evidence from Comparative Genomics Studies.

Early studies exploiting reassociation kinetics led to the realization that much of the eukaryotic genome is composed of repetitive DNA (Britten and Kohne 1968). Britten and Kohn determined that the eukaryotic genome can be divided into four major fractions, foldback DNA, highly repetitive DNA, middle-repetitive DNA, and single copy DNA, and that the majority of the genome was composed of repetitive sequences. Subsequently, several investigations have provided support for Britten and Kohn's findings. Repetitive DNA constitutes approximately 80% of angiosperm genomes with haploid DNA content greater than 5.0 pg (Flavell et al. 1974). Approximately 60% or more of the maize (Meyers et al. 2001; SanMiguel and Bennetzen 1998; SanMiguel et al. 1996), wheat (Wicker et al. 2001) and barley (Shirasu et al. 2000; Vicient et al. 1999) genomes are made up of transposable elements. Nearly 25% of the maize genome is composed of 5 classes of Long Terminal Repeat (LTR) retrotransposons alone (SanMiguel et al. 1996). Approximately 80% of the wheat genome is repetitive DNA, mainly LTR retrotransposons (Kumar and Bennetzen 1999). These observations have led to an increased interest in the importance of repetitive DNA on genome size variation and the effect of this variation on plant fitness.

Comparative genomic studies suggest extensive conservation of gene content in addition to significant microcolinearity of protein coding sequences between closely related species, and support the hypothesis that the majority of sequence variation (both sequence types in addition to number of nucleotides) occurs within the intergenic, repetitive regions of the genome (Chen et al. 1998; Chen et al. 1997; Dubcovsky et al.
2001; Feuillet and Keller 1999; Tarchini et al. 2000; Tikhonov et al. 1999). Tikhonov et al. compared 225 kb of the maize *adh* region containing 9 candidate genes to the orthologous 78 kb *adh* region in sorghum, whose genome is 3.5-fold smaller than that of maize (Tikhonov et al. 1999). Homologs of these 9 maize genes were found in colinear order in the *adh* region of sorghum. In addition, 5 other sorghum genes were identified, three of which were deleted in the maize region but present elsewhere in the maize genome. Therefore, the protein coding sequences of the *adh* region in sorghum and maize are largely colinear with a few minor rearrangements. As expected, the major unconserved regions were the intergenic regions between the genes. Many mobile elements, such as retrotransposons and MITEs, in addition to simple sequence repeats were abundant in the maize *adh* region, whereas in sorghum, LTR-retrotransposon presence was limited. Similarly, Tarchini et al. sequenced 340 Kb of the *adh1-adh2* region in rice, and the sequences were subsequently cross-hybridized to maize genomic DNA (Tarchini et al. 2000). A high degree of microcolinearity was observed between the maize and rice regions with the exception of the *adh1* gene itself. In rice, 35 kb separates the *adh1* and *adh2* loci, whereas in maize the two loci are found on separate chromosomes. Many transposable elements were identified in the rice sequence, comprising 14.4% of the segment. In contrast, retrotransposons make up approximately 60% of the maize genome. In light of the number of these elements found in maize compared to rice, the authors suggest that retrotransposons may greatly contribute to genome expansion.

In a comparison of the sh2/a1 homologous regions of maize, rice, and sorghum, the sh2 region of maize was 75% homologous over 256 bp to rice and 92% homologous over 365 bp to sorghum (Chen et al. 1998; Chen et al. 1997). The *a1* region of maize revealed 82% identity over 600 bp to rice. The genes in rice and sorghum were in the same order and orientation as their homologs in maize. Therefore, although these species have undergone millions of years of independent evolution, they have maintained gene content and order in this region of the genome. The coding sequences were highly conserved, but none of the repetitive DNA of maize cross-hybridized with sorghum or
rice, suggesting that variation between these species occurs in the intergenic, repetitive regions of the genome.

A study comparing a 23 Kb fragment surrounding the Lrk10 sequence in wheat to the Lrk regions in barley, maize, and rice also suggests conservation of gene order and content among closely related grass species (Feuillet and Keller 1999). The Lrk10 sequence in wheat was identified adjacent to the Tak10 sequence in opposite orientation. A pseudogene homologous to the LRR10 gene was identified upstream of the Lrk and Tak sequences. The position and orientation for these sequences was conserved in all 4 species. In addition, few mobile elements were detected in the sorghum, rice, and barley genomes relative to the maize genome, providing further evidence that mobile elements are the agents of marked genome size differences among these closely related grasses.

More recently, several comparative genomics studies have provided new insight into the mechanisms that contribute to genome size expansion and contraction. A particularly eloquent example is that of Ma and Bennetzen (2004), in which the authors compared >1 Mb of orthologous sequence from two Oryza sativa subspecies, japonica and indica, and polarized all identifiable changes by including an outgroup, O. glaberrima (Ma and Bennetzen 2004). Although indica and japonica diverged ~0.44 mya, rapid genomic changes were detected. Both subspecies experienced significant genome size increase (indica = 2%, japonica = 6%) due primarily to new insertions of LTR-retrotransposons. Additionally, evidence of DNA loss via illegitimate recombination and intra-strand homologous recombination was apparent in both genomes. Indeed, the number of deletions far outweighed the number of insertions. Nevertheless, these small deletions were not enough to counter genome size expansion through transposon insertion, resulting in a net increase in genome size.

Another recent study by Bruggman et al. compared duplicated regions of the maize genome with their orthologous region in rice (Bruggman et al. 2006). Approximately 7.8 Mb of maize chromosome 1 and its orthologous region on chromosome 9 (6.6 Mb) were compared to a 4.9 Mb region on rice chromosome 3. The authors found multiple mechanisms of genome evolution that contribute to genome expansion and contraction in these regions. The maize region on chromosome 1 had
expanded significantly (by a factor of 5.25) due to transposon insertion relative to its syntenic chromosome 9 counterpart. Additionally, genes in maize were typically larger than those of rice due to larger introns, and 9% of the genes in this region were missing in the syntenic rice region. The authors conclude that the C-value paradox results from a combination of forces that act to expand and contract genome size.

Transposable elements.

Transposable elements have been found in all plants investigated to date (Bennetzen 2000a). These elements are divided into two major classes based on the structure of their coding sequences and the mechanism of transposition (Capy et al. 1997; Finnegan 1989). The Class II plant transponson superfamilies are classified as Activato/Dissociator-(Ac/Ds), Enhancer/Suppressor-mutator-(En/Spm) or Mutator-(Mu) systems based on the families first described in the maize genome and the Tam elements of Snapdragon (reviewed in Kunze et al. 1997). En/Spm elements are also referred to as CACTA elements because they share a common sequence (5'-CACTA-3') at their TIR termini. An abundant group of small transposable elements, MITEs (Miniature Inverted Transposable Elements), also have a structure that suggests they are members of the Class II DNA elements (Bennetzen 2000b; Kidwell 2002). DNA elements were the first types of transposable elements identified due to their mutagenic effects observed in maize (McClintock 1949). DNA elements characteristically contain terminal inverted repeats (TIR) ranging from 11 to a few hundred base pairs in length, and families of elements are defined by these TIR sequences (Bennetzen 2000b). DNA elements are often found preferentially in coding regions and non-methylated regions of the genome where there is a high potential for active transcription to occur (Kumar and Bennetzen 1999). MITEs display a particular affinity for genic regions (Jiang and Wessler 2001) and have primarily been described in monocotyledonous plants although their existence in dicotyledonous plants, humans, and insects has been documented (Amrani et al. 2002; Bureau et al. 1994; Feschotte et al. 2002; Tikhonov et al. 1999; Tu 2001; Wessler et al. 1995; Yang and Hall 2003; Zhang et al. 2000). Because DNA elements transpose via a conservative mechanism in which the element is excised and inserted into a new area of
the genome, they are probably not responsible for the large variation in genome size observed across the plant kingdom (Kunze et al. 1997).

Class I elements contain the retroelements, by far the most abundant transposable element in the plant kingdom (Bennetzen 2002; Bennetzen et al. 1993; Grandbastien 1998; Kumar and Bennetzen 1999). LTR-retroelements are ubiquitous, having been described in all plant species studied to date. LTR-retroelements are divided into 2 classes, gypsy and copia-like, based on the position of the integrase domain within the element. Non-LTR retroelements consist of Long Interspersed Nuclear Elements (LINEs) and Small Interspersed Nuclear Elements (SINEs). LINEs and SINEs may also be found in high copy number in plants and have been identified throughout the plant kingdom. Although some retroelements are found associated with genes (Bennetzen et al. 1993; Grandbastien 1998; Hirochika et al. 1996; Johns et al. 1985; Pouteau et al. 1994), they are observed most frequently in methylated, presumably heterochromatic regions of the genome and often nested within one another (Bennetzen et al. 1994; Kumar and Bennetzen 1999; SanMiguel et al. 1998; SanMiguel et al. 1996).

Retroelements transpose via a duplicative mechanism, in which an RNA intermediate formed from the parental copy is reverse transcribed, and the newly translated copies are inserted into supplementary positions in the genome. This type of transpositional mechanism may lead to the accumulation of large quantities of repetitive sequences, likely contributing to the wide range in genome size variation observed across eukaryotic organisms.

**Gossypium as a Model System for the Study of Genome Size Evolution**

*Gossypium* (Malvaceae) is a monophyletic genus comprising approximately 50 species of small trees and shrubs that are widespread throughout tropic and sub-tropical regions of the world (Cronn et al. 2002; Fryxell 1992; Seelanan et al. 1997; Wendel and Cronn 2002). Diploid members of the genus are divided into 8 groups based on 2C DNA content and chromosome pairing behavior, in addition to fertility in interspecific hybrids (Beasley 1941; Endrizzi et al. 1985). All diploid members of the genus have 13 chromosomes, yet genome sizes range approximately 3 fold, from 885 Mb per haploid
nucleus in the American D genome species, to 2572 Mb per haploid nucleus in the Australian K genome species add (Endrizzi et al. 1985; Fryxell 1979; Hendrix and Stewart 2005; Wendel et al. 1999). An even larger range in genome size is observed across the tribe, from only 590 Mb in Gossypioideae kirkii and Kokia drynarioides to 4018 Mb per haploid nucleus in Thespesia populnea (Wendel et al. 2002). Gossypium diverged from the G. kirkii + K. drynarioides lineage approximately 12.5 mya (Cronn et al. 2002; Seelanen et al. 1997). A single polyploidization event between the African A and New World D genomes approximately 1-2 mya gave rise to the 5 currently described allopolyploid species (reviewed in Wendel and Cronn 2002). Much is known concerning the taxonomic relationships within the genus Gossypium. The closest extant relatives of the Gossypium allotetraploid species have been identified and the phylogeny of the genus is well described. The wide range in genome size observed across closely related diploid species (3-fold) and well-established phylogeny make Gossypium an excellent system for the study of genome size evolution.

**Literature Cited**


CHAPTER THREE

Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*

A paper published in the journal *Genome Research*\(^1\)

Jennifer S Hawkins\(^2\), HyeRan Kim\(^3\), John D Nason\(^4\), Rod A Wing\(^5\),
and Jonathan F Wendel\(^6\)

Abstract

The DNA content of eukaryotic nuclei (C-value) varies approximately 200,000-fold, but there is only an approximate 20-fold variation in the number of protein coding genes. Hence, most C-value variation is ascribed to the repetitive fraction, although little is known about the evolutionary dynamics of the specific components that lead to genome size variation. To understand the modes and mechanisms that underlie variation in genome composition, we generated sequence data from whole genome shotgun (WGS) libraries for three representative diploid (n =13) members of *Gossypium* that vary in genome size from 880 to 2460 Mb (1C) and from a phylogenetic outgroup, *Gossypioides kirkii*, with an estimated genome size of 588 Mb. Copy number estimates including all dispersed repetitive sequences indicate that 40-65% of each genome is composed of transposable elements. Inspection of individual sequence types revealed differential,

---

1 Reprinted with permission of Genome Research, 2006, 16(10), 1252-1261.
2 Graduate student, primary researcher and author, EEOB department, Iowa State University.
3 Graduate student, GSS library construction, Arizona Genomics Institute, University of Arizona.
4 Xxx professor, derivation of statistical equations for sampling effort and copy number estimation, EEOB department, Iowa State University.
5 Xxx professor, GSS library constructions, Arizona Genomics Institute, University of Arizona.
6 Principal investigator and corresponding author, EEOB department, Iowa State University.
lineage-specific expansion of various families of transposable elements among the different plant lineages. *Copia*-like retrotransposable element sequences have differentially accumulated in the *Gossypium* species with the smallest genome, *G. raimondii*, while *gypsy*-like sequences have proliferated in the lineages with larger genomes. Phylogenetic analyses demonstrated a pattern of lineage-specific amplification of particular subfamilies of retrotransposons within each species studied. One particular group of *gypsy*-like retrotransposon sequences, *Gorge3* (*Gossypium* retrotransposable *gypsy*-like element), appears to have undergone a massive proliferation in two plant lineages, accounting for a major fraction of genome-size change. Like maize, *Gossypium* has undergone a 3-fold increase in genome size due to the accumulation of LTR retrotransposons over the 5-10 my since its origin.

The sequence data described in this paper have been submitted to the GSS Division of GenBank under accessions DX390732 - DX406528.

**Introduction**

Genomes of eukaryotic organisms vary over 200,000-fold in size, ranging from 2.8 Mb in *Encephalitozoon cuniculi* (Biderre et al. 1998) to over 690,000 Mb in the diatom *Navicola pelliculosa* (Cavalier-Smith 1985; Li and Graur 1991). Among angiosperms, genome sizes range from approximately 108 Mb for *Fragaria viridis* (Bennett and Leitch 2005) to over 120,000 Mb in some members of the Liliaceae (Bennett and Leitch 1995; Bennett and Leitch 1997; Bennett and Smith 1991; Flavell et al. 1974; Leitch et al. 1998). Not only is wide variation in genome size common among distantly related organisms, but it also is unexceptional even among closely related species. For example, genome sizes range approximately 6-fold among members of the genus *Vicia* (Chooi 1971), and 9-fold within the genus *Crepis* (Jones and Brown 1976). Some portion of this genome size variation may be ascribed to differences in gene number amplification due to gene, chromosome segment, and whole-genome duplication, as well as to gene loss (Bancroft 2001; Bennetzen and Ramakrishna 2002; Blanc et al. 2000; Grant et al. 2000; Ku et al. 2000; Tikhonov et al. 1999; Vision et al. 2000; Wendel
Nevertheless, greater than 90% of plant genes possess close homologs within other plant species, indicative of highly conserved gene content (Bennetzen 2000a).

There appears to be no correlation between the amount of DNA per cell and organismal advancement or genetic complexity (Price 1988; Sparrow et al. 1972). This well-documented lack of correspondence between genome size and morphological or physiological complexity of an organism has been historically termed the "C-value paradox" (Thomas 1971). Since the discovery of non-coding DNA and its impact on genome size variation, "paradox" has been replaced by "enigma" in an attempt to more appropriately identify the topic as a "perplexing subject" made up of several independent components (Gregory 2002; Gregory 2004). It is now generally agreed that the C-value enigma can be largely explained by the differential amplification and proliferation among organisms of the repetitive fraction of the genome (Bennetzen 2000b; Bennetzen 2002; Kidwell 2002).

In plants, amplification and insertion of newly activated long terminal repeat (LTR) retrotransposable elements appears to be a major contributor to genome size expansion. For example, approximately 70% of the maize nuclear genome is composed of LTR-retrotransposons (SanMiguel and Bennetzen 1998). In the span of just a few million years, the maize genome doubled in size due to transposable element activity (SanMiguel and Bennetzen 1998). These transposable elements are often found in nested arrangements located between "gene islands", and often are associated with centromeres (SanMiguel et al. 1996). To date, little is known regarding the extent to which various transposable elements contribute to genome size variation or how transposable element (TE) types are distributed among closely related species. Other mechanisms posited to be responsible for genome size expansion include variation in intron size (Deutsch and Long 1999), expansion of tandemly repetitive DNA sequences (Ellegren 2002; Morgante et al. 2002), segmental duplication (Bancroft 2001; Blanc et al. 2000; Ku et al. 2000; Vision et al. 2000; Wendel 2000), accumulation of pseudogenes (Zhang 2003), and transfer of organellar DNA to the nucleus (Adams and Palmer 2003; Shahmuradov et al. 2003). However, these mechanisms generally do not appear to have a large impact on genome size differences among closely related species.
Although it has been suggested that organisms may have a "one-way ticket to genomic obesity" (Bennetzen and Kellogg 1997), it is possible that differences in genome size are not only the outcome of an organism's tolerance for accrual of non-genic DNA, but also its efficiency in removal of non-essential DNA (Petrov 2002a; Petrov and Hartl 1997; Petrov et al. 2000; Wendel et al. 2002b). Many organisms with smaller genomes are striking in their relatively small proportion of non-genic DNA. Evidence of a deletional bias among organisms with smaller versus larger genomes (Bennett and Leitch 1997; Kirik et al. 2000; Petrov and Hartl 1997; Petrov et al. 2000) has led to the "mutational equilibrium model" of DNA loss (Petrov 2002b). Other suggested mechanisms of DNA loss include unequal intrastrand homologous recombination between two tandem repeats in the same orientation, such as the LTRs of retrotransposable elements (Bennetzen 2002; Chen et al. 1998; SanMiguel et al. 1996; Shepherd et al. 1984; Vicient et al. 1999), illegitimate recombination (Bennetzen et al. 2005; Devos et al. 2002; Ma et al. 2004; Wicker et al. 2003) and double-stranded break repair (Filkowski et al. 2004; Kirik et al. 2000; Orel and Puchta 2003).

To effectively study genome size evolution from a phylogenetic perspective, it is necessary to exploit a system in which the closely related species vary widely in genome size and for whom phylogenetic relationships are well understood. A good example in this respect is the monophyletic genus *Gossypium* (Malvaceae), which is composed of approximately 50 species of small trees and shrubs with an aggregate distribution that encompasses many tropical and subtropical semi-arid regions of the world (Cronn et al. 2002; Fryxell 1992; Seelanan et al. 1997; Wendel and Cronn 2003). Diploid members of the genus are divided into eight groups based on chromosome pairing behavior and fertility in interspecific hybrids (Beasley 1941; Endrizzi et al. 1985). All diploid members of the genus have 13 chromosomes, yet genome sizes range approximately threefold, from a median estimate of 885 Mb per haploid nucleus in the American D-genome species, to 2572 Mb per haploid nucleus in the Australian K-genome species (Figure 1) (Hendrix and Stewart 2005). An even larger range in genome size is observed in the tribe to which *Gossypium* belongs (the Gossypieae), from only 590 Mb in *Gossypioides kirkii* and *Kokia drynarioides* to 4018 Mb in *Thespesia populnea* (Wendel ...
et al. 2002b). The wide range in genome size observed across closely related diploid species and the well-established phylogeny makes *Gossypium* an excellent system for the study of genome size evolution.

To better appreciate the relevance of genome size variation to organismal fitness and evolution, it is first necessary to enhance our understanding of the quantity and quality of the genomic components that distinguish two or more genomes, as well as the modes and mechanisms by which these differences arise. This insight may derive from comparative sequence analysis of specific genomic regions or from using more global approaches. An example of the former is the recent study by Grover et al. (Grover et al. 2004), who compared approximately 104 kb of aligned sequence surrounding the *CesA1* gene from the D- and A- genomes of tetraploid cotton. In this case both gene content and intergenic regions were largely conserved, and hence there was no evidence of the mechanisms responsible for the twofold size variation that characterizes these genomes.

Here we employ the second approach, utilizing whole genome shotgun (WGS) libraries constructed for three members of *Gossypium* that range threefold in genome size, and one outgroup species, *Gossypiodies kirkii*. Copy number estimates for several *Gossypium* transposable elements suggest that different types of repetitive sequences have accumulated at different rates in different plant lineages. Additionally, the results suggest that different families within a repetitive sequence type proliferate differentially. Indeed, the major fraction of the genome size variation observed in *Gossypium* is largely due to recent, lineage-specific amplification of one particular group of gypsy-like retrotransposon sequences, *Gorge3* (*Gossypium* retrotransposable gypsy-like element), within the larger-genome *Gossypium* species.

**Methods**

*Construction and Sequencing of WGS libraries*

WGS libraries were constructed according to Meyers et al. (Meyers et al. 2001) with minor modifications and sequenced at the Arizona Genomics Institute, University of Arizona. Briefly, total genomic DNA extracted from young leaves of a single individual was randomly sheared using a Hydroshear (Thorstenson et al. 1998) (GeneMachine), an
automated hydrodynamic point-sink based DNA shearing device (Oefner et al. 1996), at speed code 13 for 25 cycles at room temperature to obtain fragments from *G. herbaceum* (JMS), *G. raimondii* (JFW stock), *G. exiguum* (Gos 5184), and *G. kirkii* (JFW stock) (Fig. 1). Sheared fragments between 2500 and 6000 bp were excised and converted to bunt-ended DNA fragments using the "End-it" DNA end repair kit (Epicentre) containing T4 DNA polymerase (for 5' → 3' polymerase and 3' → 5' exonuclease activities) and T4 Polynucleotide Kinase (for phosphorylation of 5' -ends of blunt DNA), followed by ligation into pBluescriptII KS+ (Strategene) and electroporation into *E. coli* strain DH10B T1 phase resistant electrocompetent cells (Invitrogen). WGS library clones were sequenced from one direction using the T7 primer (5' TAA TAC GAC TCA CTA TAG GG 3') and BigDye Terminator v3.1 (Applied Biosystems, ABI) according to manufacturer’s instruction. Cycle sequencing was performed using PTC-200 thermal cyclers (MJ Research, Waltham, MA) in a 384-well format with the following regime: 35 cycles of 30 sec at 96°C, 20 sec at 50°C, and 4 min at 60°C. After the cycle-sequencing step, the DNA was purified by ethanol precipitation.

Samples were eluted into 20 ul of water and separated using ABI 3730x1 DNA sequencers (ABI, USA). Sequence data was collected and extracted using sequence analysis software (ABI, USA). The sequencing data was base-called using the program Phred (Ewing et al. 1998). Vector and low quality sequences were removed by the program Lucy (Chou and Holmes 2001) and then submitted to the GSS division of GenBank under accessions DX390732 - DX406528.

**Analytical framework**

The number of sequences needed to generate 95% confidence that at least one member of a given class of sequences will be sampled was determined for each species using the following equation:

\[ N_{95} = \ln(0.05) / \ln \{1 - [n \ (l - 2m + e)/(G - e)]\} \]
where $N_{95}$ is the sampling effort required to be 95% confident that at least one target sequence will be sampled, $n$ is the number of targets present in the genome, $l$ is the length of the target sequence, $m$ is the estimated minimum length required to identify the sequence in a BLAST search, $e$ is the number of base pairs sequenced from each insert, and $G$ is genome size. By using this equation we were able to estimate the sampling intensity needed to detect at least one repetitive sequence of an estimated length and copy number in each of the four genomes. Published data for diverse types of repetitive elements, such as Ty3-gypsy, Ty1-copia, LINE retroposons, SINEs, and MITEs of various estimated lengths ($l$) and copy numbers ($n$) were used to calculate $N_{95}$ in order to determine how many clones should be sequenced from each library. The value for $m$ was conservatively estimated at $m=200$ bp, which in a BLASTX analysis would equal sequence similarity over 66 or more amino acids. We estimated $e$ to be 700 bp, based on the average high-quality sequencing read length reported from the Arizona Genomics Institute. Based on these estimates, libraries were constructed that contain 1.5% (based on ~5 kb plasmid insert length) of the genome from each species. One pass sequencing from one end of the insert ($e \sim 700$ bp) was performed, which, when totaled across the number of clones sequenced, yielded sequence data for approximately 0.2% of each haploid genome (Table 1).

Data analysis and copy number estimation

Because of rapid sequence divergence of repetitive DNA and the limited database of repetitive sequences available in GenBank for plants closely related to Gossypium, sequences from the WGS libraries were subjected to BLASTX (amino acid) in addition to BLASTN (nucleotide) analyses at the NCBI web site. Hits of $e$-5 or better were retained for further analysis. In addition, libraries were queried against themselves in an attempt to identify families of repetitive elements not recognized in the initial search. In this self-BLAST analysis, sequences with greater than 80% identity over 100 bp were considered related. Clones were assigned to a general category according to their best BLAST hit. These general categories were 1) nuclear, 2) chloroplast, 3) mitochondrial, 4) repetitive, and 5) unknown.
Plant transposable elements are broadly divided into 3 main lineages: the "Transposons" consisting of the Class II DNA elements, the "Retrotransposons" containing the LTR Class I elements, and the "Retroposons" consisting of the non-LTR Class I elements (Eickbush and Malik 2002). Class I elements transpose via a duplicative mechanism, in which an RNA intermediate formed from the parental copy is reverse transcribed, and the newly translated copies are inserted into new positions in the genome. Class I LTR retrotransposons are subdivided into two classes, gypsy and copia-like, based on the position of the integrase coding domain. The non-LTR retroposons consist of autonomous Long Interspersed Nuclear Elements (LINEs) and the non-autonomous Small Interspersed Nuclear Elements (SINEs). Class II DNA elements transpose via a cut-and-paste mechanism in which the element is excised and inserted into a new area of the genome. DNA elements characteristically contain terminal inverted repeats (TIR) ranging from 11 to a few hundred base pairs in length, and families of elements are defined by these TIR sequences (Bennetzen 2000b). Class II transposons can be divided into three main superfamilies: hAT (hobo from Drosophila, Activator of maize, and _Tam_ from Snapdragon), Mutator and En/Spm, both first described in maize (Kidwell 2002). Therefore, dispersed repetitive sequences recovered from the WGS libraries were placed into the specific categories 1) gypsy-like, 2) copia-like, 3) LINE-like, 4) hAT-like, 5) En/Spm-like, and 6) Mutator-like. DNA sequence alignments were performed with published sequences of the same type to confirm sequence identity.

Tandem repeats were identified using the program **Tandem Repeat Finder** (Benson 1999). Searches were performed using the default settings. Any tandem repeat present in more than three clones with a score greater than 500 were retained for further analysis. These sequences were queried against GenBank using BLASTN to search for sequence similarity to known sequences deposited in GenBank. Sequences were queried against one another to identify sequences that were shared among the libraries.

Copy numbers (\(n\)) for various repetitive elements recovered from the WGS libraries were estimated according to the following equation:
\[ n = \left( \frac{X_{obs}}{N} \right)(G - e)(1 / l - 2m + e) \]

where \( X_{obs} \) is the observed number of copies, \( N \) is the total number of sequence reads and the other variables are as before: \( n \) = number of targets in the genome; \( l \) = length of target sequence; \( m \) = estimated minimum length required to identify sequence in a BLAST search; \( e \) = number of bp sequenced from each insert; and \( G \) = genome size. Published sequences for various repetitive elements were used to estimate \( l \). Similar to average copia sequences in rice (5-6 kb) (McCarthy et al. 2002), an \( l \) of 5.3 kb was used for copia-like sequences based on published data from Gossypium (Grover et al. 2004). Also in agreement with rice data for gypsy-like sequences (11-13 kb) (McCarthy et al. 2002), \( l \) for gypsy-like sequences in Gossypium was set at 9.7 kb (Grover unpublished). Because no data exist for other dispersed repetitive sequences in Gossypium, the estimated lengths for the remaining repetitive sequences were established according to their closest BLAST hit from GenBank and are assigned as follows: LINE retroposon 3.5 kb (GenBank accession gi|37536056|ref|NP_922330.1| from O. sativa); En/Spm with high identity to Tam1 (Nacken et al. 1991) (GenBank accession X57297) 15.2 kb; and hAT with high identity to Tam3 (Hehl et al. 1991) (GenBank accession X55078) 3.6 kb. Published lengths for Mutator-like sequences are highly variable (in Arabidopsis these range from 444-19,397 bp) (Yu et al. 2002); therefore, we did not attempt to estimate Mutator-like copy numbers. A recent manuscript by Rabinowicz et al. (2005) used WGS libraries to estimate gene number in various plant species. When using their data for Arabidopsis and rice in our equation, we recover comparable results to those published for these two sequenced genomes, suggesting that our equation results in reasonably accurate estimates of copy numbers.

**Phylogeny reconstruction**

Sequences were queried against coding domains of various repetitive sequences from Arabidopsis thaliana and Brassica oleracea obtained from S. Wessler and F. Zhang (Univ. Georgia). Amino acid sequences with an e-value of e-5 or better were imported
into BioEdit (Hall 1999) and aligned using ClustalW (Johnson et al. 1994). Neighbor-Joining analysis was performed in Paup* (Swofford 2001) using the default settings.

Results

Library construction and sequence analysis

The *Gossypioides kirkii* (outgroup), *Gossypium raimondii* (D), *G. herbaceum* (A), and *G. exiguum* (K) libraries contained 1920, 3072, 6048, and 10368 clones, of which 1464, 2722, 4864, and 6747 were successfully sequenced, respectively (Table 1). The percent of each genome sequenced (0.19 – 0.24%) was determined by multiplying the number of successfully sequenced clones by the average high quality sequencing read length, divided by the estimated genome size. Sequences were queried against GenBank using BLASTX and against each other using BLASTN, and sequences were classified as described (see Methods). All types of dispersed repetitive sequences identified via this procedure were categorized into 1) *gypsy*-like, 2) *copia*-like, 3) LINE-like, 4) *Mutator*-like, 5) *hAT*-like, 6) *En/Spm*-like, and 7) unknown repetitive sequences. *Gypsy*- and *copia*-like LTR retrotransposons, in addition to LINE-like retroposons, were abundant in all four species (Table 2 and Figure 2). Class II DNA sequences and tandem repeats were less abundant. Some classes of dispersed repetitive sequences, such as MITEs and SINEs, were not identified in the libraries. However, because of the lack of conserved domains for these two types of sequences, they may be present in *Gossypium* and unidentifiable via BLAST. Copy number estimates suggest a minimum of 44%, 54%, 52% and, 60% of the *G. kirkii*, *G. raimondii*, *G. herbaceum* and *G. exiguum* genomes, respectively, are occupied by repetitive sequences alone.

Several conserved coding domains for diverse repetitive sequences were recovered from the WGS libraries when queried against *Arabidopsis* and *Brassica* databases. A total of 427 *gypsy*-like reverse transcriptase sequences were identified. Phylogenetic analysis of 373 of these sequences confirmed the existence of three distinct classes of *gypsy*-like retrotransposons among the four libraries identified in the initial BLAST search (Figure 3, and see below). Reverse transcriptase sequences from *copia*-like retrotransposons (n = 71) and LINE-like retroposons (n = 20) in addition to
transposase sequences from hAT-like (n = 2), Mutator-like (n = 1), and En/Spm-like (n = 15) transposable elements were also retained for further analysis.

Tandem repeats were identified using Tandem Repeat Finder (Benson 1999). Sequences identified by TRF as tandemly repetitive were queried against GenBank using BLASTN in an attempt to assign sequence identity. Gossypium 5SrDNA sequences and a previously published Gossypium sequence, pXP1-80 (Zhao et al. 1998), were recovered from all four of the WGS libraries (see below).

**Copy number estimates and lineage specific amplification**

**Tandem repeats.** Sequences with high identity to previously described Gossypium 5SrDNA repeats were identified in all four libraries. Estimates for D- (7675 ± 3826) and A- (5073 ± 3379) genome 5SrDNA copy numbers are in agreement with previously published estimates (Cronn et al. 1996) of 4730 ± 893 for G. raimondii and 3415 ± 807 for G. herbaceum, those of the latter study being based on Southern hybridization data. Estimated copy numbers for 5SrDNA sequences among the four libraries fall well within the same 95% CI (Table 2). Several other tandem repeats were recovered. One of these tandem repeats was identified as a previously published Gossypium repeat, pXP1-80 (Zhao et al. 1998). This 170 - 172 bp repeat was present in all four of the WGS libraries. Similar to the 5SrDNA repeats, copy number estimates for pXP1-80 were comparable between three of the four species (G. kirkii - 12263 ± 6098; G. raimondii - 6573 ± 3956; G. herbaceum - 10101 ± 5391) but elevated in G. exiguum (23795 ± 8528). It may be that pXP1-80 is a centromere repeat, given that it is present in all of the WGS libraries, and its length is similar to that of published centromere repeats from Arabidopsis (178 bp), wheat (192 bp), rice (155 bp), and maize (156 bp) (Ananiev et al. 1998; Hall et al. 2003; Ito et al. 2004; Nagaki et al. 2004). Several other tandem repeats of unknown identity were identified by TRF. However, none of the remaining tandem repeats were shared among the WGS libraries, and all were present in low copy number.

**Class II transposons.** The three major superfamilies of Class II DNA transposons present in the WGS libraries are members of the En/Spm, Mutator, and hAT
DNA transposon families. Class II sequences identified were few in number, with copy number estimates suggesting that, taken as a whole, these sequences occupy less than 2% of the *Gossypium* genome (Table 2). *En/Spm*-like sequences occupy less than one percent of the genome in each of the four species, comprising approximately 0.2% of the *G. kirkii* (~120 copies) and *G. herbaceum* (~343 copies) genomes, but increasing in copy number in the smallest (*G. raimondii* - 0.9%, ~835 copies) and largest (*G. exiguum* - 1.0%, ~2515 copies) genomes. Similarly, *hAT*-like sequences occupy less than one percent of the genome in each of the four species. *hAT*-like sequence comprise only 0.2% of the *G. kirkii* (~300 copies) genome, and an even smaller portion of the *G. raimondii* (0.03%, ~80 copies) and *G. herbaceum* (0.06%, ~260 copies) genomes. However, a large increase in copy number occurred in the K genome lineage (0.4%, ~2600 copies). *Mutator*-like sequences were identified in the WGS library, but because of the large range in published lengths for these sequences and the absence of a described *Mutator*-like transposon for *Gossypium*, it was not possible to estimate their copy numbers with confidence. Additionally, because of the degenerate nature of the identified Class I sequences, it is likely that there are other undetected sequences of this type in *Gossypium*. There was no evidence of MITEs, TRIMs, LARDs, or Helitrons in the WGS libraries.

**Class I retrotransposons.** The most highly represented group of repetitive sequences within all four WGS libraries are the Class I elements (Figure 2). Estimated total Class I copy numbers range 4.4-fold, from 45515 ± 9241 in *Gossypiodes kirkii* to approximately 197294 ± 18935 in the K genome species, *Gossypium exiguum*. When multiplied by an average size of 9.7 kb per gypsy, 5.3 kb per *copia*, and 3.5 kb per LINE sequence, we estimate that Class I elements occupy a minimum of 45% to 60% of the genome for each of these species, suggesting they have amplified in each lineage approximately in proportion to genome size. However, differential proliferation among species for each group of retrotransposons is evident from the copy number estimates. Copy number estimates for *copia*-like retrotransposons increase proportionally with genome size, with the exception of those from the D genome, which are much higher than expected (Table 2). *Copia*-like sequences occupy 10% - 20% of the *G. kirkii*, *G.
herbaceum (A) and G. exiguum (K) genome, but have reached considerably higher density in the species with the smallest genome size, G. raimondii (D) (28% to 39%). LINE-like retroposons are present in similar copy number in the D-genome (13011; 4.3%) and outgroup G. kirkii (16006; 7.9%) species, but have reached notably higher copy numbers in the A (30000; 5.3%) and K (27563; 3.3%) species, both of which contain much larger genomes (Table 2). We were unable to identify SINE-like retroposons (the non-autonomous counterpart of LINEs) in the WGS libraries.

The most striking example of differential lineage-specific amplification of specific groups of repetitive sequences is found among the gypsy-like sequences. BLAST analysis led to the discrimination of three different types of gypsy-like sequences present in the WGS libraries, and copy-number estimates for each of these are shown separately in Table 2. Phylogenetic analysis of 373 gypsy-like reverse transcriptase sequences assembled from all four of the WGS libraries confirmed the existence of these three distinct classes, here designated Gorge1, Gorge2 and Gorge3, for Gossypium retrotransposon gypsy-like elements (Fig. 2). The Gorge1 group is similar to the Arabidopsis gypsy sequence athila, Gorge2 is similar to maize cinfu1, and Gorge3 is similar to del1-46 from Lilium henryi and deal from Ananas comosus. Copy number calculations for the three types of sequences revealed relatively stable copy numbers for Gorge1 and Gorge2 across all four species, although the copy number estimate for Gorge1 in the D genome (1971 ± 1762) is somewhat lower than that of the other three species and Gorge2 copy number is slightly elevated in the K genome (8220 ± 3983) (Table 2). In contrast to this relative stability for Gorge1 and Gorge2, there is a profound increase in copy number of Gorge3 gypsy elements in the larger-genome species. Whereas copy numbers for Gorge3 are similar in G. kirkii and G. raimondii (5502 ± 3305 and 8674 ± 3683, respectively), a striking increase in copy number has taken place in both the A (48181 ± 9257) and K (88492 ± 12904) genome lineages. There is a six-fold increase of Gorge3 copy number from the D to the A genome, and copy number in the A genome is nearly doubled in the K genome. The impact of this proliferation on genome size is apparent from density calculations: Gorge3 occupies approximately 9% of both
the *G. kirkii* and D-genomes, but 27.3% and 33.7% of the A and K genomes, respectively.

*Unidentified repetitive fraction*

Clones with no similarity to any sequence deposited in GenBank were placed in a separate database as the "unidentified fraction" of each of the WGS libraries. These sequences were queried against each other using BLASTN to identify repetitive sequences that were missed during the initial BLAST search. Any sequence with greater than 80% similarity to at least three other clones from the same library was considered repetitive. A total of 43, 129, 364, and 603 clones from the *G. kirkii*, D, A and K libraries, respectively, were considered repetitive under this criteria. The percentage of each library composed of these unidentified repetitive sequences is as follows: *G. kirkii* (OG) - 3%; *G. raimondii* (D) - 5%; *G. herbaceum* (A) - 7.5%; and *G. exiguum* (K) - 9%.

**Discussion**

Variation in nuclear DNA content observed within and between organisms has been a topic of interest dating back to the early 1900s, but was specifically defined and named the "C-value paradox" by Thomas in 1971 (Thomas 1971). Investigations over the past half century have revealed multiple sources of genome size variation, most commonly the differential accumulation or deletion of transposable elements. Repetitive DNA constitutes 80% of angiosperm genomes with haploid DNA content greater than 5.0 pg (Flavell et al. 1974). Approximately 60% or more of the maize (Meyers et al. 2001; SanMiguel et al. 1998; SanMiguel et al. 1996), wheat (Wicker et al. 2001) and barley (Shirasu et al. 2000; Vicent et al. 1999) genomes are composed of transposable elements. Nearly 25% of the maize genome consists of five classes of LTR retrotransposons alone (SanMiguel et al. 1996), and LTR retrotransposon accumulation is responsible for nearly doubling the maize genome in as little as three million years (SanMiguel and Bennetzen 1998). Roughly 80% of the wheat genome is repetitive DNA, mainly LTR retrotransposons (Kumar and Bennetzen 1999). These observations have led to an interest in the effects of repetitive DNA on genome size variation and its significance to
plant fitness. Relatively little is known, however, about the evolutionary dynamics of transposable element accumulation among closely related species and how this varies among TE classes.

We constructed WGS libraries for three *Gossypium* and one outgroup species that range approximately 4-fold in genome size in order to describe their overall genomic composition and to determine the sequences that contribute to genome size variation. Congruent with results from taxa studied to date, we found that the majority of the *Gossypium* genome consists of dispersed repetitive sequences. Density estimates based on previously reported repetitive sequence lengths suggests that the *Gossypium* genome is composed of approximately 45-60% repetitive sequences when considering only those sequences with positive BLAST matches to previously identified repetitive elements in *Gossypium* or in other species. This number is in agreement with estimates from other species with large genomes, such as maize, barley, and wheat (Kumar and Bennetzen 1999; Meyers et al. 2001; SanMiguel and Bennetzen 1998; SanMiguel et al. 1996; Shirasu et al. 2000; Vicent et al. 1999; Wicker et al. 2001), but differs from estimates for smaller-genome species such as *Arabidopsis* (~14%) and rice (~26%) (Jiang et al. 2004; Jiang and Wessler 2001; The Arabidopsis Genome Initiative 2000). Additionally, given the number of repetitive sequences of unknown identity recovered in the self-BLAST searches, 45-60% clearly is an underestimate of the actual repetitive fraction.

Also in agreement with results from other well-studied taxa, the majority of the identified repetitive fraction consists of Class I retrotransposon sequences. As expected based on reported estimates from many grasses and a few well-studied eudicots, Class II sequences were less abundant and constituted a minor fraction of the *Gossypium* genomes (2%). This estimate is comparable to that from *Arabidopsis* (2-3%) (The Arabidopsis Genome Initiative 2000), whose genome is almost five times smaller that that of *G. kirkii* (588 Mb), as well as maize (2%) (Meyers et al. 2001), whose genome is only slightly larger than that of *G. exiguum* (2460 Mb). However, this result is in contrast with that from *Brassica* and rice, whose genomes harbor approximately 6%, and 12% Class II DNA transposons, respectively (Jiang et al. 2004; Jiang and Wessler 2001).
**Lineage-specific transposition**

A key conclusion of the present study is that genome size variation in a single genus of plants reflects not only the differential amplification of diverse types of repetitive sequences, but that specific families within a repetitive sequence type proliferate differentially as well. From a purely quantitative standpoint, much of the genome size variation observed in *Gossypium* is a consequence of the propagation of one particular family within the larger class of gypsy-like retrotransposons, i.e., *Gorge3*. Recently, a gypsy-like retrotransposon (“G45” and “G84”) that is transcriptionally active was reported in the tetraploid *G. barbadense* (Zaki and Ghany 2004). Comparisons of this active gypsy with sequences recovered in the present study revealed a maximum of 96.1% and 80.9% amino acid sequence identity between the best BLAST hit to an A-genome *Gorge3* and G45 and between A-genome *Gorge3* and G84, respectively. Additionally, 165 out of 150322 *Gossypium* ESTs show greater than 60% sequence similarity over more than 75 bp to *Gorge3* (e-value cut-off of e-20, data not shown). Based on this high level of sequence identity to G45 and G84, presence in the *Gossypium* EST libraries, and overabundance of *Gorge3* in the WGS libraries, we believe *Gorge3* is a recently active, major constituent of the cotton genome that, like LTR retrotransposons in maize, has triggered a 3-fold increase in genome size over the 5-10 million years since the diversification of the major *Gossypium* clades following the origin of the genus (Cronn et al. 2002).

It is interesting to note that other repetitive sequences that are less common than *Gorge3* have also been subject to lineage-specific amplification during diversification of the genus. For example, little amplification of LINE retrotransposons has occurred in the D genome lineage, but these sequences have proliferated in the A and K genome species. Similarly, accumulation of *copia*-like retrotransposons has occurred in the D genome lineage, yet these repetitive elements have been suppressed in the remainder of the genus, with the proportion of the genome occupied by *copias* in the remaining three species being between 10 and 20%. Indeed, *G. raimondii* is the only studied *Gossypium* species in which there are more *copia*-like than gypsy-like sequences (Figure 2).
The most parsimonious interpretation of the *copia* data would invoke differential amplification in the D genome lineage. However, we cannot discount the possibility of unequal rates of DNA loss. Some species appear to be more efficient at removal of non-essential DNA, such that genome size may reflect, at least in part, differential rates of DNA loss (Kirik et al. 2000; Orel and Puchta 2003; Petrov and Hartl 1997, Petrov et al. 2000). With respect to the present study, LINE-like sequences recovered in the WGS libraries are often highly degraded and hence difficult to identify. Although the most parsimonious interpretation of our copy-number estimates is a single amplification event in the common ancestor of the A- and K- genome lineages, a formal alternative is that LINE-like sequences existed at an ancestrally high copy number and have subsequently been differentially eliminated from the species with smaller genomes (D genome and the outgroup *G. kirkii*).

*Genome size evolution in Gossypium*

At present relatively little is known about the genomic locations at which genome size evolution takes place in *Gossypium*. The data presented here show that specific families and classes of dispersed repetitive elements have differentially proliferated in different *Gossypium* lineages. Given the propensity of many high copy number LTR retrotransposons to accumulate in heterochromatic regions of the genome (Kumar and Bennetzen 1999), we suspect that much of the evolutionarily rapid genome size change that has arisen during the global radiation of *Gossypium* has occurred in these gene poor regions. Consistent with this notion, Grover et al. (2004) investigated genome size evolution in 104 kb of contiguous sequence surrounding the *CesA1* gene in the *Gossypium* A and D genomes from tetraploid cotton. Within this genic region of the *Gossypium* genome, no evidence of genome size variation was apparent, suggesting that genome size evolution in *Gossypium* takes place in heterochromatic regions located between highly conserved, euchromatic gene islands. Evaluation of this hypothesis will require additional comparative sequence and mapping data, the latter including visualization techniques such as florescent *in situ* hybridization (FISH) of various transposable elements.
In addition to transposable element accumulation, other suggested mechanisms of genome size change include variation in intron length, expansion/contraction of tandem repeats, illegitimate recombination, indel bias, and unequal intrastrand homologous recombination (Petrov and Wendel 2004). Contrary to suggestions that plants with smaller genomes carry smaller introns (Deutsch and Long 1999; Vinogradov 1999), there is no apparent correlation between genome size and intron length in *Gossypium* (Grover et al. 2004; Wendel et al. 2002a). In fact, intron length has been shown to be highly stable across 28 orthologous sets of genes from A and D genome diploid species and the outgroup species, *G. kirkii*. In the present study, we find no major difference between copy numbers for tandem 5SrDNA and pXP1-80 repeats, although there is a small increase in copy number in larger genomes. However, Cronn et al. (1996) reported a twenty-fold variation in 5SrDNA copy number among *Gossypium* species, reflecting both array expansion and contraction. Grover et al. (2004) found no evidence of an indel bias, and although there was some evidence of illegitimate recombination marked by flanking repeats of 2-15 bp in length, the resulting deletions encompass approximately the same proportion of sequence in each genome. Similar studies from other genomic locations in *Gossypium* will be necessary to determine if this result is a local or global occurrence.

**Concluding Remarks**

Comparative studies of genome size variation among phylogenetically characterized and closely related species serve an important role in clarifying the patterns and processes that underlie the striking genome size variation that characterizes eukaryotes in general and plants in particular. With respect to the latter, we note that the genomic architecture of most plant species remains to be elucidated, and hence mechanisms that characterize one group of plants may not be universal to, say, angiosperms in general. Our data, demonstrating that different families of different classes of TEs have differentially accumulated among closely related clades of a single plant genus, underscores what we believe will be a generality, namely, that mechanisms of genome size evolution are highly variable among even closely related lineages. Our
appreciation of plant genomic architecture will continue to be enhanced as comparable studies in other plant groups accumulate. These investigations will generate a deeper understanding of the genomic landscape of different plant lineages, the scale, scope and pace of evolutionary change responsible for the observed patterns, and insights into the mechanisms the underlie the differential accumulation of different sequence types among genomes.

Acknowledgments
We thank Curt Brubaker for providing *Gossypium exiguum* seeds and DNA, S. Wessler and F. Zhang for sharing a repetitive sequence database, Jordan Swanson for assistance with data analysis, Ryan Percifield, Corrinne Grover and Ryan Rapp for helpful discussion, and the anonymous reviewers for their comments. This work was funded by the National Science Foundation Plant Genome program.

Literature Cited


*Genome Res.* **12**: 1075-1079.


Table 1. Library construction and sequencing effort for three species representing different *Gossypium* genomes and one phylogenetic outgroup.

<table>
<thead>
<tr>
<th>Taxon/genome group</th>
<th>Genome size (in Mb)</th>
<th>No. clones in library</th>
<th>Successfully sequenced</th>
<th>Average read (bp)</th>
<th>% genome sequenced</th>
<th># Mb sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Gossypiooides kirkii</em></td>
<td>588</td>
<td>1920</td>
<td>1464</td>
<td>753</td>
<td>0.19%</td>
<td>1.10</td>
</tr>
<tr>
<td>Outgroup</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Gossypium raimondii</em></td>
<td>880</td>
<td>3072</td>
<td>2722</td>
<td>770</td>
<td>0.24%</td>
<td>2.10</td>
</tr>
<tr>
<td>D genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. herbaceum</em></td>
<td>1667</td>
<td>6048</td>
<td>4864</td>
<td>704</td>
<td>0.21%</td>
<td>3.42</td>
</tr>
<tr>
<td>A genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. exiguum</em></td>
<td>2460</td>
<td>10368</td>
<td>6747</td>
<td>704</td>
<td>0.19%</td>
<td>4.75</td>
</tr>
<tr>
<td>K genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TOTAL 11.4

*a*Genome size from Wendel et al. 2002 for *G. kirkii* and Hendrix and Stewart 2005 for *G. raimondii, G. herbaceum,* and *G. exiguum.*
Table 2. Repetitive element copy number and density estimates.

<table>
<thead>
<tr>
<th></th>
<th>G. kirkii Outgroup 588 Mb</th>
<th>G. raimondii D genome 880 Mb</th>
<th>G. herbaceum A genome 1667 Mb</th>
<th>G. exiguum K genome 2460 Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tandem Repeats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5SrRNA</td>
<td>4279 ± 3227</td>
<td>7675 ± 3826</td>
<td>5073 ± 3379</td>
<td>10794 ± 5082</td>
</tr>
<tr>
<td>pXP1-80</td>
<td>12264 ± 6098</td>
<td>6573 ± 3956</td>
<td>10101 ± 5392</td>
<td>23795 ± 8528</td>
</tr>
<tr>
<td>Class II transposons</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>En/Spm-like</td>
<td>120 ± 138</td>
<td>835 ± 326</td>
<td>343 ± 216</td>
<td>2514 ± 602</td>
</tr>
<tr>
<td>~0.2%</td>
<td>~0.9%</td>
<td>~0.2%</td>
<td>~1.0%</td>
<td></td>
</tr>
<tr>
<td>hAT-like</td>
<td>305 ± 352</td>
<td>81 ± 163</td>
<td>263 ± 304</td>
<td>2615 ± 986</td>
</tr>
<tr>
<td>~0.2%</td>
<td>&lt;0.1%</td>
<td>&lt;0.1%</td>
<td>~0.4%</td>
<td></td>
</tr>
<tr>
<td>Class II TOTAL</td>
<td>3.5 Mb</td>
<td>12 Mb</td>
<td>5 Mb</td>
<td>42 Mb</td>
</tr>
<tr>
<td>&lt;0.1%</td>
<td>1.0%</td>
<td>&lt;0.1%</td>
<td>~1.4</td>
<td></td>
</tr>
</tbody>
</table>
## Class I

**retrotransposons**

<table>
<thead>
<tr>
<th>Family</th>
<th>Copy Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>copia-like</strong></td>
<td>17006 ± 5765</td>
<td>9.7% - 19.7%</td>
</tr>
<tr>
<td><strong>LINE</strong></td>
<td>16006 ± 5597</td>
<td>5.1% - 10.6%</td>
</tr>
<tr>
<td><strong>GORGE1 gypsy-like</strong></td>
<td>4502 ± 2992</td>
<td>2.4% - 11.9%</td>
</tr>
<tr>
<td><strong>GORGE2 gypsy-like</strong></td>
<td>2500 ± 2233</td>
<td>0.4% - 7.5%</td>
</tr>
<tr>
<td><strong>GORGE3 gypsy-like</strong></td>
<td>5502 ± 3305</td>
<td>3.5% - 13.9%</td>
</tr>
<tr>
<td><strong>Class I TOTAL</strong></td>
<td>255 Mb</td>
<td>42%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>Copy Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>copia-like</strong></td>
<td>57956 ± 9300</td>
<td>28% - 38.7%</td>
</tr>
<tr>
<td><strong>LINE</strong></td>
<td>13011 ± 4502</td>
<td>2.8% - 5.7%</td>
</tr>
<tr>
<td><strong>GORGE1 gypsy-like</strong></td>
<td>1971 ± 1762</td>
<td>0.2% - 3.9%</td>
</tr>
<tr>
<td><strong>GORGE2 gypsy-like</strong></td>
<td>3154 ± 2227</td>
<td>1.0% - 5.7%</td>
</tr>
<tr>
<td><strong>GORGE3 gypsy-like</strong></td>
<td>8674 ± 3683</td>
<td>5.3% - 13.0%</td>
</tr>
<tr>
<td><strong>Class I TOTAL</strong></td>
<td>465 Mb</td>
<td>53%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>Copy Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>copia-like</strong></td>
<td>43181 ± 8774</td>
<td>10.7% - 16.1%</td>
</tr>
<tr>
<td><strong>LINE</strong></td>
<td>30000 ± 7335</td>
<td>4.0% - 6.5%</td>
</tr>
<tr>
<td><strong>GORGE1 gypsy-like</strong></td>
<td>5909 ± 3273</td>
<td>1.5% - 5.2%</td>
</tr>
<tr>
<td><strong>GORGE2 gypsy-like</strong></td>
<td>3181 ± 2403</td>
<td>0.4% - 3.2%</td>
</tr>
<tr>
<td><strong>GORGE3 gypsy-like</strong></td>
<td>48181 ± 9257</td>
<td>22.0% - 32.6%</td>
</tr>
<tr>
<td><strong>Class I TOTAL</strong></td>
<td>865 Mb</td>
<td>52%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>Copy Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>copia-like</strong></td>
<td>67700 ± 11324</td>
<td>11.7% - 16.5%</td>
</tr>
<tr>
<td><strong>LINE</strong></td>
<td>27563 ± 7271</td>
<td>2.4% - 4.1%</td>
</tr>
<tr>
<td><strong>GORGE1 gypsy-like</strong></td>
<td>5319 ± 3205</td>
<td>0.8% - 3.2%</td>
</tr>
<tr>
<td><strong>GORGE2 gypsy-like</strong></td>
<td>8221 ± 3983</td>
<td>1.6% - 4.7%</td>
</tr>
<tr>
<td><strong>GORGE3 gypsy-like</strong></td>
<td>88492 ± 12904</td>
<td>28.8% - 38.6%</td>
</tr>
<tr>
<td><strong>Class I TOTAL</strong></td>
<td>1400 Mb</td>
<td>58%</td>
</tr>
</tbody>
</table>
Figure 1. Evolutionary relationships among diploid members of Gossypium. *Gossypium* is a monophyletic genus composed of approximately 50 species that are widely distributed throughout many tropical and subtropical regions. Diploid species have a haploid complement of 13 chromosomes. *Gossypium* is divided into eight genome groups based on cytogenetic data and level of fertility in interspecific hybrids (Endrizzi et al. 1985). Multiple molecular datasets support the phylogenetic relationships indicated, including the outgroup relationship of *Gossypioides kirkii* (Seelanan et al. 1997, Small et al. 1998, Small et al. 1999). Despite conservation of chromosome number among the diploids, genome size varies 3-fold, from an average of 885 Mb in the New World D-genome species to an average of 2576 Mb in the Australian K-genome species (Hendrix and Stewart 2005).
Figure 2. Copy number estimates for repetitive sequences in *Gossypium*. Copy numbers for repetitive sequences recovered in the WGS libraries were estimated as described (see Methods). The majority of repetitive sequences are LTR retrotransposons, particularly in the larger-genome species. In both the A and K genomes, massive amplification of *Gorge3* gypsy-like sequences has occurred, contributing predominantly to genome size expansion in these two lineages. In the smallest *Gossypium* genome, *G. raimondii* (D-genome), *copia*-like sequences have proliferated and are primarily responsible for genome size expansion in this lineage. Class II sequences were less abundant and appear to contribute little to genome size evolution in the genus. Tandem repeats are approximately evenly distributed among all four species, with pXP1-80 sequences slightly elevated in *G. exiguum* (K-genome).
Figure 3. Neighbor-joining analysis of *Gossypium* gypsy-like Gorge1, 2, and 3 reverse transcriptase sequences. Unrooted Neighbor-joining analysis of 373 *Gossypium*, 24 *Arabidopsis*, and 36 *Brassica* gypsy reverse transcriptase sequences provides support for the three distinct classes of gypsy-like sequences in *Gossypium*. Gorge1 is similar to *Arabidopsis* gypsy sequence *athila*, Gorge2 is similar to maize *cinful1*, and Gorge3 is similar to dell-46 from *Lilium henryi* and *deal* from *Ananas comosus*. Bootstrap values for the deeper nodes are shown.
CHAPTER FOUR

Rapid DNA loss as a counterbalance to retrotransposon proliferation in plant genome size evolution

A paper submitted to Proceeding of the National Academy of Sciences of the USA

Jennifer S Hawkins¹, Ryan A Rapp², and Jonathan F Wendel³

Abstract

Nuclear DNA content in plants varies over 1,000 fold, from the miniscule 108 Mb genome of Fragaria viridis to the enormous 120,000 Mb genome of some members of the Liliaceae. Transposable elements, particularly LTR-retrotransposons, comprise the primary vehicle for genome size expansion, but the counteracting mechanisms and relative importance of genome downsizing as determinants of genome size are unclear. In the genus Gossypium (cotton), the 3-fold genome size variation among diploids is due largely to copy number variation of the gypsy-like retrotransposon Gorge3. Here we combine comparative sequence analysis with a modeling approach to study the evolutionary history of DNA loss/gain dynamics for Gorge3 in Gossypium. Our analysis reveals recent, lineage-specific and episodic amplification of Gorge3 in each genome. Additionally, we demonstrate that the rate of DNA removal in the smaller genomes is sufficient to counteract genome expansion through Gorge3 proliferation. These data suggest that rates of DNA loss can be highly variable even within a single plant genus, and that genome contraction may be powerful determinant of genome size in plants.

¹ Graduate student, primary researcher and author, EEOB department, Iowa State University, Ames, Iowa.
² Graduate student, statistical analysis and model design, EEOB department, Iowa State University, Ames, Iowa.
³ Principal investigator and corresponding author, EEOB department, Iowa State University, Ames, Iowa.
Introduction

Plant genomes vary enormously in size, from ~108 Mb in *Fragaria viridis* (Bennett and Leitch 2005) to >120,000 Mb in members of the Liliaceae (Bennett and Smith 1991; Flavell et al. 1974). While the genesis of this extraordinary variation has been of interest for over half a century (Mirsky and Ris 1951; Thomas 1971), numerous studies have shown that most genome size variation can be ascribed to differential accumulation of the repetitive fraction of the genome, particularly long terminal repeat (LTR) retrotransposons (Hawkins et al. 2006; Hill et al. 2005; Meyers et al. 2001; Neumann et al. 2006; SanMiguel et al. 1998; SanMiguel et al. 1996; Shirasu et al. 2000; Vicient et al. 1999; Vitte and Bennetzen 2006; Wicker et al. 2001). Additionally, TE proliferation is a dynamic process, occurring repeatedly over short evolutionary timescales. For example, studies in maize suggest a doubling of its genome over as little as three million years due to TE accumulation alone (SanMiguel et al. 1998; SanMiguel et al. 1996). The same pattern has been shown in *Oryza australiensis*, where three types of LTR-retrotransposons have proliferated recently and rapidly, in episodic bursts that have doubled the genome within the last two million years (Piegu et al. 2006).

Several mechanisms of DNA loss have been proposed as possibilities for attenuating genomic expansion through TE proliferation. One is intra-strand homologous recombination, thought to occur predominantly between the directly repeated LTRs of retrotransposons, and evidenced by a remaining solo LTR (Lim and Simmons 1994; Shirasu et al. 2000). As shown by Devos et al. (Devos et al. 2002), the impact of LTR recombination on genome size can be inferred from the frequency of solo LTRs remaining in a genome. A second mechanism is illegitimate recombination, which generally takes place via non-homologous end-joining (NHEJ) or slip-strand mispairing, resulting in small deletions. Kirik and Puchta have shown that NHEJ results in larger and more frequent deletions in *Arabidopsis* than *Nicotiana*, the latter having a genome 20 times larger than the former (Kirik et al. 2000). Similarly, comparisons of internally deleted LTR-retrotransposons from rice and *Arabidopsis* suggest that illegitimate recombination may be the driving force behind DNA removal in these smaller-genome taxa (Devos et al. 2002; Ma et al. 2004). These observations led to the hypothesis of a
“small indel bias”, in which smaller genomes are derived by being more strongly impacted by deletional forces than are larger genomes (Petrov 2002; Petrov et al. 2000). However, Vitte and Bennetzen have shown that across a diverse group of angiosperms, the rate of DNA loss through illegitimate recombination is not correlated with phylogenetic relatedness or genome size (Vitte and Bennetzen 2006). The impact of a small indel bias remains controversial (Gregory 2004), and further investigations are required in order to assess its relative importance in shaping modern plant genomes.

Given the rapid and recent accumulation of TEs in many plant genomes, combined with a short half-life for LTR-retrotransposons (Ma et al. 2004), insights into deletion dynamics and their impact on plant genome size are likely to emerge from studies of relatively recently diverged taxa. The cotton genus, *Gossypium*, is an especially good model in this respect. *Gossypium* (Malvaceae) is a monophyletic genus comprising approximately 50 diploid species of small trees and shrubs that are distributed throughout the world (Cronn et al. 2002; Fryxell 1992; Seelanan et al. 1997; Wendel and Cronn 2002). Diploid members contain 13 chromosomes, and are divided into eight (A-G, K) genome groups based on chromosome pairing behavior and interspecific fertility in hybrids (Beasley 1941; Endrizzi et al. 1985). Haploid nuclear content ranges three-fold, from an average 885 Mb in the New World, D-genome species, to 2572 Mb in the Australian, K-genome species (Hendrix and Stewart 2005). This wide range in genome sizes and a well-established phylogeny make *Gossypium* an excellent model for studying the impact and dynamics of DNA removal as an evolutionary determinant of genome size.

Here, we focus on the abundant *gypsy*-like LTR-retrotransposon, *Gorge3* (Hawkins et al. 2006). Using degenerate primers for the reverse transcriptase (RT) region of *Gorge3*, we amplified and performed phylogenetic analysis of 724 sequences from three *Gossypium* species that range three-fold in genome size, and from a phylogenetic sister group (Seelanan et al. 1997) to *Gossypium*, i.e., *Gossypioides kirkii*. Consistent with expectations from other studies in angiosperms, we show that recent episodic bursts of transposition have in fact occurred in each lineage, and that the magnitude of each burst is in direct positive correlation with genome size. In addition,
however, we use a novel modeling approach to show that species with small genomes have experienced a faster rate of Gorge3 sequence removal than those with large genomes. The implication is that DNA removal is a powerful determinant of genome size variation among plants, and that it can be a sufficiently strong force to overwhelm expansion through transposition.

Methods

Plant material, PCR amplification, and phylogenetic analysis

Total genomic DNA from the A-genome species *G. herbaceum* (JMS; 1C=1667Mb), the D-genome species *G. raimondii* (JFW; 1C=880Mb), the K-genome species *G. exiguum* (Gos 5184; 1C=2460Mb), and the phylogenetic outgroup *Gossypioidea kirkii* (JMS stock; 1C=588Mb) was extracted using the Plant DNeasy mini kit (Qiagen Inc., Valencia, CA). The Gorge3 reverse transcriptase (RT) region was amplified using primers obtained from M. Ungerer at Kansas State Univ. Primer sequences are as follows: Gorge3F: 5’GGA CCT GCT GGA CAA GGG NTW YAT HMG 3’, and Gorge3R 5’CAG GAA GCC CAC CTG CCK NWR CCA RAA 3’. PCR products were amplified in 20 µl reactions containing 1X PCR buffer (BioLine USA Inc., MA), 1.8 mM MgCl2, 500 µM dNTPs, 3.75 µM each primer, and 2.5 U Taq DNA polymerase (BioLine USA Inc., MA). The amplification profile was as follows: 94°C for 3 min, followed by 35 cycles of 94°C for 1 min, 45°C for 2 minutes, and 72°C for 1 min, and a final elongation step of 72°C for 6 min. PCR products were excised from 1.0% agarose gels, purified using the Qiaquick gel purification kit (Qiagen Inc., Valencia, CA), ligated into the pGEMT-easy vector (Promega, WI), and subsequently transformed into Mach1 chemically competent cells (Invitrogen, CA). Cloned PCR products were sequenced from one direction using the T7 primer and BigDye Terminator v3.1 (Applied Biosystems, CA). Sequencing products were separated on an ABI 3730xl DNA analyzer (Applied Biosystems, CA) at the Iowa State University DNA sequencing facility. Vector sequence was removed with Crossmatch (Green 1999) and primer sequences were removed by hand. Sequences were aligned using MUSCLE (Edgar 2004) and subsequently inspected by eye. Neighbor-joining analyses were performed on the aligned sequences using PAUP* (Swofford
Distances were uncorrected (“p”) and missing data were ignored for affected pairwise comparisons.

**Identification and dating of lineage-specific transposition events**

Evidence suggests that an average of 10% sequence divergence has accumulated between orthologous A and D transposable elements since their divergence from a common ancestor in *Gossypium* (Grover et al. 2007). Therefore, monophyletic RT clades sharing an average of 90% or greater sequence identity were considered to be “lineage-specific”. Based on sequences from 48 nuclear genes, the genetic distance between *Gossypium* and the outgroup *Gossypioides kirkii* is approximately twice that of the A-D divergence (Senchina et al. 2003), which represents the basal-most split in the evolutionary history of *Gossypium* (estimated at 5-7 mya; Cronn et al. 2002; Wendel and Cronn 2002). Accordingly, RT lineages in *Gossypium* that shared an average of ≥80% sequence identity were considered to be *Gossypium*-specific. Using molecular clock-based estimates of divergence times these lineages were inferred to have transposed within the last 10-12 million years (my). RT sequences more dissimilar than 80% from one another were considered to have originated from transposition events that occurred prior to the origin of the genus.

To elucidate the timing of TE proliferation events, we constructed midpoint-rooted, neighbor joining trees based on pairwise distances generated using Felsenstien’s 84 model. An orthonormal basis was used for node numbering and the node lineage from each tip sequence to the root was extracted. Pairwise nucleotide diversity (π) was calculated for all comparisons of daughter sequences at each node, and kernel density estimates were made on the frequency of nucleotide divergences to assess whether transposition occurred stochastically or in bursts. Statistical work was done in R with the base package, and using the phylogenetic package ape (Paradis et al. 2004; R developmental core team 2005).

*Estimated amount of post-speciation Gorge3 accumulation and deletion*
Genomic shotgun sequences (GSS) comprise a putatively unbiased random sample of a genome. To identify the proportion of extant *Gorge*3 sequences originating at various evolutionary times, the *Gorge*3 PCR sequences were queried against previously generated (Hawkins et al. 2006) *Gossypium* GSS libraries using TBLASTX with an e-value cut-off of 1e-10. The 294 *Gorge*3 GSS sequences recovered were aligned with the PCR sequences in MUSCLE (Edgar 2004). The combined data set of 1018 sequences was subjected to neighbor-joining analysis and the approximate timing of the origin of each clade was determined (described above). The number of GSS sequences originating within the three separate time points (lineage-specific, *Gossypium* specific, pre-*Gossypium*) was determined, and these numbers were used to estimate the copy numbers of *Gorge*3 from each of these three time points (Hawkins et al. 2006). The total number of *Gorge*3 Mb from each time point was estimated by multiplying the copy number by 9.7 Mb, the average length of *Gorge*3 in *Gossypium* (Grover et al. 2007; Hawkins et al. 2006).

We wished to ascertain whether specific portions of the lineage (i.e. internodes) experienced biased gain/loss ratios of *Gorge*3. To this end, we used the GSS copy numbers from each time period combined with a modeling approach to construct possible *Gorge*3 gain/removal ratios on each internode of the species tree. Specifically, we began by considering two components: 1) the system of equations that describes the rate of change of *Gorge*3 DNA, and 2) the error associated with our empirical observations (i.e., estimated number of Mb from each time point). To model the change in *Gorge*3 abundance across the trees, it is necessary to consider the unique components (lineage-specific) and the shared components (internal internodes), and their relative implications on rate estimations. A system of linear equations was developed for this purpose, wherein each equation represents a specific extant lineage in the phylogenetic tree, and each term is composed of a scalar, the standardized branch length, and an unknown rate. Branch lengths are standardized by dividing by the total length of the tree, thereby making the scalars equivalent with respect to time. Thus, for the matrix:
$r$ represents the rate of gain to loss unique to the branch, and the subscripts denote the nomenclature of the internode. The right-hand side is generated by calculating the percent change of Mb of DNA in each lineage (simply dividing the extant amount by the ancestral state). This construction is applicable to a tree approach, as the scalar now represents the percentage of the organism’s life history spent under the influence of a given gain/loss ratio.

The second component to be considered is the error associated with the empirical measurements of DNA and the estimations of ancestral amounts. Using the 95% confidence interval for these values, it is possible to construct a parametric solution space for the system of equations, and describe the bounds of possibility, given our model and the empirical error. To construct the solution space, Givens Rotations were used to formulate the QR decomposition of the system of equations, and this was used to generate a least-squares fit to our underdetermined system. Because this is a tree-driven structuring of a linear system, terminal branches can be solved while the internal nodes retain an infinite number of solutions. The unsolved internal internodes form the basis for a parametric solution space. Although interesting, we are concerned with the lineage specific effects, and the impact of error from the manner in which the data were collected. To assess this, we explored the possible set of QR decompositions that result from the range of variability being sampled. Specifically, we ran 100,000 simulations where extant MB of DNA were randomly chosen from inside the 95% confidence interval for each species and the ancestral state. From these random values, we resolved the linear system, and constructed a kernel density estimate to understand the density variations of the space occupied by the solution set.

**Results**

*Phylogenetic analysis and timing of transposition events*
A total of 724 unique and diverse reverse transcriptase (RT) sequences from *G. herbaceum* (A), *G. raimondii* (D), *G. exiguum* (K), and *Gossypioides kirkii* (outgroup) were subjected to phylogenetic analysis using neighbor-joining (Figure 3, supplemental). The resulting phylogram contained two "cotton-specific" clades consisting of sequences from all three *Gossypium* species. Lineage-specific sequences from the A- and K-genome species, which have the larger genomes, formed distinct clusters with short to medium branch-lengths, while sequences from the D-genome and *G. kirkii* presented longer branch-lengths. However, recent amplification of *Gorge3* even in the two species with small genomes, *G. kirkii* and *G. raimondii* (D-genome) was evidenced by small clusters with very short branch-lengths present at the tips of multiple longer branches. Few non-lineage specific sequences were recovered from the taxa with the larger genomes (*G. herbaceum* and *G. exiguum*).

Evaluation of the lineage-specific transpositional nature and timing of *Gorge3* in each genome revealed episodic bursts of activity since divergence from a common ancestor in all species, at different points in their respective evolutionary histories (Figure 1). All A-genome pairwise comparisons among lineage-specific clades cluster at 95% sequence identity, suggesting a sudden burst of transposition approximately 2-3 million years ago, followed by relative quiescence. Similarly, the K-genome appears to have experienced a burst of *Gorge3* transposition at approximately the same time as the A-genome. Although few lineage-specific D-genome sequences were sampled, most share greater than 99% sequence identity, suggesting very recent transpositional activity, perhaps within the last million years. Similarly, *G. kirkii* sequences clustered at 99% sequence identity, but also appear to have experienced a burst of transposition between 7 and 12 million years ago.

**Lineage-specific rates of expansion and contraction**

*Gorge3* sequences from previously constructed random GSS libraries were combined with the PCR-amplified sequences to estimate the amount of *Gorge3* gain and loss occurring in each lineage. A second round of neighbor-joining analysis was performed on the 724 PCR amplified sequences plus 294 *Gorge3* GSS sequences. As
with PCR sequences alone, the GSS sequences were partitioned into the three time points (lineage-specific, *Gossypium* specific, and pre-*Gossypium*) and the copy number and total number of *Gorge3* Mb originating at each time point was estimated (Table 1).

Surprisingly, *Gorge3* copy numbers were higher from the ancient, pre-*Gossypium* time point than any of the other time points, for all taxa. Copy number estimates from this oldest time point in *G. kirkii* (3001±2445) and the D genome (4731±2725) were not significantly different from one another, but many retained ancient copies of *Gorge3* were identified in the A genome (22272±6331) and twice as many ancient copies were recovered from the K genome (43037±9063). Copy number estimates for *Gossypium* -specific and lineage-specific time points were so low in the *G. kirkii* and D genomes that they cannot be accurately estimated at this level of sampling. However, a consistent decrease in the number of copies originating during these two time points is observed in the A and K genomes. Approximately 16363±5434 *Gossypium*-specific and only 6818±3515 lineage-specific copies were recovered from the A genome. Similarly, 27563±7271 *Gossypium*-specific and 12089±4827 lineage-specific copies were identified in the K genome. These copy numbers were subsequently used to estimate the total number of Mb from each time point in each genome, assuming the average *Gorge3* is 9.7 kb in length. While approximately 831 total Mb of *Gorge3* resides in the 2460 Mb genome of *G. exiguum*, only 111 Mb originated specifically within the lineage. The same trend is observed in all of the genomes, with approximately 223 Mb pre-*Gossypium* and 68 Mb lineage-specific in the A genome, and 47 Mb pre-*Gossypium* and only a few Mb lineage-specific in the D genome.

A linear programming approach was taken to address the underdetermined system of equations that represents the lineage-specific components of gain/loss ratios for *Gorge3* DNA in extant taxa (see methods). Considering the empirical estimates of extant *Gorge3* and their 95% confidence intervals, the lineage-specific relative rates of gain/loss of *Gorge3* DNA are as follows: A-genome: 6.13, D-genome: 0.78, K-genome: 16.60 *G. kirkii* genome: 0.26 (Figure 2). Values under 1 imply a loss of overall DNA, while values over 1 indicate a gain. When we examined the robustness of these inferences with respect to the parameter space as described above, we found that the D-genome and *G.*
kirkii densities were biased below 1 (DNA loss) with 70% of the possible values occurring below 1 for G. raimondii and 100% of the values occurring below 1 for G. kirkii. In contrast, the two species with larger genomes (A- and K-genome), both were biased toward ratios greater than 1 (DNA gain), with all possible values for both species being much larger than 1 (Figure 4, supplemental). Thus, it appears that genome contraction through deletion of Gorge3 elements has played a dominant role in shaping the G. raimondii and Gossypioides kirkii lineages, whereas genome expansion through Gorge3 proliferation is implicated in the other two lineages.

Discussion

That TE proliferation leads to genome expansion is widely recognized (SanMiguel and Bennetzen 1998; Vicent et al. 1999), but attempts to identify the counteracting forces that lead to genome downsizing have proven more elusive. Here, we investigated both the quantitative and temporal nature of gypsy-like Gorge3 evolution in Gossypium. Previous results indicate that copy number variation of this particular LTR-retrotransposon family is primarily responsible for the three-fold variation in genome size observed among diploid members of the genus (Hawkins et al. 2006). Congruent with these findings, we show here that Gorge3 has amplified differentially and independently in each of the lineages studied, with the highest copy number of new sequences in the largest (K) genome and the lowest in the smallest (D) genome. However, the transpositional history in each lineage is distinctive and different. While lineage-specific transposition is episodic in nature in all genomes investigated, transpositional events occurred at different times in the evolutionary history of each clade. Episodic bursts of transposition have also been demonstrated in Oryza australiensis (Piegu et al. 2006), a relative of rice with a large genome, suggesting that episodic, transpositional bursts may be a general phenomenon in angiosperm evolution. To the extent that this pattern holds, it raises intriguing questions about the mechanisms that govern relatively long periods of evolutionary stasis, as well as the nature of the “triggers” that release TEs from suppression. Stress and interspecific gene flow are known to disrupt epigenetic regulation, and hence these factors may well be involved; in
this respect it is noteworthy that *Gossypium* contains many documented examples of interspecific hybridization (Cronn and Wendel 2004).

*Evidence for genome downsizing in Gossypium*

Comparisons between orthologous BACs from the A and D genomes have provided insight into the mechanisms and rates of DNA loss in *Gossypium* (Grover et al. 2004; Grover et al. 2007). In a gene-rich region surrounding the *CesA* gene, both the genic and intergenic regions were highly conserved, but in the *AdhA* region this was not the case. Specifically, in this region the A-genome contained unique transposable element insertions and the D-genome exhibited a two-fold higher rate of indels, most containing hallmarks of illegitimate recombination, suggesting a higher rate of deletion in the smaller genome. Solo LTRs, indicative of DNA loss through intra-strand homologous recombination, were also evident, suggesting that both mechanisms are operating to remove DNA in *Gossypium*.

Evidence presented here supports the interpretation that genome downsizing occurred in the D-genome lineage since its origin and in spite of TE proliferation. Our combined empirical and modeling approaches suggest that there is enormous lineage-specific variation in the gain/loss ratio of *Gorge3* retrotransposons. The sequencing data highlight an ancient and massive retrotranspositional event in the common ancestor of all *Gossypium* species as well in the outgroup, *Gossypioides kirkii*. It is apparent that the A and K lineages have been unable to purge this ancient *Gorge3* DNA and have concomitantly accumulated more lineage-specific *Gorge3* copies. In contrast, the D genome not only has discarded much of its ancient *Gorge3* complement, but has also suppressed other rounds of TE proliferation. Our modeling results highlight the robust nature of this conclusion. Under most reasonable scenarios, the gain/loss ratios are significantly biased for loss in the taxa having smaller genomes.

One caveat of the current study is that of sequence identification. The GSS sequences used to estimate copy numbers from each genome were identified via similarity searches, and if in fact the D- and *G. kirkii* genomes posses a higher rate of small deletions, then the more degenerate *Gorge3* sequences will be difficult to identify.
This would lead to an underestimation of the total number of ancient copies residing in the smaller genomes. Every effort was made to avoid this potential pitfall by performing iterative blast searches within each GSS library to identify degenerate sequences with low sequence identity to *Gorge3*. Additionally, the GSS libraries represent a minimum level of sampling from each genome, so some of the paralogs for a particular transposition event may not be sampled.

One may question whether the observed evolutionary trends for *Gorge3* are representative of the entire genome or if *Gorge3* is subject to evolutionary pressures unique to its particular genomic milieu. For example, *gypsy*-like retrotransposons have been shown to preferentially insert into pericentromeric heterochromatin in *A. thaliana* (Peterson-Burch et al. 2004). It is unknown whether *Gorge3* exhibits similar insertion preferences or other biases, but the possibility remains that the inferences drawn here for rates of DNA loss and gain are not reflective of the genome overall. The veracity of the conclusions reached here, both with respect to *Gossypium* and other angiosperms (and perhaps other eukaryotes), will only become clear following comparable studies in other genera, using multiple sequence types and with experimental designs that include sampling a diversity of taxa whose phylogenetic relationships are clear, as exemplified in the present study.

**Concluding Remarks**

The present study demonstrates that genome size is a dynamically changing feature of plant genomes, even among recently diverged taxa within a single genus. As expected, some but not all genomes have a “one-way ticket to genomic obesity” (Bennetzen and Kellogg 1997), such as *G. exiguum*, whose very low deletion rate and highly proliferative native *gypsy* indicate that it “feasts” without purging over long evolutionary timeframes. On the opposite extreme, *G. raimondii* tolerates little TE proliferation, and, like the small-genome plants *Arabidopsis* and rice, which exhibit high rates of DNA removal, seems to be on the fast track to a more streamlined genome. The whys and wherefores of these variations remain mysterious however; Why is genome expansion easily tolerated in some lineages and not others? Why are some TE families
more successful than others, even among closely related taxa? What internal factors and external forces induce or prevent TE proliferation? Further comparative studies may help elucidate the particular species-specific attributes that allow surplus transposition leading to genome expansion. Similarly, a phylogenetic perspective is not only desired, but required to establish the specific effects of deletional mechanisms on shaping plant genomes.

**Acknowledgements**

The authors would like to thank Jordan Swanson for computational help, Mark Ungerer for providing primer sequences, Ashley Davidson for generation of some sequence data, and Corrinne Grover for providing molecular clock data. This work was funded by the National Science Foundation.

**Literature Cited**


Table 1. Estimated copy number and total number of Mb of Gorge3 from various time points.

<table>
<thead>
<tr>
<th></th>
<th>G. kirkii Outgroup</th>
<th>G. raimondii D-genome</th>
<th>G. herbaceum A-genome</th>
<th>G. exiguum Kgenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lineage-specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;90% seq ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>copies</td>
<td>500 ± 1000 ‡</td>
<td>789 ± 1114</td>
<td>6818 ± 3515</td>
<td>12089 ± 4827</td>
</tr>
<tr>
<td>Mb</td>
<td>*</td>
<td>*</td>
<td>68 ± 35</td>
<td>111 ± 46</td>
</tr>
<tr>
<td>Gossypium specific</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;80&lt;90% seq ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>copies</td>
<td>789 ± 1114</td>
<td>16363 ± 5434</td>
<td>27563 ± 7271</td>
<td></td>
</tr>
<tr>
<td>Mb</td>
<td>*</td>
<td>163 ± 54</td>
<td>290 ± 74</td>
<td></td>
</tr>
<tr>
<td>Pre-Gossypium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;80% seq ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>copies</td>
<td>3001 ± 2445</td>
<td>4731 ± 2725</td>
<td>22272 ± 6331</td>
<td>43037 ± 9063</td>
</tr>
<tr>
<td>Mb</td>
<td>30 ± 24</td>
<td>47 ± 27</td>
<td>223 ± 64</td>
<td>430 ± 90</td>
</tr>
</tbody>
</table>

‡ Lineage-specific estimate for G. kirkii includes all sequences with an average of greater than 80% sequence identity instead of 90%. * Unable to estimate number of Mb with this data.
Figure 1. Lineage-specific nature and timing of Gorge3 transposition in Gossypium. A) Neighbor-joining analyses for PCR amplified Gorge3 sequences are presented, with lineage-specific sequences in color and sequences originating before diversification in black. B) The curves represent the distribution of pairwise comparisons among lineage-specific sequences for each genome. The bottom axis represents the percent divergence, the top axis is the estimated transposition time, and the y-axis is the density of pairwise comparisons at a given time point.
Figure 2. Phylogenetic relationships and estimated rates of Gorge3 gain and loss among diploid members of Gossypium. Branch lengths are to scale. Numbers above the branches represent most likely rates of gain to loss ratios of Gorge3 DNA. Taxa are shown at tips with entire genome size as well as the amount (in Mb) of DNA from Gorge3 elements.
Figure 4, supplemental. Density distributions of the parametric solution space possible for the model based on empirical sampling error. Values for each element of the system of linear equations were sampled at random from the 95% confidence intervals and the QR decomposition was calculated for 100,000 replicates. Each curve represents the range and probability of a given solution to the system of linear equations. Blue curve: *G. raimondii*, orange curve: *G. kirkii*, green curve: *G. herbaceum*, purple curve: *G. exiguum*. The vertical red line denotes the ratio of 1, where gain/loss is equal.
CHAPTER FIVE

Phylogenetic determination of the pace of transposable element proliferation in plants: \textit{copia} and \textit{LINE}-like elements in \textit{Gossypium}

A paper prepared for submission to the journal \textit{Genome}

Jennifer S Hawkins\(^1\), Ryan A Rapp\(^2\), Guanjing Hu\(^3\), Jessie Grafenberg\(^4\), and Jonathan F Wendel\(^5\)

Abstract

Transposable elements contribute significantly to plant genome evolution in myriad ways ranging from local mutations via novel insertions to the global effects exerted on genome size through accumulation. With respect to the latter, it is now well known that differential accumulation and deletion of transposable elements can have a profound effect on genome size among members of the same genus and over short evolutionary time scales. One example is that of \textit{Gossypium} (cotton) where genome size ranges 3-fold among diploid members that diverged 5-7 mya. Much of the genome size variation in \textit{Gossypium} is due to differential accumulation of one particular \textit{gypsy}-like LTR-retrotransposon, \textit{Gorge3}. \textit{Gorge3} has been shown to be recently active in all genomes, regardless of genome size, and to proliferate in an episodic manner, followed by extended periods of quiescence. \textit{Copia} and non-LTR \textit{LINE} retrotransposons are also major components of the \textit{Gossypium} genome, and, unlike \textit{Gorge3}, their extant copy

\________\________\________\________\________
\(^1\) Graduate student, primary researcher and author, EEOB department, Iowa State University, Ames, Iowa.
\(^2\) Graduate student, statistical analysis, EEOB department, Iowa State University, Ames, Iowa.
\(^3\) Graduate student, collection and analysis of copia data, EEOB department, Iowa State University, Ames, Iowa.
\(^4\) Undergraduate student, collection and some analysis of LINE data, EEOB department, Iowa State University, Ames, Iowa.
\(^5\) Principal investigator and corresponding author, EEOB department, Iowa State University, Ames, Iowa.
numbers do not correlate with genome size. In the present study, we describe the nature and timing of transposition for *copia* and LINE retrotransposons in *Gossypium*. Our findings indicate that, similar to Gorge3, *copia* retrotransposons have been active in each lineage since divergence from a common ancestor, and that they proliferate in punctuated manner. However, the evolutionary history of LINEs contrasts markedly with that of the LTR-retrotransposons. Although LINEs have also been active in each lineage, they accumulate in a stochastically regular manner, and phylogenetic analysis suggests that extant LINE populations in *Gossypium* are dominated by ancient insertions. Interestingly, the magnitude of transpositional bursts in each lineage corresponds directly with extant estimated copy number.

**Introduction**

Transposable elements (TE) are extraordinarily diverse and prominent components of plant genomes. Originally thought to be "junk" or "selfish" DNA (Doolittle and Sapienza 1980; Orgel and Crick 1980), TEs are now recognized to play an important role in genome evolution via disruptive insertions and TE-induced gene duplication (Brunner et al. 2005; Hoen et al. 2006; Jiang et al. 2004; Lai et al. 2005; Morgante et al. 2005; Wang et al. 2006; Zabala and Vodkin 2005; Zhang et al. 2005), effects on gene expression, transposon-mediated chromosomal rearrangements (Gray 2000; Zhang and Peterson 1999), and their pronounced effects on genome size, (Hawkins et al. 2006; Hill et al. 2005; Holligan et al. 2006; Neumann et al. 2006; Piegu et al. 2006; SanMiguel and Bennetzen 1998; SanMiguel et al. 1996). TEs, particularly LTR-retrotransposons, may massively amplify and achieve extraordinarily high copy number within plant genomes over very short periods of time. Over the past decade, it has become evident that most genome size variation in plants is due to the dynamic activity of LTR-retrotransposons, in terms of both differential rates of transposition (Hawkins et al. 2006; Piegu et al. 2006; SanMiguel and Bennetzen 1998) and variation in deletion rates through illegitimate recombination (Devos et al. 2002; Kirik et al. 2000; Ma et al. 2004) and intra-strand homologous recombination (Devos et al. 2002; Shirasu et al. 2000). It is now evident that transposable elements are significant contributors to plant
gene and genome evolution, with respect to local mutagenic effects, genome architecture, and total nuclear DNA content.

TEs are divided into two major classes based on their genetic structure and method of transposition (Capy et al. 1997; Finnegan 1989). Class II DNA elements consist of the superfamilies \textit{hAT}, \textit{En/Spm} and Mutator, the non-autonomous MITEs, and the recently described helitrons. These TEs transpose via a conservative cut and paste mechanism, and accordingly are often found in lower copy numbers in plant genomes. Class I elements replicate via a duplicative mechanism, in which an RNA intermediate formed from the parental copy is reverse transcribed, and the newly translated copies are inserted into supplementary positions in the genome. These retrotransposons are divided into two major types based on the presence or absence of long terminal repeats (LTRs). Non-LTR retroelements consist of the autonomous Long Interspersed Nuclear Elements (LINEs) and non-autonomous Small Interspersed Nuclear Elements (SINEs). The LTR-retrotransposons are sub-divided into 2 classes, \textit{gypsy} and \textit{copia}, with respect to the placement of the integrase domain within the \textit{pol} gene. LTR-retrotransposons are by far the most abundant TE type found in plants, comprising large portions of all genomes investigated to date (Hawkins et al. 2006; Holligan et al. 2006; Meyers et al. 2001; Neumann et al. 2006; Piegu et al. 2006; Vitte and Bennetzen 2006).

Consistent with the aforementioned studies, data from cotton (\textit{Gossypium}) indicates that LTR-retrotransposons comprise a major fraction of all diploid genome species, regardless of total genome size (Hawkins et al. 2006). Analysis of genomic survey sequences (GSS) from three \textit{Gossypium} species, whose genome sizes range approximately 3-fold (Hendrix and Stewart 2005), suggests that most genome size variation among diploid members of the genus is due to differential accumulation of LTR-retrotransposons and that different sequence types have proliferated in different genomes. Further analysis of the highly abundant \textit{gypsy}-like retrotransposon family \textit{Gorge3} showed that recent lineage-specific amplification has occurred in each of the species studied, although to various magnitudes and at different time points in each of their evolutionary histories (Hawkins et al., 2007, submitted). \textit{Gorge3} proliferation was shown to occur in a punctuated manner in each of the genomes surrounded by periods of
quiescence. However, as this study evaluated only the Gorge3 family of LTR-retrotransposons, it is at present unknown whether the results are applicable globally or if they are particular to this specific sequence type.

To address this question, we investigate the evolutionary dynamics of the copia LTR-retrotransposon and non-LTR LINE retrotransposons in Gossypium. Through PCR amplification of a portion of the reverse transcriptase domain from three representative diploid Gossypium species that range 3-fold in genome size and one outgroup species, Gossypioides kirkii (Figure 1), we evaluate the pace and tempo of transposition for each sequence type in each lineage. Results indicate that these LTR-retrotransposons accumulate in a punctuated manner at different times in different lineages. In contrast, LINEs appear to experience a low level of stochastically regular amplification with a few instances of elevated activity. We further show that the magnitude of transpositional bursts corresponds with previously estimated copy numbers for each of the sequence types.

Methods

Plant materials, DNA extraction, and phylogenetic analysis

Total genomic DNA from the A-genome species G. herbaceum (JMS; 1C=1667Mb), the D-genome species G. raimondii (JFW; 1C=880Mb), the K-genome species G. exiguum (Gos 5184; 1C=2460Mb), and the phylogenetic outgroup Gossypioides kirkii (JMS stock; 1C=588Mb). DNA was extracted from young leaves using the Plant DNeasy mini kit (Qiagen Inc., Valencia, CA). Degenerated copia primers were designed by hand from aligned Gossypium copia GSS sequences, and subsequently tested to ensure amplification of the appropriate sequence type with an acceptable range in sequence diversity (Hawkins et al. 2006). LINE primers were those described in Alix and Heslop-Harrison (Alix and Heslop-Harrison 2004). Primers sequences are as follows: copia Reverse 5’-TGN TCC CAA ATC TTT NAT CTC -3’, copia Forward 5’-GCN ATG NAN GAN GAG ATG GA -3’, LINE Forward 5’-RVN RAN TTY CGN CCN ATH AG -3’, and LiNE Reverse 5’-GAC ARR GGR TCC CCC TGN CK-3’. The PCR reaction conditions were 94°C for 3 min, followed by 30 cycles of 94°C for 1 min,
copia 48°C/ LINE 45°C for 2 minutes, and 72°C for 1 min, ending with a final elongation step of 72°C for 6 min. PCR products were purified and sequenced as described (Hawkins et al. 2006). To confirm their identity, PCR sequences were queried using BLASTX to an Arabidopsis copia and LINE database provided by S. Wessler and F. Zhang (University of Georgia) and using TBLASTX to an in-house Gossypium database composed of previously identified GSS sequences. Confirmed sequences were aligned using MUSCLE (Edgar 2004) and were subsequently manually inspected. Neighbor-joining analyses were performed on the aligned sequences using PAUP* (Swofford 2003). Distances were set to the uncorrected (“p”) DNA/RNA distances and missing data was ignored for affected pairwise comparisons.

**Timing of lineage-specific transposition events**

TEs with 90% or greater similarity are inferred to be related by duplication (transposition) events that occurred subsequent to divergence of the A-genome and D-genome lineages of plants (Fig. 1) (Grover et al. 2007). Based on sequences from 48 nuclear genes, the genetic distance between Gossypium and the outgroup Gossypioides kirkii is approximately twice that of the A - D divergence (Senchina et al. 2003), which represents the basal-most split in the evolutionary history of Gossypium (estimated at 5 to 7 mya; Wendel and Cronn 2002, Cronn et al. 2002). Accordingly, TE lineages in Gossypium that share an average of ≥ 80% sequence identity were considered to be Gossypium-specific. All other, more divergent sequences are considered to have originated before the origin of the genus. Determination of "lineage-specific" and "Gossypium-specific" clades and estimation of their transpositional activity was performed in R with the base package, and with the phylogenetic package, ape, as described in Hawkins et al., 2007 (Hawkins et al. submitted). Briefly, nucleotide diversity at each node was calculated, and the daughter sequences of all nodes with <0.1 average pairwise divergence were extracted. Pairwise nucleotide diversity ($\pi$) was calculated among all daughter sequences of each lineage-specific node, and density functions were plotted for the frequency of nucleotide divergences.
Results

Copia evolution in Gossypium

PCR amplification and neighbor-joining analysis. A total of 563 copia reverse transcriptase sequences of approximately 600 bp in length from the A, D, K, and G. kirkii genomes were PCR amplified, aligned, and subjected to neighbor-joining analysis. The resulting phylogram (Figure 2a) contained distinct species-specific clusters of sequences for each genome in addition to several basal clades containing sequences from various, primarily Gossypium, genomes. There is high bootstrap support for the G. kirkii-specific clade (BS=83) and the large Gossypium-specific clade (BS=82). Additionally, one divergent group of D-genome copia sequences clustered distantly from the other RT sequences and shared 100% bootstrap support. Inspection of sequences belonging to this clade revealed a single synapomorphic amino acid deletion.

Pace and timing of transposition. To evaluate the lineage-specific transpositional nature and timing of copia retrotransposition in each genome, we employed a cotton-specific TE molecular clock estimated from orthologous transposable elements in Gossypium (Grover et al. 2007). We calculated π among all pairwise comparisons of “lineage-specific” PCR sequences and translated the sequence identity between copia paralogs into an estimated transposition date (Figure 3b). Only monophyletic lineages in which members share an average of 90% identity or greater across unambiguously aligned sequence at the node representing the most recent common ancestor (MRCA) were considered lineage-specific clades. Similar to our previous findings for the gypsy retrotransposon family Gorge3 (Figure 3a), copia retrotransposons proliferate in a mostly episodic fashion, that is, in temporally compressed periods at various points in the history of the plant lineages (Figure 3b). A burst of copia transposition occurred approximately 4 mya in G. kirkii and in the largest Gossypium genome (K), after which time there appears to have been little copia activity. In contrast, D-genome copias appear to have amplified recently (within the last 1 my), but also rather continuously over the last 4-5 my, as judged from the broad, flat peak extending from about 0.02 to 0.09 sequence diversity. Pairwise comparisons among A-genome lineage-specific clusters peak at approximately 2-3 mya, but as with the D-genome, there appears to have been a lower
level of *copia* activity since the origin of this lineage throughout most of its history. Interestingly, the transpositional timing for the D and A-genome *copias* is similar to that found for *Gorge3* (Figure 3a), although the magnitude of the transpositional bursts as measured by peak height (pairwise density) varies among sequence types.

**LINE evolution in Gossypium**

*PCR amplification and neighbor-joining analysis.* Degenerate primers amplified a 380 bp region of the LINE reverse transcriptase domain. A total of 299 LINE RT sequences from the A, D, K and *G. kirkii* genomes were aligned and subjected to neighbor-joining analysis (Figure 2b). In contrast with the *copia* and *Gorge3* phylogenies, LINE RTs cluster into what appears to be three major clades, each containing sequences from each of the four species studied. Two of the three clades contained bootstrap support of 100%, while the third large clade contained several internally supported groups. Branch lengths range from very long to small short terminal clusters, regardless of the genome from which they originated. Several small species-specific clusters for each genome are evident in all three major clades.

*Pace and timing of transposition.* The temporal profile of LINE transpositional events among lineage-specific sequences suggests, in contrast to that for LTR-retrotransposons, stochastically regular transposition in all genomes, with a few punctuated periods of proliferation since divergence from a common ancestor (Figure 3c). All genomes seem to have experienced recent transposition of LINEs, and, unlike that of the LTR-retrotransposons, the peak heights are mostly uniform among taxa, with the exception of some recent elevated activity in the A-genome.

**Discussion**

Most of the genomic components responsible for the extraordinary genome size variation in plants are now clear, and the once apparently contradictory "C-value paradox" has graduated to the now perplexing "C-value enigma" (Gregory 2002; Gregory 2004). The significant impact of TE proliferation on genome size growth has been thoroughly documented, particularly with respect to large-scale sequencing projects in
the grasses and a few model dicots. Recent, massive genomic bombardment by transposons has doubled the *Oryza australiensis* (Piegu et al. 2006) and maize (Meyers et al. 2001; SanMiguel and Bennetzen 1998; SanMiguel et al. 1996) genomes over only a few million years, and less than half of the TEs in the small genome of *O. sativa* originated before its origin, less than 680,000 years ago (Gao et al. 2004). Indeed, lineage-specific transposition has been observed in every plant genome investigated to date. However, relatively little information exists with respect to the evolutionary dynamics of retrotransposon proliferation among closely related genomes over short evolutionary time scales.

We show here that analysis of TE sequences within a phylogenetic framework yields novel temporal insights into patterns of transpositional activity for all major classes of retrotransposons. Combined with our previous analysis of the *gypsy* LTR-retrotransposon family *Gorge3* (Hawkins et al., 2007 submitted), we characterize the lineage-specific pace and timing of retrotransposon evolution among diploid members of *Gossypium*. These studies show that there has been recent, lineage-specific LTR-retrotransposon activity, and that retroelements have proliferated in a punctuated fashion in all species studied, but in an idiosyncratic, lineage-specific fashion. Genome-specific clusters with short branch lengths bespeak recent proliferation subsequent to divergence from a common ancestor. Interestingly, bursts of *copia* and *Gorge3* amplification within the A and D genomes appear to have occurred at similar but perhaps different time points during each species’ evolutionary history. For example, the A-genome experienced a burst of both *copia* and *Gorge3* activity approximately 2-3 mya. However, *copia* activity in the smallest *Gossypium* genome (D) occurred twice since its divergence from the remainder of the genus (Figure 1), once within the last million years and an older amplification approximately 4 mya, but only the recent (~1 my) amplification event is evident in the A-genome.

The pattern that emerges is one where there is transposition of different sequence types in each plant lineage, at different time points in each respective evolutionary history, and to varying levels of proliferation. The underlying causes behind punctuated proliferation are unclear, but it is commonly thought that bursts of transposition occur
due to some form of biotic or abiotic stress, as organisms capable of genetic
diversification under stress-conditions are more likely to survive and reproduce (Wessler
1996). Although transposition events in the outgroup, *G. kirkii*, and largest *Gossypium*
genome (K) were episodic in nature, correlated LTR-retrotransposition was not observed.
It is unclear why some species would experience correlated transposition of two similar
sequence types, while others do not.

The transpositional history of non-LTR LINEs contrasts remarkably with that of
the LTR-retrotransposons. LINEs appear to have experienced stochastically regular
accumulation in each lineage, with a few peaks representing points of increased
accumulation. No apparent periods of transpositional inactivity were evident. This
would suggest that LINEs propagate regularly over long evolutionary time periods in
*Gossypium*, albeit at low levels, consistent with earlier results that showed relatively low
copy numbers (Hawkins et al., 2006). In this respect, it is of interest to note that LINEs
are more highly expressed in the *Gossypium* EST libraries than would be expected based
on their estimated copy number (JS Hawkins, personal observation). One possible
explanation for this persistent, low-level proliferation is the tendency for LINEs to be
found in close association with genes, i.e., in transcriptionally active areas of the
genome.

An interesting observation is the correlation between apparent magnitude of the
transposition, as measured by peak height (density function), and estimated copy number
from the whole-genome shotgun sequencing surveys. *Gorge3* copy numbers were found
to be greatest in the two largest *Gossypium* genomes (A=48181 ± 9257, and K=88492 ±
12904), but very low and not significantly different from one another in the D-genome
(8674 ± 3683) and the outgroup *G. kirkii* (5502 ± 3305). Consistent with this
observation, the magnitude of lineage-specific transposition is greatest in the A- and K-
genomes, with little activity in the smaller genomes (Figure 3a). Similarly, *copia* copy
numbers were found to be highest in the largest and smallest *Gossypium* genomes
(D=57956 ± 9300, and K= 67700 ± 11324), slightly lower in the A genome (43181 ±
8774) and lowest in the outgroup (17006 ± 5765), while LINE copy numbers were
greatest in the A (30000 ± 7335) and K (27563 ± 7271) genomes, and slightly lower in
the D (13011 ± 4503) and G. kirkii (16006 ± 5597) genomes. Peak heights associated with the magnitudes of transpositional bursts correspond with these previously estimated copy numbers (Figure 3). This correlation was also reported for the rice species Oryza australiensis (Piegu et al. 2006).

Of particular relevance to the present work is the example of gypsy and LINE evolution in three diploid member of Vicia that vary in genome size (Hill et al. 2005). Combined with previous work describing copia retrotransposons in Vicia, the authors concluded that LTR-retrotransposons experienced recent proliferation, as demonstrated by low levels of sequence diversity among PCR-amplified paralogs, but that LINEs were highly heterogeneous, indicating that LINE populations were dominated by ancient insertions. Additionally, copia copy number in Vicia is not correlated with genome size (Pearce et al. 1996), similar to our results in cotton. These similarities between Gossypium and Vicia with respect to retrotransposon evolution are remarkable, and suggest that there may be a degree of generality to our conclusions. The mirror images in two separate, phylogenetically distant plant groups suggests that TE life-history traits, such as intragenomic location and insertional target site preference, may play important roles with respect to the evolutionary pace of transposition and ability to accumulate.

Comparative studies such as this provide the underlying framework for understanding why some TEs are more successful than others and why some genomes are more permissive of TE proliferation than others. Given the rapid and recent nature of TE proliferation observed in various angiosperm systems, comparisons among closely related, recently diverged taxa are likely to provide the most precise information with respect to global patterns of TE evolution.

Acknowledgments

We would like to thank R. Percifield for performing some data analysis, J. Swanson for help with computational analysis, and S. Wessler and F. Zhang for providing copia and LINE RT databases. This work was funded by the National Science Foundation.
Literature Cited


Figure 1. Genome sizes and evolutionary relationships among diploid members of *Gossypium* and a phylogenetic outgroup, *Gossypioides kirkii*. The indicated phylogenetic relationships are supported by several lines of molecular evidence (Seelanan et al. 1997; Small et al. 1998, 1999). *Gossypium* diverged from a common ancestor approximately 10-12 mya and experienced rapid radiation 5-7 mya (Wendel and Cronn 2002; Cronn et al. 2002). Genome sizes are those reported by Hendrix and Stewart 2005.
Figure 3. **Timing of lineage-specific retrotransposon proliferation in *Gossypium***.
Curves represent the distribution of pairwise comparisons among lineage-specific sequences for each genome. The bottom axis shows the percent sequence divergence between paralogs, the y-axis is the density of pairwise comparisons, and the top axis is the estimated transposition time.
CHAPTER SIX

General Conclusions

Current evidence from a wide range of diverse organisms indicates that genome size evolution is a highly dynamic process where total nuclear DNA content is often correlated with transposon proliferation. As reviewed in Chapter 1, recent, sometimes massive, transposon proliferation has been shown in every plant genome investigated and is often interpreted as evidence for recent genome size expansion. However, a paucity of information exists with respect to counteracting forces that attenuate genome expansion through TE proliferation, primarily because significant deletional events are less easily observed due to the need for comparisons between long orthologous tracts of DNA sequence. Therefore, our current understanding of the mechanisms, pace and directionality of genome size change is heavily limited to the effects of transposon proliferation among model organisms. Additionally, these comparisons are performed frequently among quite divergent taxa, thereby risking misinterpretation of overlapping mutational events. Most comparative studies among closely related taxa have focused on members of the Poaceae, biasing the available information to short-lived (annual) species. The goal of this dissertation work was to extend our understanding of genome size evolution to phylogenetically well-informed, closely related taxa outside the grasses, with respect to sequence types responsible for genome size variation, and to the pace, tempo and directionality of genome size change over a short evolutionary time scale.

In chapter 3, we described the various types of repetitive sequences present in Gossypium that are responsible for the wide range in genome size by constructing genomic survey sequence (GSS) libraries for three diploid species that range 3-fold in total nuclear DNA content, and by including a close phylogenetic outgroup, Gossypioides kirkii (Hawkins et al. 2006). By sample-sequencing the same proportion of each genome, we were able to statistically estimate the number of copies of various types of repetitive sequences present in each of the species studied. Our findings indicate that most of the genome size differences in Gossypium are due to variation in the copy number of one
particular gypsy-like LTR-retrotransposon, which we named \textit{Gorge3} (\textit{Gossypium} retrotransposable gypsy element). \textit{Gorge3} was found in relatively low copy number in both the smallest \textit{Gossypium} genome and in the outgroup species, but increased by 5-fold in the A genome, and doubled that of the A in the large K genome. However, \textit{Gorge3} was the only TE whose copy number was found to significantly correlate with genome size. \textit{Copia}-like LTR-retrotransposons were found to be most abundant in the smallest (D) and largest (K) genomes. Additionally, the density of \textit{copia}-like sequences was found to be highest in the smallest \textit{Gossypium} genome. All other gypsy-like sequences were present in equal copy numbers in all genomes. Non-LTR LINEs were present in slightly higher copy number in the two larger genomes, although these two estimates did not differ significantly from one another. These data provided evidence that, although the total fraction of each genome comprised of transposable elements was similar among the four species, different TEs behave idiosyncratically in different genomes, and that this behavior need not be correlated with genome size.

From the work described in chapter 3, we were compelled to more thoroughly investigate the evolutionary history of the gypsy-like LTR-retrotransposon, \textit{Gorge3}, whose copy numbers varied widely across the genus, and were found to be in direct correlation with genome size. In chapter 4, we used degenerate primers to PCR amplify a portion of the \textit{Gorge3} reverse transcriptase (RT) domain from the same individuals used for GSS library construction. Phylogenetic analysis of these RT sequences produced a topology of mostly species-specific clades, especially in the larger genomes, providing support for recent lineage-specific amplification of \textit{Gorge3} in each of the studied genomes. To estimate the approximate timing and nature (episodic vs. stochastic) of the lineage-specific transpositional events, pairwise comparisons among all paralogs belonging to clades that share an average of 90\% or greater sequence identity at the node representing their most recent common ancestor were plotted against a \textit{Gossypium}-specific molecular clock. These analyses revealed that \textit{Gorge3} amplified recently in each genome, although at different time points in each of the species, and in an episodic manner.
We wished to investigate *Gorge3* dynamics further by exploring the possibility of differential *Gorge3* removal rates among these four genomes. To this end, we combined our *Gorge3* PCR sequences with the previously generated GSS sequences to determine the approximate date of origin for each of the GSS sequences. The GSS sequences were divided into three categories of origination (lineage-specific, *Gossypium*-specific, or pre-origin of the genus), and the extant number of Mb remaining from each time point was estimated. By employing a unique modeling approach that estimates rates of gain and loss of *Gorge3* along each branch of the *Gossypium* phylogeny, we show that, although lineage-specific amplification did in fact occur in each lineage, the strength of *Gorge3* removal in the smaller genomes outweighs that of *Gorge3* gain, potentially leading to a decrease in genome size. Additionally, we found a surprisingly high amount of ancient, retained *Gorge3* in the larger genome species, suggesting genome downsizing through DNA loss plays a larger role in shaping genome size in *Gossypium* than previously anticipated.

In chapter 5, we extended the detailed analysis of the evolutionary dynamics of retrotransposons in *Gossypium* to the *copia* LTR-retrotransposons and the non-LTR LINEs. We performed the same phylogenetic analyses and estimation of the nature and timing of transposition among the same species for these additional two sequence types. Here we found that *copia* retrotransposons also proliferate in a punctuated manner and have amplified in each of the genomes since their divergence from a common ancestor. In two of the genomes, their proliferation was correlated with that of *Gorge3*, while in the other two genomes it was not. Also, a divergent and recently transposed group of D genome specific *copia* sequences was recovered in the phylogenetic analysis, providing an explanation for their higher density in the smallest *Gossypium* genome. LINE patterns of evolution in *Gossypium* were strikingly different than that of the LTR-retrotransposons. Phylogenetic analysis revealed three distinct groups of LINEs with high bootstrap support composed of primarily ancient sequence indicated by long branches. However, evidence for recent lineage-specific LINE accumulation was apparent by clusters of species-specific sequences with short branch lengths throughout the phylogeny. LINEs appear to accumulate in a more stochastically regular manner in
Gossypium, with a few instances of increased activity, but no apparent periods of quiescence.

In summary, this work provides evidence that genome size evolution in Gossypium is a dynamic process, molded by the net effects of DNA gain through transposon proliferation in addition to rapid rates of DNA loss in the taxa with smaller genomes. A key conclusion of this work is that transposon proliferation does not always result in genome size expansion. To our knowledge, this is the first study in plants to show that genome downsizing though DNA loss may be greater than that of gain mediated by transposon proliferation. Key to all of this work has been the unique application of phylogenetic methodologies to entire classes of retroelement evolution.

Our work in Gossypium presents some similarities to the findings in rice and maize, such as evidence for recent, lineage-specific amplification of transposable elements, particularly LTR-retrotransposons, in each of the taxa studied, in addition to the episodic nature through which these amplification events occur. However, some notable differences exist. One example is that of the ancient origin of most extant LTR-retrotransposons in Gossypium. Recent research in rice estimates a half-life of less than 6 my (Ma et al. 2004) for LTR-retrotransposons. Similarly, evidence from maize indicates that its genome size has doubled over the last 3 my through transposon proliferation alone (SanMiguel and Bennetzen 1998). Our results in Gossypium contrast in this respect with those of maize and rice, where the majority of the extant transposable elements appear to be of ancient origin. Perhaps this is due to the fact that Gossypium is a long-lived perennial, in which fixation of mutational event would takes much longer to occur relative to short-lived annuals such as these members of the Poaceae.

This work has provided insight into the evolutionary dynamics of transposable elements in Gossypium. However, many questions remain to be addressed. Why are some transposable elements more successful in some genomes than others, even when these genomes are very closely related? Why are some organisms more tolerant of transposable element accrual? Are different regions of the genome affected differently, and, if so, how does this contribute to overall genome size? Given the rapidity of transposable element proliferation and deletion demonstrated here and in other diverse
organismal systems, answers to these questions will be most easily facilitated through studies among closely related organisms within a well-defined phylogenetic context.

**Literature Cited**


ACKNOWLEDGMENTS

I wish to begin by expressing my sincerest gratitude to all of the members of the Wendel lab, both past and present, for their direct impact on my success over the past several years. First and foremost, I would like to thank Jonathan Wendel, who provided ongoing intellectual stimulation, undying support and an occasional kick in the pants, without which I would have never completed this work. A special thanks goes out to Ryan Rapp, a kindred spirit in many ways (do you know where your keys are?), and a wonderful collaborator. I would also like to express my appreciation toward Corrinne Grover, Lex Flagel, Ran Hovav, and Guanjing Hu for being wonderful friends and colleagues.

I would especially like to thank my husband, Ryan Percifield, for putting up with me during the more stressful times: I know it wasn’t easy. I would also like to mention my mother, Linda Potter; I can’t remember a time in my life that she wasn’t there for me, always sure I would succeed and always providing words of comfort and support. Thanks, Mom.