Evolutionary dynamics of mechanisms that affect genome size in the cotton genus (*Gossypium*)

by

Corrinne Elaine Grover

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Jonathan Wendel, Major Professor
Lynn Clark
Thomas Peterson
Randy Shoemaker
Daniel Voytas

Iowa State University

Ames, Iowa

2007

UMI Number: 3289384

UMI Microform 3289384

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

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 3
- Literature Cited 4

CHAPTER TWO. GENOME SIZE EVOLUTION 5
- Introduction 5
- Patterns of genome size variation in plants 5
- Mechanisms that cause genome size growth 6
- Mechanisms that cause genome size contraction 9
- Polyploids and the evolution of genome size 11
- *Gossypium* as a model for genome size evolution 12
- Literature Cited 13

CHAPTER THREE. INCONGRUENT PATTERNS OF LOCAL AND GLOBAL GENOME SIZE EVOLUTION IN COTTON 20
- Abstract 20
- Introduction 21
- Methods 23
- Results 26
- Discussion 36
- Conclusions 40
- Acknowledgments 40
- References 41

CHAPTER FOUR. MICROCOLINEARITY AND GENOME EVOLUTION IN THE ADHA REGION OF DIPLOID AND POLYPLOID COTTON (GOSSYPIUM) 52
- Summary 52
- Introduction 53
- Results 56
- Discussion 69
- Methods 74
- Acknowledgments 76
- References 76

CHAPTER FIVE. A PHYLOGENETIC ANALYSIS OF INDEL DYNAMICS IN THE COTTON GENUS 86
Abstract 86
Introduction 87
Methods 90
Results 93
Discussion 106
Concluding Remarks 112
Acknowledgments 113
Literature Cited 113

CHAPTER SIX. GENERAL CONCLUSIONS 118
Literature Cited 123

ACKNOWLEDGMENTS 124

LIST OF FIGURES

The evolutionary history of diploid and tetraploid *Gossypium*, as inferred by numerous chloroplast and nuclear data sets

Pairwise alignment of *CesA* homoeologous BACs, A_T and D_T, to scale.

Nested insertions of retroelements in the A_T BAC of *Gossypium hirsutum*

The spectrum of small indels inferred from sequence alignment of the A_T and D_T *CesA1* BACs.

The evolutionary history of diploid and tetraploid *Gossypium* species groups (n = 13 and 26, respectively), as inferred from multiple molecular datasets.

Multiple alignment of orthologous *AdhA* BACs from four different genomes (A, D, A_T, and D_T; the latter two are co-resident in the nucleus of polyploid cottons

Illegitimate recombination represents several different mechanisms leading to the deletion of a sequence bounded by small repeats (only 1bp of homology required), as well as one of the bounding repeats, or, less commonly, the addition of an intervening sequence

Possible splicing of the putative caffeic acid encoding genes in the *AdhA* region

Evolutionary history of and rates of genome loss and gain in four *Gossypium* genomes

LIST OF TABLES

Gene features predicted along homoeologous A_T and D_T BACs surrounding the *CesA1* gene in allopolyploid cotton

Spectrum of small indels in the comparison between A_T and D_T homoeologous BACs of *Gossypium hirsutum*

Gene features predicted in the *AdhA* regions of diploid and tetraploid cotton

Types and frequency of mechanisms contributing to genome size change in the *AdhA* region

Repetitive element lengths in diploid and polyploid cotton

Types and frequency of mechanisms contributing to genome size change in the AdhA region

Rates of insertions and deletions in the *AdhA* and *CesA* regions of the cotton genome

Insertions and deletions by region and mechanism

Rates of small insertions and deletions (<400nt) in the *AdhA* and *CesA* regions of the cotton genome

Average insertion and deletion rates and sizes for indels < 400nt

ABSTRACT

Eukaryotic genomes vary remarkably in size even between closely related species. This variation reflects a balance between mechanisms that expand and contract genomes, and which vary in their magnitude during evolution. While much is known about mechanisms that affect genome size expansion, particularly the effects of transposable elements (TEs), less is known concerning deletional mechanisms and the rates and scales at which they operate. The goal of this thesis was to extend our understanding of genome size evolution by studying diploid *Gossypium* species that vary twofold in genome size as well as their polyploid derivative, and using a phylogenetic approach employing as an outgroup *Gossypioides kirkii*. We assessed the rates and mechanisms operating in four *Gossypium* genomes: the two co-resident genomes of the allopolyploid *G. hirsutum* and its model diploid progenitors, *G. arboreum* and *G. raimondii*. Two BAC-sized regions of the cotton genome were sequenced and analyzed with respect to the mechanisms that alter genome size, and rates of sequence change (insertions, deletions, and net) were calculated for each region and genome. These regions were similar in that they both represent gene islands with extraordinary conservation of intergenic space; however, the regions did differ in terms of amount of genome size change. Whereas the first region showed no signs of the twofold genome size difference characterizing the species, the second region mirrored this difference, as the smaller genomes were represented by half the amount of sequence as the larger genomes. Notably, while still gene dense, this region had nearly half the gene density of the previous region. Analysis of the mechanisms responsible for shaping these regions led to several

conclusions. First, genome size change is attributable to many mechanisms, some of which are unknown. Second, while TEs had the greatest impact on genome size differences, other mechanisms, such as intra-strand homologous recombination, played key roles as well. Finally, genomes of diploid *Gossypium* species have experienced growth, whereas the polyploid has experienced contraction; however, the rates and direction of change vary between regions and over time.

CHAPTER ONE

GENERAL INTRODUCTION

Description of Research Objectives

Originally termed the “c-value paradox” (Thomas 1971), the disconnect between genome size and organism complexity has been noted for over half a century (Mirsky and Ris 1951). It soon became apparent that the observed differences in genome size reflected not polyploidy or polyteny, as once thought, but differences in the amount of non-coding DNA (Flavell et al. 1974). Extraordinary variation in eukaryotic genome size has since been observed, with some estimates placing the range in genome size to be over 200,000-fold (Gregory 2001). This remarkable variation is not limited to comparisons between widely divergent taxa; the variation among land plants alone exceeds 2300-fold and members within a single genus can vary over 60-fold (Bennett and Leitch 2005).

The resolution of the c-value paradox offered by the observed differences in non-coding DNA content changed the nature of the question from how an organism’s DNA content could change without regard to complexity to the causes and consequences of genome size variation. Renamed the “c-value enigma” to reflect the myriad questions surrounding genome size evolution and the multi-dimensional nature of these questions (Gregory 2005), research now seeks not only to describe the extent of genome variation among species, but also to address the internal and external factors that lead to the enormous range in genome size observed. As the non-coding fraction of many plant genomes is largely composed of transposable elements, whose amplification is not difficult to detect, it is not surprising that this mechanism of genome size change has been the most widely evaluated. The apparent inattention to mechanisms of genome size contraction is, in part, what led to the controversial idea that plants have a “one-way ticket to genomic obesity” (Bennetzen and Kellogg 1997). This proposition stimulated additional effort to

finding and evaluating mechanisms capable of contracting genomes; however, much of this effort has been limited to certain types of sequences (pseudogenes and transposable elements) or evolutionarily distant taxa, precluding a thorough evaluation of deletional mechanisms leading to genome size change. Moreover, of the research that addresses genome size evolution phylogenetically over a smaller timescale, none have addressed the rates of insertion and deletion attributable to specific mechanisms of genome size change and how those rates contribute to overall genome size change.

The purpose of my doctoral research is to evaluate, using a phylogenetically informed approach, the mechanisms that operate to expand and contract genomes and their relative rates and impact, by employing large insert sequencing in the cotton genus, *Gossypium*. Toward this end, I describe research that addresses the mechanisms of genome size change that have impacted two separate regions of five genomes, the relative impact each mechanism has had in shaping those regions and genomes, and the dynamics of genome size evolution as inferred from these regions. Specifically, the following questions are addressed:

1. What mechanisms contribute to genome size differences?
2. What is the relative impact each mechanism has had on shaping the current genomes?
3. At what rates have the genomes of *Gossypium* expanded and contracted due to specific mechanisms and how have these rates changed over time?
4. At what rates have the genomes of *Gossypium* expanded and contracted overall and how have these rates changed over time?
5. In the 5-10 million years since divergence, have the genomes of *Gossypium* expanded, contracted, or both?
6. Is genome size evolution an even, global phenomenon or is it influenced by genomically local dynamics?

Dissertation Organization

This dissertation is organized into six chapters. Chapter two, entitled “Genome size evolution” provides a review of current literature addressing the extent of genome size variation, the mechanisms that contribute to the observed variation, and the possible impact genome size has on organisms. The three chapters that follow represent original research concerning the direction of genome size change, the mechanisms responsible for genome size change, and the relative impacts of each mechanism on extant genome size in *Gossypium*. Chapter three, a research paper published in *Genome Research* entitled “Incongruent patterns of local and global genome size evolution in cotton”, describes a comparison of the genome size mechanisms operating in 100kb of sequence surrounding the gene encoding *cellulose synthase* (*CesA*) in the two co-resident genomes of the allotetraploid, *Gossypium hirsutum*, which differ approximately twofold in size. Chapter four, a research paper published in *The Plant Journal* entitled, “Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*)”, details a thorough evaluation and comparison of the mechanisms influencing genome size change in the region surrounding the gene encoding *alcohol dehydrogenase A* (*AdhA*) in four *Gossypium* genomes and provides insight into the biased accumulation of small deletions between differently sized genomes, as well as the change in frequency of small deletions upon polyploidization. Chapter five, a research paper entitled, “A phylogenetic analysis of indel dynamics in the cotton genus” and prepared for submission to *Molecular Biology and Evolution*, builds upon this work by adding sequence for the diploid progenitors (previously unrepresented in the *CesA* region), as well as the outgroup, *Gossypioides kirkii*, for both the *CesA* and *AdhA* regions, to provide a more in-depth analysis of the mechanisms responsible for genome size change and their relative impacts on these regions

(and, by extension, the genome). The last chapter summarizes the results of this dissertation and places them in a larger context.

Literature Cited

Bennett, M. D., and I. J. Leitch. 2005. Plant DNA C-values Database (release 4.0, October 2005). http://www.rbgkew.org.uk/cval/homepage.html.

Bennetzen, J. L., and E. A. Kellogg. 1997. Do plants have a one-way ticket to genomic obesity? The Plant Cell **9**:1509-1514.

Flavell, R. B., M. D. Bennett, J. B. Smith, and D. B. Smith. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochemical Genetics **12**:257-269.

Gregory, T. R. 2005. The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. Ann Bot **95**:133-146.

Gregory, T. R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. Camb. Philos. Soc. **76**:65-101.

Mirsky, A. E., and H. Ris. 1951. The DNA content of animal cells and its evolutionary significance. J. Gen. Physiol. **34**:451-462.

Thomas, C. A. 1971. The genetic organization of chromosomes. Annu. Rev. Genet. **5**:237-256.

CHAPTER TWO

GENOME SIZE EVOLUTION

Introduction

Although extraordinary variation in genome size marks eukaryotic genomes, over 200,000-fold different by some estimates (Gregory 2001), the range in number of protein coding genes varies only about 20-fold (Li 1997). Originally termed the "c-value paradox", this phenomenon was thought to reflect polyploidy or polyteny (Thomas 1971); however, subsequent research revealed that most the genome is comprised of non-coding, repetitive DNA (Flavell 1974) providing a partial resolution to this paradox. Retermed the "c-value enigma" to reflect the difficult, yet not unexplainable, nature of the problem (Gregory 2005), this curiosity has stimulated research on two fronts: (1) cataloging the extent of the variation and how it relates to organismal traits and (2) defining and understanding the mechanisms responsible for the exceptional ranges in genome size.

Patterns of Genome Size Variation in Plants

Although first described in animals (Mirsky & Ris 1951), the problem of genome size has seen the greatest gains in plants (Gregory 2005). Wide, targeted surveys across all of land plants have revealed some of the variation that exists among members (Bennett 2005). The variation among land plants as a whole exceeds 2300-fold, ranging from 54 Mbp in the pteridophyte *Selaginella caulescans* to 124,852 Mbp in the tetraploid angiosperm *Fritillaria assyriaca* (Bennett and Leitch 2005). Extensive variation marks individual groups as well, such as the pteridophytes (>1300-fold) and the angiosperms (>1270-fold). Genera themselves are not exempt from significant variation; of the genera with more than one genome size estimate, the average within-genus size variation is 3-fold and the upper bound on the range is more than 63-fold (Bennett and Leitch 2005). Adding further intrigue is the notion of intraspecific genome size variation, which may not be quite as

extensive as once believed due to measurement errors (Greilhuber 2005), but has been well-demonstrated in some species (Kalendar 2000, Baack et al 2005, Leong-Skornickova et al 2007).

The variation observed between organisms does not exist in a vacuum and can be influenced by myriad organismal and ecological traits that are subject to evolutionary constraints and pressures. Genome size has been linked to cell volume (Cavalier-Smith 1985, Cavalier-Smith 2005), seed mass (Beaulieu et al 2007), leaf size (Chung et al 1998, Wakamiya et al 1993, Ceccarelli et al 1993), annual/weedy lifestyle (Chase et al 2005, Bennett 1987, Albach et al 2004), extremity of environment (Knight et al 2005), time to maturity (Greilhuber and Obermayer 1997), cell-cycle duration (Bennett 1972), endangered status (Vinogradov 2003), drought tolerance (Castro-Jiminez et al 1989, Wakamiya et al 1993, Wakamiya et al 1996), frost tolerance (MacGillivray and Grime 1995), and altitude (Knight et al 2002), among others. Many of these correlations have been in conflict with regard to the strength and direction of the correlation, which may be the result of environmental influences on the molecular mechanisms that are responsible for creating genome size differences and not on genome size as a character.

Mechanisms that cause genome size growth

The mechanisms responsible for increasing DNA amounts are more studied, and thus better understood, than the mechanisms of contraction. One of the first explanations for the differences in genome size, polyploidy, is still often cited as a mechanism of genome size growth. As it represents the merger of two genomes, however, there has been some debate concerning the validity of including polyploidy as a mechanism of genome size growth and whether, in considering polyploid genome size, to consider the polyploid genome as a whole or the genomes that comprise it individually (Bennett and Leitch 2005, Gregory 2005, Greilhuber et al

2005). Regardless, given that polyploidy is fairly common in plants and polyploids can subsequently be returned to a diploid state (e.g. maize; Ilic et al 2003), it is clear that polyploidy is a mechanism by which genomes can grow substantially, albeit, as argued by some, extremely slowly (Gregory 2005).

A related phenomenon, though perhaps less common, that is also capable of increasing genome size is the large-scale duplication of part of a chromosome. Genetic maps of many plant species reveal large scale duplications (Bennetzen 2000), some of which are as large as whole chromosome arms. While some of these maps may represent ancient polyploidy (Wang et al 2005), still other evidence exists for large scale genomic duplications. Rice, for example, has experienced a large segmental duplication (3 Mb, involving chromosomes 11 and 12) independent of its ancient polyploid state (Wang et al 2005). Due to dosage concerns for the genes contained within duplicated blocks, this phenomenon is likely less common than polyploidy.

Aside from large-scale and whole-genome duplications, transposable elements (TEs) are considered the primary mechanism capable of increasing genome size. The contribution of TEs to genome size and structure was first described in the grasses (San Miguel et al 1996, SanMiguel and Bennetzen 1998, Chen et al 1998, Tikhonov et al 1999) and current evidence remains largely restricted to this family. From these examples, and other, are borne several concepts surrounding TE effects on genome size. First, TEs make at least 60-80% of the total DNA content of angiosperm species considered to have large genomes (Flavell 1974; SanMiguel and Bennetzen 1998, SanMiguel et al 1996, Meyers et al 2001, Wicker et al 2001, Shirasu et al 2000, Vicient et al 1999), thus underscoring their relationship to genome size. Second, TE proliferation can act to rapidly expand genome size, as exemplified by maize where TE insertions doubled the genome size in as little as

three million years. Third, TE proliferation of a single or few families may be responsible for that rapid genome growth and those families that prove successful need not be the same for all species (Piegu et al 2006, Hawkins et al 2006). These observations and similar others, have led some to conclude that genomes have a "one-way ticket to genomic obesity" fueled by transposable element activity from which there may be no return (Bennetzen and Kellogg 1997, Vitte and Bennetzen 2006).

In general, less is known about other, smaller scale mechanisms, in part due to their less dramatic effects on genome size. Repeated organelle to nuclear transfers have been demonstrated in many plants (Blanchard and Schmidt 1995, The Arabidopsis Genome Initiative 2000, Adams and Palmer 2003, Shahmuradov et al 2003), and the complete genomes of rice and *Arabidopsis* show that some organisms transfer far more organellar DNA to the nucleus than others (94kb versus 20kb in rice and *Arabidopsis*, respectively; Shahmuradov et al 2003). The overall impact of these transfers in different plant systems, while likely less than that made by TEs, is yet unknown. Similarly, increased intron size has also been thought to correspond to genome size; however, little evidence exists for this beyond broad phylogenetic comparisons (Deutsch and Long 1999, Vinogradov 1999, Bruggmann et al 2006). Expansion (as well as contraction) or tandemly repeated arrays (e.g. rDNA or satellite sequences) also has the potential to contribute to genome size; however, the evidence for this is extremely limited. Finally, duplication and subsequent pseudogenization of genes may, in principle, contribute to genome size expansion, although there is no evidence to suspect that plant species vary dramatically in their rates of duplication and pseudogenization.

Mechanisms that cause genome size contraction

The mechanisms capable of contracting genome size have proved more elusive, and thus are less-understood than are mechanisms responsible for genome size expansion. Whereas the effects of insertions of a known type (e.g. TE, organellar, etc) on genome size can be gauged by simply calculating the fraction of the genome attributable to that sequence type, the effects of deletional mechanisms can only be evaluated in comparison to non-deleted sequence. Rapid evolution also quickly becomes the enemy when attempting to detect the small footprints that are characteristic of deletional mechanisms, such as small (2-15 nt) illegitimate recombination associated direct repeats. The relative dearth of information regarding mechanisms of contraction led Bennetzen and Kellogg (1997) to posit the "one-way ticket to genomic obesity" notion, partly to bring attention to the then current data from the grass family suggesting an upward spiral of ever-increasing genome size and partly as a challenge to find mechanisms capable of shrinking genomes. If, as the placement of taxa with small genomes suggests (Wendel et al 2002, Bennett and Leitch 1997, Leitch et al 2005), some genomes are capable of overall contraction, then what are the underlying mechanisms responsible for this contraction?

Intra-strand homologous recombination was the first deletional mechanism in plants to gain popularity as playing a significant role in genome size reduction. Although the notion that an LTR-retrotransposon could undergo homologous recombination between its LTRs had been previously noted, Vicient et al (1999) was the first to report an abundance of solo-LTRs in a plant genome (barley). This "partial return ticket from genomic obesity" (Vicient et al 1999) stimulated others to evaluate the extent of intra-strand homologous recombination in various genomes and its potential impact on genome size. Subsequent research continued to evaluate the extent of intra-strand homologous recombination in various genomes: rice (Vitte and

Panaud 2003, Ma et al 2004), *Arabidopsis* (Devos et al 2002), barley (Shirasu et al 2000), maize (Meyers et al 2001), wheat (Wicker et al 2001) and others. Comparative analyses across genomes of species varying in relatedness both herald and question the potential for intrastrand homologous recombination to be significant in genome size reduction, especially when compared to the dramatic effects of the TEs themselves (Bennetzen 2002, Devos et al 2002, Ma et al 2004, Vitte and Bennetzen 2006, Piegu et al 2006, and others).

Illegitimate recombination (i.e. *RecA* independent recombination involving regions of microhomologous) has also been posited, as well as questioned, as a potential mechanism capable of contracting genome size. Devos et al (2002) suggested that illegitimate recombination may be the primary mechanism capable of counteracting genome size expansion in *Arabidopsis*. This trend was not upheld for rice (Ma et al 2004); however, illegitimate recombination has been demonstrated to remove large blocks of DNA in wheat (Chantret et al 2005) and considered an active force in genomic reshaping in this genome (Wicker 2003, Gu 2006). Further, one of the two mechanisms attributed to this overarching category, non-homologous end-joining (NHEJ; the other being slipstrand mispairing), has been demonstrated empirically to influence genome size using the contrasting genomes of *Arabidopsis thaliana* and *Nicotiana tabacum*, a 40-fold range in size (Kirik et al 2000, Orel and Puchta 2003). These studies demonstrated that the repair of double-stranded breaks via NHEJ in *A. thaliana* resulted in deletions that were larger and more frequent than occurred in the larger genome of *N. tabacum*. A more recent, if broad, comparison raises questions about the potential of illegitimate recombination to affect genome size in the long term and in the face of imminent TE proliferation (Vitte and Bennetzen 2006).

Biased illegitimate recombination, such as observed between *Arabidopsis* and tobacco, may in part be responsible for the observed bias in small indel formation that led to the "mutational equilibrium model" of genome size evolution (Petrov 2002). This theory posits that genome size expands or contracts until DNA loss through small deletions is offset by DNA gain through larger insertions. Once this equilibrium is achieved, the model suggests that organisms with smaller genomes experience more frequent and larger deletions than those with larger genomes. The evidence for this model is limited, arising mostly from animal data (Petrov and Hartl 1999, Petrov 2000), and some have questioned the fundamental suppositions that this model was based upon (Gregory 2003, Gregory 2004); thus, further evaluation is required before weighing in on the impact it has on genome size.

Polyploids and the evolution of genome size

As mentioned previously, the manner with which polyploid genomes are treated in regards to size (one genome or multiple, co-resident genomes) is a point of some contention; however, the argument concerning how polyploidy affects the mechanisms that effect genome size change is conspicuously lacking. The effects of polyploidization on the genome have been demonstrated to be many and diverse (reviewed in Adams and Wendel 2005, Chen and Ni 2006), some of which may include changes in the mechanisms of genome size evolution. A striking trend across polyploid species is the non-additivity of neopolyploid genomes with respect to their diploid progenitors (Leitch and Bennett 2004), which may reflect immediate changes in the mechanisms controlling genome size, not unlike or perhaps related to, the immediate epigenetic changes linked to polyploidization (Adams and Wendel 2005). This phenomenon of non-additivity, dubbed polyploid genome down-sizing (Leitch and Bennett 2004), has been poorly evaluated across species. Evidence in *Nicotiana* polyploids suggests that the genome polyploid tobacco is volatile and subject to high sequence turnover, even 5 my post-polyploidization, indicating the

potential for accelerated genome size change. Some evidence in wheat (Chantret et al 2005, Gu et al 2006) suggests illegitimate recombination plays a large role in polyploid genome down-sizing; however, evidence outside of wheat is limited, thus the generality of these studies cannot be established.

Gossypium as a model for genome size evolution

Gossypium is a young, monophyletic genus whose members range 3-fold in genome size (Cronn, Small et al. 2002; Wendel and Cronn 2003). The genus diverged from a common ancestor (*Gossypioides kirkii*, genome size 590Mbp) approximately 10 - 15 mya and member species began to diverge 5 – 10 mya, ultimately resulting in the recognized eight diploid and single polyploid genome groups. During these 5 - 10 my, the diploid genome groups acquired a nearly 3-fold range in genome size, from 885 Mbp in the New World D-genome species to 2572 Mbp the K-genome Australian species (Hendrix and Stewart 2005). Two of the diploid genome groups involved in that basal split, designated A-genome (1697 Mbp) and D-genome (885 Mbp), subsequently became reunited in a common nucleus approximately 1 – 2mya through allopolyploidization, leading to the evolution of modern polyploid cotton species, including *Gossypium hirsutum* (Wendel and Cronn 2003). The resulting genome size of the polyploid species was slightly less than the sum of the model diploid progenitors (reduced by approximately 180 Mbp; Hendrix and Stewart 2005). The wide range in genome size, well established outgroup, and robust phylogeny makes *Gossypium* an excellent system for genome size evolution studies. Furthermore, as a young dicot genus, *Gossypium* will provide much needed perspective on genome size evolution outside of the grass family and on a shorter time scale than previously investigated.

Literature cited

Adams, K. L., and J. D. Palmer. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution **29**:380-395.

Adams, K. L., and J. F. Wendel. 2005. Polyploidy and genome evolution. Curr Opin Plant Biol. **8**:135-141.

Albach, D. C., and J. Greilhuber. 2004. Genome Size Variation and Evolution in *Veronica*. Ann Bot **94**:897-911.

Baack, E. J., K. D. Whitney, and L. H. Rieseberg. 2005. Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. New Phytologist **167**:623-630.

Beaulieu, J. M., A. T. Moles, I. J. Leitch, M. D. Bennett, J. B. Dickie, and C. A. Knight. 2007. Correlated evolution of genome size and seed mass. New Phytologist **173**:422-437.

Bennett, M. D. 1972. Nuclear DNA content and minimum generation time in herbaceous plants. Proceedings of the Royal Society of London, Series B **181**:109-135.

Bennett, M. D. 1987. Variation in genomic form in plants and its ecologoical implications. New Phytol. **106**:177-200.

Bennett, M. D., and I. J. Leitch. 1997. Nuclear DNA amount in angiosperms. Phil. Trans. Royal Soc. London B **334**:309-345.

Bennett, M. D., and I. J. Leitch. 2005. Plant DNA C-values Database (release 4.0, October 2005). http://www.rbgkew.org.uk/cval/homepage.html.

Bennetzen, J. L. 2000. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. Plant Cell **12**:1021-1029.

Bennetzen, J. L., and E. A. Kellogg. 1997. Do plants have a one-way ticket to genomic obesity? The Plant Cell **9**:1509-1514.

Blanchard, J., and G. Schmidt. 1995. Pervasive migration of organellar DNA to the nucleus in plants. Journal of Molecular Evolution **41**:397-406.

Bruggmann, R., A. K. Bharti, H. Gundlach, J. Lai, S. Young, A. C. Pontaroli, F. Wei, G. Haberer, G. Fuks, C. Du, C. Raymond, M. C. Estep, R. Liu, J. L. Bennetzen, A. P. Chan, P. D. Rabinowicz, J. Quackenbush, W. B. Barbazuk, R. A. Wing, B. Birren, C. Nusbaum, S. Rounsley, K. F. X. Mayer, and J. Messing. 2006. Uneven chromosome contraction and expansion in the maize genome. Genome Research:gr.5338906.

Castro-Jimenez, Y., R. Newton, H. J. Price, and R. Halliwell. 1989. Drought stress responses of *Microseris* species differing in nuclear DNA content. American Journal of Botany **76**:789-795.

Cavalier-Smith, T. 1985. The evolution of genome size. Pp. 523. John Wiley & Sons Ltd.

Cavalier-Smith, T. 2005. Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. Ann Bot **95**:147-175.

Ceccarelli, M., S. Minelli, M. Falcinelli, and P. G. Cionini. 1993. Genome size and plant development in hexaploid *Festuca arundinaceae*. Heredity **71**:555-560.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M.-F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. 2005. Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). Plant Cell **17**:1033-1045.

Chase, M. W., L. Hanson, V. A. Albert, W. M. Whitten, and N. H. Williams. 2005. Life History Evolution and Genome Size in Subtribe Oncidiinae (*Orchidaceae*). Ann Bot **95**:191-199.

Chen, M., P. SanMiguel, and J. L. Bennetzen. 1998. Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. Genetics **148**:435-443.

Chen, Z. J., and Z. Ni. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. Bioessays **28**:240-252.

Chung, J., J. H. Lee, K. Armumuganathan, G. L. Graef, and J. E. Specht. 1998. Relationships between nuclear DNA content and seed and leaf size in Soybean. Theor. and Appl. Gen. **96**:1064-1068.

Cronn, R. C., R. L. Small, T. Haselkorn, and J. F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium* : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany **89**:707-725.

Deutsch, M., and M. Long. 1999. Intron-exon structure of eukaryotic model organisms. Nuc. Acids Res. **27**:3219-3228.

Devos, K. M., J. Brown, and J. L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Research **12**:1075-1079.

Flavell, R. B., M. D. Bennett, J. B. Smith, and D. B. Smith. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochemical Genetics **12**:257-269.

Gregory, T. R. 2004. Insertion-deletion biases and the evolution of genome size Gene **324**:15-34.

Gregory, T. R. 2003. Is small indel bias a determinant of genome size? Trends in Genetics **19**:485-488.

Gregory, T. R. 2005. The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. Ann Bot **95**:133-146.

Gregory, T. R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. Camb. Philos. Soc. **76**:65-101.

Greilhuber, J. 2005. Intraspecific Variation in Genome Size in Angiosperms: Identifying its Existence. Ann Bot **95**:91-98.

Greilhuber J, Dolezel J, Lysak MA, Bennett MD. 2005. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot 95:255–260.

Greilhuber, J., and R. Obermayer. 1997. Genome size and maturity group in *Glycine max* (soybean). Heredity **78**:547-551.

Gu, Y. Q., J. Salse, D. Coleman-Derr, A. Dupin, C. Crossman, G. R. Lazo, N. Huo, H. Belcram, C. Ravel, G. Charmet, M. Charles, O. D. Anderson, and B. Chalhoub. 2006. Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes. Genetics **174**:1493-1504.

Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing, and J. F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. %R 10.1101/gr.5282906:gr.5282906.

Hendrix, B., and J. M. Stewart. 2005. Estimation of the Nuclear DNA Content of Gossypium Species. Annals of Botany 95:789-797.

Ilic K., SanMiguel P.J., Bennetzen J. L. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci USA 100:12265–12270.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo, and A. Schulman. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceeding of the National Academy of Science **97**:6603-6607.

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

Knight, C. A., and D. D. Ackerly. 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecology Letters **5**:66-76.

Knight, C. A., N. A. Molinari, and D. A. Petrov. 2005. The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype. Annals of Botany **95**:177-190.

LEITCH, I. J., D. E. SOLTIS, P. S. SOLTIS, and M. D. BENNETT. 2005. Evolution of DNA Amounts Across Land Plants (Embryophyta). Ann Bot **95**:207-217.

Leong-Skornickova, J., O. Sida, V. Jarolimova, M. Sabu, T. Fer, P. Travnicek, and J. Suda. 2007. Chromosome Numbers and Genome Size Variation in Indian Species of Curcuma (Zingiberaceae). Annals of Botany **100**:505-526.

Li, W.-H. 1997. Molecular Evolution. Sinauer Associates, Inc., Sunderland, MA.

MacGillivray, C., and J. Grime. 1995. Genome size predicts frost resistance in British herbaceous plants: implications for rates of vegetation response to global warming. Functional Ecology **9**.

Meyers, B. C., S. V. Tingey, and M. Morgante. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11**:1660-1676.

Mirsky, A. E., and H. Ris. 1951. The DNA content of animal cells and its evolutionary significance. J. Gen. Physiol. **34**:451-462.

Orel, N., and H. Puchta. 2003. Differences in the processing of DNA ends in Arabidopsis thaliana and tobacco: possible implications for genome evolution. Plant Molecular Biology **51**:523-531.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, H. Kim, K. Collura, D. S. Brar, S. Jackson, R. A. Wing, and O. Panaud. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. %R 10.1101/gr.5290206 **16**:1262-1269.

SanMiguel, P., and J. L. Bennetzen. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany **82**:37-44.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P. S. Springer, K. J. Edwards, M. Lee, Z. Avramova, and J. L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science **274**:765-768.

Shahmuradov, I. A., Y. Y. Akbarova, V. V. Solovyev, and J. A. Aliyev. 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol. Biol. **52**:923-934.

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

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796-815.

Thomas, C. A. 1971. The genetic organisation of chromosomes. Annual Review of Genetics **5**:237-256.

Tikhonov, A., P. SanMiguel, Y. Nakajima, N. M. Gorenstein, J. L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. USA **96**:7409-7414.

Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav, N. E., and A. H. Schulman. 1999. Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. The Plant Cell **11**:1769-1784.

Vinogradov, A. E. 2003. Selfish DNA is maladaptive: evidence from the plant Red List. Trends Genet **19**:609-614.

Vinogradov, A. E. 1999. Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol. **49**:376-384.

Vitte, C., and J. L. Bennetzen. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceeding of the National Academy of Science **103**:17638-17643.

Vitte, C., and O. Panaud. 2003. Formation of Solo-LTRs Through Unequal Homologous Recombination Counterbalances Amplifications of LTR Retrotransposons in Rice Oryza sativa L. Mol Biol Evol **20**:528-540.

Wakamiya, I., H. J. Price, M. Messina, and R. Newton. 1996. Pine genome size diversity and water relations. Physiologia Plantarum **96**:13-20.

Wang, X., X. Shi, B. Hao, S. Ge, and J. Luo. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytologist **165**:937-946.

Wendel, J. F., and R. C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Adv. Agron. **78**:139-186.

Wendel, J. F., R. C. Cronn, J. S. Johnston, and H. J. Price. 2002. Feast and famine in plant genomes. Genetica **115**:37-47.

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

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.-D. Liu, J. Dubcovsky, and B. Keller. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell **15**:1186-1197.

.

CHAPTER THREE

INCONGRUENT PATTERNS OF LOCAL AND GLOBAL GENOME SIZE EVOLUTION IN COTTON

A paper published in the journal *Genome Research*[1]

Corrinne E. Grover[2], HyeRan Kim[3], Rod A. Wing[4], Andrew H. Paterson[5], Jonathan F. Wendel[6]

Abstract

Genome sizes in plants vary over several orders of magnitude, reflecting a combination of differentially acting local and global forces such as biases in indel accumulation and transposable element proliferation or removal. To gain insight into the relative role of these and other forces, approximately 105 kb of contiguous sequence surrounding the cellulose synthase gene *CesA1* was compared for the two co-resident genomes (A_T and D_T) of the allopolyploid cotton species, *Gossypium hirsutum*. These two genomes differ approximately two-fold in size, having diverged from a common ancestor ~5-10 million years ago (mya) and been reunited in the same nucleus at the time of polyploid formation, ~1-2 mya. Gene content, order and spacing is largely conserved between the two genomes, although a few transposable elements and a single cpDNA fragment distinguish the two homoeologues. Sequence conservation is high in both intergenic and genic regions, with 14 conserved genes detected in both genomes yielding a density of 1 gene every 7.5 kb. In contrast to the two-fold overall difference in DNA content, no disparity in size was observed for this 105 kb region. 555 indels were detected that distinguish the two homoeologous BACs, approximately equally distributed between A_T and D_T in number and aggregate size. The data demonstrate that genome size

---

[1] Reprinted with permission of Genome Research, 2004, 14(8), 1474-1482.
[2] Graduate student, primary researcher and author, EEOB department, Iowa State University
[3] Graduate student, BAC sequencing and finishing, Arizona Genomics Institute, University of Arizona
[4] Director, BAC sequencing and finishing, Arizona Genomics Institute, University of Arizona
[5] Professor and Director, BAC selection and hybridizations, Plant Genome Mapping Laboratory, University of Georgia
[6] Principal investigator and corresponding author, EEOB department, Iowa State University

evolution at this phylogenetic scale is not primarily caused by mechanisms that operate uniformly across different genomic regions and components; instead, the two-fold overall difference in DNA content must reflect locally operating forces between gene islands or in largely gene-free regions.

Introduction

The lack of correlation between genome size and organism complexity, known as the “C-value paradox” (Thomas 1971) or “G-value/N-value paradox” (Bertran and Long 2002; Claverie 2000), has been recognized for over half a century (Mirsky and Ris 1951). Genome size in eukaryotes varies more than 200,000 fold, from approximately 2.8Mbp in *Encephalitozoon cuniculi* (Biderre et al. 1998) to greater than 690,000Mbp in the diatom *Navicola pelliculosa* (Cavalier-Smith 1985; Li and Graur 1991). Even within various eukaryotic groups, there are remarkable differences in genome size. Protozoans display a 5800-fold genome size variation, vertebrates a 330-fold variation, and angiosperms display a 2300-fold variation in genome size (Bennett and Leitch 2003; Cavalier-Smith 1985; Gregory 2001). Significant genome size variation has also been observed among closely related species; for instance, the plant genus *Crepsis* displays a 9-fold variation (Jones and Brown 1976), while another plant genus, *Vicia*, displays a 6-fold variation in genome size (Chooi 1971). Despite this impressive variation in genome size, the amount of variation in the numbers of protein coding genes is only about 20-fold (Li 1997).

Though it is generally agreed that the majority of genome size variation can be accounted for by differences in the amount of non-coding DNA, the relative importance of mechanisms that generate genome size variation are not well-understood. In plants, the most prominent forces involved in genomic expansion are acknowledged to be polyploidy (Wendel 2000) and transposable element (TE) amplification (Bennetzen 2002), complemented by smaller scale processes such as

increases in pseudogene number (Zhang 2003), intron size (Deutsch and Long 1999; Vinogradov 1999), and incorporation of organellar genome fragments into the nucleus (Adams and Palmer 2003; Shahmuradov et al. 2003). Taken alone, these forces would cause an upward spiral toward bloated genomes (Bennetzen and Kellogg 1997). This one-way ticket to obesity is contraindicated by the phylogenetic distribution of plants with smaller genomes (Bennett and Leitch 1995; Bennett and Leitch 1997; Leitch et al. 1998; Wendel et al. 2002b), as well as by the existence of many plants, such as *Arabidopsis* (Vision et al. 2000) and maize (Ilic et al. 2003) , that clearly have eliminated massive amounts of DNA following polyploidization. Less well understood are evolutionary mechanisms that reduce genome size. Global mechanisms, such as small indel (<400bp) mutational bias (Petrov 2002b) and species-specific differences in non-homologous end joining (Kirik et al. 2000; Orel and Puchta 2003), have the potential to stochastically and differentially contract genomes. Sequence-specific mechanisms, such as LTR recombination (Shirasu et al. 2000; Vitte and Panaud 2003), ectopic recombination (Bennetzen 2000b; Langley et al. 1988; Petrov et al. 2003) and illegitimate recombination (Devos et al. 2002; Ma et al. 2004), have been shown to be capable of removing larger segments of DNA. Superimposed on these internal molecular genetic mechanisms are external factors and selective forces that may mold genome size; cell size limitations and cell division rate selection, for example, may constrain genome size (Gregory 2002).

Some mechanisms of genome size evolution, such as polyploidy and global deletional biases, are expected to affect all genomic constituents approximately equally, whereas others, such as proliferation of transposable elements, are likely to be more heterogeneous in their impacts on various genomic regions. To evaluate these alternatives it may be informative to compare closely related species that differ dramatically in genome size. Here we demonstrate this approach using model species from the genus *Gossypium*. Despite its relatively young age (5-10 million

years old) (Cronn et al. 2002) and conserved complement of genes, DNA content varies more than three-fold within the genus, from 980 to 3425 Mbp per 1C nucleus (Wendel and Cronn 2003). Two diploid groups of species, designated A-genome and D-genome, diverged from a common ancestor about 5-10 mya and acquired genomes that differ approximately two-fold in size. Approximately 1-2 mya, these two genomes became reunited in a common nucleus through allopolyploidization, leading to the evolution of the modern polyploid cotton species, including *G. hirsutum*, the primary cotton of commerce. Backed by a well-studied phylogeny (Fig. 1), we have embarked on comparative BAC sequencing to illuminate the patterns and processes responsible for modern-day genome size differences. For our initial study, we compared 100kb+ of homoeologous sequence surrounding a cellulose synthase gene (*CesA1*) from the two genomes (designated A_T and D_T) that comprise the allotetraploid, *G. hirsutum,* and which differ overall in genome size by a factor of two (1C = 980Mbp and 1860Mbp for D_T and A_T respectively) (Endrizzi et al. 1985). Remarkably, sequence conservation between the A_T and D_T genomes is shown to be high, even in intergenic regions. No evidence of mechanisms that underlie the 2-fold genome size difference is observed within this genomic region, where even the more than 550 small indels detected are evenly divided among the two genomes. The results show that genome size evolution operates regionally rather than globally at this phylogenetic scale, perhaps largely between gene islands.

Methods

BAC library screening and BAC selection

A cotton (*Gossypium hirsutum* L.) BAC library (Tomkins et al. 2001) was screened for clones containing a gene encoding cellulose synthase *(CesA1)*, as previously reported (Tomkins et al. 2001). This gene was previously isolated and its sequenced determined from A- and D-genome diploids cottons as well as from both genomes of polyploid cotton (Senchina et al. 2003), which facilitated identification of

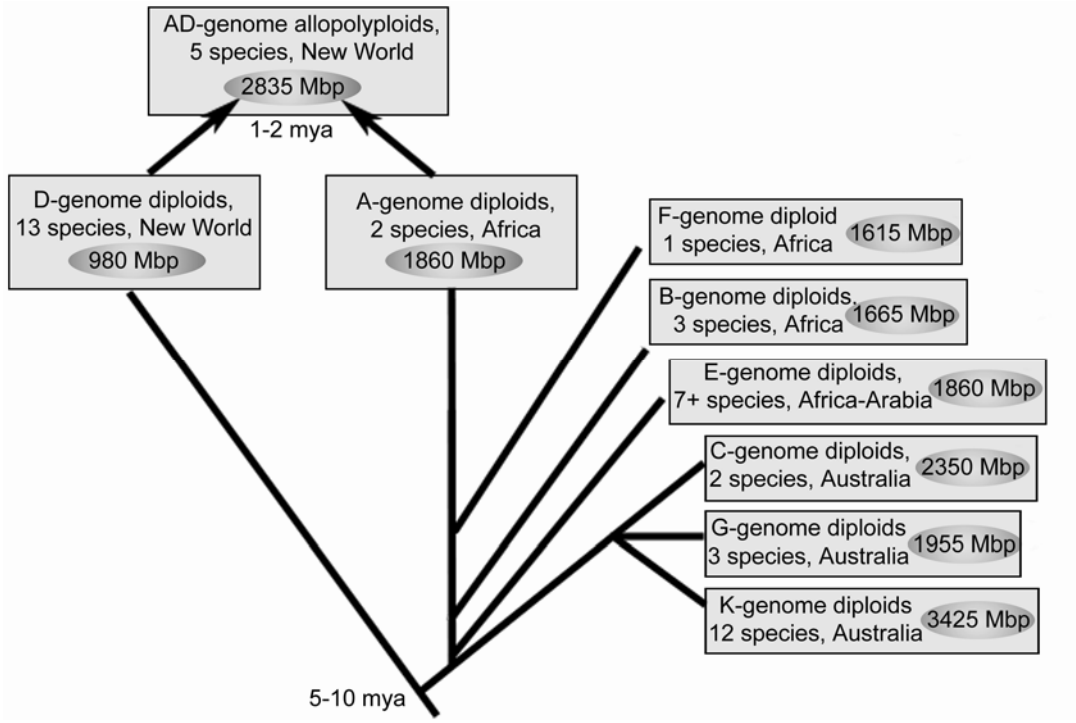


Figure 1. The evolutionary history of diploid and tetraploid *Gossypium*, as inferred by numerous chloroplast and nuclear datasets (Cronn et al. 2002; Seelanan et al. 1997; Small et al. 1998). Genome groups designate closely related species, as determined by interspecific meiotic pairing and chromosome size (Endrizzi et al. 1985). All diploid species have the same base chromosome number (n=13); however, each genome group varies in genome size (1C content indicated in circles). Polyploid species are thought to have originated 1-2 mya, following divergence of their diploid progenitors 5-10 mya.

the genomic origin of each BAC. PCR and sequencing were used to verify the presence of *CesA1* and to determine which homoeologue of the tetraploid (A_T or D_T) was represented by each BAC screened. The largest clone from the D_T genome (BAC clone=106I22) was sequenced to completion first. Following contig assembly, candidate A_T BACs for comparison were evaluated for maximal overlap with the sequenced D_T BAC, using a combination of PCR screening of inferred genes (3 and 11; see Fig. 2) as well as BAC-end sequencing. Because the *G. hirsutum* BAC library was created using partially digested (*HindIII*) genomic DNA, some BAC ends were conserved and shared among homoeologues. Thus, an A_T clone that shared a BAC end sequence and tested genes with 106I22 (A_T BAC clone=155C17) was

verified as providing maximum overlap for the region. This clone was then sequenced as described below.

Shotgun sequencing, assembly and analysis

BAC DNA was sheared using a HydroShear (GeneMachines) DNA shearing device at speed code 12 with 25 cycles at room temperature. Fragmented DNA was end-repaired using the 'End-it' DNA end repair kit (Epicentre), separated on an agarose gel, and size selected for a range of 2 – 6Kb. This prepared insert DNA was randomly cloned into a pBluescript II KS+ vector (Strategene) and sequenced with the universal vector primers T7 and T3 to an average depth of 8x (approximately 1152 clones in A_T and 1920 clones in D_T). The resulting sequences were base-called using the program Phred (Ewing and Green 1998; Ewing et al. 1998), vector sequences were removed by CROSS_MATCH (Ewing and Green 1998; Ewing et al. 1998), and assembled by the program Phrap (Green 1999). Contigs were visualized and edited with CONSED (Gordon et al. 1998). Potential genes were predicted by three independent programs: FGENESH (http://www.softberry.com/), GENEMARK.HMM (Lukashin and Borodovsky 1998), and GENSCAN+ (Burge and Karlin 1997). Predicted proteins were used as input for BLASTP searches against the non-redundant GENBANK protein database. To further investigate potential genes in the assembled sequence, 500bp segments of each assembled BAC were subjected to BLASTX queries against the non-redundant GENBANK protein database and BLASTN queries against the cotton EST database.

Alignment of the homoeologous BACs to each other was accomplished using LAGAN (Brudno et al. 2003). The resulting alignment was checked manually for errors using BIOEDIT (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Preliminary mining for repetitive elements was accomplished through RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html), CENSOR (Jurka et al. 1996), and BLAST homology to known elements in RepBase (version 8.5)(Jurka 2000). MITEs were mined using the program FINDMITE (Tu 2001) and querying the results for repetitiveness in the genome (Hawkins et al, unpublished), as well as by searching for conserved *Arabidopsis* TIR and TSD sequences. Each potential MITE was inspected manually to insure that the predicted TIRs were not composed primarily of simple sequence repeats that would generate a false prediction. In addition, each BAC was queried against itself in 500 bp fragments to reveal potentially missed repetitive elements. Finally, each BAC was again queried in 500 bp fragments against whole-genomic shotgun sequences characterized by an ongoing study (Hawkins et al, unpublished).

Results

General sequence comparison of the homoeologous BACs

The *CesA1* BACs from the A_T and D_T genomes were shotgun sequenced and assembled, giving a total of 2311 sequence reads and 4019 sequence reads, respectively. The overall gapped, aligned length of A_T with D_T is 123.8 kb. The ungapped aligned length of the A_T BAC is 103.9kb, and the ungapped aligned length of the D_T BAC is 107.9kb. Thus, for the *CesA1* region in *G. hirsutum,* there is only a 4kb difference in length between the A_T and D_T genomes. Both BACs are equal in GC content (33%GC). Database searches led to the inference of fourteen genes in the *CesA1* region, shared by both genomes. The total length of these genes was calculated to be 29.2kb, or about one third of the sequence. Excluding the 555 gapped positions (see below), which collectively exclude 36 kb and distinguish the two homoeologues, sequence identity over the aligned, ungapped positions was extraordinarily high (95%).

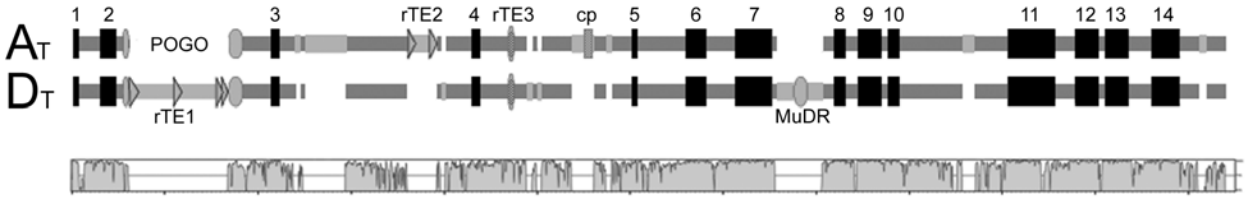


Figure 2. Pairwise alignment of *CesA1* homoeologous BACs, A_T and D_T, to scale. A_T and D_T are shown as block diagrams: numbered boxes are predicted genes corresponding to the list presented in Table 1; rTE1, rTE2, and rTE3 represent the 3 largely intact retrotransposons identified (rTE1 encompasses two predicted *copia* elements); the POGO and MuDR-like TEs are indicated individually, as is the *ycf2* fragment of plastidial origin. The lowermost panel indicates a continuous window of sequence identity between the two BACs, scaled from 50% - 100%.

Analysis of potential genes

Fourteen genes were predicted along the colinear segment (Fig. 2), giving an average density of 1 gene per 7.5kb of sequence. This is slightly less than the average *Arabidopsis* gene density of 1 gene per 4.5kb of sequence (The *Arabidopsis* Genome Initiative 2000) and similar to the average gene density in rice (The Rice Chromosome 10 Sequencing Consortium 2003). The *CesA1* region appears to be part of a gene island, as the gene density is fairly high and the non-genic DNA content low. The predicted genes (Table 1) range in size from a partial 244bp fragment of a putative ABC-transporter to 4.3kb in a predicted gene which is similar to an expressed *Arabidopsis* protein (gi:18396997). Silent and replacement site substitutions were calculated for each gene (Table 1). Synonymous substitution rates between homoeologous genes vary over a 10-fold range, from 0.008 to 0.084, with a weighted mean of 0.037; this value is identical to the weighted mean of 0.037 that was previously reported for a set of approximately 40 homoeologous genes in polyploid *Gossypium* (Senchina et al. 2003).

We searched a growing collection of cotton EST data sets for evidence of transcription of the predicted genes. To date, approximately 150,000 ESTs have been generated from various tissues and organs of diploid and polyploid cotton (Udall et al., unpublished). Searches of these data sets revealed evidence for expression of 5 of the 14 genes inferred to reside on the *CesA1* BACs. This, in

addition to the sequence divergence evidence and low levels of replacement substitutions (Table 1), lends support to the gene predictions.

Analysis of potential transposable elements

Differential insertions of transposable elements (TEs) are recognized as a prominent force in genome size expansion. Thus, we examined the *CesA1* BACs for evidence of transposable elements. A total of six largely intact TEs were detected in the two *G. hirsutum* homoeologues, two that are shared, one that is unique to A_T, and three that are unique to D_T. The two genomes also share a highly degraded *EnSpm* (class 2) remnant, identified by CENSOR (Jurka et al. 1996), as well as a potential and highly degraded retrotransposon, identified by BLAST homology to elements characterized in an ongoing study (Hawkins et al, unpublished). The A_T genome has a series of potential, highly degraded retrotransposons of indeterminate number, again identified by BLAST homology. Additionally, 9 potential miniature inverted-repeat TEs (MITEs) were predicted in the *CesA1* region, 7 shared between A_T and D_T and 2 that are unique to A_T. Overall, transposable elements (including remnants) account for 28.5kb of sequence in the region, 10.8kb in A_T and 17.7kb in D_T.

The two shared intact transposable elements belong to different classes. One of the shared transposable elements has similarity to known POGO elements from *Arabidopsis*. The putative POGO is flanked by 15 bp terminal inverted repeats (TIR), which have 73% identity (5' TIR versus 3' TIR) and which retain the typical TA dinucleotide target site duplication. Each *Gossypium* POGO element retains ~90% identity over the TIR to several *Arabidopsis* POGO elements (Feschotte and Mouches 2000) and 35% identity over the entire element. Compared with each other, the A_T POGO (1940bp) and the D_T POGO (2150bp) have 83% sequence identity, including gaps, and 92% sequence identity when gaps are excluded.

The other shared intact transposable element is a retrotransposon of unidentified type. This element was identified through its BLASTX identity to known reverse transcriptase (RT) sequences (40% identity and 65% similar over 100 amino acids to numerous RT sequences from *Arabidopsis*). There is evidence that this RT sequence may have been derived from a degraded non-LTR retroelement, as a few BLASTX hits were to non-LTR RT sequences and no vestige of ancient LTRs was identified.

The A_T and D_T genomes also share what appears to be a 45bp remnant of a highly degraded *EnSpm* transposon. This remnant was identified by CENSOR as having identity to the described *EnSpm* element ATENSPM5 from *Arabidopsis* (Jurka 2000). Sequence identity between A_T and D_T over the remnant is 100%, and the sequence identity between either *Gossypium* remnant and ATENSPM5 is 82%.

Finally, the A_T and D_T genomes also share a potential, highly degraded *gypsy* retrotransposon. The D_T element shows 164bp identity to *Gossypium gypsy* elements, whereas 204bp of identity was observed for the A_T element.

The A_T BAC sequence contains only one identified largely intact transposable element that is not shared with the D_T genome. This element is a predicted long terminal repeat (LTR) retrotransposon of unknown type. The element contains 612 bp LTRs, which retain 98% sequence identity with each other. The element is 3138 bp in length and contains homology to identified tomato (gi: 4235644) and *Arabidopsis* pol proteins of unspecified type.

The A_T BAC sequence contains two potential and extremely degraded retroelement clusters. The first retroelement cluster spans 7.5kb of sequence, though only 1573

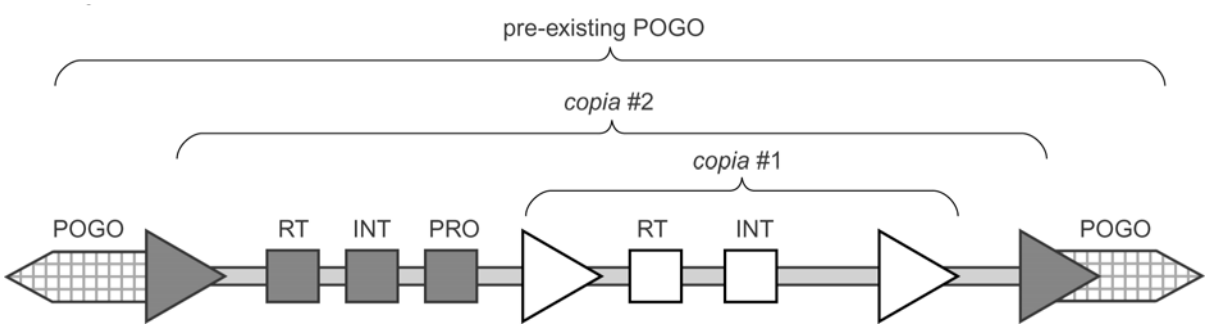
Table 1. Gene features predicted along homoeologous A_T and D_T BACs surrounding the *CesA1* gene in allopolyploid cotton.

		Length (bp) [b]			Total Length, Exons[b]			Total Length, Introns		Length (a.a.)		Divergence[c]		
Gene	Putative Function[a]	D_T	A_T	Exons	D_T	A_T	Introns	D_T	A_T	D_T	A_T	Ks	Ka	Ksil
1	ABC transporter[d]	244	244	*	244	244	*	0	0	81	81	0.035	0.000	0.035
2	GTP binding protein	1909	1925	2	657	657	1	1252	1268	218	218	0.054	0.006	0.047
3	WRKY TF	1024	1013	3	822	819	2	202	194	273	274	0.064	0.014	0.041
4	*Arabidopsis* hypothetical protein	994	992	3	471	471	2	523	521	156	156	0.020	0.017	0.061
5	no BLAST homology	519	519	1	519	519	0	0	0	172	172	0.084	0.026	0.084
6	G protein B	2376	2376	9	1068	1068	8	1308	1308	355	355	0.043	0.012	0.029
7	CesA	4080	4083	12	2925	2925	11	1155	1158	974	974	0.041	0.004	0.033
8	LeuRR	1308	1308	1	1308	1308	0	0	0	435	435	0.020	0.012	0.020
9	PRR/Se-binding protein	2628	2628	3	2418	2418	2	210	210	805	805	0.019	0.013	0.018
10	Ribosomal protein L11	1273	1275	4	519	519	3	754	756	172	172	0.008	0.000	0.029
11	LeuRR transmembrane or kinase	3189	3197	11	1857	1857	10	1332	1340	618	618	0.042	0.006	0.031
12	growth regulator	2631	2637	10	1797	1800	9	834	837	598	599	0.050	0.015	0.037
13	permease	2633	2629	13	1575	1575	12	1058	1054	524	524	0.016	0.006	0.021
14	*Arabidopsis* expressed protein	4349	4325	8	1287	1278	7	3062	3047	425	422	0.037	0.014	0.029
										Weighted Average		**0.037**	**0.010**	**0.032**

[a] Putative function is assigned by BLAST homology to genes in Genbank. The locations of genes along the BAC contigs are represented in Figure 2.
[b] Total length including stop codon.
[c] K_s, K_a, and K_{sil} denote rate of substitution across all sites, substitutions at nonsynonymous sites, and synonymous sites within codons plus all noncoding positions, respectively.
[d] This predicted gene is fragmented in the BAC; a start codon was identified, but no intronic sequences or stop codon was found. This gene would presumably full length, were it not fragmented in the generation of the BAC library.

bp can be identified as belonging to degraded TEs. This cluster may have contained 2-3 *gypsy* elements, one shared with the D_T BAC sequence (204 bp; mentioned above), and contains moderate sequence identity (60% - 70%) to previous reported A-genome specific repetitive sequences (Zhao et al. 1998). The second degraded retroelement cluster contains a potential *gypsy* remnant and a potential *copia* remnant. These two remnants are situated on either side of a cpDNA insertion (see below), and were likely genomic neighbors before being separated by the cpDNA insertion.

The D_T BAC contains three transposable elements (one DNA element and two *copia* retrotransposons) that are not shared with the A_T genome. The DNA element has



homology to several *Oryza mutator* (MuDR) elements, as well as some limited homology to the *Arabidopsis Vandal12* DNA element (Jurka et al. 1996). The element appears to be degraded, as the protein alignment generated by BLASTX shows only 26% identity (44% similarity) over 536 amino acids.

Figure 3. Nested insertions of retroelements in the A_T BAC of *Gossypium hirsutum*. The outer *copia* is shown in grey and the inner *copia* in black. Four LTRs, corresponding to the two *copia* insertions, are shown as triangles. The three coding domains of the *copias*, reverse transcriptase (RT), integrase (INT), and protease (PRO), are designated by the labeled boxes within the LTRs. Surrounding the *copia* nest is a single POGO element that is shared by A_T and D_T, and which was split in two when the *copias* inserted.

The two *copia* insertions that are D_T specific for this BAC are nested within the POGO insertion (Fig. 3). The outer *copia* has 200 bp LTRs that are 97% identical. The

element is 5.3 kb in length and has well-defined reverse transcriptase, integrase and protease coding domains. The inner *copia* has 561bp LTRs that are 99.7% identical. This element also is 5.3 kb in length and has well-defined reverse transcriptase and integrase coding domains. The protease coding domain for this element could not be identified. The inner *copia* inserted between the protease coding domain and LTR of the outer *copia,* after the outer one had inserted. These copia insertions share no identity with each other; thus, they probably belong to different families. Retrotransposon insertions can be dated based on LTR divergence (SanMiguel et al. 1998), although these estimates provide only approximations, given the unknown absolute rate of mutation. Previous data on sequence divergence in *Gossypium* (Senchina et al. 2003) can be used to infer the relative insertion times of each *copia*. The percent divergence between LTRs of the outer *copia* (3%) is similar to that estimated for divergence of A- and D-genome diploids, suggesting transposition shortly after the divergence of these two species groups. Similarly, LTR divergence of the inner *copia* (0.3%) is slightly less than that estimated for comparisons between model diploid progenitors and their counterparts in the polyploid, suggesting insertion of the internal *copia* subsequent to polyploidization.

Miniature inverted-repeat transposable elements (MITEs) are a common feature of gene rich regions (Feschotte et al. 2002). Although they are categorized as class 2 elements, these non-autonomous TEs do not encode a transposase or transposase remnant; thus, the prediction and classification of potential MITEs is primarily achieved through terminal inverted repeat (TIR) and target site duplication (TSD) identification (Feschotte et al. 2002). Considering this, two approaches were employed to predict MITEs in the A_T and D_T *CesA* BACs. The first approach, which attempted to predict MITEs from known families (*Stowaway, Tourist, etc.*) by searching for similarity to TIRs from known MITEs in *Arabidopsis*, *Brassica*, and the

grasses, did not reveal any known MITEs in the *CesA* BACs. The second approach employed a *de novo* search method (Tu 2001), which inspects the sequence for potential TIRs that also have a TSD. Although this method predicted many MITEs in both A_T and D_T, subsequent inspection revealed that a majority of the predicted TIRs and TSDs contained simple sequence repeats (SSRs). Predicted MITEs whose TIRs were comprised mostly of SSRs were considered probable artifacts. In total, 16 MITEs were predicted in the *CesA1* region, 7 shared and 2 unique to A_T, accounting for 2.5kb and 2kb in A_T and D_T, respectively.

Other potential mechanisms of genome size evolution

In addition to TE insertions, the *CesA1* BAC alignments were examined for other distinctions. Most prominent among these is a 900 bp fragment of the plastid gene *ycf2*, which was inserted into a non-coding region of the A_T genome (Fig. 2; 90% identity over 897bp to *ycf2* from *Arabidopsis*) and was flanked by 1.5kb of A_T-specific sequence of undetermined identity. While accounting for a mere 0.86% of the total aligned length of the homoeologous BACs, the 900bp *ycf2* fragment accounts for 5.6% of the A_T-specific sequence.

Intron sizes for each gene were compared for all inferred genes on the homoeologous BACs to evaluate their potential contribution to the genome size variation. Intron sizes deviated by an average of 4.3 bp per gene, with a range of 0-16 bp. The total contribution of intron size differences to the size difference of the region was a mere gain of 3 bp in A_T. This result provides a striking contrast to reports of intron sizes contributing to genome size differences over much longer evolutionary timescales (Deutsch and Long 1999; Vinogradov 1999). The present study concurs with previous data on *Gossypium* intron size variation, which suggested that there exists little intron size variation among *Gossypium* species, irrespective of genome size (Wendel et al. 2002a).

Evidence for a bias in small indel number and length was also examined for the homoeologous sequences (Fig. 4; Table 2). The frequency of small indels was computed for any gapped position smaller than 400 bp in length. A total of 555 small gaps were scored in the two BACs, approximately equally distributed between A_T and D_T in number and aggregate size. Of the 269 indels in A_T and 286 indels in D_T, 264 and 279 were classified as small indels, respectively. Moreover, small indels account for 2777bp of missing sequence in A_T and 2897bp of missing sequence in D_T, a difference of only 120bp. In addition to similarities in number and aggregate size, the frequency spectrum of small indels is similar in shape and position between the A_T and D_T BACs; that is, the number of indels of any length is similar between A_T and D_T (Fig. 4). Overall, small indels account for 14% and 18% of the total length in A_T and D_T, respectively, but fail to contribute significantly to the overall size difference in the aligned region.

Table 2. Spectrum of small indels[a] in the comparison between A_T and D_T homoeologous BACs of *Gossypium hirsutum*.

	A_T genome		D_T genome	
	# indels	bp	# indels	bp
1 - 10bp	210	593	207	489
11 - 20bp	25	369	34	506
21 - 30bp	10	260	17	414
31 - 40bp	8	274	5	239
41 - 50bp	1	49	8	346
51 – 100	5	353	5	351
101 – 200	4	665	2	321
200 - 400	1	214	1	231
Small indels	264	2777	279	2897
All indels	269	19977	286	16009

[a]Indels are binned in multiples of 10bp up to an indel length of 50bp; the last three bins span 50bp, 100bp, and 200bp, respectively, due to the infrequency of these larger indels in either genome. The last two rows tally totals for the number and amount of sequence accounted for by small indels (<400bp) and all indels, respectively.

One hallmark of illegitimate recombination is the presence of direct flanking repeats 2-15 bp in size (Ma et al. 2004). We searched all indels discovered here for flanking

repeats, restricting our attention to the 144bp indels that were at least 10bp in length (Ma et al. 2004). Of these, 55 (38%) showed flanking repeats of 2-15bp (excluding possible mono- or dinucleotide and microsatellite expansion/contraction events). These flanking repeats were unequally distributed in number between the AT and DT genomes (19 versus 36), but encompassed approximately the same amount of sequence (11,720 and 13,164, respectively). These data suggest that illegitimate recombination is a common mechanism of sequence evolution in Gossypium, and that it may play a role in genome size evolution. Additional analyses that include outgroups for phylogenetic polarization of indels will shed light on the extent and importance of this mechanism.

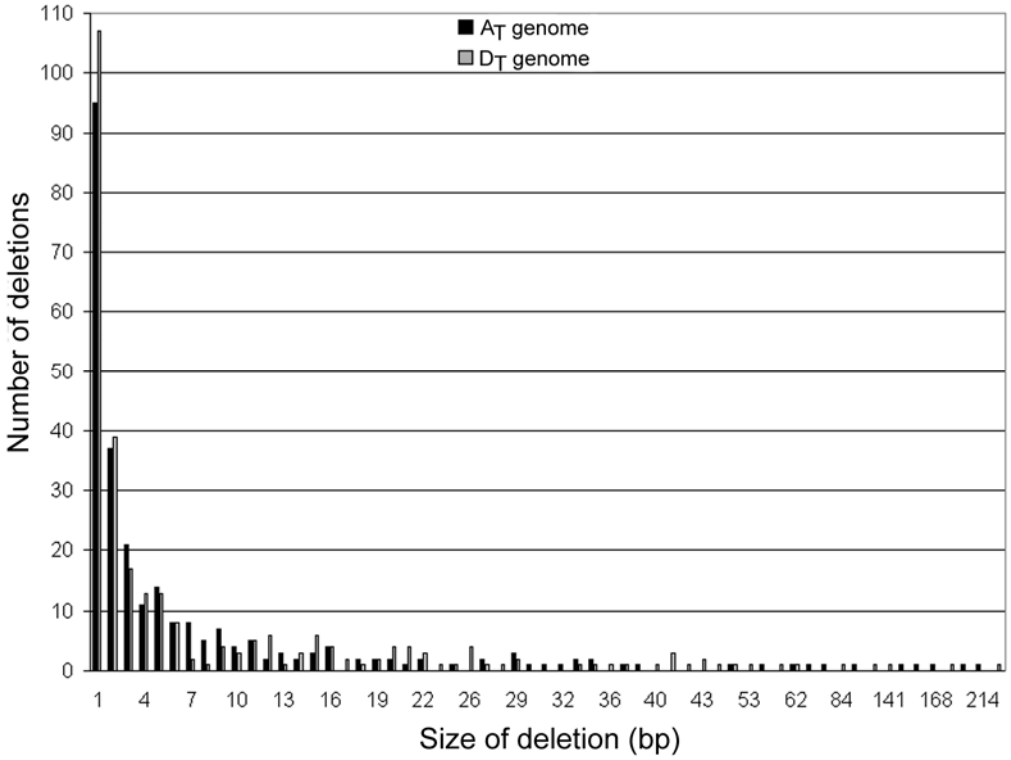


Figure 4. The spectrum of small indels inferred from sequence alignment of the A_T and D_T *CesA1* BACs. For A_T (solid bars), "differences" are gapped positions relative to D_T, whereas for D_T (open bars), differences reflect gaps relative to A_T. These indels are not phylogenetically polarized, although the spectrum of indels is equivalent in the two genomes.

Discussion

In recent years there has been a rapidly accumulating literature focused on comparative analyses of contiguous, homologous stretches of genomic sequence in plants. Stimulated by the seminal investigations of Bennetzen and colleagues on the maize, rice, and sorghum sh2/a1 and *Adh* regions (Chen et al. 1997; Chen et al. 1998; SanMiguel et al. 1996; Tikhonov et al. 1999) and the increasing accessibility of genomic tools, "microcolinearity" has been studied for numerous other genomic regions and taxa (Chantret et al. 2004; Dubcovsky et al. 2001; Fu and Dooner 2002; Ku et al. 2000; Ramakrishna et al. 2002a; Rossberg et al. 2001; Tarchini et al. 2000; van Leeuwen et al. 2003; Vandepoele et al. 2002; Wicker et al. 2001). Among the generalizations and insights that emerged from these analyses is the concept that gene order and content may be conserved over long periods of evolutionary time (Bennetzen 2000a; Gale and Devos 1998), that polyploidy may lead to a rapid decay in synteny and gene content preservation among homoeologues (Ilic et al. 2003; Kellis et al. 2004; Langham et al. 2004), and that intergenic regions may be subject to more dramatic and rapid evolutionary alterations. The latter in particular has led to the notion that much of the genome size evolution that takes place in plant genomes is due to differential accumulation of retroelements in intergenic regions (Bennetzen 2000b), although it also is evident from the draft *Oryza sativa* genome sequence (Goff et al. 2002; Yu et al. 2002) that retroelements may be concentrated near centromeres and other largely heterochromatic regions. Superimposed on these ideas has been the concept that genome size itself may have biological significance and be visible to natural selection (Bennett 1985; Bennett 1987; Gregory and Hebert 1999), applying directional pressure on all genomic constituents simultaneously, perhaps through molding genome-specific mutational processes that determine the frequency and spectrum of deletions or insertions (Kirik et al. 2000; Orel and Puchta 2003; Petrov 2002b). Based on the foregoing, we anticipated that the two-fold genome size difference that exists between A- and D-genome cotton

species might reflect similar phenomena of either differential intergenic retroelement accumulation or perhaps a more globally operating bias in the prevalence and size of insertions and deletions. Neither of these expectations was realized, however, and in addition, a remarkable degree of conservation of the entire *CesA1* region was observed, including the size and sequence of most intergenic regions.

Genome evolution in the CesA1 *region of polyploid cotton*

The most likely process responsible for the two-fold genome size difference between the A_T and D_T genomes is differential accumulation or retention of transposable elements, particularly retroelements. In the region studied here, however, relatively few TEs were detected and their differential presence does not correspond with the genome size difference; three of the four unique and intact TE insertions are found in the smaller (D_T) of the two genomes, accounting for 15.5kb in D_T versus 5.8kb in A_T. The presence of unique MITEs in A_T did little to counteract the disparity, only accounting for +500bp in A_T. Thus, while transposable element amplification may have contributed to the 2-fold genome size difference, this phenomenon is not evidenced in this genomic region.

The *CesA1* region was examined for evidence of ectopic recombination among retroelements. If ectopic recombination has played a role in shaping the *CesA1* region, then footprints of the recombined elements should be apparent, such as solo LTRs resulting from recombination between LTRs of individual retroelements or between LTRs of distinct but linked elements (Devos et al. 2002; Kalendar et al. 2000; Shirasu et al. 2000; Vicient et al. 1999; Vitte and Panaud 2003). In the *CesA1* region, however, all elements are either fully situated within a span of unique non-coding DNA or are identifiably full-length. Thus, while ectopic recombination may play a role in shaping the genome and genome sizes in *Gossypium*, no evidence of that role was seen here.

Similarly, illegitimate recombination has recently been shown to have the ability to reduce genome size more than was previously anticipated (Ma et al. 2004). The current comparison does not distinguish insertions from deletions, and thus we are unable to accurately gauge the extent to which illegitimate recombination has shaped this region. However, as slightly more than a third of the indels larger than 10 bp in size had flanking repeats, illegitimate recombination may prove to be an active force contributing to genome size evolution in *Gossypium*. Follow-up studies that distinguish insertions from deletions will further enable an evaluation of the importance of illegitimate recombination in cotton.

Analysis of the sequenced *Arabidopsis* and rice genomes showed that organelle-nuclear transfers (fragmented or full length) can be common in some genomes (rice) and relatively infrequent in others (*Arabidopsis)* (Shahmuradov et al. 2003). In the present study, one chloroplast DNA (cpDNA) fragment was found in the A_T BAC, nestled among unique non-coding DNA. This fragment was only 900bp in length, however, so it does not contribute significantly to genome size evolution in this region.

Despite evidence from broader phylogenetic surveys and some other systems that intron size may be correlated with genome size (Deutsch and Long 1999; McLysaght et al. 2000; Moriyama et al. 1998; Vinogradov 1999), this is not true for genes in the *CesA1* region . The average intron size deviation was 4.3 bp per gene (60 bp total). Intron size deviation was not biased with respect to genome; the A_T BAC sequence contains a total of only three bp more intronic sequence than does its homoeologue. This result is not surprising for *Gossypium;* a previous study reported for 40 nuclear genes that there exists no significant size variation between *Gossypium* species groups with varying genome sizes (Senchina et al. 2003). Thus, while intron size

expansion/contraction may play a role in shaping the size of other genomes, evidence from *Gossypium* indicates that it has not played a similar role at the phylogenetic scale encompassed by this genus.

One of the attractive proposals that attempts to account for genome size variation is that there exist biases in the frequency and size of insertions and deletions (Bensasson et al. 2001; Petrov 2002a; Petrov 2002b). To evaluate this possibility, we tabulated the spectrum of small indels in the *CesA1* region of the A_T and D_T genomes. The data reveal no evidence of an indel bias (Table 2, Fig. 3). For each indel bin, there was approximately the same number of indels accounting for a similar total of nucleotides. The maximum difference for any bin was 344bp, which was counteracted through indels in other bins. Overall, the total difference in genome size attributable to small indels is a scant 120 bp (in D_T). These observations demonstrate the absence of a globally operating indel bias in *Gossypium*, despite evidence to the contrary in some other plants (Kirik et al. 2000; Orel and Puchta 2003).

Remarkable conservation of intergenic space

Aside from several TE insertions and a single chloroplast insertion, most intergenic space between the A_T and D_T genomes is highly conserved. This contrasts with most other studies of microcolinearity, reflecting both the absence of major structural alterations in this genomic region (as discussed above) and perhaps the amount of time that has elapsed since the A and D genomes diverged from their last common ancestor. Yet reports from grasses suggest that ~11 million years is sufficient to remove homology outside of genes (Ramakrishna et al. 2002b; SanMiguel et al. 2002) and, in some cases, only 0.5 - 1 million years is required (Wicker et al. 2003). Since the A_T and D_T genomes evolved in isolation on different continents for 5-10 million years prior to becoming reunited by polyploidy ~ 1mya (Cronn et al. 2002;

Senchina et al. 2003), one might not have been surprised by detecting a larger amount of intergenic divergence and lessened sequence identity. The remarkable conservation we observed indicates that the evolutionary forces and molecular mechanisms responsible for rapid intergenic divergence in other plant systems do not operate similarly in this region of *Gossypium*.

Concluding remarks

A large body of empirical evidence has demonstrated that myriad external forces and internal molecular genetic mechanisms are involved in the complex suite of phenomena that collectively mold genome size (Petrov 2001; Petrov and Wendel 2004). Recent technological advances in large insert libraries and high throughput sequencing have made genomic comparisons accessible and feasible, thereby promising increasing application to non-model organisms. These comparisons will enable insights into the organization of genomes and their evolution, and are likely to be more informative when conducted within well-understood phylogenetic frameworks. The research described here represents a first step in this direction for *Gossypium*, which contains, in addition to the A and D genomes, other diploid groups (Fig. 1) whose genome sizes span an even greater range than the two-fold size difference studied here. Extension of the present study to include more of this diversity, as well as to additional genomic regions will enable us to more critically evaluate the suggestion of relative stasis in gene islands and conservation of intergenic sequence reported here. In this regard, the recent publication of a high-density genetic map for *Gossypium* (Rong et al. 2004) will facilitate targeted selection of genomic regions for analysis.

Acknowledgments

We thank Trent Grover, Jamie Estill, and Jordan Swanson for technical assistance, Anna Gardner for help with Figure 3, Jennifer Hawkins for helpful discussion and

assistance with some analyses, and Cedric Feschotte for guidance concerning mining for MITEs. This work was funded by the National Science Foundation Plant Genome program, whose support we gratefully acknowledge.

Literature Cited

Adams, K.L. and J.D. Palmer. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* **29:** 380-395.

Bennett, M.D. 1985. Intraspecific variation in DNA amount and the nucleotypic dimension in plant genetics. In *Plant Genetics* (ed. M. Freeling), pp. 283-302. A. R. Liss, NY.

Bennett, M.D. 1987. Variation in genomic form and its ecological implications. *New Phytol.* **106:** 177-200.

Bennett, M.D. and I.J. Leitch. 1995. Nuclear DNA amounts in angiosperms. *Annals Bot.* **76:** 113-176.

Bennett, M.D. and I.J. Leitch. 1997. Nuclear DNA amount in angiosperms. *Phil. Trans. Royal Soc. London B* **334:** 309-345.

Bennett, M.D. and I.J. Leitch. 2003. Plant DNA C-values Database. *http://www.rbgkew.org.uk/cval/homepage.html*.

Bennetzen, J.L. 2000a. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* **12:** 1021-1029.

Bennetzen, J.L. 2000b. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* **42:** 251-269.

Bennetzen, J.L. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* **115:** 29-36.

Bennetzen, J.L. and E.A. Kellogg. 1997. Do plants have a one-way ticket to genomic obesity? *The Plant Cell* **9:** 1509-1514.

Bensasson, D., D.A. Petrov, D.X. Zhang, D.L. Hartl, and G.M. Hewitt. 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Molecular Biology and Evolution* **18:** 246-253.

Bertran, E. and M. Long. 2002. Expansion of genome coding region by acquisition of new genes. *Genetica* **115:** 65-80.

Biderre, C., G. Metenier, and C.P. Vivares. 1998. A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*. *Mol Biochem Parasitol* **74:** 229-231.

Brudno, M., C. Do, G. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large scale multiple alignment of genomic DNA. *Genome Research* **13:** 721-731.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78-94.

Cavalier-Smith, T. 1985. *The Evolution of Genome Size.* John Wiley, New York.

Chantret, N., A. Cenci, F. Sabot, O. Anderson, and J. Dubcovsky. 2004. Sequencing of the *Triticum monococcum Hardness* locus reveals good microcolinearity with rice. *Molecular Genetics and Genomics* **Online First:** 1617-4623.

Chen, M., P. SanMiguel, A.C. de Oliveira, S.-S. Woo, H. Zhang, R.A. Wing, and J.L. Bennetzen. 1997. Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. *Proceeding of the National Academy of Science* **94:** 3431-3435.

Chen, M.S., P. SanMiguel, and J.L. Bennetzen. 1998. Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* **148:** 435-443.

Chooi, W.Y. 1971. Variation in nuclear DNA content in the genus *Vicia*. *Genetics* **68:** 195-211.

Claverie, J.-M. 2000. What if there are only 30,000 human genes? *Science* **291:** 1255-1257.

Cronn, R.C., R.L. Small, T. Haselkorn, and J.F. Wendel. 2002. Rapid diversification of the cotton genus (Gossypium : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany* **89:** 707-725.

Deutsch, M. and M. Long. 1999. Intron-exon structure of eukaryotic model organisms. *Nuc. Acids Res.* **27:** 3219-3228.

Devos, K.M., J.K.M. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Research* **12:** 1075-1079.

Dubcovsky, J., W. Ramakrishna, P.J. SanMiguel, C.S. Busso, L.L. Yan, B.A. Shiloff, and J.L. Bennetzen. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiology* **125:** 1342-1353.

Endrizzi, J.D., E.L. Turcotte, and R.J. Kohel. 1985. Genetics, cytology, and evolution of *Gossypium*. *Advances in Genetics* **23:** 271-375.

Ewing, B. and P. Green. 1998. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequences traces using phred. I. Accuracy assessment. *Genome Research* **8:** 175-185.

Feschotte, C. and C. Mouches. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution* **17:** 730-737.

Feschotte, C., X. Zhang, and S.R. Wessler. 2002. Miniature inverted-repeat transposable elements and their relationship to established DNA transposons. In *Mobile DNA II* (ed. N.L. Craig). ASM Press, Washington D. C.

Fu, H.H. and H.K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 9573-9578.

Gale, M.D. and K.M. Devos. 1998. Plant comparative genetics after 10 years. *Science* **282:** 656-659.

Goff, S.A., D. Ricke, T.-H. Lan, and 52 coauthors. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296:** 92-100.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8:** 195-202.

Green, P. 1999. Phrap documentation. *http://www.phrap.org/phrap.docs/phrap.html*.

Gregory, T.R. 2001. Animal genome size database. *http://www.genomesize.com*.

Gregory, T.R. 2002. Genome size and developmental complexity. *Genetica* **115:** 131-146.

Gregory, T.R. and P.D.N. Hebert. 1999. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* **9:** 317-324.

Ilic, K., P.J. SanMiguel, and J.L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *PNAS* **100:** 12265-12270.

Jones, R.N. and L.M. Brown. 1976. Chromosome evolution and DNA variation in *Crepsis*. *Heredity* **36**.

Jurka, J. 2000. Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **9:** 418-420.

Jurka, J., P. Klonowski, V. Dagman, and P. Pelton. 1996. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry* **20:** 119-122.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo, and A.H. Schulman. 2000. Genome evolution in wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci. USA* **97:** 6603-6607.

Kellis, M., B.W. Birren, and E. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **Advane online publication, 7 March 2004:** doi:10.1038/nature02424.

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

Ku, H.M., T. Vision, J. Liu, and S.D. Tanksley. 2000. Comparing sequenced segments of tomato and *Arabidopsis* genomes: large scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97:** 9121-9126.

Langham, R.J., J. Walsh, M. Dunn, C. Ko, S.A. Goff, and M. Freeling. 2004. Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics* **166:** 935-945.

Langley, C.H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52:** 223-235.

Leitch, I.J., M.W. Chase, and M.D. Bennett. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals Bot. (Suppl. A)* **82:** 85-94.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Inc., Sunderland, MA.

Li, W.-H. and D. Graur. 1991. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Lukashin, A. and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nuc Acids Res* **26:** 1107-1115.

Ma, J., K.M. Devos, and J.L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Gen. Res.* **14:** 860-869.

McLysaght, A., A.J. Enright, L. Skrabanek, and K.H. Wolfe. 2000. Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast* **17:** 22-36.

Mirsky, A.E. and H. Ris. 1951. The DNA content of animal cells and its evolutionary significance. *J. Gen. Physiol.* **34:** 451-462.

Moriyama, E.N., D.A. Petrov, and D.L. Hartl. 1998. Genome size and intron size in Drosophila. *Molecular Biology and Evolution* **15:** 770-773.

Orel, N. and H. Puchta. 2003. Differences in the processing of DNA ends in Arabidopsis thaliana and taobacco: possible implications for genome evolution. *Plant Molecular Biology* **51:** 523-531.

Petrov, D.A. 2001. Evolution of genome size: new approaches to an old problem. *Trends in Genetics* **17:** 23-28.

Petrov, D.A. 2002a. DNA loss and evolution of genome size in Drosophila. *Genetica* **115:** 81-91.

Petrov, D.A. 2002b. Mutational equilibrium model of genome size evolution. *Theoretical Population Biology* **61:** 531-544.

Petrov, D.A., Y.T. Aminetzach, J.C. Davis, D. Bensasson, and A.E. Hirsh. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Molecular Biology and Evolution* **20:** 880-892.

Petrov, D.A. and J.F. Wendel. 2004. Evolution of eukaryotic genome structure. In *Evolutionary Genetics: Concepts and Case Studies* (eds. C.W. Fox and J.B. Wolf), pp. in press. Oxford Univ. Press.

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

Ramakrishna, W., J. Emberton, M. Ogden, P. SanMiguel, and J.L. Bennetzen. 2002b. Structural analysis of the maize Rp1 complex reveals numerous sites

and unexpected mechanisms of local rearrangement. *Plant Cell* **14:** 3213-3223.

Rong, J., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. DelMonte, X. Ding, J.J. Garza, B.S. Marler, C. Park, G.J. Pierce, K.M. Rainey, V.K. Rastogi, S.R. Schulze, N.L. Trolinder, J.F. Wendel, T.A. Wilkins, T.D. Williams-Coplin, R.A. Wing, R.J. Wright, X. Zhao, L. Zhu, and A.H. Paterson. 2004. A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission, and evolution of cotton (*Gossypium*). *Genetics* **166:** 389-417.

Rossberg, M., K. Theres, A. Acarkan, R. Herrero, T. Schmitt, K. Schumacher, G. Schmitz, and R. Schmidt. 2001. Comparative sequence analysis reveals extensive microcolinearity in the Lateral suppressor regions of the tomato, Arabidopsis, and Capsella genomes. *Plant Cell* **13:** 979-988.

SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20:** 43-45.

SanMiguel, P., W. Ramakrishna, J.L. Bennetzen, C.S. Busso, and J. Dubcovsky. 2002. Transposable elements, genes, and recombination in a 215kb contig from wheat chromosome $5A^m$. *Funct. Integr. Genomics* **2:** 70-80.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765-768.

Seelanan, T., A. Schnabel, and J.F. Wendel. 1997. Congruence and consensus in the cotton tribe*. Syst. Bot.* **22:** 259-290.

Senchina, D.S., I. Alvarez, R.C. Cronn, B. Liu, J.K. Rong, R.D. Noyes, A.H. Paterson, R.A. Wing, T.A. Wilkins, and J.F. Wendel. 2003. Rate variation

among nuclear genes and the age of polyploidy in Gossypium. *Molecular Biology and Evolution* **20:** 633-643.

Shahmuradov, I.A., Y.Y. Akbarova, V.V. Solovyev, and J.A. Aliyev. 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Plant Mol. Biol.* **52:** 923-934.

Shirasu, K., A.H. Schulman, T. Lahaye, and P. Schulze-Lefert. 2000. A Contiguous 66-kb Barley DNA Sequence Provides Evidence for Reversible Genome Expansion. *Genome Res.* **10:** 908-915.

Small, R.L., J.A. Ryburn, R.C. Cronn, T. Seelanan, and J.F. Wendel. 1998. The tortoise and the hare: Choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* **85:** 1301-1315.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey, and A. Rafalski. 2000. The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. *The Plant Cell* **12:** 381-391.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. *Nature* **408:** 796-815.

The Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity, and evolution or rice chromosome 10. *Science* **300:** 1566-1569.

Thomas, C.A. 1971. The genetic organisation of chromosomes. *Annu. Rev. Genet.* **5:** 237-256.

Tikhonov, A.P., P.J. SanMiguel, Y. Nakajima, N.M. Gorenstein, J.L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proceedings of the National Academy of Sciences of the United States of America* **96:** 7409-7414.

Tomkins, J.P., D.G. Peterson, T.J. Yang, D. Main, T.A. Wilkins, A.H. Paterson, and R.A. Wing. 2001. Development of genomic resources for cotton (*Gossypium hirsutum* L.): BAC library construction, preliminary STC analysis, and

identification of clones associated with fiber development. *Mol. Breed.* **8:** 255-261.

Tu, Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proceeding of the National Academy of Science* **98:** 1699-1704.

van Leeuwen, H., A. Monfort, H.B. Zhang, and P. Puigdomenech. 2003. Identification and characterisation of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microcolinearity between Cucumis melo and Arabidopsis thaliana. *Plant Molecular Biology* **51:** 703-718.

Vandepoele, K., Y. Saeys, C. Simillion, J. Raes, and Y. Van de Peer. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Research* **12:** 1792-1801.

Vicient, C.M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav, N. E., and A.H. Schulman. 1999. Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. *The Plant Cell* **11:** 1769-1784.

Vinogradov, A.E. 1999. Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* **49:** 376-384.

Vision, T.J., D.G. Brown, and S.D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290:** 2114-2117.

Vitte, C. and O. Panaud. 2003. Formation of Solo-LTRs Through Unequal Homologous Recombination Counterbalances Amplifications of LTR Retrotransposons in Rice Oryza sativa L. *Mol Biol Evol* **20:** 528-540.

Wendel, J.F. 2000. Genome evolution in polyploids. *Plant Molecular Biology* **42:** 225-249.

Wendel, J.F. and R.C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **78:** 139-186.

Wendel, J.F., R.C. Cronn, I. Alvarez, B. Liu, R.L. Small, and D.S. Senchina. 2002a. Intron size and genome size in plants. *Molecular Biology and Evolution* **19:** 2346-2352.

Wendel, J.F., R.C. Cronn, J.S. Johnston, and H.J. Price. 2002b. Feast and famine in plant genomes. *Genetica* **115:** 37-47.

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

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.-D. Liu, J. Dubcovsky, and B. Keller. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell* **15:** 1186-1197.

Yu, J., S. Hu, J. Wang, G.K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79-92.

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18:** 292-298.

Zhao, X.P., Y. Si, R.E. Hanson, C.F. Crane, H.J. Price, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Research* **8:** 479-492.

Web Site References

http://www.phrap.org/, The Phred/Phrap/Consed System Home Page

http://www.softberry.com/, Softberry homepage

http://lagan.stanford.edu/, LAGAN alignment toolkit website

http://www.mbio.ncsu.edu/BioEdit/bioedit.html, Biological sequence alignment editor

http://ftp.genome.washington.edu/RM/RepeatMasker.html, RepeatMasker web server

CHAPTER 4

MICROCOLINEARITY AND GENOME EVOLUTION IN THE ADHA REGION OF DIPLOID AND POLYPLOID COTTON (GOSSYPIUM)

A paper published in The Plant Journal[1]

Corrinne E. Grover[2], HyeRan Kim[3], Rod A. Wing[4], Andrew H. Paterson[5], Jonathan F. Wendel[6]

Summary

Genome sizes vary by several orders of magnitude, driven by mechanisms such as illegitimate recombination and transposable element proliferation. Prior analysis of the *CesA* region in two cotton genomes that diverged 5-10 million years ago (mya) and acquired a 2-fold difference in genome size revealed extensive local conservation of genic and intergenic regions, with no evidence of the global genome size difference. The present study extends the comparison to include BAC sequences surrounding the gene encoding alcohol dehydrogenase A (*AdhA*) from four cotton genomes: the two co-resident genomes (A_T and D_T) of the allotetraploid, *Gossypium hirsutum*, as well as the model diploid progenitors, *G. arboreum* (A) and *G. raimondii* (D). In contrast to earlier work, evolution in the *AdhA* region reflects, in a microcosm, the overall difference in genome size, with a nearly twofold difference in aligned sequence length. Most size difference may be attributed to differential accumulation of retroelements during divergence of the genome diploids from their common ancestor, but in addition there has been a biased accumulation of small deletions, such that those in the smaller D genome are on average twice as large as

---

[1] Reprinted with permission of The Plant Journal, 2007, 50(6), 995-1006.

[2] Graduate student, primary research and author, EEOB department, Iowa State University
[3] Post-doctoral associate, BAC sequencing and finishing, Arizona Genomics Institute, University of Arizona
[4] Director, BAC sequencing and finishing, Arizona Genomics Institute, University of Arizona
[5] Professor and director, BAC selection and hybridizations, Plant Genome Mapping Laboratory, University of Georgia
[6] Principal investigator and corresponding author, EEOB department, Iowa State University

those in the larger A genome. The data also provide evidence for the global phenomenon of “genomic downsizing” in polyploids shortly after formation. This in part reflects a higher frequency of small deletions post-polyploidization, and increased illegitimate recombination. In conjunction with previous work, the data here confirm the conclusion that genome size evolution reflects many forces that collectively operate heterogeneously among genomic regions.

Keywords: genome size, genome evolution, transposable elements, c-value, *Gossypium*, cotton

Introduction

The observation that genome sizes vary tremendously among eukaryotes, and are largely uncorrelated with organismal complexity, has generated substantial interest over the last half-century. This interest has stimulated numerous genome size surveys for diverse organisms (Bennett and Leitch, 2005a; Gregory, 2006) as well as discussion of the modes and mechanisms responsible for the observed variation (Flavell *et al.*, 1974; Bennetzen, 2000; Gregory, 2001; Petrov, 2001; Bennetzen, 2002; Gregory, 2005). Once thought to result mostly from polyploidy or polyteny (Thomas, 1971), genome size evolution is now recognized as reflecting the net effects of a suite of mechanisms that sometimes work antagonistically to expand and contract the genome. Best understood are the array of mechanisms responsible for genome size expansion, most prominently polyploidy (Wendel, 2000) and transposable element amplification (Bennetzen, 2000, 2002; Kidwell, 2002; Piegu *et al.*, 2006), but also smaller-scale processes such as tandem repeat expansion (Ellegren, 2002; Morgante *et al.*, 2002), gene duplication and pseudogenization (Zhang, 2003), organellar transfer to the nucleus (Shahmuradov *et al.*, 2003), and intron size expansion (Deutsch and Long, 1999; Vinogradov, 1999). Less is known about mechanisms of genome size contraction, of which unequal intrastrand

homologous recombination (Shepherd *et al.*, 1984; SanMiguel *et al.*, 1996; Chen *et al.*, 1998; Vicient *et al.*, 1999; Shirasu *et al.*, 2000), double-strand break repair (Kirik *et al.*, 2000; Orel and Puchta, 2003), and illegitimate recombination (Wicker *et al.*, 2001; Devos *et al.*, 2002; Ma *et al.*, 2004; Bennetzen *et al.*, 2005) are thought to be important. Processes such as replication error and recombination in regions of tandem repeats may further contribute to genome size contraction through biases favoring small deletions over insertions (Petrov, 1997; Petrov, 2002). Superimposed on these "internal" molecular and genetic mechanisms that contribute to genome size differences are myriad "external" biological and ecological factors that may potentially influence, or be influenced by, genome size (Bennett *et al.*, 1998; Vinogradov, 2003; Cavalier-Smith, 2005; Knight *et al.*, 2005; Petrov and Wendel, 2006), although in most cases these relationship remain unclear.

Comparative approaches offer numerous opportunities for advancing our understanding of genome size evolution, including those that involve detailed study of microcolinearity among phylogenetically well-understood species. Previously, we reported a comparison of 100+ kb of homoeologous sequence surrounding a cellulose synthase gene (Grover *et al.*, 2004) from the two genomes that coexist in the allotetraploid nucleus of the cultivated cotton species *Gossypium hirsutum.* These two genomes differ by twofold in size, despite having originated from diploid species that have the same chromosome number and suite of life-history features (Wendel and Cronn, 2003). Analysis of the *CesA* region demonstrated that the twofold difference in overall genome size is differentially distributed among genomic regions. Furthermore, the *CesA* region displayed extraordinary conservation in both gene content and intergenic sequence, which was unexpected given prior comparisons in plants.

To continue to investigate the patterns and processes responsible for genome size evolution in *Gossypium*, we report further comparative sequencing using both diploid and allopolyploid cotton species. *Gossypium* is an approximately 5-10 million year old genus whose members have genomes that range 3-fold in size, from the D-genome diploids in the New World to the Australian K-genome diploids (Hendrix and Stewart, 2005). Approximately 5-10 Ma, two diploid groups, designated A-genome and D-genome, diverged and subsequently acquired genomes that differ approximately twofold in size. Allopolyploidization reunited these two genomes approximately 1-2 Ma (Figure 1), generating five species, including the agronomically important *G. hirsutum*, the genomes of which are slightly less than additive with respect to their diploid progenitors (Hendrix and Stewart, 2005).

We present here an analysis of comparative sequencing of a BAC-sized region surrounding the alcohol dehydrogenase A gene (*AdhA*), from two diploid species representing the closest living relatives of the A- and D-genome species involved in the allopolyploidization event (reviewed in Wendel and Cronn, 2003), as well as from both homoeologous genomes (A_T and D_T) from the tetraploid, *Gossypium hirsutum*. In contrast to the previously sequenced *CesA* region, the sequence composition of the *AdhA* region mirrors the overall pattern of genome size evolution in the diploid genomes. While still retaining a high level of intergenic sequence conservation, the *AdhA* region in the A and A_T genomes is disrupted by the presence of many *gypsy* elements, which serve to expand the region in a manner that reinforces the conclusions reached following analysis of sequences from whole-genome shotgun libraries (Hawkins *et al.*, 2006). In addition to describing this phenomenon, the data presented here reveal details of “genomic downsizing” in polyploids shortly after their formation, suggest an indel bias leading to frequent and larger deletions in smaller genomes, and provide evidence that increased illegitimate recombination that may lead to genome size contraction.

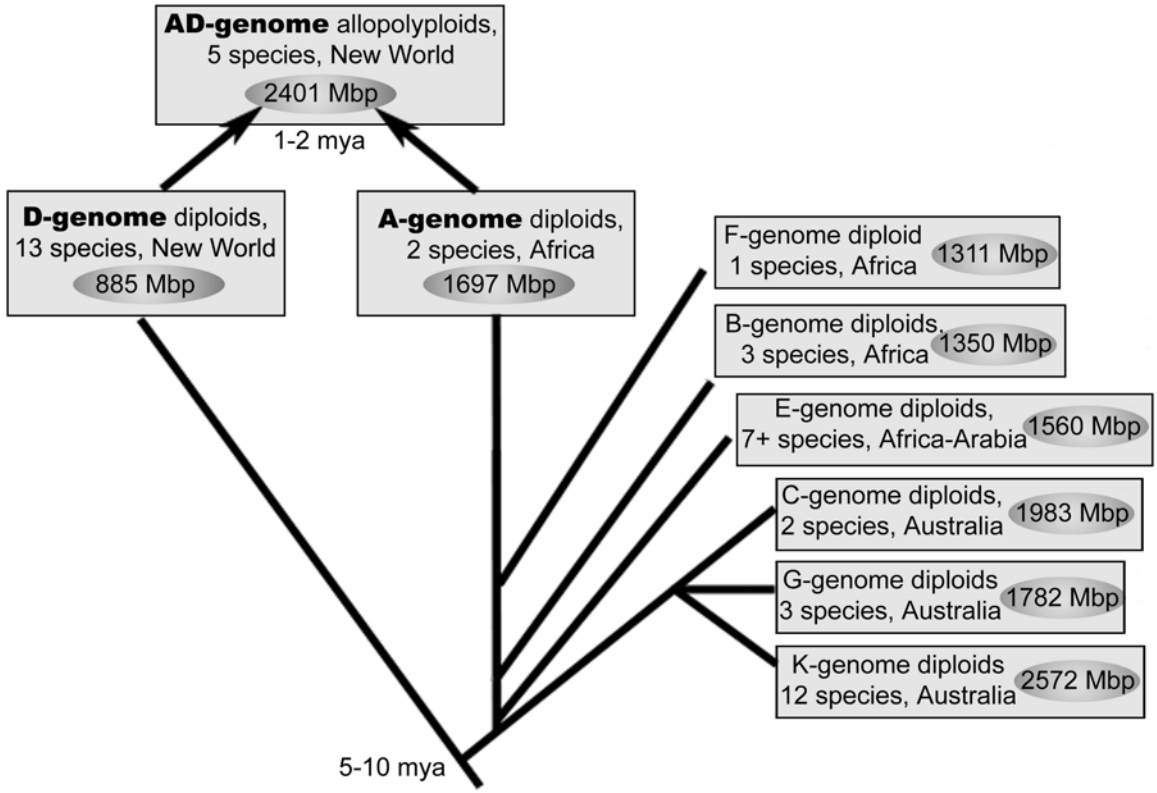


Figure 1. The evolutionary history of diploid and tetraploid *Gossypium* species groups (n = 13 and n = 26, respectively), as inferred from multiple molecular datasets (Seelanan *et al.*, 1997; Small *et al.*, 1998; Cronn *et al.*, 2002). The eight diploid genome groups, determined by interspecific meiotic pairing and chromosome size (Endrizzi *et al.*, 1985), range in size from an average of 885 Mbp in the D genome diploids to an average of 2576 Mbp in the K genome diploids (Hendrix and Stewart, 2005). Polyploid species are thought to have originated 1-2 Ma, following divergence of their diploid progenitors 5-10 Ma, and have an average genome size that is slightly less than additive with respect to their diploid progenitors (Hendrix and Stewart, 2005). The model diploid species used here, *Gossypium raimondii* (D) and *Gossypium arboreum* (A), represent the closest extant relatives of the polyploid genome donors (estimated 0.68% and 1.05% sequence divergence from the polyploid *G. hirsutum* to *G. raimondii* and *G. arboreum*, respectively; Cronn *et al.* 1999).

Results

Sequence comparison between BACs from diploid and polyploid genomes: A versus A_T

The *AdhA* BACs from the A genome diploid (112.3 kb) and the A_T genome from the allotetraploid (195.3 kb) were shotgun sequenced and assembled. The aligned

length of the two BACs was 117.3 kb, accounting for the full 112.3 kb in A and 101.7 kb in A_T, with the elongated alignment reflecting gaps between the diploid and polyploid sequences (Figure 2).

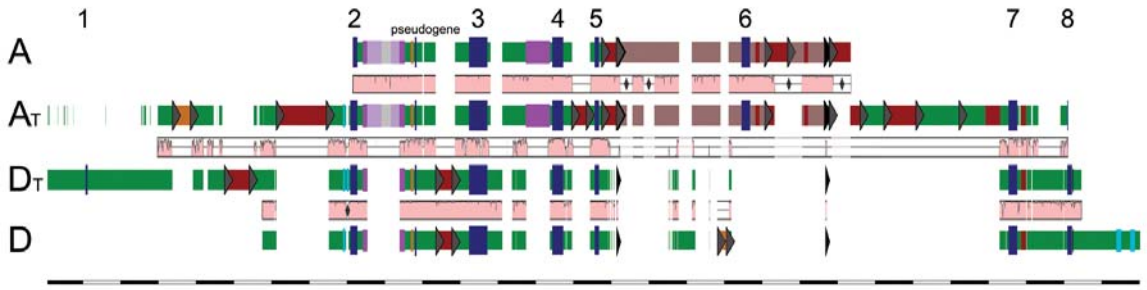


Figure 2. Multiple alignment of orthologous *AdhA* BACs from four different genomes (A, D, A_T and D_T; the latter two are co-resident in the nucleus of polyploid cottons). Numbered blue boxes are predicted genes corresponding to the list presented in Table 1; *copia* elements are in orange, *gypsy* elements in red, and LINE elements in pink. Identifiable LTRs are depicted by triangles. Continuous windows of sequence identity are shown between each pair of BACs, with that in the middle illustrating sequence identity between the two BAC-pairs (A and A_T vs. D and D_T); all are scaled from 50% to 100%. Grey diamonds on the identity plots denote the location of large (>400bp), unpolarized indels between the diploid progenitor and respective polyploid genome. The scale bar at the bottom indicates increments of 10 Kb.

Database searches led to the inference of five shared genes and one shared pseudogene (Table 1), giving gene densities of one gene per 22 kb and one gene per 20 kb for A and A_T, respectively (19 kb and 17 kb if the pseudogene is included). Collectively, the five genes account for approximately 12.8 kb of sequence in each BAC, or approximately 10-12% of each BAC. Both BACs have a GC content of 34% and were determined to be 98.5% identical in sequence (81.28% including gaps). A total of 122 gaps appear in the alignment of the A and A_T sequences; these are unequally distributed as 28 gaps in the A sequence (151bp) and 64 gaps in the A_T sequence (15,548bp). When large indels (>400 bp) are removed, the number and length of gaps in A remains the same, but diminishes in A_T to 60 gaps (449 bp). As these gaps are inferred to have evolved subsequent to

Table 1: Gene Features predicted in the *AdhA* region of diploid and tetraploid cotton

		Nucleotide length					Total Length in Exons					Total Length in Introns				Protein Length				
Gene #	Putative Function	A	A_T	D_T	D	Number of Exons	A	A_T	D_T	D	Number of Introns	A	A_T	D_T	D	A	A_T	D_T	D	EST
1	Ca-binding	*	*	519	*	1	*	*	519	*	0	*	*	0	*	*	*	172	*	No
2	alcohol dehydrogenase A	1223*	1844	1833	1834	8	885*	1149	1149	1149	7	338*	695	684	685	407*	382	382	382	Yes
3	FAD-dependent oxidoreductase	4759	4766	4403	4415	4	1272	1272	1272	1272	3	3487	3494	3131	3143	423	423	423	423	Partial
4	protein disulfide isomerase	2846	2843	2870	2866	10	1488	1488	1488	1488	9	1358	1355	1382	1378	495	495	495	495	Yes
5	integral membrane	1110	1110	1110	1110	1	1110	1110	1110	1110	0	0	0	0	0	369	369	369	369	Partial
6	caffeic acid O-methyltransferase A	2271	2276	---	---	3	1077	1077	---	---	2	1194	1199	---	---	358	358	---	---	Yes
7	caffeic acid O-methyltransferase B	*	2302	2274	2278	3	*	1035	1077	1077	2	*	1267	1197	1201	*	344	358	358	Yes (b)
8	photosystem II	*	178*	1203	1205	4	*	178*	837	837	3	*	0*	366	368	*	58	278	278	Yes

* truncated by BAC end or does not exist because BAC ends

A has internal stop in D_T…maybe pseudogenized

B EST support for D/D_T mRNA only

the origin of the polyploids about 1-2 Ma, the foregoing numbers reflect the differential accumulation of indels subsequent to polyploid formation. Also distinguishing the two genomes is a single retrotransposon insertion in the A_T genome (between genes 4 and 5, Figure 2), accounting for 4799 bp, which by its exclusivity is inferred to have been inserted since the origin of the polyploids.

Sequence comparison between BACs from diploid and polyploid genomes: D versus D_T

The *AdhA* BACs from the D genome diploid (101.3 kb) and the D_T genome from the allotetraploid (130.9 kb) were also shotgun sequenced and assembled. The aligned length of the two genomes was 86.7 kb, accounting for 85.7 kb in D and 80 kb in D_T, again indicating a size differential between the diploid and polyploid that most likely reflects evolution since polyploidization. Database searches led to the inference of six shared genes (Table 1), one of which may recently be pseudogenized, and one shared pseudogene, giving gene densities of one gene per 14 kb and one gene per 13 kb for D and D_T, respectively (12 kb and 11 kb, if the ancient and recent pseudo genes are included). The six shared genes account for 13.7 kb of sequence in each BAC, or approximately 16-17% of each BAC. The D and D_T genome BACs had GC contents of approximately 33.6% and were determined to be 98.2% identical in sequence (89.38% including gaps). A total of 121 phylogenetically unpolarized gaps (i.e. gaps that were not distinguishable as insertions or deletions, see methods) differentiate the D and D_T genomes, distributed as 57 gaps in D (943 bp) and 64 gaps in D_T (699 bp), and again reflecting indels that arose since polyploidization. When large gaps are excluded (>400 bp; Figure 2), the number and length of gaps in D reduces to 56 gaps (309 bp), whereas the number and length in D_T remains the same. A single *copia* insertion in the D genome (between genes 5 and 6, Fig. 2) also distinguishes the two genomes, accounting for 2348 bp.

Sequence comparison between BACs from all diploid and polyploid genomes

The aligned length of the *AdhA* BACs from all four genomes was 132.8 kb, accounting for 112.3 kb of sequence in A, 101.7 kb in A_T, 55 kb in D, and 49 kb in D_T. The size differential between the A/A_T genomes and D/D_T genomes is approximately 50%, which mirrors their relative difference in overall genome size (885 vs. 1697 Mbp; Figure 1). All predicted genes and pseudogenes were shared, with the exception of a putative caffeic acid O-methyltransferase encoding gene, which was duplicated in the A_T genome (Table 1; Figure 2). The pairwise comparison of A BACs with D BACs, irrespective of origin (diploid versus tetraploid), gave an average of 92% sequence identity (91.97% to 92.01%; 28.6% to 32.9% if including gaps).

As previously reported for *Gossypium* (Grover *et al.*, 2004), the intergenic space was remarkably conserved between the A and D genomes, which diverged 5-10Ma, as well as between the diploid and tetraploid genomes. Interestingly, the conserved intergenic space was mostly represented by DNA of unidentified origin or function. These sequences could represent TEs degraded beyond the point of recognition, unidentified regulatory elements, functionally constrained sequences, or more likely, a combination of these and other sequences.

Gap polarization and analysis

Overall, of the phylogenetically polarized gaps (i.e. gaps clearly identifiable as insertions or deletions, see methods) that were inferred to have arisen from mechanisms other than TE insertion or deletion, there were significantly (chi^2; $p < 0.0059$) more identifiable deletions than insertions (50 versus 26). Excluding a single large insertion, small insertions ranged in size from 1 – 13 nt (range in average size = 1 - 2.8 nt/insertion). The range in deletion size was larger (1 – 32 nt; average = 1.92 - 5.27 nt/deletion), and the average deletion size in the A and A_T BACs was approximately half of that observed in the D and D_T BACs.

Each indel was assigned a probable mechanism of origin (Table S1), regardless of whether the indel was polarized. Transposable element insertions and probable insertions account for the majority of sequence difference between the four genomes, representing over half of the alignment for the A genomes. Illegitimate recombination, a RecA independent form of recombination involving regions of microhomology (2-15 bp) flanked by short direct repeats, incorporates a variety of mechanisms, most notably double-stranded break (DSB) repair and slipstrand annealing. Both were common in the alignment, occurring mostly within transposable elements. A majority (50.42%) of indels were classified as having been generated by an "unknown mechanism", due to the absence of mechanistic hallmarks. Finally, a portion (9.2%) of indels were classified as having arisen from "illegitimate recombination-double strand break repair or illegitimate recombination-slipstrand mispairing", as it often is not possible to tell these two mechanisms apart (Figure 3).

Aside from transposable elements, and a single 3.6 kb insertion in D, all phylogenetically polarized indels were less than 400 bp in size, the limit considered here to be a "small indel". The number of insertions and deletions that differentiate the D and D_T genomes, determined by using A and A_T to approximate the ancestral state, was similar for each genome (six insertions and 13 deletions versus six insertions and 20 deletions in the D and D_T genomes, respectively); however, this was not the case for the A and A_T genomes (where D and D_T represent the ancestral state), where we inferred 12 insertions and four deletions in the A genome but two insertions and 13 deletions in the A_T genome. Thus, there were more insertions and fewer deletions in the A genome than in other genomes studied, and a similar number of deletions in both genomes of the allopolyploid (13 vs. 20 for the D_T and A_T, respectively).

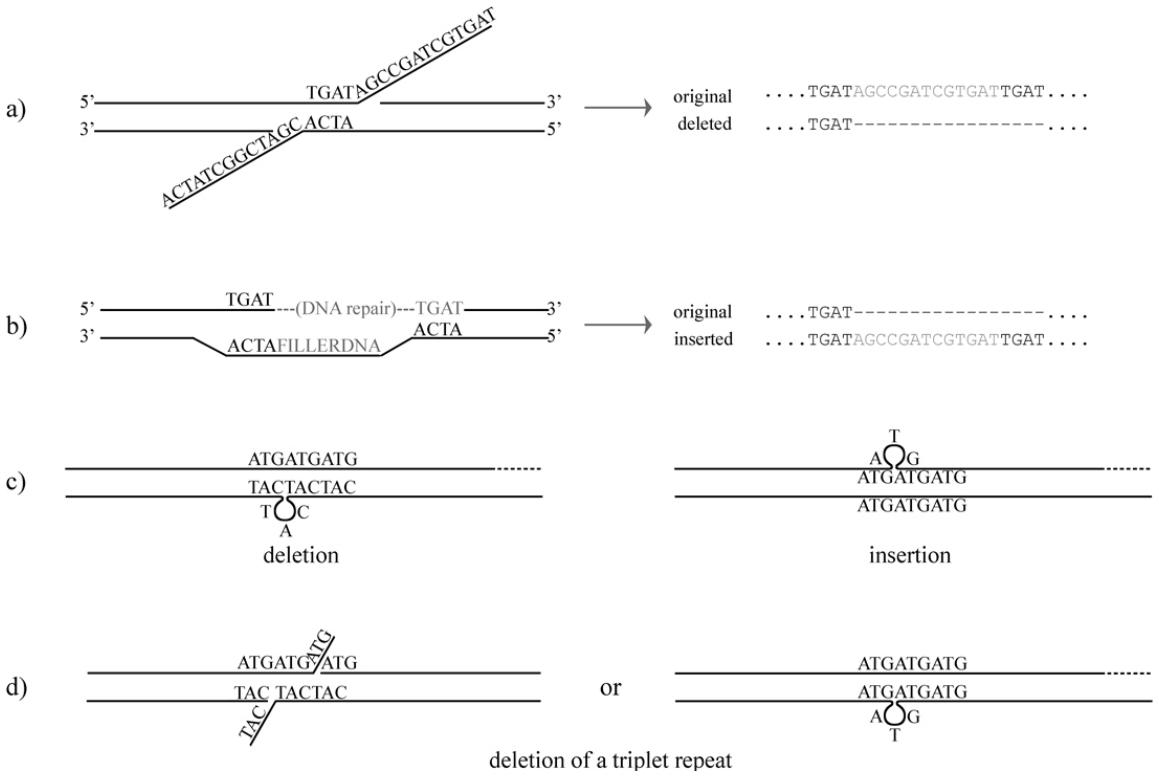


Figure 3. Illegitimate recombination represents several different mechanisms leading to the deletion of a sequence bounded by small repeats (only 1bp of homology required), as well as one of the bounding repeats, or, less commonly, the addition of intervening sequence. Three mechanisms are thought to be involved in illegitimate recombination, including two (panels a and b) that involve double strand break (DSB) repair.
(a) Single strand annealing, as shown, leads to deletion of sequence between short repeated motifs;
(b) synthesis-dependent strand annealing leads to insertion of filler DNA from diverse potential templates until a matching motif anneals to the other strand, which is then repaired complementary to the inserted foreign DNA;
(c) slipstrand mispairing may lead to either sequence insertion or deletion;
(d) In some cases, as in the example illustrated here, it is not possible to confidently distinguish DSB repair from slipstrand mispairing.

When the amount of sequence is considered and the gap data are normalized (e.g., per 100 kb; Table 2), the disparity in insertion rates among genomes largely disappears, whereas the disparity in deletion amounts increases. In addition, the average insertion size in the A genome, excluding TEs and the single large D genome insertion, was slightly larger than in the other genomes (2.8 nt in A versus 1 nt for both A_T and D, and 1.8 nt for D_T), whereas the average deletion size in the A

genome mirrored the average deletion size in A_T (2 nt and 1.9 nt, respectively) and was approximately half the average deletion size in D/D_T (5.3 nt and 4.2 nt, respectively). Thus, the data of Table 2 highlight two salient features of genome size evolution in the *AdhA* region: (1) the higher frequency and size of deletions in the D genome than in the A genome, consistent with their global difference in genome size; and (2) the higher rate of deletion in polyploid *Gossypium* than in its diploid antecedent genomes, consistent with the phenomenon of "genomic down-sizing" following polyploid formation.

Table 2: Types and frequency of mechanisms contributing to genome size change in the *AdhA* region

Mechanism	Type	*G. herbaceum* *A*	*G. hirsutum* A_T	*G. hirsutum* D_T	*G. raimondii* *D*
Overall	# deletions	4	13	20	13
	nt deletions	8	29	77	91
	# insertions	12	2 (1)	6	6 (4)
	nt insertions	34	4799 (1)	11	5971 (4)
	# unknown gaps (excluding TEs)	159	169	154	152
	nt missing (excluding TEs)	7809	21715	15440	15404
Small indels (< 400 bp) per 100 kb in the *AdhA* region	# deletions	3.57	12.78	36.36	26.53
	nt deletions	7.14	28.52	140	185.71
	# insertions	10.71	0.99	10.91	8.16
	nt insertions	30.28	0.99	20	8.16

Numbers in parentheses refer to the number and length of insertions, excluding large insertions (> 400 bp)

Analysis of putative genes

Six genes and one pseudogene are predicted to occur ancestrally in the *AdhA* region (Table 1). These six genes range in size from a 1.1 kb putative integral membrane protein-encoding gene to a 4.9 kb putative FAD-dependent oxidoreductase protein-encoding gene. The structures of four of the six genes were confirmed fully by EST evidence (Table 1), and the other two were partially

confirmed by incomplete EST evidence (Udall *et al.*, 2006; http://www.genome.arizona.edu/genome/cotton.html).

The putative integral membrane protein-encoding gene, partially confirmed by EST evidence, may be recently pseudogenized in the D_T genome. The matching EST is derived from a D-genome library, indicating transcription at the diploid level, and extends past the point in which the D_T genome has acquired a stop codon. This pseudogenization is inferred to be relatively recent, as no acceleration in non-synonymous mutations is observed (Ka A-D = 0.0024, Ka A-DT = 0.0024). A conserved-domain search (Marchler-Bauer and Bryant, 2004) indicated that this unknown gene bears a slight similarity (E-value=$2e^{-6}$) to nucleotide-sugar transporters.

A single gene duplication, involving a putative caffeic acid O-methyltransferase-encoding gene, differentiates the *AdhA* BACs of the A/A_T genomes from those of the D/D_T genomes. By virtue of its shared presence in the former two genomes, and its absence from the latter two genomes, we infer that the duplication event happened subsequent to the divergence of the A and D genome diploids from their common ancestor 5-10 Ma, but prior to polyploid formation 1-2 Ma. The duplicate falls within a block of several nested *gypsy* elements (full length as well as remnant) present in both the A and A_T genomes. Interestingly, the predicted intron/exon structure of the duplicate in the A genomes more closely resembles the structure found in the D genomes than its syntenic copy, primarily because of the predicted compensatory intron/exon boundary changes in the original A_T copy necessary to restore function in response to a 22 bp frame-shifting insertion (Figure 4). Alternatively, the original copy of the caffeic acid O-methyltransferase encoding gene may be pseudogenized by the 22bp insertion in the A genomes.

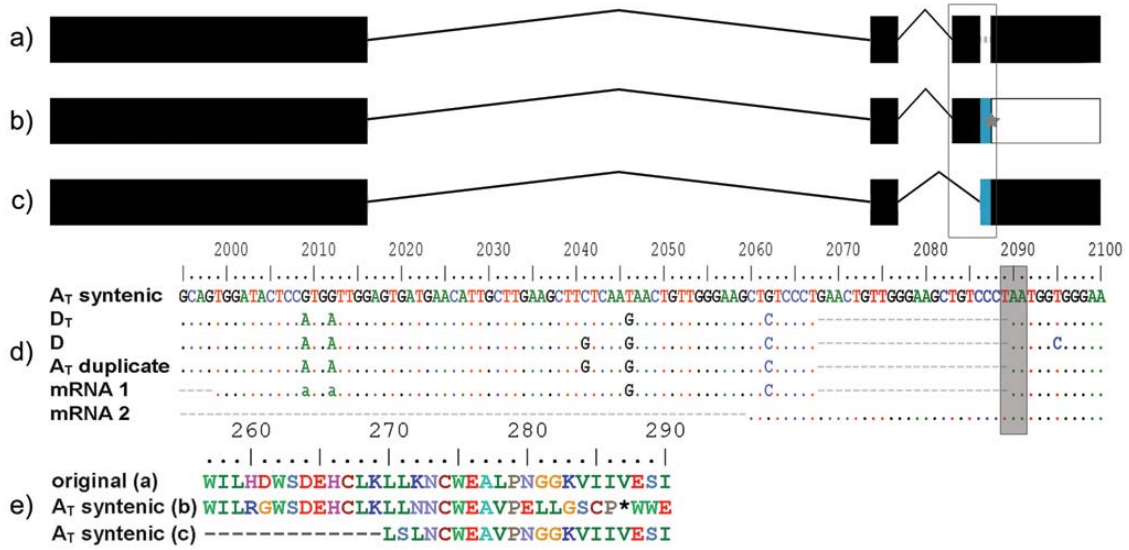


Figure 4: Possible splicing of the putative caffeic acid encoding genes in the *AdhA* region.
a) Structure of the putative caffeic acid encoding gene in D, D_T, and the duplicate (non-syntenic) copy in A_T. The location of the insertion in the syntenic A_T copy is noted by a dotted line.
b) Structure of the original putative caffeic acid encoding gene (syntenic copy) in A_T with conserved splice sites. The 22bp insertion is shown in blue and occurs fully within exon 3 in this model. This results in a frame-shifting, nonsense mutation that ultimately leads to a premature stop codon (star) and truncated protein.
c) Possible alternate structure of the original putative caffeic acid encoding gene in A_T to compensate for the 22bp insertion. The new splice site falls just before the insertion and restores the reading frame to create a nearly full-length protein (345 vs 358 amino acids).
d) Sequence alignment from both copies of the A_T genome and the D/D_T genome copies of the region containing the insertion (open box in a, b, and c). mRNA1 represents the putative mRNA from the D genome copies and the A_T duplicate copy, and mRNA2 represents the putative mRNA from the A_T syntenic copy. The grey box denotes the stop codon that would occur if the A_T syntenic copy follows the splicing depicted by mRNA1.
e) The resulting protein for the boxed region from each line drawing above.

Analysis of potential transposable elements and intergenic space

Differential accumulation of transposable elements was evaluated for the four genomes (Table 3). All four genomes share a LINE element (LINE1, approximately 4 kb, contains a second LINE insertion in the A/A_T genomes), a *copia*-like pol fragment (820 bp), and possible long terminal repeats (LTRs) of an ancient retroelement, representing transposable element insertions that occurred prior to or concurrent with the origin of the genus (and hence are not relevant to genome size evolution within the genus). Two TEs in the D/D_T genomes (one shared, one

unique) and six TEs in the A/A_T genomes (five shared, one unique) differentiate the region, in concordance with global differences in genome size.

The D and D_T genomes share a *gypsy* element insertion of approximately 5 kb in length (*gypsy*D2; 4.8 kb in D and 5.1 kb in D_T; between genes 2 and 3 of Fig. 2). The LTRs are approximately 98% and 97% identical (excluding gaps) in D and D_T, respectively. Of the 11 mutations in the LTRs, 10 have occurred since the D/D_T divergence (3 mutations in D and 7 in D_T), indicating the element likely inserted just prior to that divergence, approximately 1-2 Ma. The D genome also has a unique *copia* (*copia*D; Tnt-94-like) insertion of approximately 2.3 kb (between genes 5 and 6, Fig. 2), whose 420 bp LTRs are 96% identical (excluding gaps). The small size of the element indicates possible decay or internal deletions.

The A and A_T genomes share two LINE elements (LINE1b and LINE2), apart from the one shared with the D/D_T genomes, each approximately 6.3 kb. LINE1b occurs within the LINE1 element shared by all four genomes (between genes 2 and 3, Figure 2). This element is 98.7% identical (excluding gaps) and contains a 2.3 kb insertion of a repetitive sequence of unknown type. The second A/A_T LINE element, LINE2 (between genes 3 and 4, Fig. 2), exhibits 98.5% sequence identity (excluding gaps) between these two genomes.

The A and A_T genomes share three discrete *gypsy* elements and an undetermined number of fragmented elements found in a large "*gypsy* landing pad". The three discrete *gypsy* elements include one full-length element, one truncated by the end of the A genome BAC but presumed to be full-length (*gypsy*A4; possibly *Gorge1* or *Gorge3*), and one full-length element in A that is represented by only a solo-LTR in A_T (*gypsy*A6; possibly *Gorge3*). Characterization of the particular family to which each element belongs was made possible by a larger survey of cotton repetitive

sequences (Hawkins *et al.*, 2006). The range in full-length element size is from 8.1 kb to 16.7 kb, and all LTRs are approximately 95% identical (within elements). For the three discrete *gypsy* elements examined, the orthologous elements in the A_T genome were smaller than their A genome counterparts by a minimum of 20%. In addition, the A_T genome contains a 4.8 kb unique LTR-retrotransposon of probable *gypsy* origin (*gypsy*A2). Overall, aside from being more abundant in the A genomes, intact *gypsy* elements were larger than those found in the D genomes. The largest *gypsy* represented in the D genomes was still smaller than the smallest intact *gypsy* in the A genomes and less than half the size of the largest (7.5 kb in D versus 8.1 kb and 16.7 kb in A/A_T).

In intergenomic comparisons, the divergence between transposable elements, which were identical at the point of divergence, ranged from approximately 9% (when comparing either A genome to either D genome) to approximately 1% in the LINEs shared by the A-A_T genome. The two TEs shared between either A versus either D genome and the three TEs shared between D and D_T showed less than 1% variation in sequence divergence between the different elements, whereas the seven shared TEs between A and A_T varied 2.5% in sequence divergence, from 1.2% -3.6% divergence. These values were invariably larger than when comparing unassigned intergenic space between genomes, most likely due to a combination of factors, including conserved regulatory elements in the unassigned intergenic space and the rapid mutation of TE sequences (SanMiguel and Bennetzen, 1998). The divergence of the unassigned intergenic space between the diploid and polyploid genomes closely mirrored the values obtained from 48 nuclear genes in *Gossypium* [0.008 intergenic versus 0.007 nuclear (Senchina *et al.*, 2003) A-A_T; 0.0142 intergenic versus 0.010 nuclear (Senchina *et al.*, 2003) D-D_T]. The divergence of the unassigned intergenic space between the A and D genomes was nearly identical, regardless of which A and D genome were compared (0.058 for A-D, A-D_T, A_T-D_T

and 0.059 for A_T-D), and these values were over double the divergence calculated from nuclear genes (0.022 A-D and 0.024 A_T-D_T; (Senchina *et al.*, 2003)), possibly indicating the presence of previously (and perhaps currently) rapidly evolving, severely degraded TEs that are unrecognizable.

INTRASTRAND HOMOLOGOUS RECOMBINATION: The *AdhA* region was evaluated for the hallmark of intrastrand homologous recombination, namely, solo-LTRs. A single solo-LTR (see above) was detected in the A_T genome, reducing a 10 kb *gypsy* element in the A genome to a single 2.6 kb LTR, a reduction of 74%. The solo-LTR belongs to a group of *gypsy* elements (*Gorge3*) shown elsewhere (Hawkins *et al.*, 2006) to have recently expanded in certain *Gossypium* lineages.

UNIDENTIFIED REPETITIVE DNA: Repetitive sequences not assigned to a class were uncovered through BLAST identity to repetitive whole-genome shotgun sequences of unknown origin. These did not substantially contribute to the alignment, representing approximately 5.7 kb and 4.4 kb in A/A_T and D/D_T, respectively.

INTRON SIZE BIAS*:* The predicted genes were evaluated for possible bias in intron size that correlates with genome size (Wendel *et al.*, 2002). The four shared genes contained introns that ranged in size from 684 bp to 3494 bp. There was no significant difference between introns from either polyploid genome versus its progenitor diploid (9 bp and 13 bp for A/A_T and D/D_T, respectively); however, unlike previous reports for intron size in *Gossypium* (Wendel *et al.*, 2002; Grover *et al.*, 2004), there was an substantial difference (approximately 350 bp) in comparing the A genomes to the D genomes. This difference is mainly due to the 3'-most intron of a single gene, the predicted protein disulfide isomerase encoding gene (Table 1). As previously reported for *Gossypium*, no other gene shows significant intron size variation.

SMALL SCALE INSERTIONS: The data were evaluated for possible evidence of pseudogene formation and organellar transfer to the nucleus, other mechanisms that may contribute in a minor way to genome size evolution. No unshared pseudogenes were detected, save for the potentially recently pseudogenized integral membrane protein encoding gene discussed above, and no organellar transfers were detected.

Discussion

Mechanisms of genome evolution in the AdhA *region*

In an earlier analysis of BAC sequences surrounding the *CesA* region in the A_T and D_T genomes of tetraploid cotton (Grover et al., 2004), the most striking conclusion was that this region revealed no evidence of the two-fold size difference that characterizes these genomes. In addition, not only was the genic portion highly conserved, but intergenic regions were also more highly conserved than in comparable studies in other plant groups, most notably from models from the grasses (Chen *et al.*, 1997; Chen *et al.*, 1998; Ramakrishna *et al.*, 2002; SanMiguel *et al.*, 2002; Wicker *et al.*, 2003). Based on these observations, Grover et al. (2004) concluded that the mechanisms that underlie the two-fold difference in genome size operate heterogeneously among genomic regions, leaving some regions relatively unchanged while more dynamically affecting others. In the present study we confirm and extend these earlier conclusions, and in addition provide glimpses into the modes and mechanisms that on a local scale generate the global patterns.

A primary difference between the present and earlier studies is that unlike the *CesA* region, the *AdhA* region mirrors, within the span of just over 100 kb, the twofold overall size difference that characterizes the 1697 and 885 Mbp genomes of the A and D genome lineages. In accordance with other plant systems and the repeat

analysis of whole genome shotgun libraries of the *Gossypium* genus (Hawkins *et al.*, 2006), the primary force responsible for the size difference between the A and D genomes in the *AdhA* region was differential accumulation of *gypsy* transposable elements. Accumulation of *gypsy* elements in each genome accounts for >32.7kb, 25.3kb, 5.1kb, and 7.1kb in the A, A_T, D_T, and D genomes, respectively. Thus, as expected based on studies in other groups (SanMiguel and Bennetzen, 1998; Bennetzen, 2002; Kidwell, 2002; Ramakrishna *et al.*, 2002), differential TE accumulation appears to account for a large fraction of genome size evolution.

In addition to genome expansion via TE activity, genomes may contract via several different mechanisms, including intrastrand homologous recombination, illegitimate recombination, and biased distribution of insertions and deletions. With respect to the former, homologous recombination between the LTRs of a single or adjacent retrotransposable elements leaves characteristic footprints in the form of solo-LTRs (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Shirasu *et al.*, 2000; Devos *et al.*, 2002; Vitte and Panaud, 2003). For genomes with relatively poorly characterized LTR-retrotransposon data, many solo-LTRs may go undetected; however, the comparative approach, as used here, provides a more robust means of identifying solo-LTRs. In the present comparison, a single solo-LTR was detected in the A_T genome through comparison to the A genome. This recombination event represents a significant reduction in the overall TE length for the A_T genome, accounting for over half of the total difference.

Illegitimate recombination has been demonstrated to have a profound effect counteracting genome size expansion in certain plants (Devos *et al.*, 2002; Ma *et al.*, 2004), and has been suggested to have influenced *Gossypium* genomes (Grover *et al.*, 2004). Although the present study was able to polarize only a small number of indels as insertions or deletions via illegitimate recombination, a substantial body of

unpolarized sequence data reveals the hallmarks of illegitimate recombination, particularly in the A_T genome. The gaps represented by these events contribute, in a large part, to the total *gypsy* element length difference between A and A_T.

A bias in the formation of small indels has been implicated in genome size differences (Petrov *et al.*, 1996; Kirik *et al.*, 2000; Petrov *et al.*, 2000; Petrov, 2002; Orel and Puchta, 2003), but has not been demonstrated to date for cotton (Grover *et al.*, 2004). The limited polarized indel data available indicate a possible insertional bias which suggests that the A genome is more prone to insertions than the other genomes and that it is the only genome where small insertions outweigh small deletions. Furthermore, the polarized deletions suggest that a deletional bias exists between A/A_T and D/D_T, with small deletions occurring more frequently and of greater average length in the D genomes. The polarized indels represent insertion and deletion events occurring since polyploid formation and, when extrapolated to the entire genome, indicate that a bias in small indels could be responsible for adding several hundred kb to the A genome and removing several hundred kb (in increasing amounts) from the A_T, D, and D_T genomes in the last 1-2 my. A larger data set of polarized indels, involving more genomic regions and additional outgroups such that events distinguishing diploid genomes may be polarized, is required to confirm the link to genome size evolution suggested here. We do point out, though, that the deletional bias is mirrored in the distribution of unpolarized gaps between the four genomes. The A genome had approximately 2-fold fewer unpolarized gaps then the D genome, representing a propensity for insertions in A, deletions in D, or, a combination of these two processes, as reflected in the polarized gap data.

While the polarized and unpolarized gap data suggest an indel bias exists in *Gossypium*, this bias cannot currently be described as acting homogeneously in all

genomic regions. In particular we note that in our previous study involving the *CesA* region (Grover *et al.*, 2004), comparative sequencing of ~100 kb found the distribution of indels, with respect to size and frequency, to be equivalent for the A_T and D_T genomes. Thus, the mechanisms involved in generating the indel bias in *Gossypium* do not act homogeneously among genomic regions, but instead appear to be affected by regional dynamics. Certain mechanisms that have the ability to generate small indels, such as illegitimate recombination, may be modulated by locally operating genomic forces such as recombination rate or degree chromatin condensation, thus possibly explaining a locally operating indel bias.

Genome evolution in polyploid cotton

Polyploid formation is known to be accompanied by myriad genomic and genetic alterations, which have been the subject of a number of recent reviews (Adams and Wendel, 2005; Chen and Ni, 2006). Evidence suggests that polyploid genomes need not be additive with respect to parental genome sizes, but instead are often slightly less than the combined parental genome size (Soltis and Soltis, 1999; Ozkan *et al.*, 2003; Bennett and Leitch, 2005b). To date, there is little information on the dynamics of genomic down-sizing in polyploid genomes (Chantret *et al.*, 2005; Gu *et al.*, 2006).

A conclusion of the present study is that in the *AdhA* region there has been genomic down-sizing in the polyploid relative to its diploid progenitors. Of 121 “small” gaps in the alignment, a greater number were in A_T than in A (64 vs. 28; $p < 0.0002$) as well as in D_T than in D (64 vs. 56), though in the latter comparison the difference is not statistically significant. In addition, the total amount of sequence attributable to transposable elements in the BACs from the polyploid was less than the sum from the homologous regions in the diploid progenitors. This was primarily due to the insertion of a unique *copia* element in the D genome, but was counteracted in the A_T

genome by a unique *gypsy* insertion. Excluding the unique *gypsy* insertion, the solo-LTR, and the region of the third *gypsy* truncated by the end of the A genome BAC, the total length of *gypsy* elements in A_T remains only approximately 65% the length of the A genome *gypsy* elements. This is largely due to several large gaps in the A_T *gypsy* elements, many of which had the hallmarks of illegitimate recombination. This mirrors the results of several studies in wheat, which suggest that the evolution of genomic structures observed in polyploid wheats are largely due to the opposing influences of insertions caused by TE activity and deletions mediated through illegitimate recombination (Chantret *et al.*, 2005)(Gu *et al.*, 2006). Taken together, these studies suggest that increased illegitimate recombination may be a general consequence of polyploidization. Additional studies of *Gossypium* as well as other plant polyploids will be necessary to test the generality of this conclusion.

Finally, the present study provides an example of pseudogenization following polyploid formation in cotton, a rare fate for genes duplicated by polyploidy in the cotton genome (Cronn *et al.*, 1999). A mutation in the D_T copy of the integral membrane protein-encoding gene caused a premature stop codon to arise halfway through the coding region, resulting in a truncated protein (182aa versus 368aa). Interestingly, this pseudogene was not the only one uncovered in the region. An ancient myosin pseudogene was shared between all genomes, and the original caffeic acid encoding gene in A_T (Table 1, gene 7) may also be silenced as a pseudogene (versus possessing an altered intron/exon structure for the last intron/exon junction). Nonetheless, the pseudogene discovered here adds a genomic example of gene silencing to an accumulating data set demonstrating expression-level changes and subfunctionalization of duplicated genes in *Gossypium* polyploids (Adams et al., 2003, 2004; Udall et al., 2006).

Experimental Procedures

BAC library screening and BAC selection

Three *Gossypium* BAC libraries (Tomkins *et al.*, 2001) were screened, as previously reported (Grover *et al.*, 2004), for clones containing the gene encoding alcohol dehydrogenase A. This gene was previously isolated and sequenced from A- and D-genome diploid cottons, as well as both genomes of polyploid cotton (Small *et al.*, 1999), which facilitated identification of the genomic origin of each BAC. PCR and sequencing were used to verify the presence of *AdhA* and, in the case of *G. hirsutum*, to determine which homoeolog of the tetraploid (A_T or D_T) was represented by each BAC screened. The largest clone from the A_T genome was sequenced to completion first. Following contig assembly, candidate A, D, and D_T BACs were evaluated for maximal overlap with the sequenced A_T BAC via PCR screening of inferred genes from various positions along the contig. BACs from the A, D, and D_T libraries that shared the most PCR markers were selected for sequencing.

Shotgun sequencing, assembly and analysis

E. coli genomic DNA free BAC plasmid DNA was sheared using a HydroShear (GeneMachines) DNA shearing device at speed code 12 with 25 cycles at room temperature. Fragmented DNA was end repaired using the 'End-it' DNA end repair kit (Epicentre), separated on an agarose gel, and size-selected for a range of 2 – 6Kb. This prepared insert DNA was randomly cloned into a pBluescript II KS+ vector (Strategene) and sequenced with the universal vector primers T7 and T3 to an average depth of 8x. The resulting sequences were base-called using the program Phred (Ewing and Green, 1998; Ewing *et al.*, 1998), vector sequences were removed by CROSS_MATCH (Ewing and Green, 1998; Ewing *et al.*, 1998), and assembled by the program Phrap (Green, 1999). Contigs were visualized and edited with CONSED (Gordon *et al.*, 1998). The output from three *ab initio* gene prediction programs, FGENESH (http://www.softberry.com/), GENEMARK.HMM

(Lukashin and Borodovsky, 1998), and GENSCAN+ (Burge and Karlin, 1997), was used as input for BLASTP(Altschul *et al.*, 1997) searches against the non-redundant GenBank protein database. In addition, 500bp segments of the sequence were subjected to BLASTX queries against the non-redundant GenBank protein database and BLASTN queries against the cotton EST database (Udall *et al.*, 2006). Repetitive element prediction was accomplished through RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html), CENSOR (Jurka *et al.*, 1996), and BLAST identity to known elements in RepBase (version 8.5) (Jurka, 2000) and GenBank. Each BAC was again queried in 500 bp fragments against whole-genomic shotgun (WGS) sequences representing approximately 0.1% of the each of four cotton genomes to uncover repetitive sequences of unknown origin (Hawkins *et al.*, 2006).

Alignment of the homologous BACs to each other was accomplished using Multi-LAGAN (Brudno *et al.*, 2003) with the input tree of ((A A_T) (D D_T)) and *Arabidopsis* repeatmasking. The resulting alignment was checked manually for errors using BIOEDIT (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Gap polarization

Polarization of indels as either insertions or deletions is necessary to evaluate possible bias in indel directionality and for comparisons of bias among genomes. Sequence from an outgroup is the best method for determining the ancestral state and polarizing indels; however, when outgroup sequence is unavailable, phylogenetics provides the capacity to polarize a fraction of the indels. For this comparison, any indel that occurred subsequent to the divergence of the diploid and polyploid genomes can be polarized as an insertion or deletion. That is, if three of the genomes share sequence where the fourth has a gap, the shared state is assumed to be ancestral and a deletion is assigned to the genome with the gap.

Likewise, if three of the genomes share a gap where the fourth has sequence, an insertion is assigned to that genome. For indels that are shared by only two genomes, polarization requires an outgroup.

Acknowledgements

We thank Trent Grover, Jamie Estill, and Jordan Swanson for technical assistance and Jennifer Hawkins for helpful discussion. This work was funded by the National Science Foundation Plant Genome program, whose support we gratefully acknowledge.

References

Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution. Curr Opin Plant Biol., 8, 135-141.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389-3402.

Bennett, M.D., Leitch, I.J. and Hanson, L. (1998) DNA amounts in two samples of angiosperm weeds. Ann Bot, 82, 121-134.

Bennett, M.D. and Leitch, I.J. (2005a) Plant DNA C-values Database (release 4.0, October 2005). http://www.rbgkew.org.uk/cval/homepage.html.

Bennett, M.D. and Leitch, I.J. (2005b) Genome size evolution in plants. In The evolution of the genome (Gregory, T.R., ed. San Diego: Elsevier, pp. 89-162.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. Plant Molecular Biology, 42, 251-269.

Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica, 115, 29-36.

Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of Recent Genome Size Variation in Flowering Plants. Ann Bot, 95, 127-132.

Brudno, M., Do, C., Cooper, G., Kim, M.F., Davydov, E., Green, E.D., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large scale multiple alignment of genomic DNA. Genome Research, 13, 721-731.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268, 78-94.

Cavalier-Smith, T. (2005) Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. Ann Bot, 95, 147-175.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.-F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P. and Chalhoub, B. (2005) Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). Plant Cell, 17, 1033-1045.

Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A. and Bennetzen, J.L. (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. Proceeding of the National Academy of Science, 94, 3431-3435.

Chen, M., SanMiguel, P. and Bennetzen, J.L. (1998) Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. Genetics, 148, 435-443.

Chen, Z.J. and Ni, Z. (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays, 28, 240-252.

Cronn, R., Small, R.L. and Wendel, J.F. (1999) Duplicated genes evolve independently following polyploid formation in cotton. Proc. Natl. Acad. Sci. USA, 96, 14406-14411.

Cronn, R.C., Small, R.L., Haselkorn, T. and Wendel, J.F. (2002) Rapid diversification of the cotton genus (Gossypium: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany, 89, 707-725.

Deutsch, M. and Long, M. (1999) Intron-exon structure of eukaryotic model organisms. Nuc. Acids Res., 27, 3219-3228.

Devos, K.M., Brown, J. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Research, 12, 1075-1079.

Ellegren, H. (2002) Mismatch repair and mutational bias in microsatellite DNA. Trends in Genetics, 18, 552.

Endrizzi, J.D., Turcotte, E.L. and Kohel, R.J. (1985) Genetics, cytology, and evolution of Gossypium. Advances in Genetics, 23, 271-375.

Ewing, B. and Green, P. (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research, 8, 186-194.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequences traces using phred. I. Accuracy assessment. Genome Research, 8, 175-185.

Flavell, R.B., Bennett, M.D., Smith, J.B. and Smith, D.B. (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochemical Genetics, 12, 257-269.

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. Genome Research, 8, 195-202.

Green, P. (1999) Phrap documentation. http://www.phrap.org/phrap.docs/phrap.html.

Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. Camb. Philos. Soc., 76, 65-101.

Gregory, T.R. (2005) The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. Ann Bot, 95, 133-146.

Gregory, T.R. (2006) Animal genome size database. http://www.genomesize.com.

Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H. and Wendel, J.F. (2004) Incongruent Patterns of Local and Global Genome Size Evolution in Cotton. Genome Res., 14, 1474-1482.

Gu, Y.Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G.R., Huo, N., Belcram, H., Ravel, C., Charmet, G., Charles, M., Anderson, O.D. and Chalhoub, B. (2006) Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes. Genetics, 174, 1493-1504.

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

Hendrix, B. and Stewart, J.M. (2005) Estimation of the Nuclear DNA Content of *Gossypium* Species. Ann Bot %R 10.1093/aob/mci078, 95, 789-797.

Jurka, J., Klonowski, P., Dagman, V. and Pelton, P. (1996) CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry, 20, 119-122.

Jurka, J. (2000) Repbase Update: a database and an electronic journal of repetitive elements. Trends in Genetics, 9, 418-420.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A. (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceeding of the National Academy of Science, 97, 6603-6607.

Kidwell, M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica, 115, 49-63.

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

Knight, C.A., Molinari, N.A. and Petrov, D.A. (2005) The Large Genome Constraint Hypothesis: Evolution, Ecology and Phenotype. Annals of Botany, 95, 177-190.

Lukashin, A. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. Nuc Acids Res, 26, 1107-1115.

Ma, J., Devos, K.M. and Bennetzen, J.L. (2004) Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. Genome Res., 14, 860-869.

Marchler-Bauer, A. and Bryant, S. (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Research, 32, 327-331.

Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics, 30, 194-200.

Orel, N. and Puchta, H. (2003) Differences in the processing of DNA ends in Arabidopsis thaliana and tobacco: possible implications for genome evolution. Plant Molecular Biology, 51, 523-531.

Ozkan, H., Tuna, M. and Arumuganathan, K. (2003) Nonadditive changes in genome size during allopolyploidization in the wheat group (Aegilops-Triticum) group. Journal of Heredity, 94, 260-264.

Petrov, D. (1997) Slow but steady: Reduction of genome size through biased mutation. Plant Cell, 9, 1900-1901.

Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. (1996) High intrinsic rate of DNA loss in Drosophila. Nature, 384, 346-349.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L. and Shaw, K.L. (2000) Evidence for DNA loss as a deteminant of genome size. Science, 287, 1060-1062.

Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. Trends in Genetics, 17, 23-28.

Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. Theoretical Population Biology, 61, 531-544.

Petrov, D.A. and Wendel, J.F. (2006) Genome evolution in eukaryotes: the genome size perspective. In Evolutionary Genetics: Concepts and Case Studies (Fox, C.W. and Wolf, J.B., eds): Oxford University Press, pp. 144-156.

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

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

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science, 274, 765-768.

SanMiguel, P. and Bennetzen, J.L. (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany, 82, 37-44.

SanMiguel, P., Ramakrishna, W., Bennetzen, J.L., Busso, C.S. and Dubcovsky, J. (2002) Transposable elements, genes, and recombination in a 215kb contig from wheat chromosome $5A^m$. Funct. Integr. Genomics, 2, 70-80.

Seelanan, T., Schnabel, A. and Wendel, J.F. (1997) Congruence and consensus in the cotton tribe. Syst. Bot., 22, 259-290.

Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J.K., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A. and Wendel, J.F. (2003) Rate variation among

nuclear genes and the age of polyploidy in *Gossypium*. Molecular Biology and Evolution, 20, 633-643.

Shahmuradov, I.A., Akbarova, Y.Y., Solovyev, V.V. and Aliyev, J.A. (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol. Biol., 52, 923-934.

Shepherd, N.S., Schwarz-Sommer, Z., Blumberg vel Spalve, J., Gupta, M., Wienand, U. and Saidler, H. (1984) Similarity of the Cin1 repetitive family of *Zea mays* to eukaryotic transposable elements. Nature, 307, 185-187.

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

Small, R.L., Ryburn, J.A., Cronn, R.C., Seelanan, T. and Wendel, J.F. (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogenetic reconstruction in a recently diverged plant group. American Journal of Botany, 85, 1301-1315.

Small, R.L., Ryburn, J.A. and Wendel, J.F. (1999) Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Molecular Biology and Evolution, 16, 491-501.

Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol. Evol., 9, 348-352.

Thomas, C.A. (1971) The genetic organisation of chromosomes. Annual Review of Genetics, 5, 237-256.

Tomkins, J.P., Peterson, D.G., Yang, T.J., Main, D., Wilkins, T.A., Paterson, A.H. and Wing, R.A. (2001) Development of genomic resources for cotton (*Gossypium hirsutum* L.): BAC library construction, preliminary STC analysis, and identification of clones associated with fiber development. Mol. Breed., 8, 255-261.

Udall, J.A., Swanson, J.M., Haller, K., Rapp, R.A., Sparks, M.E., Hatfield, J., Yu, Y., Wu, Y., Dowd, C., Arpat, A.B., Sickler, B.A., Wilkins, T.A., Guo, J.Y., Chen, X.Y., Scheffler, J., Taliercio, E., Turley, R., McFadden, H., Payton, P., Klueva, N., Allen, R., Zhang, D., Haigler, C., Wilkerson, C., Suo, J., Schulze, S.R., Pierce, M.L., Essenberg, M., Kim, H., Llewellyn, D.J., Dennis, E.S., Kudrna, D., Wing, R., Paterson, A.H., Soderlund, C. and Wendel, J.F. (2006) A global assembly of cotton ESTs. Genome Research, 16, 441-450.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., E., N. and Schulman, A.H. (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. The Plant Cell, 11, 1769-1784.

Vinogradov, A.E. (1999) Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol., 49, 376-384.

Vinogradov, A.E. (2003) Selfish DNA is maladaptive: evidence from the plant Red List. Trends Genet, 19, 609-614.

Vitte, C. and Panaud, O. (2003) Formation of Solo-LTRs Through Unequal Homologous Recombination Counterbalances Amplifications of LTR Retrotransposons in Rice *Oryza sativa* L. Mol Biol Evol, 20, 528-540.

Wendel, J.F. (2000) Genome evolution in polyploids. Plant Molecular Biology, 42, 225-249.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L. and Senchina, D.S. (2002) Intron size and genome size in plants. Molecular Biology and Evolution, 19, 2346-2352.

Wendel, J.F. and Cronn, R.C. (2003) Polyploidy and the evolutionary history of cotton. Adv. Agron., 78, 139-186.

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

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.-D., Dubcovsky, J. and Keller, B. (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell, 15, 1186-1197.

Zhang, J. (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18, 292-298.

Supplementary Table 1: Types and frequency of mechanisms contributing to genome size change in the *AdhA* region

Mechanism	Type	*G.herbaceum*	*G hirsutum*A	*G.hirsutum*D	*G.raimondii*
Overall	# deletions	4	13	20	13
	nt deletions	8	29	77	91
	# insertions	12	2 (1)	6	6 (4)
	nt insertions	34	4799 (1)	11	5971 (4)
	# unknown gaps (excluding TEs)	159	169	154	152
	nt missing (excluding TEs)	7809	21715	15440	15404
illegitimate recombination	# deletion events	0	2	1	1
	nt deleted	0	12	10	14
	# insertion events	1	0	0	1
	nt inserted	13	0	0	3619
	# unclear events	22		8	
	nt unclear	3432	14384	697	697
slipstrand	# deletion events	2	6	8	4
	nt deleted	2	10	10	4
	# insertion events	8	1	6	3
	nt inserted	12	1	11	3
	# unclear events; A/At, D/Dt	42		30	
	nt unclear	64	116	39	43
	# unclear events; complex	18 (8 involve A, 17 involve A_T, 15 involve D_T, 13 involve D)			
	nt unclear	47	71	77	63
slipstrand or IR/SSA	# deletion events	0	0	3	2
	nt deleted	0	0	10	35
	# insertion events	1	0	0	0
	nt inserted	7	0	0	0
	# unclear events; A/At, D/Dt	11		5	
	nt unclear	37	84	46	71
	# unclear events; complex	2 (only 1 involves *G. raimondii*)			
	nt unclear	11	16	2572	2564
Unknown mechanism	# deletion events	2	5	8	6
	nt deleted	6	7	47	38
	# insertion events	2	0	0	1
	nt inserted	2	0	0	1
	# unclear events	60		62	
	nt unclear	3999	6779	6025	5985
	# unclear events; complex	5 (A_T,D_T,D; A_T,D_T,D; A_T,A_T,D_T,D; A,D_T,D; A,A_T,D)			
	nt unclear; complex	16	62	3920	3918
	# unclear events (within poor alignment)	11		28	
	nt unclear (within poor alignment)	203	203	2064	2063
Transposable elements	# polarized insertions	0	1	0	1
	nt polarized insertions	0	4799	0	2348
	# probable insertions	3+		1	
	nt probable insertions	47448*	35336*	5103	4758

* minimum estimates

\+ numbers in parantheses refer to the number and length of insertions excluding large insertions (>400bp)

CHAPTER FIVE

A PHYLOGENETIC ANALYSIS OF INDEL DYNAMICS IN THE COTTON GENUS

A paper prepared for submission to *Molecular Biology and Evolution*

Corrinne E. Grover[1], Rod A. Wing[2], Andrew H. Paterson[3], Jonathan F. Wendel[4]

Abstract

Genome size evolution is a dynamic process composed of counterbalancing mechanisms whose actions vary throughout the genomic landscape, across lineages, and over time. While the primary mechanism of expansion, transposable element (TE) amplification, has been widely documented, the evolutionary dynamics of genome contraction have been less thoroughly explored. To evaluate the relative impact and evolutionary stability of the mechanisms that affect genome size, we conducted a phylogenetic analysis of indel rates for two genomic regions in four *Gossypium* genomes: the two co-resident genomes (A_T and D_T) of tetraploid cotton and its model diploid progenitors, *Gossypium arboreum* (A) and *Gossypium raimondii* (D). From this analysis, we determined the rates of sequence gain or loss along each branch, partitioned by mechanism, and how these changed during species divergence. In general, there has been a propensity toward growth of the diploid genomes and contraction in the polyploid. Most of the size difference between the diploid species occurred prior to polyploid divergence, and was largely attributable to TE amplification in the A/A_T genome. After separating from the polyploid genomes, both diploid genomes experienced slower sequence gain than their respective ancestors, due to fewer TE insertions in the A genome and a combination of increased deletions and decreased insertions in the D genome. Both genomes of the polyploid displayed increased rates of deletion and decreased rates

[1] Graduate student, primary research and author, EEOB department, Iowa State University
[2] Director, BAC sequencing and finishing, Arizona Genomics Institute, University of Arizona
[3] Professor and director, BAC selection and hybridizations, Plant Genome Mapping Laboratory, University of Georgia
[4] Principal investigator and corresponding author, EEOB department, Iowa State University

of insertion, leading to a rate of near stasis in D_T and overall contraction in A_T resulting in polyploid genome contraction. As expected, TE insertions contributed significantly to the genome size differences; however, several conclusions were drawn for the other mechanisms of change. Intra-strand homologous recombination, although rare, had the most significant impact on the rate of deletion when present. Small indel data for the diploids suggest the possibility of a bias, as the smaller genomes tend to add less or delete more sequence through small indels than the larger genomes, whereas data for the polyploid suggests increased sequence turnover in general (both as small deletions and small insertions). Illegitimate recombination, although not demonstrated to be a dominant mechanism of change, did experience a biased shift in the polyploid toward deletions, which may provide a partial explanation of polyploid genomic downsizing.

Introduction

In recent years there has been considerable interest in the evolutionary forces and mechanisms that underlie the extraordinary genome size variation observed within and among various groups of organisms. The primary mechanism of genome expansion, transposable element (TE) amplification, has been documented in broad surveys across angiosperm lineages (SanMiguel and Bennetzen 1998; Vitte and Bennetzen 2006) and within genera or closely related species (Hill et al. 2005; Hawkins et al. 2006; Piegu et al. 2006; Petit et al. 2007); however, the evolutionary dynamics and primary mechanisms of genome size contraction have been less thoroughly explored. As mechanisms of deletion are more challenging to study, requiring orthologous sequence from closely related species, evidence for genome size contraction as a whole has been limited to phylogenetic inferences based on the placement of taxa having small genomes (Bennett and Leitch 1995; Leitch, Chase, and Bennett 1998; Wendel et al. 2002; Bennett and Leitch 2005a). Analyses of deletional mechanisms thought to be most important, i.e., intra-strand

homologous recombination and illegitimate recombination, have produced conflicting results concerning their relative importance and whether either can affect genome size as dramatically as TE proliferation (Wicker et al. 2001; Devos, Brown, and Bennetzen 2002; Wicker et al. 2003; Ma, Devos, and Bennetzen 2004; Vitte and Bennetzen 2006).

To evaluate the relative impact of mechanisms of genome size change and their evolutionary stability, we conducted a phylogenetic analysis of indels in two regions of the cotton (*Gossypium*) genome for which we had previously generated data for several species. By including orthologous sequence from the phylogenetic outgroup, *Gossypioides kirkii*, we were able to partition indels into insertions and deletions and study their relative rates, both overall and with respect to contributing mechanism. *Gossypium* is a 5-10 million year old (myo) genus whose genomes range nearly threefold in size, from 885 Mb in the New World diploids to over 2570 Mb in the Australian diploids (Hendrix and Stewart 2005). Early in the history of the genus, D-genome diploids and A-genome diploids diverged, subsequently acquiring a twofold difference in genome size (Figure 1). These divergent genomes later became reunited with allopolyploid formation approximately 1-2 myo, leading to extant allopolyploid species that have a genome size that is slightly less than the sum of their model diploid progenitors (Hendrix and Stewart 2005). *Gossypium* as a genus diverged from its closest extant relative, *Gossypioides kirkii*, approximately 15 myo; the latter species has the smallest genome of the species studied here (590 Mb).

Here we present perhaps the first phylogenetic analysis of indel rates in plants, using two BAC-sized genomic regions surrounding the genes cellulose synthase (*CesA*) (Grover et al. 2004) and alcohol dehydrogenase A (*AdhA*) (Grover et al. 2007), by studying two diploid species representing the closest living ancestors of

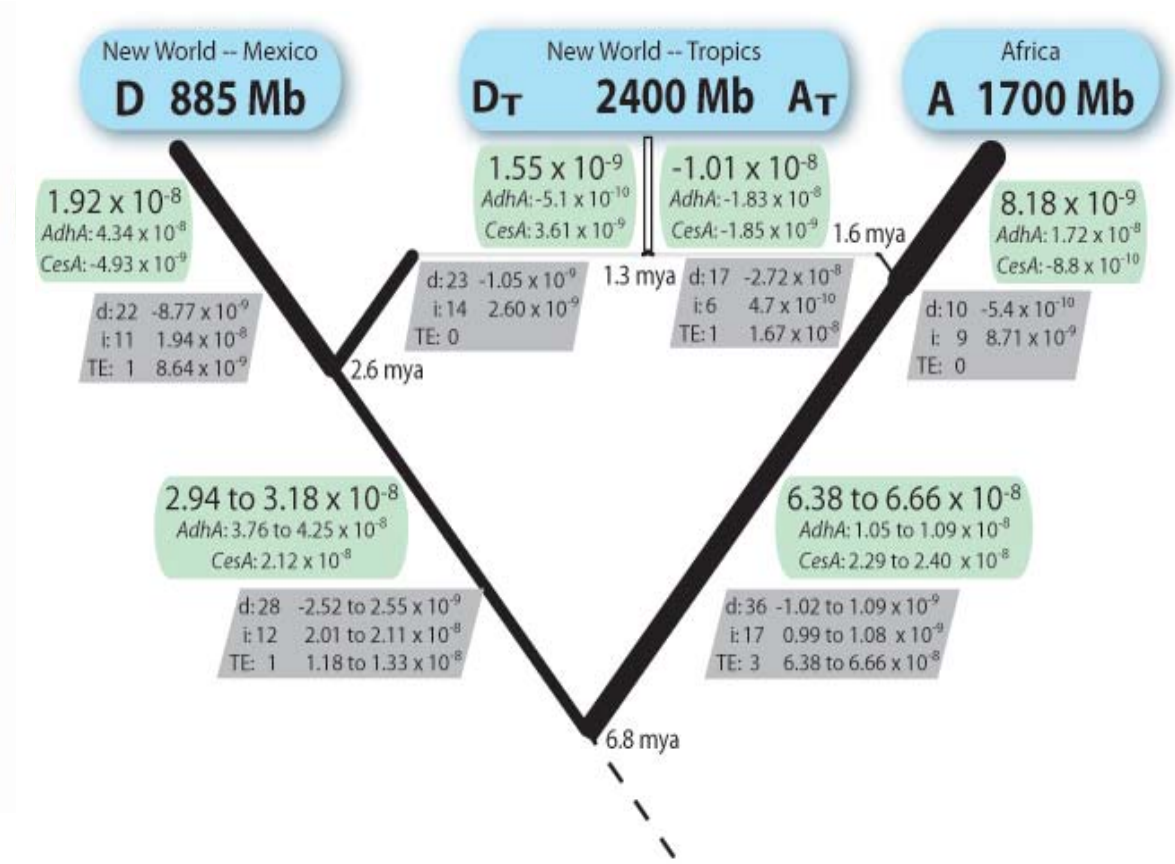


Figure 1: Evolutionary history of and rates of genome loss and gain in four *Gossypium* genomes. The evolutionary relationship and times of divergence between the model diploid progenitors for the A and D genomes (*Gossypium arboreum* and *G. raimondii*, respectively), the true parents to the polyploid, and their subsequent reunion in the polyploid (AD) are shown. Branch lengths reflect time, and branch thickness indicates change in genome size (solid denotes sequence gain; open indicates sequence loss). *Gossypium* diverged from the outgroup (*Gossypioides kirkii*, 1C=590Mb) approximately 10-15 mya and A- and D-genome cottons diverged from each other approximately 6.8 mya. The genome groups evolved independently for 5.2 and 4.2 my, respectively, before the model diploid progenitors diverged from the actual (and extinct) parents of the polyploid 1.6 and 2.6 mya for the A- and D-genomes, respectively. Approximately 1.3 mya, the A and D genomes were reunited in a polyploid nucleus, whose genome size is slightly less than the sum of the two model parents. Overall rates of genome size change are represented by the first line in the green boxes, while the individual regional rates are listed independently underneath. Rates of deletion (d), non-TE insertions (i), and TE insertions (TE) are also listed in the grey boxes.

polyploid *Gossypium*, both of its genomes, and the outgroup *Gossypioides kirkii*. We focus on the mechanisms that gave rise to insertions and deletions and their evolutionary dynamics. Using this quintet of genomes, the direction and timing of each indel (insertion or deletion; pre- or post-polyploidization) was determined, and the rate and direction of overall genomic change for each branch were calculated.

From this curated analysis among closely related species, we determined rates of sequence gain or loss along each branch and assessed rate change during species divergence.

Methods

BAC library screening and BAC selection

Gossypium arboreum, G. raimondii, and *Gossypioides kirkii* BAC libraries were screened, as previously reported (Grover et al. 2004), for clones containing the gene encoding cellulose synthase a1 (*CesA*), and several other predicted genes in the previously sequenced region (Grover et al. 2004). The same treatment was applied to the *G. kirkii* library with respect to the BAC containing the alcohol dehydrogenase A gene (*AdhA*) (Grover et al. 2007). The resulting positive clones for each marker were evaluated to facilitate selection of clones that provided maximum overlap with the previously generated BAC sequences. PCR and sequencing were used to verify the presence of the desired markers on the selected BACs prior to shotgun sequencing.

Shotgun sequencing, assembly and analysis

BAC plasmid DNA was sheared at room temperature using a HydroShear (GeneMachines) DNA shearing device at speed code 12 for 25 cycles. The resulting DNA fragments were end-repaired using the 'End-it' DNA end repair kit (Epicentre) and subsequently separated on an agarose gel for size selection (range 2 – 6Kb). These fragments were cloned into a pBluescript II KS+ vector (Strategene) and sequenced with universal vector primers (T7 and T3) to an average depth of 8x. Each sequence was base-called using the program Phred (Ewing and Green 1998; Ewing et al. 1998) and vector sequences were masked by CROSS_MATCH (Ewing and Green 1998; Ewing et al. 1998). Trimmed sequences

were assembled by the program Phrap (Green 1999), and contigs were visualized and edited with CONSED (Gordon, Abajian, and Green 1998).

The newly sequenced and previously published (*CesA* gi: AY632359-60; *AdhA* gi: EF457751-54) BACs were aligned using Multi-LAGAN (Brudno et al. 2003) with *Arabidopsis thaliana*-based repeat masking and with the input tree ((A A_T) (D D_T) Gk), where A and D refer to sequences from the diploids, A_T and D_T designate their counterparts in the allopolyploid G hirsutum, and Gk refers to the outgroup species. The resulting alignment was checked manually for errors using BIOEDIT (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Predicted features from the previously sequenced genomes (Grover et al. 2004; Grover et al. 2007) were mapped onto the new alignment, and novel sequence (i.e. sequence unique to the newly sequenced genomes) was analyzed as previously described (Grover et al. 2004; Grover et al. 2007).

Gap polarization and analysis

Indels were partitioned into insertions or deletions and phylogenetically placed using outgroup polarization. Thus, where sequence for *Gossypioides kirkii* existed, if two genomes (A/A_T or D/D_T) shared sequence with the outgroup, then the gap was considered a deletion that arose along the branch leading to the other two genomes during divergence of the A and D genome groups; an insertion during diploid divergence, prior to polyploid formation, was called when the outgroup shared a gap with either A/A_T or D/D_T. If three genomes shared sequence or a gap with the outgroup, then a deletion or insertion, respectively, was inferred to have occurred in the remaining genome after polyploid formation. If only one genome shared sequence or a gap with the outgroup and the other three existed in the opposite state, that event was labeled as unknown, since two separate events occurring in the outgroup and the genome sharing its state are equally parsimonious as the

opposite occurring separately in one genome post-polyploidization and in the pre-polyploidization lineage for the other two genomes. For regions where the outgroup lacked homologous sequence, only those gaps arising since polyploidization could be polarized. In this case, if three of the genomes share sequence while the fourth has a gap, the shared sequence is assumed to be plesiomorphic and the gap is characterized as a deletion; likewise, if three genomes share a gap while the fourth has intervening sequence, an insertion is inferred. Indels that occurred between A/A_T and D/D_T in regions without outgroup sequence were not polarized.

Number of nucleotides (nt) added, deleted, or missing were standardized to nt per year based on previously estimated organismal divergence times for the diploid divergence (6.8 my since A-D divergence (Cronn et al. 2002; Wendel and Cronn 2003) and polyploid formation, as estimated based on multiple nuclear gene sequences (Senchina et al. 2003). Because modern A genome diploids are a closer model of the actual A-genome donor to the polyploids than are D genome species, by about 50%, the divergence of the extant diploid species from the polyploid ancestor genomes was calculated. Specifically, branch lengths were apportioned based on the ratio of the diploid branch length over the total branch length (Senchina et al. 2003) to the calculated time since divergence over 6.8my (time since A-D divergence; Figure 1). These values were used to estimate rates of indel evolution along each branch of the phylogeny.

Mechanisms responsible for indel formation were hypothesized based on the sequence within and surrounding alignment gaps. Inserted sequence deemed by sequence homology to be transposable element in origin was considered the result of TE amplification; deletion via intra-strand homologous recombination was assumed when one genome shared a single LTR in an orthologous position with one or more other genomes, but no internal sequence or second LTR. Single nucleotide

gaps were classified into a category of the same name. Illegitimate recombination represents a group of mechanisms that often, but not always, are associated with short, direct repeats. Several molecular mechanisms are encompassed by illegitimate recombination, including double- stranded break repair and slipstrand mispairing. For the purpose of this study, illegitimate recombination was subdivided into three categories based on the hallmarks of the associated gap: (1) illegitimate recombination via double-stranded break repair, (2) illegitimate recombination via double-stranded break repair or slipstrand mispairing, and (3) illegitimate recombination via slipstrand mispairing. Double-stranded break repair was inferred when the gap was flanked by short (<15 nt), direct repeats; slipstrand mispairing was inferred when the sequence in the gap was directly repeated; and, in cases where a gap met both criteria, it was placed in a separate category.

Results

General description of the sequenced regions

A summary of the results for both regions for all genomes is in Tables 1 and 2. The *AdhA* region includes approximately twice as much aligned sequence in the A genomes as the D genomes (94kb in A and 89kb in A_T versus 52kb in D and 46kb in D_T), largely from multiple TE insertions in A/A_T during its divergence from D/D_T (see below and Table 1). Both genomes of the tetraploid are represented by less sequence than their diploid counterparts due to a combination of net growth in both of the diploids, as well as diminished growth in D_T and net loss in A_T (Table 1). The nearly twice as large sequence for the A and A_T genomes, as well as the overall contraction of the polyploid genome, is congruent with expectations due to genome size (A = 1697 Mbp; D = 885 Mbp; AD = 2401 Mbp, whereas A+D = 2582 Mbp).

As detailed in Grover et al. (2004), the *CesA* region in *Gossypium* is rather different from the *AdhA* region in that the representative sequence for the A genomes is only slightly larger than that of the D genomes (57kb in A and 51kb in A_T versus 49kb in

Table 1: Rates of insertions and deletions in the *AdhA* and *CesA* regions of the cotton genome*

		A/A_T		D/D_T		A		D		A_T		D_T	
		Events	Rate	Events	Rate	Events	Rate	Events	Rate	Events	Rate	Events	Rate
AdhA	Deleted	41	1.42 to 1.48 x 10^{-9}	16	3.7 to 4.2 x 10^{-10}	4	3 x 10^{-11}	10	5.0 x 10^{-10}	12	5.17 x 10-^{8}	14	5.8 x 10^{-10}
	Inserted	17	1.10 to 1.15 x 10^{-9}	11	1.44 to 1.63 x 10^{-10}	10	1.73 x 10^{-8}	7	2.66 x 10^{-8}	1	1 x 10^{-11}	7	7 x 10^{-11}
	TE insertions	5	1.05 to 1.10 x 10^{-7}	1	2.36 to 2.66 x 10^{-8}	0	0	1	1.78 x 10^{-8}	1	3.33 x 10^{-8}	0	0
	Unknown	30	3.9 to 4.1 x 10^{-10}	31	3.16 to 3.57 x 10^{-8}	20	1.06 x 10^{-9}	3	2.1 x 10^{-10}	30	9.77 x 10^{-9}	10	3.2 x 10^{-10}
	Rate	**1.05 to 1.09 x 10^{-8}**		**3.76 to 4.25 x 10^{-8}**		**1.72 x 10^{-8}**		**4.34 x 10^{-8}**		**-1.83 x 10^{-8}**		**-5.1 x 10^{-10}**	
	Range	**1.04 to 1.45 x 10^{-8}**		**0.60 to 4.29 x 10^{-8}**		**0.16 to 2.70 x 10^{-8}**		**4.32 to 4.37 x 10^{-8}**		**-2.81 to -1.73 x 10^{-8}**		**-8.3 to -3.0 x 10^{-10}**	
CesA	Deleted	31	6.1 to 6.9 x 10^{-10}	40	4.66 to 4.68 x 10^{-9}	15	1.04 x 10^{-9}	33	1.70 x 10^{-8}	22	2.77 x 10^{-9}	32	1.52 x 10^{-9}
	Inserted	16	0.88 to 1.00 x 10^{-9}	13	2.58 to 2.59 x 10^{-8}	7	1.6 x 10^{-10}	15	1.21 x 10^{-8}	11	9.2 x 10^{-10}	20	5.13 x 10^{-9}
	TE insertions	1	2.27 to 2.37 x 10^{-8}	0	0	0	0	0	0	0	0	0	0
	Unknown	46	3.16 to 3.59 x 10^{-9}	45	2.10 to 2.11 x 10^{-8}	5	1.06 x 10^{-8}	6	4.7 x 10^{-10}	3	4 x 10^{-11}	11	9 x 10^{-11}
	Rate	**2.29 to 2.40 x 10^{-8}**		**2.12 x 10^{-8}**		**-8.8 x 10^{-10}**		**-4.93 x 10^{-9}**		**-1.85 x 10^{-9}**		**3.61 x 10^{-9}**	
	Range	**1.98 to 4.50 x 10^{-8}**		**0.019 to 2.48 x 10^{-8}**		**-1.15 to -0.084 x 10^{-8}**		**-5.40 to -4.84 x 10^{-9}**		**-1.89 to 8.74 x 10^{-9}**		**3.52 to 4.08 x 10^{-9}**	
Average	Deleted	36	1.02 to 1.09 x 10^{-9}	28	2.52 to 2.55 x 10^{-9}	10	5.4 x 10-^{10}	22	8.77 x 10^{-9}	17	2.72 x 10^{-8}	23	1.05 x 10^{-9}
	Inserted	17	0.99 to 1.08 x 10^{-9}	12	2.01 to 2.11 x 10^{-8}	9	8.71 x 10^{-9}	11	1.94 x 10^{-8}	6	4.7 x 10^{-10}	14	2.60 x 10^{-9}
	TE insertions	3	6.38 to 6.66 x 10^{-8}	1	1.18 to 1.33 x 10^{-8}	0	0	1	8.64 x 10^{-9}	1	1.67 x 10^{-8}	0	0
	Unknown	38	1.78 to 2.00 x 10^{-9}	38	2.63 to 2.84 x 10^{-8}	13	5.83 x 10^{-9}	5	3.4 x 10^{-10}	17	4.91 x 10^{-9}	11	2.1 x 10^{-10}
	Rate	**6.38 to 6.66 x 10^{-8}**		**2.94 to 3.18 x 10^{-8}**		**8.18 x 10^{-9}**		**1.92 x 10^{-8}**		**-1.01 x 10^{-8}**		**1.55 x 10^{-9}**	
	Range	**6.20 to 9.50 x 10^{-8}**		**0.310 to 3.38 x 10^{-8}**		**0.235 to 1.31 x 10^{-8}**		**1.89 to 1.94 x 10^{-8}**		**-1.50 to -0.43 x 10^{-8}**		**1.30 to 1.85 x 10^{-9}**	

* Rates of insertion and deletion were standardized to nt per 100kb per my based on the sequenced length of the region and the time since divergence (5.2my for the A/A_T branch, 4.2my for D/D_T branches, 1.6my for A and A_T, and 2.6my for the D and D_T branches; Figure 1). Rates are parsed into contribution by deletions, insertions, and TE insertions, and the overall rate of gain or loss is represented. In addition, a range representing the minimum and maximum rate was calculated by considering the unknown indels of a branch pair (A/A_T and D/D_T, A, and A_T, D and D_T) to represent only deletions and only insertions, respectively. Gain is indicated by positive values and loss is indicated by negative values.

Table 2: Insertions and deletions by region and mechanism*

	AdhA	A/A_T			D/D_T				A			A_T			D			D_T		
		Events	nt per year		%	Events	nt per year		%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%
Deletions	SN	13	2.70E-11	2.80E-11	1.9%	10	4.50E-11	5.10E-11	12.3%	4	2.70E-11	100.0%	8	5.60E-11	0.1%	5	3.70E-11	7.4%	3	2.30E-11	4.0%
	IR	11	3.19E-10	3.33E-10	22.5%	2	1.36E-10	1.54E-10	37.1%				1	4.20E-11	0.1%	2	3.59E-10	71.9%	2	6.20E-11	10.8%
	IR/SS	4	1.80E-11	1.90E-11	1.3%											1	2.20E-11	4.4%	4	1.01E-10	17.6%
	SS	1	8.00E-12	9.00E-12	0.6%								1	1.40E-11	0.0%						
	Unknown	12	1.04E-09	1.09E-09	73.7%	4	1.86E-10	2.10E-10	50.6%				1	2.10E-11	0.0%	2	8.10E-11	16.2%	5	3.89E-10	67.7%
	LTR-recombination												1	5.15E-08	99.7%						
	Total	**41**	**1.41E-09**	**1.48E-09**		**16**	**3.67E-10**	**4.15E-10**		**4**	**2.70E-11**		**11**	**5.17E-08**		**10**	**4.99E-10**		**14**	**5.75E-10**	
Insertions	SN	7	1.36E-11	1.50E-11	0.0%	5	2.30E-11	2.60E-11	0.1%	4	2.70E-11	0.2%	1	7.00E-12	0.0%	4	2.90E-11	0.1%	5	3.90E-11	55.7%
	IR	4	1.02E-10	1.07E-10	0.1%	2	1.42E-08	1.61E-08	37.5%	2	1.71E-08	99.3%				1	2.65E-08	60.5%			
	IR/SS	2	3.90E-11	4.10E-11	0.0%	2	1.13E-10	1.28E-10	0.3%	4	8.70E-11	0.5%				2	4.40E-11	0.1%	2	3.10E-11	44.3%
	SS	1	7.94E-10	8.29E-10	0.7%	1	1.40E-11	1.50E-11	0.0%												
	Unknown	3	1.47E-10	1.54E-10	0.1%	1	2.30E-11	2.60E-11	0.1%												
	polarized retro	1	1.30E-08	1.36E-08	12.3%	1	2.36E-08	2.66E-08	62.0%							1	1.73E-08	39.4%			
	unpolarized retro	4	9.20E-08	9.60E-08	86.7%								1	3.33E-08	100.0%						
	Total	**22**	**1.06E-07**	**1.11E-07**		**12**	**3.80E-08**	**4.29E-08**		**10**	**1.73E-08**		**2**	**3.33E-08**		**8**	**4.39E-08**		**7**	**7.00E-11**	

	CesA	A/A_T			D/D_T				A			A_T			D			D_T		
		Events	nt per year		%	Events	nt per year		%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%
Deletions	SN	16	5.30E-11	6.10E-11	8.8%	19	9.10E-11	9.20E-11	2.0%	8	8.70E-11	8.3%	14	1.72E-10	6.2%	7	5.50E-11	0.3%	15	1.17E-10	7.7%
	IR	3	1.80E-10	2.05E-10	29.5%	7	3.18E-10	3.18E-10	6.8%	1	4.78E-10	45.8%	3	2.29E-09	82.7%	8	1.47E-09	8.7%	5	7.93E-10	52.0%
	IR/SS	3	2.00E-11	2.30E-11	3.3%	2	3.40E-11	3.40E-11	0.7%	2	5.40E-11	5.2%	4	1.72E-10	6.2%	2	3.10E-11	0.2%	1	1.60E-11	1.0%
	SS																				
	Unknown	9	3.58E-10	4.05E-10	58.4%	12	4.22E-09	4.23E-09	90.5%	4	4.24E-10	40.7%	1	1.35E-10	4.9%	16	1.55E-08	90.8%	11	5.99E-10	39.3%
	LTR-recombination																				
	Total	**31**	**6.11E-10**	**6.94E-10**		**40**	**4.66E-09**	**4.68E-09**		**15**	**1.04E-09**		**22**	**2.77E-09**		**33**	**1.70E-08**		**32**	**1.53E-09**	
Insertions	SN	9	3.00E-11	3.40E-11	0.1%	5	2.40E-11	2.40E-11	0.1%	4	4.30E-11	26.5%	4	4.90E-11	5.3%	8	6.20E-11	0.5%	5	8.00E-12	0.2%
	IR	2	7.92E-10	8.98E-10	3.6%								2	1.85E-10	20.0%	1	1.19E-08	98.7%	3	1.24E-10	2.4%
	IR/SS	4	3.70E-11	4.20E-11	0.2%	4	6.70E-11	6.80E-11	0.3%	3	1.19E-10	73.5%	3	6.16E-10	66.7%	5	7.00E-11	0.6%	6	2.41E-10	4.7%
	SS																		1	2.30E-11	0.4%
	Unknown	1	2.30E-11	2.70E-11	0.1%	4	2.58E-08	2.58E-08	99.6%				2	7.40E-11	8.0%	1	2.30E-11	0.2%	5	4.73E-09	92.3%
	polarized retro																				
	unpolarized retro	1	2.27E-08	2.37E-08	95.9%																
	Total	**17**	**2.35E-08**	**2.47E-08**		**13**	**2.59E-08**	**2.59E-08**		**7**	**1.62E-10**		**11**	**9.24E-10**		**15**	**1.21E-08**		**20**	**5.13E-09**	

	Overall	A/A_T			D/D_T				A			A_T			D			D_T		
		Events	nt per year		%	Events	nt per year		%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%	Events	nt per year	%
Deletions	SN	15	4.00E-11	4.45E-11	4.0%	15	6.80E-11	7.15E-11	2.7%	6	5.70E-11	10.7%	11	1.14E-10	0.4%	6	4.60E-11	0.5%	9	7.00E-11	6.7%
	IR	7	2.50E-10	2.69E-10	24.6%	5	2.27E-10	2.36E-10	9.0%	1	2.39E-10	44.7%	2	1.17E-09	4.3%	5	9.16E-10	10.5%	4	4.28E-10	40.7%
	IR/SS	4	1.90E-11	2.10E-11	1.9%	1	1.70E-11	1.70E-11	0.7%	1	2.70E-11	5.0%	2	8.60E-11	0.3%	2	2.65E-11	0.3%	3	5.85E-11	5.6%
	SS	1	4.00E-12	4.50E-12	0.4%								1	7.00E-12	0.0%						
	Unknown	11	7.00E-10	7.47E-10	69.1%	3	2.20E-09	2.22E-09	87.6%	2	2.12E-10	39.6%	1	7.80E-11	0.3%	9	7.78E-09	88.7%	8	4.94E-10	47.0%
	LTR-recombination												1	2.58E-08	94.7%						
	Total	**36**	**1.01E-09**	**1.09E-09**		**23**	**2.52E-09**	**2.55E-09**		**10**	**5.35E-10**		**17**	**2.72E-08**		**22**	**8.76E-09**		**23**	**1.05E-09**	
Insertions	SN	8	2.18E-11	2.45E-11	0.0%	5	2.35E-11	2.50E-11	0.1%	4	3.50E-11	0.4%	3	2.80E-11	0.2%	6	4.55E-11	0.2%	5	2.35E-11	0.9%
	IR	3	4.47E-10	5.03E-10	0.7%	1	7.12E-09	8.04E-09	22.3%	1	8.57E-09	98.4%	1	9.25E-11	0.5%	1	1.92E-08	68.7%	2	6.20E-11	2.4%
	IR/SS	3	3.80E-11	4.15E-11	0.1%	3	9.00E-11	9.80E-11	0.3%	4	1.03E-10	1.2%	2	3.08E-10	1.8%	4	5.70E-11	0.2%	4	1.36E-10	5.2%
	SS	1	3.97E-10	4.15E-10	0.6%	1	7.00E-12	7.50E-12	0.0%										1	1.15E-11	0.4%
	Unknown	2	8.50E-11	9.05E-11	0.1%	3	1.29E-08	1.29E-08	40.4%				1	3.70E-11	0.2%	1	1.15E-11	0.0%	3	2.37E-09	91.0%
	polarized retro	1	6.50E-09	6.79E-09	10.0%	1	1.18E-08	1.33E-08	36.9%							1	8.64E-09	30.9%			
	unpolarized retro	3	5.73E-08	5.99E-08	88.4%								1	1.67E-08	97.3%						
	Total	**20**	**6.48E-08**	**6.77E-08**		**13**	**3.19E-08**	**3.44E-08**		**9**	**8.71E-09**		**7**	**1.71E-08**		**12**	**2.80E-08**		**14**	**2.60E-09**	

* Polarized insertions and deletions for each region of each genome are broken down by mechanism. The number of events attributed to each mechanism is listed, followed by the contribution of that mechanism to the region (as a rate in nt per year and a percentage). The three illegitimate recombination sub-categories are shaded grey and collectively represent illegitimate recombination as a whole.

both D and D_T). The length of this region in the A_T genome was, as expected, slightly shorter than the region in the A genome; however, this was not true for D and D_T, whose lengths were nearly identical. This region had far fewer TEs than *AdhA*, marked only by a single TE insertion that arose in A/A_T during its divergence from D/D_T (Table 1).

Indel dynamics during the evolution of A/A_T

The period of evolution between the divergence of A- D diploids and the divergence of the diploid A genome from the ancestor to the polyploid A_T is marked by several transposable element insertions, particularly in the *AdhA* region, that dramatically skew the nearly 1:1 ratio of deletions:insertions toward insertions (Table 1). The rate of deletion for the two regions combined ranged from 1.02×10^{-9} – 1.09×10^{-9} nt per year, virtually identical to the non-TE insertion rate of 9.9×10^{-10} – 1.08×10^{-9} nt year. Thus, the bulk of the size change during the 5.2 my post A-D divergence and pre A-A_T divergence was due to TE insertions in the *AdhA* region, leading to an overall gain of 6.38×10^{-8} to 6.66×10^{-8} nt per year. Gain via TE insertions (99% of sequence added; Table 2) was distantly followed by illegitimate recombination, which contributed 1.9% of the total sequence added, although 92.9% of the non-TE insertion total. With respect to deletions, illegitimate recombination was second to the unknown mechanism category in terms of relative importance (27.4% versus 68.5% of sequence removed, respectively), removing approximately less than half the amount of sequence in comparison to the unknown category and only about ¼ of the total sequence removed. These trends in mechanistic preference were common to the *AdhA* and *CesA* regions.

The two analyzed regions were similar in number of deletions and insertions; however, the *AdhA* region experienced slightly more deletions than did *CesA* (Table 1) and the average sizes of deletions and insertions (non-TE) were larger, which led

to a slight bias toward deletions in the *AdhA* region. The rate of insertion and deletion was slightly higher for the *AdhA* region than for *CesA*; however, this contribution to sequence turnover was minimal in comparison to the increased TE activity.

Indel dynamics during the evolution of D/D_T

The period of evolution between the divergence of the A - D diploids and the diploid D from the polyploid D_T is marked by a single transposable element insertion (*AdhA*) and two large insertions (one 3.1kb illegitimate recombination associated insertion in *AdhA* and one 5.1kb insertion of unknown mechanism in *CesA*; Table 2). These 3 insertions alone comprise over 97% of the sequence added to the regions (approximately 13.4 kb out of 13.7 kb total), and shifts what would be a 1.05:1 deletion:insertion ratio to 0.3:1, leading to a net gain of $2.94 \times 10^{-8} – 3.18 \times 10^{-8}$ nt per year. Aside from the single TE, insertion via illegitimate recombination and via unknown mechanisms had the greatest impact, contributing 29.1% and 23.1%, respectively, to the total inserted. DNA removal via unknown mechanisms had the largest deletional impact on the region (87.3%), followed by illegitimate recombination at 10.5%.

As with the A/A_T genome, the *AdhA* and *CesA* displayed a similar number of insertions (Table 1); however, unlike the A/A_T genome, deletions were more frequent (over two times) in the *CesA* region than in the *AdhA* region. While the number of insertions for *AdhA* and *CesA* were similar, the amount of sequence inserted was nearly two times greater in *CesA*. These two factors led to a greater rate of non-TE sequence turnover in the *CesA* region. The deletion mechanism that had the greatest impact (unknown category) was common to both regions; however, the insertion mechanism having the greatest impact varied across the regions (TE insertion in *AdhA* and unknown in *CesA*).

Indel dynamics during the evolution of A alone

The period of evolution in the A genome after its divergence from the polyploid A_T genome saw a reduction in average rate of deletion compared to the prior 5.2 my (from 1.02×10^{-9} – 1.09×10^{-9} nt per year to 5.4×10^{-10} nt per year; Table 1), while the average non-TE insertion rate rose from 9.9×10^{-10} – 1.08×10^{-9} nt per year to 8.71×10^{-9} nt per year. No transposable elements were polarized in the aligned region, thus the insertion rates derived for this genome were purely from non-TE mechanisms. As on the previous branch (A/A_T), the deletion and insertion rates varied between the two loci; however, the amount of variation dramatically increased after the divergence of A from A_T. Whereas the deletion rates between *AdhA* and *CesA* in A/A_T varied slightly over twofold and the non-TE insertion rates 1.2-fold, the deletion rates in A varied nearly 35-fold and the insertion rates over 100-fold. The combined deletion and insertion rates yield a net gain of over 8.18×10^{-9} nt per year. IR played a large role in this region, as it was the largest contributor to deletions overall (Table 2; 49.8% versus 39.6% for unknown mechanisms, the second largest) and contributed nearly all inserted DNA (99.6%). This was true for both deletions and insertions in both regions with the exception of deletions in *AdhA*, which consisted solely of four single nucleotide deletions.

The two sequenced regions differed both in rate of sequence turnover, as well as direction of change. The *AdhA* region experienced far greater sequence gain than *CesA* without much loss, leading to a net gain of 1.72×10^{-8} nt per year (Table 1). The *CesA* region, in contrast, experienced more loss than gain, leading to a net loss of 8.8×10^{-10} nt per year, a rate attributed to a small amount of gain outweighed by a nearly as small amount of loss.

Indel dynamics during the evolution of A_T alone

The period of evolution in the A_T genome after its divergence from the diploid A genome saw a reversal from net gain (6.38×10^{-8} to 6.66×10^{-8} nt per year) to net loss (-1.01×10^{-8} nt per year; Table 1), a reversal that was mirrored in both the *AdhA* and *CesA* regions with each experiencing sequence loss. When TE insertions are excluded from both branches, as they are episodic in nature and may not have had as much opportunity to affect the A_T genome in the last 1.6 my as the A/A_T branch over the previous 5.2 my, the A_T genome shows an even more dramatic bias toward DNA removal (adjusted loss of 2.68×10^{-8} nt per year in A_T versus adjusted loss of $1 \times 10^{-11} – 3 \times 10^{-11}$ nt per year in A/A_T). TEs contributed the most to sequence gain and loss (via intra-strand homologous recombination; Table 2) in this genome, despite their action being limited to the *AdhA* region.

As on the previous branch, the insertion and deletion rates varied between the two loci; however, the amount of variation dramatically increased after the divergence of A_T from A. Whereas the deletion rates between *AdhA* and *CesA* for A/A_T varied slightly over twofold and the non-TE insertion rates 1.3-fold, the deletion rates in A_T varied nearly 19-fold (more deletions in *AdhA*) and the insertion rate varied 92-fold (more insertions in *CesA*). The major mechanisms contributing to sequence loss and gain in the *AdhA* region were removal and insertion of TEs, as noted above and in Table 2; the *CesA* was not subject to either TE loss or gain, thus the largest contributor to sequence change in this region was non-TE in nature (illegitimate recombination for both loss and gain). As in the A genome, the *AdhA* region experienced far greater sequence turnover, in terms of nucleotides deleted, due to the large actions of the TE deletion and insertion.

Overall indel dynamics during the evolution of D alone

The period of evolution in the D genome after its divergence from the polyploid D_T genome saw an increase in average rate of deletion compared to the prior 4.2 my (from 2.52 x 10^{-9} – 2.55 x 10^{-9} nt per year to 8.77 x 10^{-9} nt per year; Table 1), as well as decreases in the both the non-TE insertion and TE insertion rates (by 7.7 nt x 10^{-10} – 1.73 x 10^{-9} nt per year and 3.15 x 10^{-9} – 4.67 x 10^{-9} nt per year, respectively). Insertion and deletion rates varied between the two regions, with the *CesA* region experiencing 34-fold more nucleotides deleted, but less than half the amount of non-TE insertions. This difference in insertion and deletion rates led to a net gain in *AdhA* and a net loss in *CesA*. Combined, the regions experienced a slightly smaller net gain than experienced in D/D_T (1.92 x 10^{-8} nt per year). This rate was largely attributed to the combined effects of a single TE insertion (47.1%; Table 2) and two large illegitimate recombination associated insertions (52.5% together), but was slightly relieved by the longer and more frequent deletions (primarily of unknown mechanism) in the *CesA* region. Overall, deletion via unknown mechanisms (88.7%) and insertion via illegitimate recombination (52.5%) had the greatest effects, an observation common to both regions for insertions, but not deletions (deletion via illegitimate recombination had the most impact in *AdhA*).

Indel dynamics during the evolution of D_T alone

The period of evolution in the D_T genome since divergence from the D genome saw a slight decrease in the rate of deletion (1.09 x 10^{-9} nt per year in D_T versus 2.52 x 10^{-9} - 255 x 10^{-9} nt per year in D/D_T; Table 1) and a substantial reduction of the insertion rate (by nearly 90%). The two regions displayed opposite changes in rates, with *AdhA* experiencing loss and *CesA* experiencing gain, leading to an overall rate of sequence gain equivalent to 1.51 x 10^{-9} nt per year (down from 2.94 x 10^{-8} - 3.18 x 10^{-8} nt per year in the ancestor, D/D_T). Neither region was affected by transposable element proliferation, and thus the difference in direction of genome size change

reflects other indel dynamics. In general, this genome experienced less turnover than the ancestral D/D_T and the D genome, with the exception of slightly more deletions in the *AdhA* region than experienced by the other two genomes. Illegitimate recombination and the unknown mechanism category impacted the genome approximately equally with respect to deletions (Table 2), whereas the latter was the major contributor to insertions in those regions (91%). These regions experienced mechanistic biases in this genome as well, with the unknown mechanisms contributing the most sequence loss to *AdhA* and the most sequence gain to *CesA*, and single nucleotide insertions and deletion via illegitimate recombination having the greatest impact in *AdhA* and *CesA*, respectively.

Unpolarized Indels

The number of unpolarized gaps (between A/A_T and D/D_T, A and A_T, and D and D_T) ranged in number and size across the genomes and served to expand the range in possible rates for these genomes (Table 1). One hundred fifty two gaps were unpolarized between A/A_T and D/D_T, accounting for 1.78×10^{-9} – 200×10^{-9} nt per year and 2.63×10^{-8} – 2.84×10^{-8} nt per year missing from A/A_T and D/D_T, respectively. This increases the range in the rate of overall genome size expansion in A/A_T from 6.38×10^{-8} – 6.66×10^{-8} nt per year to 6.20×10^{-8} – 9.50×10^{-8} nt per year, a range shift that is mostly toward further expansion. Similarly, the range in the rate of genome size expansion for the D/D_T branch increased from 2.94×10^{-8} – 3.18×10^{-8} nt per year to 3.10×10^{-9} – 3.38×10^{-8} nt per year, a range shift which suggests that the rate of growth due to polarized indels is likely an overestimate for this branch. The unpolarized gaps between A and A_T represent more missing sequence in A than A_T (5.83×10^{-9} versus 4.91×10^{-9} nt per year); however, even taking this into consideration, the range in overall rate of sequence change remains positive in A (gain 2.35×10^{-9} to 1.31×10^{-8} nt per year) and negative in A_T (loss of 4.25×10^{-9} to 1.50×10^{-8} nt per year). The unpolarized gaps between D and D_T

represent slightly more sequence missing in D than in D_T (3.4×10^{-10} versus 2.1×10^{-10} nt per year), and created a small range in rates for each (1.89×10^{-8} – 1.94×10^{-8} nt per year for D and 1.30×10^{-9} – 1.85×10^{-9} nt per year for D_T).

Analysis of indels <400nt

Previously, we reported that deletions in the *AdhA* region were consistent in size and frequency with the expectations of small indel bias and genome size (i.e. more and longer deletions in the smaller D genome; (Grover et al. 2007)). Furthermore, we noted a higher rate of deletion in the polyploid compared with the diploid ancestors, noting that this observation is congruent with the idea of non-additivity of polyploid genome sizes relative to their diploid antecedents. By adding sequence from the outgroup, we are now able to evaluate, and for a much larger data set, the rate of small indel formation to include the longer period prior to polyploid formation. Contrary to expectations, deletions were more than twice as frequent in A/A_T than in D/D_T (Table 3); however, in accordance with expectations, the deletions were over 1.5-fold larger in the smaller D/D_T genome (Table 4). Small insertions in A/A_T versus D/D_T were congruent with the expectations of a small indel bias, in that they were more frequent and larger in the larger A/A_T genome.

We also previously reported that the spectra of small indel sizes, unpolarized, was nearly equivalent with respect to size and frequency between the A_T and D_T genomes in the *CesA* region (Grover et al. 2004). Upon polarization, a slight bias with regard to frequency appeared for both deletions and insertions (deletions 1.1-fold more frequent in A_T; insertions 1.1-fold more frequent in D_T; Table 3); however, the amount of sequence affected was more variable, particularly for deletions (2.6-fold more sequence deleted in A_T per my). The diploid A and D genomes displayed a similar pattern for small insertions (more in the D genome, but smaller in size);

Table 3: Rates of small insertions and deletions (<400nt) in the *AdhA* and *CesA* regions of the cotton genome*

		A/A$_T$		D/D$_T$		A		D		A$_T$		D$_T$	
		Events	Nt/100kb/my	Events	Nt/100kb/my	Events	Nt/100kb/my	Events	Nt/100kb/my	Events	Nt/100kb/my	Events	Nt/100kb/my
AdhA	Deleted	40	5.9 to 6.2 x 10^{-10}	16	3.7 to 4.2 x 10^{-10}	4	3 x 10^{-11}	10	5.0 x 10^{-10}	11	1.3 x 10^{-10}	14	5.8 x 10^{-10}
	Inserted	17	1.10 to 1.15 x 10^{-9}	10	3.2 to 3.6 x 10^{-10}	9	2.0 x 10^{-10}	6	7 x 10^{-11}	1	1 x 10^{-11}	7	7 x 10^{-11}
	Unknown	30	3.9 to 4.1 x 10^{-10}	31	2.00 to 2.26 x 10^{-9}	20	1.06 x 10^{-9}	3	2.1 x 10^{-10}	29	1.85 x 10^{-9}	10	3.2 x 10^{-10}
	Rate	**5.1 to 5.2 x 10^{-10}**		**-6 to -5 x 10^{-11}**		**1.7 x 10^{-10}**		**-4.3 x 10^{-10}**		**-1.2 x 10^{-10}**		**-5.1 x 10^{-10}**	
	Range	**0.12 to 2.79 x 10^{-9}**		**-2.32 to 0.34 x 10^{-9}**		**-0.89 to 2.02 x 10^{-9}**		**-6.4 to -1.1 x 10^{-10}**		**-1.97 to 0.94 x 10^{-9}**		**-8.3 to -3.0 x 10^{-10}**	
CesA	Deleted	31	6.1 to 6.9 x 10-^{10}	39	1.55 x 10^{-9}	15	1.04 x 10^{-9}	32	1.07 x 10^{-8}	22	2.77 x 10^{-9}	32	1.52 x 10^{-9}
	Inserted	16	0.88 to 1.00 x 10^{-9}	12	1.50 to 1.51 x 10^{-9}	7	1.6 x 10^{-10}	14	1.6 x 10^{-10}	11	9.2 x 10^{-10}	19	1.41 x 10^{-9}
	Unknown	45	1.41 to 1.60 x 10^{-9}	42	2.90 to 2.94 x 10^{-9}	4	8 x 10^{-11}	6	4.7 x 10^{-10}	3	4 x 10^{-11}	11	9 x 10^{-11}
	Rate	**2.7 to 3.1 x 10^{-10}**		**-5 to -4 x 10^{-11}**		**-8.8 x 10^{-10}**		**-1.06 x 10^{-8}**		**-1.85 x 10^{-9}**		**-1.1 x 10^{-10}**	
	Range	**-1.14 to 3.25 x 10^{-9}**		**-2.95 to 1.56 x 10^{-9}**		**-9.6 to -8.4 x 10^{-10}**		**-1.10 to -1.05 x 10^{-8}**		**-1.89 to -1.77 x 10^{-9}**		**-2.0 to 3.6 x 10^{-10}**	
Average	Deleted	36	6.0 to 6.6 x 10^{-10}	28	9.6 to 9.9 x 10^{-10}	10	5.4 x 10^{-10}	21	5.61 x 10^{-9}	17	1.45 x 10^{-9}	23	1.05 x 10^{-9}
	Inserted	17	0.99 to 1.08 x 10^{-9}	11	9.1 to 9.4 x 10^{-10}	8	1.8 x 10^{-10}	10	1.2 x 10^{-10}	6	4.7 x 10^{-10}	13	7.4 x 10^{-10}
	Unknown	38	0.90 to 1.01 x 10^{-9}	36	2.45 to 2.60 x 10^{-9}	12	5.7 x 10^{-10}	5	3.4 x 10^{-10}	16	9.5 x 10^{-10}	11	2.1 x 10^{-10}
	Rate	**3.9 to 4.2 x 10^{-10}**		**-5 x 10^{-11}**		**-3.6 x 10^{-10}**		**-5.49 x 10^{-9}**		**-9.9 x 10^{-10}**		**-3.1 x 10^{-10}**	
	Range	**-0.51 to 3.02 x 10^{-9}**		**-2.50 to 0.96 x 10^{-9}**		**-9.3 to 5.9 x 10^{-10}**		**-5.83 to -5.23 x 10^{-9}**		**-0.48 to 1.04 x 10^{-9}**		**-5.2 to 0.3 x 10^{-10}**	

* Rates of insertion and deletion were standardized to nt per year based on the sequenced length of the region and the time since divergence (5.2my for the A/AT branch, 4.2my for D/DT branches, 1.6my for A and AT, and 2.6my for the D and DT branches; Figure 1), and the range representing the minimum and maximum rate was calculated. Gain is indicated by positive values and loss is indicated by negative values.

Table 4: Average insertion and deletion rates and sizes sizes for indels < 400nt

		A/A$_T$			D/D$_T$			A		D		A$_T$		D$_T$	
		Events/my	Avg Low	Avg High	Events/my	Avg Low	Avg High	Events/my	Average	Events/my	Average	Events/my	Average	Events/my	Average
AdhA <400nt	Deleted	7.7	7.67E-11	8.06E-11	3.8	9.71E-11	1.10E-10	2.5	1.20E-11	3.8	1.30E-10	6.9	1.89E-11	5.4	1.08E-10
	Inserted	3.3	3.36E-10	3.52E-10	2.4	1.34E-10	1.51E-10	5.6	3.56E-11	2.3	3.03E-11	0.6	1.60E-11	2.7	2.60E-11
CesA <400nt	Deleted	6.0	1.02E-10	1.16E-10	9.3	1.67E-10	1.67E-10	9.4	1.11E-10	12.3	8.69E-10	13.8	2.01E-10	12.3	1.24E-10
	Inserted	3.1	2.86E-10	3.25E-10	2.9	5.25E-10	5.29E-10	4.4	3.66E-11	5.4	2.97E-11	6.9	1.34E-10	7.3	1.93E-10
<400nt	Deleted	6.9	8.67E-11	9.53E-11	6.7	1.44E-10	1.49E-10	6.3	8.64E-11	8.1	6.95E-10	10.6	1.36E-10	8.8	1.19E-10
	Inserted	3.3	3.03E-10	3.30E-10	2.6	3.47E-10	3.59E-10	5.0	3.60E-11	3.8	3.12E-11	3.8	1.25E-10	5.0	1.48E-10

however, small deletions were fully consistent with a small indel bias, with more frequent and larger deletions in the D genome. The average deletion and insertion sizes among the genomes (Table 4) did not mirror the results of *AdhA*. The A genome deletions were on average smaller than those from the D genome (6.93 vs. 33.47 nt), as expected based on the previously analyzed *AdhA* region. Deletions in the polyploid, however, did not mirror its diploid counterparts, but instead were characterized by an acceleration in deletions in the A_T genome and a deceleration in D_T. The pattern between A/A_T and D/D_T, in the case of the *CesA* region, more closely resembled what the small indel bias would predict in terms of average deletion size and frequency (deletions in D/D_T 1.5-fold as frequent and twice the size of those in A/A_T). Conversely, the pattern of insertions is contrary to what the small indel bias would predict, with insertions in A/A_T being 1.1 times as frequent as and more than 50% smaller than in D/D_T.

Analysis of illegitimate recombination

Previously we reported that illegitimate recombination may be a key player in *Gossypium* genome size evolution, particularly in the polyploid. The data reported here (Table 2) provide the ability to assess not only the rate of illegitimate recombination for each genome since polyploid formation, but also how those rates compare to the ancestral rates. The combined data suggest that every lineage (A, A_T, D, and D_T) has experienced accelerations (to varying degrees) in the rate of deletion via illegitimate recombination since the diploid - polyploid divergence, from the near doubling in D to the over 4.5-fold increase in A_T; however, this increase was not for both regions in every genome. Whereas the D and D_T genomes displayed an increase in illegitimate recombination for both regions, both A and A_T had a slight rate decrease in *AdhA* that was compensated for by the increased rate in *CesA*.

The rates attributable to insertion via illegitimate recombination display opposite effects in the diploids and polyploids (Table 2); whereas the diploids experienced an overall increase in the rate of IR-associated insertions, both genomes of the polyploid experienced decreases. Again, this overall trend was not equivalent in the two regions. While the *AdhA* region mirrored the overall results, in the *CesA* region the A genome experienced a decrease in rate, while the D_T genome experienced an increase. Overall, the amount of sequence deleted via illegitimate recombination was less than the total amount of sequence inserted for the diploid lineages (added $1{,}04 \times 10^{-9}$ – 1.84×10^{-8} nt per year) and more for the polyploid lineages (deleted -5.5×10^{-10} and -2.0×10^{-10} nt per year for A_T and D_T, respectively).

The relative impact of illegitimate recombination varied by genome and by region (Table 2). In the *AdhA* region, most polarized deletions were attributed to either the unknown mechanism category (A/A_T – 73.7%; D/D_T – 50.6%; D_T – 67.7%), single nucleotide deletions (A – 100%) or LTR-recombination (A_T – 99.7%). Only in the D genome were most deletions attributed to IR (76.3%). In the *CesA* region, the most deleted nucleotides were attributed to illegitimate recombination in some genomes (A – 51%; A_T – 88.9%; D_T – 53%), whereas along other branches the unknown mechanism was dominant (A/A_T – 58.4%; D/D_T – 90.5%; D –90.8%). When viewed together, the A genome was the only one where more nucleotides were deleted via IR than from any other mechanism (52.5%).

The relative impact of insertion via illegitimate recombination was small in comparison to the amount of sequence inserted by TEs in the A/A_T and A_T genomes (98% and 97.4%, respectively), unknown mechanisms in D_T (93.9%), and the combined efforts of TEs and unknown insertions in D/D_T (47.7% and 23.1%, respectively). Illegitimate recombination only had a major impact on insertions in the diploid A and D genomes (at 99.8% and 52.7%, respectively). This trend was

largely similar between both regions, with the only difference occurring in the *CesA* region of A_T genome, where illegitimate recombination represented 86.7% of the nucleotides inserted, a difference attributed to the lack of TE insertions in this region.

Discussion

Genome size evolution is a dynamic process, reflecting the net effects of counterbalancing mechanisms whose actions vary across a genomic landscape, across lineages, and over time. The potential for the primary mechanism of genome size change, TE proliferation, to affect genome size has become evident, although its catalysts are less clear. Mechanisms of deletion, by their nature, are more difficult to study; whereas TE proliferation can be gauged by simply evaluating the extent of TE sequence in a genome, deletional mechanisms can only be identified and evaluated by comparison to non-deleted sequence. Compounding this problem is rapid evolution, which may quickly erase by superimposed mutations the hallmarks of deletional mechanisms that leave small footprints, such as illegitimate recombination. Comparisons of long, orthologous tracts of sequence between closely related species that are polarized by an outgroup provides a potentially powerful means to evaluate the relative effects of different mechanisms influencing genome size change (both growth and reduction).

Rate of sequence loss and gain on six branches of the Gossypium *phylogeny*

The combined rates of DNA deletion and insertion are ultimately what determine genome size change. Comparisons of extant genome sizes and their transposable element contents provide important information on the probable direction and nature of genome size change, but a phylogenetic perspective adds insights into the tempo, details, and dynamics of genomic divergence. One might imagine two species with the same genome size and TE composition that have different genomic histories; for

example, one species may have acquired its genome size through slow and steady TE accumulation while the other taxon has achieved a similar genome size via rapid flux of intergenic space (nearly concurrent insertions and deletions). Comparative genomic sequencing of closely related species, as exemplified here, provides the opportunity to illuminate this history and similar nuances of genome evolution.

The general trend unveiled by comparative sequencing of the *AdhA* and *CesA* regions is that there has been an overall propensity toward growth of the diploid genomes and an overall contraction of the polyploid (Figure 1; Table 1), while each of these two regions display heterogeneous rates of sequence gain and loss at different times in the evolutionary histories of the genomes studied, possibly linked to genomically regional properties. All genomes experienced growth except for the A_T genome; however, the *AdhA* region experienced contraction in the diploid D (in addition to A_T) and the *CesA* region displayed contraction for the A and D genomes in addition to the A_T genome. In addition, there appears to be a regional bias dependent on lineage, and possibly, ploidy level. For the four purely diploid branches (A/A_T, D/D_T, A, and D), the *AdhA* region experienced more gain than did the *CesA* region (A/A_T and D/D_T) or, in the case of A and D, gain versus loss. The converse was seen for the polyploid lineages, where both A_T and D_T experienced loss, or more loss, in *AdhA* compared to *CesA* and the diploid branches experienced gain.

In general, the A/A_T branch is marked by large sequence gains, primarily TE in origin (Tables 1 & 2). The rate of deletion barely outweighs the rate of non-TE insertion, indicating that genome growth along this branch has primarily been due to the action of TEs. The *AdhA* region of the A/A_T genome gained sequence at a rate 4.5-fold higher than did the *CesA* region, due to TE proliferation and possibly indicating insertional preferences or exclusion. The D/D_T branch also experienced sequence

gain related to TE insertions, but an even higher rate of non-TE sequence gain. The combined deletion rate in D/D_T was 2.5 times the rate of A/A_T, consistent with a hypothesis that small genomes differ from large genomes in part due to their inherently higher deletion rates. We note, however, that the rate of deletion was still only about $1/13^{th}$ the total rate of insertion (versus $1/64^{rd}$ in A/A_T). These results indicate that the trend for the majority of the genome size divergence between the A and D genome species, having taken place on the A/A_T and D/D_T branches, is one of genome growth, with the rate in A/A_T 2-fold higher than in D/D_T, and both being largely dependent upon the rate of insertion.

After divergence from the polyploid A_T, the rate of sequence gain experienced in the diploid A was less than 13% of that experienced prior to polyploid formation (Figure 1; Table 1), primarily due to the lack of TE insertions. The overall rate of sequence gain in A, however, still outweighed that of deletion due to the higher number of insertions and fewer deletions found in *AdhA*. The situation for the D genomes is far less exaggerated in this respect; after divergence from the polyploid D_T, the rate of gain in the diploid D decreased to about 65% of the ancestral rate in D/D_T. This rate reflects a combination of increased deletions, and a slight decrease in non-TE insertions and TE insertions.

The impetus to change rates of indel evolution and, consequently, genome size can come from many and varied sources, one of which being the union of two divergent genomes in an allopolyploid nucleus. Polyploidization has been implicated in numerous genetic and genomic changes (reviewed in Adams and Wendel 2005 and Chen and Ni 2006), and the resulting genome size of the polyploid species is often less than the sum of the two parental genomes (Soltis and Soltis 1999; Ozkan, Tuna, and Arumuganathan 2003; Bennett and Leitch 2005b). This phenomenon of “genomic downsizing” has been explored in several other cases (Chantret et al.

2005; Gu et al. 2006), but to our knowledge this is the first phylogenetically informed evaluation of changes in deletion and insertion rate that accompany polyploidization using large contiguous tracts of orthologous and homoeologous sequence. Both genomes of the polyploid show an increase in the rate of deletion (more dramatic in the case of A_T) and a reduction in the rate of insertion when compared to their ancestral lineages (Figure 1; Table 1). The shifting balance from insertions to deletions produced a rate of near stasis in D_T and an overall rate of contraction in A_T, leading to a combined shrinkage of the polyploid genome. Inter-region variability was also present in the polyploid genomes, serving to shrink the *AdhA* region in A_T tenfold more than *CesA* and contracting the *AdhA* region in D_T (compared to the moderate gain experienced in *CesA*). Since the D_T genome spent half of its time since divergence from the diploid D as a diploid itself (Figure 1), the rate of loss in the polyploid D_T may be, in part, and underestimate masked by gains (primarily in the *CesA* region) that could have occurred during the 1.3 my spent as a diploid. These data suggests that the polyploid genome has, in fact, been experiencing genomic shrinkage in the 1-2 my post-polyploidization instead of the alternative (growth in the diploids relative to slower growth or stasis in the polyploid). Further analyses in *Gossypium* and other polyploids are required to test the generality of these observations.

Mechanisms affecting the rate of sequence loss and gain

Transposable element proliferation is thought to be responsible for most genome size growth in angiosperms. This leads to the *a priori* hypothesis that a majority of the size difference between extant genomes reflects differential proliferation of TEs in a manner congruent with genome size (i.e., the A/A_T lineage will have accrued TEs twice as fast as the D/D_T lineage, as would A versus D). The bias in TE proliferation observed between A/A_T and D/D_T is in the direction that is expected, yet the difference is even more exaggerated than expected (Figure 1; Table 1). The

A/A_T lineage gained TE sequence at a rate that was over 5-fold greater than the D/D_T lineage. In the time that the A and D genomes evolved independent from the polyploid genomes (approximately 1-2my), however, the A genome has not gained TE sequence in either of these regions, while the D genome has gained 8.64×10^{-9} nt per year (due to a single insertion). This difference in TE insertion rates may be explained by the episodic nature of TE proliferation (Hawkins et al., in press), although an interesting alternative is that while the D genome continues to gain TE sequence at a rate more similar to its ancestor lineage, the A genome has become less permissive of TE proliferation.

The exaggerated rate of TE gain in the A/A_T lineage may also reflect non-mutually exclusive factors. For example, the TE population in the D genome may be concentrated in regions that have not been surveyed, potentially due to different integrational or targeting requirements. An alternative, for which there is no evidence in these regions, is that the A/A_T lineage has simply been more dynamic in general, rapidly removing DNA while allowing TEs to proliferate. A third alternative, one that was realized in part in the data presented here, is that the D/D_T lineage experienced growth via other mechanisms than TE proliferation (Table 2). Overall, the D/D_T lineage experienced an non-TE insertion rate that was approximately 1.7-fold greater than the TE insertion rate, which consequently brought difference in rate of sequence gain between A/A_T and D/D_T down to just over 2-fold.

Just as TEs have been implicated in genome size growth, their amplification and genomic presence can also lead to genome size contraction via intra-strand homologous recombination. Intra-strand homologous recombination has been demonstrated in many systems and at various levels (Kalendar et al. 2000; Devos, Brown, and Bennetzen 2002; Vitte and Panaud 2003; Wicker et al. 2003; Vitte and Bennetzen 2006), raising questions about how constraints on or stimulation of LTR

recombination varies among species. The data from *Gossypium* indicates that intra-strand homologous recombination may be rare, as only one solo-LTR was observed (versus 13 intact elements in the 4 genomes; Table 2); however, data indicate that even a rare event can greatly impact the rate of deletion. The *AdhA* in A_T region experienced a rate of deletion that was over 397-fold greater due to the single intra-strand homologous recombination event, ultimately leading to net contraction for the region; similarly, this single deletion increased the overall deletion rate across the two regions nearly 19-fold and reversing what would be an overall net gain of 1.57 x 10^{-8} nt per year to a contraction of 1.01 x 10^{-8} nt per year.

Biased accumulation of small indels has been promoted (Petrov 1997; Petrov 2002), as well as criticized (Gregory 2003), as a solution to the discordance between the phylogenetic placement of plants possessing small genomes and the potential of deletions to shrink genomes (Vitte and Bennetzen 2006). The indel bias proposal is that, on average, smaller genomes will acquire more frequent and larger deletions (<400nt) than larger genomes, thus slowly and stochastically shrinking the size of the genome more in the smaller genomes. While the data presented here provide some support for this notion, this is limited to the period of evolution since polyploidization. The branches that we would expect to show the most bias (i.e. the branches where the most differential genome size change likely took place, A/A_T and D/D_T) were contradictory over the two regions in this respect, with the average deletion size less than twice the average size in D/D_T and the insertion size greater for A/A_T in only one of the two sequenced regions. Overall, the data support a slightly larger average deletion size for D/D_T, while also being less frequent, yet a slightly larger average insertion size for D/D_T, although also less frequent. Taken together, the data for the diploid genomes suggest the possibility of a small indel bias, as the smaller genomes tend to add less sequence (D/D_T) or delete more sequence (D) through small indels than the larger genomes, while noting regional

biases in small indel formation. The data also suggest that polyploid genome has, in general, experienced increased sequence turnover (both as small deletions and small insertions).

Illegitimate recombination is attractive as a method for genome contraction, despite its tendency to create small rather than large deletions (Petrov 2002; Bennetzen, Ma, and Devos 2005), due to its presumed global nature and the idea that the effects of a slow, consistent “genomic leak” would outweigh episodic TE amplification over time. The data presented here fail to provide support for IR as a key determinant of genome size variation in *Gossypium*. Just as with other mechanisms of genome size evolution, illegitimate recombination may operate heterogeneously within genomes, affecting some genomic regions more than others and perhaps linked to regional features such as level of chromatin unwinding. This heterogeneity is evident in the regions and genomes studied here, where illegitimate recombination added sequence in about half of the cases, but deleted sequence in the polyploid. Finally, for most regions harboring TEs, TEs played a larger role in genome size increase than could be compensated for by illegitimate recombination, and removal of a single TE (as observed in the *AdhA* region of A_T) creates a much greater sequence reduction than deletion via illegitimate recombination (over 900-fold more for this region).

Although IR does not appear to be a major mechanism of genome size change in the regions and genomes studied here, our data suggest polyploidy induced a shift in the bias of illegitimate recombination toward deletions over insertions. While this bias toward contraction may be thwarted by TE insertions, as mentioned above, it is suggestive, as previously reported (Grover et al. 2007), of a mechanism to partially explain the phenomenon of genomic downsizing in polyploids.

Concluding remarks

The data presented here highlight the instability of the rates and mechanisms of genome size change on an evolutionary timescale corresponding to divergence of species within a single angiosperm genus. While much research has focused on mechanisms of genome size change, less is known concerning rates of DNA removal and gain due to specific mechanisms, and, to our knowledge, none have addressed the issue of how rates of the various mechanisms governing genome size expansion or contraction change over time. The heterogeneous nature of genome size evolution elucidated here is underscored by both the differences in genome contraction and growth experienced by regions within a single genome and by genomes over time. The complexities revealed here underscore the dynamics of genome size evolution that may be revealed by focused phylogenetic analyses.

Acknowledgements

We thank R. Percifield for technical assistance, the National Science Foundation Plant Genome Program, and the National Science Foundation BAC Program for financial support.

References

Adams, K. L., and J. F. Wendel. 2005. Polyploidy and genome evolution. Curr Opin Plant Biol. **8**:135-141.

Bennett, M. D., and I. J. Leitch. 1995. Nuclear DNA amounts in angiosperms. Annals Bot. **76**:113-176.

Bennett, M. D., and I. J. Leitch. 2005a. Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. Ann Bot **95**:45-90.

Bennett, M. D., and I. J. Leitch. 2005b. Genome size evolution in plants. Pp. 89-162 *in* T. R. Gregory, ed. The evolution of the genome. Elsevier, San Diego.

Bennetzen, J. L., J. Ma, and K. M. Devos. 2005. Mechanisms of Recent Genome Size Variation in Flowering Plants. Ann Bot **95**:127-132.

Brudno, M., C. Do, G. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large scale multiple alignment of genomic DNA. Genome Research **13**:721-731.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M.-F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. 2005. Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). Plant Cell **17**:1033-1045.

Chen, Z. J., and Z. Ni. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. Bioessays **28**:240-252.

Cronn, R. C., R. L. Small, T. Haselkorn, and J. F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium* : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany **89**:707-725.

Devos, K. M., J. Brown, and J. L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Research **12**:1075-1079.

Ewing, B., and P. Green. 1998. Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research **8**:186-194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequences traces using phred. I. Accuracy assessment. Genome Research **8**:175-185.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Research **8**:195-202.

Green, P. 1999. Phrap documentation. http://www.phrap.org/phrap.docs/phrap.html.

Gregory, T. R. 2003. Is small indel bias a determinant of genome size? Trends in Genetics **19**:485-488.

Grover, C. E., H. Kim, R. A. Wing, A. H. Paterson, and J. F. Wendel. 2007. Microcolinearity and genome evolution in the *AdhA* region of diploid and polyploid cotton (*Gossypium*). The Plant Journal **50**:995-1006.

Grover, C. E., H. Kim, R. A. Wing, A. H. Paterson, and J. F. Wendel. 2004. Incongruent Patterns of Local and Global Genome Size Evolution in Cotton. Genome Res. **14**:1474-1482.

Gu, Y. Q., J. Salse, D. Coleman-Derr, A. Dupin, C. Crossman, G. R. Lazo, N. Huo, H. Belcram, C. Ravel, G. Charmet, M. Charles, O. D. Anderson, and B. Chalhoub. 2006. Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes. Genetics **174**:1493-1504.

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

Hendrix, B., and J. M. Stewart. 2005. Estimation of the Nuclear DNA Content of *Gossypium* Species. Annals of Botany **95**:789-797.

Hill, P., D. Burford, D. Martin, and A. J. Flavell. 2005. Retrotransposon populations of Vicia species with varying genome size. Molecular Genetics and Genomics **273**:371-381.

Kalendar, R., J. Tanskanen, S. Immonen, E. Nevo, and A. H. Schulman. 2000. Genome evolution in wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA **97**:6603-6607.

Leitch, I. J., M. W. Chase, and M. D. Bennett. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. Annals Bot. (Suppl. A) **82**:85-94.

Ma, J., K. M. Devos, and J. L. Bennetzen. 2004. Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. Genome Res. **14**:860-869.

Ozkan, H., M. Tuna, and K. Arumuganathan. 2003. Nonadditive changes in genome size during allopolyploidization in the wheat group (*Aegilops-Triticum*) group. Journal of Heredity **94**:260-264.

Petit, M., K. Y. Lim, E. Julio, C. Poncet, F. D. de Borne, A. Kovarik, A. R. Leitch, M. Grandbastien, and C. Mhiri. 2007. Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). Molecular Genetics and Genomics **278**:1.

Petrov, D. 1997. Slow but steady: Reduction of genome size through biased mutation. Plant Cell **9**:1900-1901.

Petrov, D. A. 2002. Mutational equilibrium model of genome size evolution. Theoretical Population Biology **61**:531-544.

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

SanMiguel, P., and J. L. Bennetzen. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Annals of Botany **82**:37-44.

Senchina, D. S., I. Alvarez, R. C. Cronn, B. Liu, J. K. Rong, R. D. Noyes, A. H. Paterson, R. A. Wing, T. A. Wilkins, and J. F. Wendel. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Molecular Biology and Evolution **20**:633-643.

Soltis, D. E., and P. S. Soltis. 1999. Polyploidy: recurrent formation and genome evolution. Trends Ecol. Evol. **9**:348-352.

Vitte, C., and J. L. Bennetzen. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceeding of the National Academy of Science **103**:17638-17643.

Vitte, C., and O. Panaud. 2003. Formation of Solo-LTRs Through Unequal Homologous Recombination Counterbalances Amplifications of LTR Retrotransposons in Rice Oryza sativa L. Mol Biol Evol **20**:528-540.

Wendel, J. F., and R. C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Adv. Agron. **78**:139-186.

Wendel, J. F., R. C. Cronn, J. S. Johnston, and H. J. Price. 2002. Feast and famine in plant genomes. Genetica **115**:37-47.

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

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.-D. Liu, J. Dubcovsky, and B. Keller. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell **15**:1186-1197.

CHAPTER SIX

GENERAL CONCLUSIONS

Genome size change in any given taxon reflects the net balance between mechanisms leading to expansion and those leading to contraction. As reviewed in Chapter 2, much evidence has been gathered for the effects of transposable elements on genome expansion, and, to a lesser degree, the effects of intra-strand homologous recombination and illegitimate recombination on genome contraction. Much of this effort has been concentrated in the grass family and few (Ma and Bennetzen 2004, Vitte and Bennetzen 2006) have tried to weigh the relative impact of these mechanisms against each other. In addition, no study to date has attempted to address the rates of sequence gain and loss attributable to these mechanisms and how they change over time. Thus, while significant advances have been made in understanding the mechanisms that shape genome size, there remains much to be learned about the rates at which genomes expand and contract, as well as how these rates change over time and across lineages. The goal of this thesis was to extend our knowledge in the foregoing areas by describing the rates of sequence gain and loss during the evolution of the cotton genus (*Gossypium*).

In chapter 3, we provided the first assessment of genome size evolution in *Gossypium* through comparative genomic sequencing. By evaluating the 105kb region surrounding the gene encoding cellulose synthase, we were able to begin to evaluate the mechanisms that shape the twofold difference in genome size between the genomes of the allopolyploid, *Gossypium hirsutum*. Remarkably, and in contrast to evidence derived from the grasses, the intergenic space between the two genomes was extraordinarily conserved (95%, ungapped). Contrary to our expectations, we found no evidence of the twofold genome size differences between the species. Neither the distribution of TEs in the region, the distribution of the

nearly 550 small indels, nor any of the more minor mechanisms demonstrated a bias that could account for the difference in genome size. Analysis of gaps bearing the hallmarks of illegitimate recombination, while unable to suggest a bias when unpolarized, did suggest that illegitimate recombination may be an important mechanism for *Gossypium* genome size evolution. These data suggested that genome size differences are unevenly distributed across the genome, as no evidence of the twofold genome size differenced was observed, and must be due to mechanisms that are not globally operating; instead, the twofold genome size difference between species must be due to mechanisms that experience local effects and biases.

From the work described in chapter 3, we extended our analysis of genome size evolution to another region of the genome in order to more fully understand the regional effects and biases of genome size evolution and further seek out the mechanisms that generated the observed differences in genome size. For this analysis (chapter 4), we sequenced the region surrounding the gene encoding *alcohol dehydrogenase A*, for the two co-resident genomes (A_T and D_T) of the allopolyploid, *Gossypium hirsutum,* as well as its model diploid progenitors, *Gossypium raimondii* (D) and *Gossypium arboreum* (A). This region contrasted the work described in chapter 3 by reflecting, in a microcosm, the overall twofold difference in genome size. The aligned sequence lengths of the two smaller genomes (D and D_T) versus the two larger genomes (A and A_T) themselves were nearly twofold different in size. Analysis of the transposable element content revealed that a majority of the size differences in the region could be attributed to differential TE proliferation as the A and A_T genomes contained far more TE sequence than the D and D_T genomes (32.7 and 25.3kb versus 7.1 and 5.1kb, respectively). The data also suggest, however, that the genome size difference in the region may have been further exaggerated via a biased accumulation of small

indels. The data for the A genome suggested a propensity for small insertions, as it was the only genome where small insertions outweighed small deletions. The small indel analysis also indicated a deletional bias for the D and D_T genomes, which experienced more frequent and, on average, longer deletions than the A and A_T genomes. Thus, a bias in small indels, although not globally operating, was determined to be a viable contributor to genome size differences in *Gossypium*.

The data also allowed us to draw conclusions concerning the nature of genome size evolution in a polyploid, relative to its model diploid progenitors. True to the nature of polyploids, the genomes of *Gossypium hirsutum* have experienced genomic downsizing relative to its model diploid progenitors. The indel data suggested that the polyploid genomes have experienced an acceleration in small deletions and illegitimate recombination post-polyploidization, which could provide a partial explanation for the phenomenon of polyploid “genomic downsizing”.

In chapter 5, we extended our prior two analyses (chapters 3 and 4) to include all four *Gossypium* genomes (A, A_T, D, D_T), as well as a phylogenetic outgroup, *Gossypioides kirkii*, which provided the ability to polarize indels that occurred during the evolution of A-A_T together and D-D_T together. This analysis used the combined data from the *CesA* and *AdhA* regions to determine the rates of sequence gain or loss along each branch, partitioned by mechanism, and how these changed during species divergence. The data revealed an overall trend toward growth of the diploid genomes and contraction in the polyploid. Most of the size difference between the diploid species occurred prior to polyploid divergence, and was largely attributable to TE amplification in the A/A_T genome, although slightly counteracted by increased non-TE insertions in the D/D_T genome. After separation from the polyploid genomes, both diploid genomes experienced slower sequence gain than in the ancestor, which as attributable to fewer TE insertions in the A genome and a

combination of increased loss and decreased gain in the D genome. Both genomes of the polyploid, like the D genome, displayed increased rates of deletion and decreased rates of insertion, leading to a rate of near stasis in D_T and overall contraction in A_T and ultimately resulting in polyploid genome contraction. As expected, TE insertions contributed significantly to the genome size differences. Intra-strand homologous recombination was rare, but, when present, it had the most significant impact on the rate of deletion. Small indel data for the diploids suggested the possibility of a bias, as the smaller genomes added less or deleted more sequence through small indels than did the larger genomes; however, data for the polyploid suggests increased sequence turnover in general (both as small deletions and small insertions) with no discernible bias in direction. Illegitimate recombination was not demonstrated to be a dominant mechanism of genome size change in the diploid; however, in the polyploids illegitimate recombination was biased toward deletions, which may provide a partial explanation of polyploid genomic downsizing.

In summary, this work speaks to the dynamic nature of genome size evolution and the mechanisms that effect change, which led to several key conclusions. First, genome size change is effected by many mechanisms, some of which may yet be unknown. While a seemingly intuitive notion, the present research represents one of few evaluations of genome size change that was not limited to a specific mechanism due to presumed importance. Second, while TEs had the most impact on genome size differences, as expected, other mechanisms (intra-strand homologous recombination, the combined effects of non-TE insertions) also played key roles in shaping the regions and their sizes. Third, the trend in size change for *Gossypium* typically consists of growth in the diploid genomes and contraction in the polyploid, and these rates of change vary depending on region and over time. Finally, the mechanisms that affect genome size are themselves affected by regional properties, such that even the effects of speculatively global phenomena (e.g. small indel bias)

can be enhanced or curtailed by location. These conclusions were made possible through the careful analysis and curation of indels occurring in a close phylogenetic framework.

Our work in *Gossypium* parallels other systems in that transposable elements play a large role in creating the genome size differences between species, whereas the contributions of deletional mechanisms are variable and less clear. Also, similar to findings in wheat (Chantret 2005, Gu 2006), our results indicate that increased illegitimate recombination may be partly responsible for polyploid genome down-sizing. This research, however, goes beyond what is currently known to examine the actual rates of genome size change along evolutionary branches and attributable to specific mechanisms. Through careful analysis of genomic sequence combined with the phylogenetic background in this genus, we were able to determine not only the amount of sequence attributable to specific mechanisms, but also the rate at which those mechanisms have operated over evolutionary time in ancestral, as well as extant, taxa. This type of analysis has yet to be completed, to our knowledge, for any other system.

This research has illuminated the complex and dynamic nature of genome size evolution by addressing some of the fundamental questions concerning the rates of and mechanisms involved in genome size change and, in the process, has stimulated several questions of its own. A natural follow-up question is whether the results presented here extend to other regions of the genome. The data suggest regional biases in the rate and mechanisms of genome size change that may be linked to intrinsic properties, which begs the question: what rate and mechanistic differences characterize regions of the genome that are inherently different (e.g. heterochromatic versus euchromatic regions)? For that matter, what are these genomic properties that affect the rate of and mechanisms involved in genome size

change? Comparative genomic sequencing for many regions, or whole genome sequences, that is anchored with a phylogenetic perspective provides the ability to detail the complexities that surround the notion of genome size evolution.

Literature cited

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M.-F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. 2005. Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). Plant Cell **17**:1033-1045.

Gu, Y. Q., J. Salse, D. Coleman-Derr, A. Dupin, C. Crossman, G. R. Lazo, N. Huo, H. Belcram, C. Ravel, G. Charmet, M. Charles, O. D. Anderson, and B. Chalhoub. 2006. Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes. Genetics **174**:1493-1504.

Ma, J. and J. L. Bennetzen. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences 101: 12404-12410

Vitte, C., and J. L. Bennetzen. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceeding of the National Academy of Science **103**:17638-17643.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my mentor, Dr. Jonathan Wendel, for his on-going support, for sharing his wealth of knowledge, and for much needed encouragement during the most difficult tasks. I would also like to thank my committee for their review and insightful comments on this work, and particularly Dr. Lynn Clark for helpful advice. Special thanks go out to my research colleagues, past and present, for helpful discussions and, often, a sympathetic ear. In particular, I would like to thank Dr. Jennifer Hawkins, Ryan Percifield, and Jordan Swanson for their help on this project.

I would be remiss in not thanking all of those in my family who have supported me throughout this endeavor: my husband, Trent, who never lets me quit; my son, Cillian, who brightens even the most disappointing days; my parents (Elaine and Rick), grandparents (Dan and Joyce), and a special aunt (Janet), whose belief in me is unwavering; and my sister, Beth, for always lending an ear.