Molecular evolution at homoeologous loci in allotetraploid cotton (Malvaceae: Gossypium hirsutum L)

Richard Clark Cronn

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Botany Commons, and the Genetics Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/11973
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600
Molecular evolution at homoeologous loci in allotetraploid cotton

(Malvaceae: *Gossypium hirsutum* L.)

by

Richard Clark Cronn

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Major: Genetics
Major Professor: Jonathan F. Wendel

Iowa State University
Ames, Iowa
1997
Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of
Richard Clark Cronn
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College
TABLE OF CONTENTS

CHAPTER ONE: GENERAL INTRODUCTION ...................................................... 1
  Description of Research Objectives ...................................................... 1
  Dissertation Organization ........................................................................ 2
  Literature Cited ......................................................................................... 4

CHAPTER TWO: GENE DUPLICATION AND THE IMPORTANCE OF POLYPLOIDY .......... 6
  Gene and Genome Duplication - An Introduction ...................................... 6
  Locus Duplication in Diploid Genomes .................................................... 7
  One Possible Fate - Gain of a Novel Gene Function ..................................... 8
  A Second Fate - Loss of Gene Function and Gene Silencing ......................... 9
  A Third Fate - Maintenance of Duplicate Gene Function ............................. 11
  A Final Fate - Gene Interaction through Polyploidy .................................... 13
  How Frequent Each Fate? .......................................................................... 15
  Literature Cited ........................................................................................ 16

CHAPTER THREE: POLYMORPHISM AND CONCERTED EVOLUTION IN A TANDEMLY REPEATED GENE FAMILY: 5S RIBOSOMAL DNA IN DIPLOID AND ALLOPOLYPLOID COTTONS .......... 21
  Summary ................................................................................................. 21
  Introduction ............................................................................................. 22
  Materials and Methods ............................................................................ 24
  Results ..................................................................................................... 27
  Discussion ............................................................................................... 33
  Acknowledgments .................................................................................... 43
  Literature Cited ....................................................................................... 43

CHAPTER FOUR: SIMPLE METHODS FOR ISOLATING HOMOELOGOUS LOCI FROM ALLOPOLYPLOID GENOMES ................................................................. 71
  Abstract .................................................................................................. 71
  Introduction ............................................................................................. 72
  Materials and Methods ............................................................................ 74
  Results ..................................................................................................... 78
  Discussion ............................................................................................... 80
  Acknowledgements .................................................................................. 83
  References ................................................................................................ 83

CHAPTER FIVE: MODE AND TEMPO OF SEQUENCE EVOLUTION AT HOMOELOGOUS LOCI FROM THE COTTON GENOME .......................................................... 91
  Abstract .................................................................................................. 91
  Introduction ............................................................................................. 92
  Materials and Methods ............................................................................ 93
  Results ..................................................................................................... 95
  Discussion ............................................................................................... 98
  Acknowledgements .................................................................................. 103
  Literature Cited ....................................................................................... 103

CHAPTER SIX: GENERAL CONCLUSIONS .................................................. 114

ACKNOWLEDGMENTS ............................................................................. 118
CHAPTER ONE
GENERAL INTRODUCTION

Description of Research Objectives

Polyploid formation has been a frequent process in the plant kingdom, as 50% (Grant, 1981) to 70% (Masterson, 1994) of all angiosperms show evidence of one or more rounds of genome doubling. Despite the prevalence and central importance of this process in flowering plant evolution (Soltis and Soltis, 1993), surprisingly little is known regarding the genetic consequences of allopolyploidy. One important feature of allopolyploidy is the creation of genome-wide genetic duplication and "redundancy." Theory suggests that this redundancy may lead to a relaxation of selective pressure on one of two duplicate loci, a process which can lead to an increase in the rate of DNA sequence divergence of a duplicate pair. The ultimate fate of one redundant locus may include: (i) maintenance of function due to selective pressure or concerted evolution (Hughes, 1994), (ii) loss of function due to gene silencing or pseudogene formation (Li, 1980; Walsh, 1995), and (iii) acquisition of a new function not present in the original diploid progenitors (Ohno, 1970; Ohta, 1994; Pickett and Meeks-Wagner, 1995). While each of these theoretical predictions seem sound, empirical evidence regarding the relative frequency of each of these fates is lacking, and the factors that promote a particular outcome remain obscure.

This thesis represents an initial investigation of the patterns of sequence divergence at duplicate loci that are related via recent allopolyploid formation, a relationship which is termed "homoeology" (Huskins, 1931). The experimental approach used in this study exploits a well-defined natural lineage — that of the cotton genus, *Gossypium* L. — which is composed of 45 diploid members from eight genome groups (A, B, C, D, E, F, G and K-genomes; \(2x = 2n = 26\)), and five allopolyploid species which combine one A-genome and one D-subgenome into a common nucleus (AD-genome; \(2x = 4n = 52\); reviewed in Endrizzi et al., 1985). Due to the combined efforts of a number of investigators over one-half century, the evolutionary relationships among the major genome lineages have been established (Wendel and Albert, 1992; Wendel et al., 1995; Seelanan et al., 1997), the genomes which participated in polyploid formation have been identified (Endrizzi et al., 1985; Wendel, 1989; Wendel et al., 1995; Cronn et al., 1996; Small et al., 1998), and extant examples of the diploid progenitors and their allopolyploid derivatives are readily obtainable. Of equal importance, the genomes of the putative progenitor diploid A- and D-genomes and the
allotetraploid A- and D-subgenomes have been well-characterized cytogenetically (Endrizzi et al., 1985; Price et al., 1990; Crane et al., 1993) and genetically, most recently by comparative linkage mapping approaches using a common set of molecular markers (Reinisch et al., 1994; Brubaker et al., 1998). These complementary sources of information provide the critical context for establishing that particular loci are homoeologous, as opposed to paralogous, and for evaluating the types of changes that have arisen between homoeologues in an evolutionary (phylogenetic) context.

An important product of the body of work described in this thesis is a detailed analysis of both the pattern and the degree of change in nucleotide sequences at fifteen sets of homoeologous loci in the tetraploid cotton genome. These loci range in copy numbers of several thousand per genome (5S rRNA genes) to unique sequences present in one to several copies per genome. Using an outgroup taxon (a C-genome diploid cotton from Australia) to polarize changes in the ingroup, phylogenetic and statistical methods have been used to evaluate both the patterns and rates of divergence at three hierarchical levels: between duplicate, homoeologous loci which reside on the A- and D-subgenomes of allotetraploid cotton; between orthologous loci from diploid progenitor genomes and the corresponding subgenome from allotetraploid cotton; and between orthologous loci in the A- and D-diploid genomes. Each of these sixteen loci provide an independent test of two important null hypotheses that comprise the prevailing “dogma” of polyploid evolution:

HYPOTHESIS 1: Homoeologous genes evolve independently.

HYPOTHESIS 2: Rates of sequence divergence for homoeologous loci are equivalent to each other, and to rates of the corresponding loci in parental diploids.

**Dissertation Organization**

The main body of this dissertation is organized into five chapters, one chapter that provides a review of relevant literature, three which consist of original research results on aspects of homoeologous locus evolution, and a final chapter which summarizes the results of this dissertation research.

Chapter Two of this dissertation is entitled “Gene Duplication and the Importance of Polyploidy”. Much of this chapter, indeed much of our current knowledge, draws upon the genetics of gene duplication in organisms with diploid genomes. While the events that yield locus duplications in diploid genomes are largely unrelated to the processes involved in polyploid genome duplication, previous literature concerning gene duplication at the diploid level is illuminating with regard to the theoretical frameworks which have been used to
evaluate duplicate locus evolution. In addition, results at the diploid level have led a number of investigators to make predictions about how duplicate loci may evolve in polyploid genomes. This literature is reviewed in depth, and the chapter concludes with a brief description of two examples of duplicate locus evolution in polyploids.

Chapter Three, entitled “Polymorphism and Concerted Evolution in a Tandemly Repeated Gene Family: 5S Ribosomal DNA in Diploid and Allopolyploid Cottons”, was published in the *Journal of Molecular Evolution* (Cronn et al., 1996) and provides a description of the sequence evolution of homeologous genes which encode the 5S ribosomal RNA. 5S RNA genes and their non-transcribed spacers exist in highly-repeated tandem arrays at a single locus in diploid cottons, and as two homeologous loci in allotetraploid cotton (Crane et al., 1993). In contrast to previous reports on the behavior of 45S rRNA genes from identical taxa (Wendel et al., 1995), 5S rDNA arrays have retained their subgenomic identity in allotetraploid cotton, thereby indicating that interlocus concerted evolution has not been an important factor in the evolution of 5S rDNA arrays. Furthermore, intra-locus homogenization forces have also been relatively weak. Evidence for this is provided by the high level of sequence polymorphism observed within and among arrays from a single individual. A novel and hopefully broadly-applicable model, one that incorporates the opposing forces of mutation and homogenization within a selective framework, has been outlined to account for these observations.

Chapter Four of this dissertation is entitled “Simple Methods for Isolating Homoeologous Loci from Allopolyploid Genomes” and has been submitted to the journal *Genome*. This article details specific methods developed for isolating strictly orthologous, duplicated loci from allotetraploid cotton and its diploid progenitors. The methods utilize restriction-digested, size-fractionated genomic DNA as templates for standard plasmid cloning or PCR amplification. The fractionation procedure, when combined with limited Southern hybridization and mapping information, can be used to separate homeoeologues from each other and from other divergent cross-hybridizing loci (paralogues). Even though the methods described in this chapter were specifically designed for isolating mapped, homeoeologous loci from allotetraploids, they should also be readily adaptable to a broad spectrum of loci from diploid and polyploid plants, and be useful for studying both recent and ancient gene duplications.

Chapter Five of this dissertation is entitled “Mode and Tempo of Sequence Evolution at Homoeologous Loci from the Cotton Genome” and has been prepared for submission to *Molecular Biology and Evolution*. This paper provides a glimpse into the patterns and rates
of nucleotide substitution at fourteen mapped, homoeologous loci from allopolyploid cotton and their orthologous counterparts in progenitor diploid genomes. While the majority of the loci described in this paper were revealed by anonymous \textit{PstI} mapping probes, six have been identified as genes, allowing patterns of variation to be partitioned into classes of "genic" (coding sequence) change and non-genic change. Since the loci included in this study are separated by a minimum of 20 centiMorgans, each of these loci provide an independent test of homoeologue divergence, as well as providing empirical data to test theoretical models concerning duplicate gene evolution.

Chapter Six, entitled "General Conclusions," provides a summary of the work presented in this dissertation. In addition, inferences are made regarding the general mode and tempo of homoeologous sequence evolution in polyploid cotton, and in other polyploid taxa as well.

\textbf{Literature Cited}


CHAPTER TWO

GENE DUPLICATION AND THE IMPORTANCE OF POLYPLOIDY

Gene and Genome Duplication - An Introduction

Gene duplication, first described by Sturtevant (1925) as the source of the Bar/Ultrabar phenotype in Drosophila melanogaster, has been widely recognized as an important genetic process in the evolution of eukaryotic genomes because it creates genetic redundancy, a condition required for the functional diversification of gene families (Ohno, 1970; Ohta, 1994a; Pickett and Meeks-Wagner, 1995). While genes can be duplicated by numerous mechanisms such as unequal crossing-over, reverse-transcriptase mediated events, replicative transposition and chromosomal rearrangements (Li and Grauer, 1991), the process of genome duplication ("polyploidization") is unique with respect to the extent of genic duplication it generates. Rather than giving rise to small and/or regional duplications, polyploidy multiplies the entire genic content of a diploid genome by a factor of two or more. This mode of gene duplication initially creates homoeologous loci (duplicated via polyploidy) often without disturbing gene structure, linkage relationships or gene expression status, making it possible for the vast majority of homoeologous genes to remain functional. Hence, the process of polyploidization — specifically, the generation of an entire genome-equivalent (or more) of redundant genetic information — sets in motion an astronomically large number of gene duplication “experiments” within the confines of a common nucleus.

Despite the relatively high prevalence of polyploidy in angiosperms (Grant, 1981; Soltis and Soltis, 1993; Masterson, 1994), surprisingly little is known about the mode or tempo of sequence divergence among loci which are duplicated during polyploidization. This is largely a consequence of the naturally high degree of genetic redundancy in most functionally diploid plants (Cavalier-Smith, 1985; Tanksley and Pichersky, 1988). The presence of numerous related genes in a common genome means that duplicated loci can be related either by polyploidy (a homoeologous relationship) or by other duplication processes or ancient cycles of whole-genome duplication (a paralogous relationship; Fitch, 1970). These technical hurdles have prevented detailed characterization of the modes of evolution of homoeologous loci in plants, even though these loci presumably play a central role in the evolutionary success and adaptive radiation of polyploid taxa (Spoorff, 1969; Werth and Windham, 1991; Grant, 1981). Accordingly, in order to discuss the potential importance and
ultimate fate(s) of genes duplicated by polyploidy, it can be illustrative to consider the fate of duplicate genes within the context they have been most frequently described - at the level of single gene duplications in diploid organisms.

**Locus Duplication in Diploid Genomes**

Well before the physical nature of the gene was understood, early geneticists such as J. B. S. Haldane (1932) and Calvin Bridges (1935) appreciated the potentially central role that gene duplication could play in evolution. If vital functions were assigned to a single locus, they reasoned, random mutations had a high likelihood creating deleterious alleles that would impair the function of that gene. In addition, mutation *per se* seemed an unlikely force for providing the apparently limitless variation seen in natural populations, since mutation acting alone could only replace existing functions with new functions, or (in the worst case) *result in a net loss of gene functions*. On the other hand, if vital functions could be assigned to two or more loci (which arise via gene duplication; Bridges, 1935), deleterious mutations in one gene copy could be compensated by a second, redundant copy. The existence of a redundant, duplicate gene could also allow for the addition of new or modified gene activities while maintaining the original gene function. Although these early authors disagreed about the ultimate fate of duplicated genes within populations, all agreed that evolutionary complexity (or “improvements,” according to Haldane) could be achieved most simply if genes existed at two or more loci.

To the credit of these early investigators, gene duplication is still considered the preeminent process responsible for the creation and evolution of new gene functions (Ohno, 1970; Li et al., 1985; Meeks-Wagner, 1994). The central importance of this process was summarized by Ohno (1970) who stated that “natural selection merely modified, while redundancy created”. This elegant and simple statement captures the essence of gene duplication, although only implicitly. To be more explicit, the process of gene duplication can: (1) change the complement of genes in a nucleus, creating gene families composed of only a few members (e.g., *rbcS* genes; Meagher, 1989) to arrays composed of thousands of repeated members (such as rRNA genes; Hillis and Dixon, 1991); and (2) allow for a variable relaxation of selective pressure on redundant loci, either freeing them from selection and allowing them to diversify as has happened with proteases and their inhibitors (Ohta, 1994b), or holding them under strict evolutionary constraints as seen with rRNA genes (Hillis and Dixon, 1991). Although functional divergence is clearly one of the most important outcomes of gene duplication, duplicate loci can ultimately encounter fates which range from mundane — maintenance of the original gene function — to fatal — pseudogene formation.
One Possible Fate - Gain of a Novel Gene Function

As the processes regarding the nature of the gene became more transparent, a radical benefit from gene duplication was proposed; that is, gene duplication could create a new gene with a previously unknown function (Haldane, 1932). During the process of being "ignored" by natural selection, a redundant gene will accumulate mutations at a higher rate than a locus under selection. Although the vast majority of these mutations will be deleterious (making pseudogene formation the most likely fate for a duplicated gene [Clark, 1994]), a small number of mutations may singly or in combination (Ohta, 1988) create a gene and subsequent gene product which confers a positive Darwinian advantage. Once formed, this selective advantage can drive a novel mutation quite rapidly to fixation, even if the overall mutation rate of favorable change is small and the effective population size is very large (Walsh, 1995).

The earliest evidence for this "functional divergence after gene duplication" was provided by the amino acid sequences of the serine proteases trypsin (Keil et al., 1963) and chymotrypsin (Kauffman, 1965). These proteases recognize very different substrates (trypsin cleaves the C-terminal side of basic amino acids while chymotrypsin cleaves the C-terminal side of aromatic amino acids), yet the amino acid sequence of the catalytic sites and the overall secondary structure of these proteins are highly similar (Neurath et al., 1967) and provide evidence for a common ancestry. Due to the small number of radical amino acid differences in the active sites of trypsin and chymotrypsin, these proteases likely diverged in several steps whereby one protein was rendered non-functional with respect to either its original or newly-gained catalytic function. Such a process requires "compensatory" mutation during a period of non-functionality (Ohta, 1988; Hughes, 1994) to restore the protein's original function, or to give it an altered and novel activity. Since the precursors of these proteases - trypsinogen and chymotrypsinogen - require proteolytic activation from trypsin for proper maturation, trypsinogen appears to be the most likely progenitor of these two proteases.

There are now a number of examples that invoke gene duplication, followed by positive Darwinian selection, as the cause for creating proteins with novel functions. A brief list includes stomach lysozyme of ruminants (Irwin and Wilson, 1990), hemoglobin \( \gamma \) (Fitch et al., 1991), ion channel family (Strong et al., 1993) and the growth hormone-prolactin gene family (Ohta, 1993). While no convincing examples exist at present showing that polyploidization per se has provided the genesis of a new gene activity, it is clearly an important source of genetic redundancy, the raw material from which new genes are created.
**A Second Fate - Loss of Gene Function and Gene Silencing**

While gene duplication must always precede the evolution of a new gene function, the most likely outcome of a duplication event is that one copy becomes nonfunctional (Nei, 1969). Duplicate loci can be transiently silenced by epigenetic means (such as methylation or chromatin repatterning; Flavell, 1994; Matzke et al., 1996; Chen and Pikaard, 1997), or they can be more permanently silenced by the accumulation of deleterious mutations, such as insertions and deletions that change the reading frame or introduce a premature stop codon (Quigley et al., 1989; Gottlieb and Ford, 1997). In either case, the relaxation of selective pressure at a silenced locus can hasten its 'erosion' by allowing further mutation, ultimately assisting in the permanent (fixed) silencing of the gene (Walsh, 1995).

Isozyme analyses in particular have provided a wealth of data regarding the frequency of silencing at duplicate loci in diploid and polyploid organisms (Allendorf et al., 1984; Voelker et al., 1980; Wilson et al., 1983; Gastony, 1991). While this technique alone lacks the power to reveal the exact change responsible for locus silencing, it has been effectively used to identify null (inactive) alleles, which can then be evaluated via molecular dissection. The recent characterization of an allozyme null locus, cytosolic phosphoglucose isomerase (**PgiC**; Ford et al., 1995; Gottlieb and Ford, 1996; Gottlieb and Ford, 1997) from *Clarkia* provides a detailed example of the evolution of one pseudogene, and yields greater insight into selective forces that may influence pseudogenization. Three general findings of this paper can be compared with previous theoretical predictions on pseudogenization:

(i) The upstream promoter and the first five exons (of 23 total) of **γPgiC2** were deleted. An additional nine deletions were observed in the eighteen remaining exons, which combined to form six frameshift mutations, one premature stop codon, and the elimination of four intron splice junctions. Hence, the silencing of **PgiC2**, while it may have initially been due to a single severe mutation, has hastened the accumulation of deleterious mutations, any combination of which seem likely to maintain the inactivated state of this gene. This finding conforms to theory which predicts that a gene will rapidly degrade beyond recognition after pseudogenization due to the relaxation of purifying selection (Li et al., 1985; Walsh, 1995).

(ii) Despite the large number of insertion/deletions observed, the remaining coding sequence of **γPgiC2** showed only four replacement substitutions (out of ~390 amino acids) relative to functional **PgiC** genes, and the rates of sequence divergence between **γPgiC2** and functional **PgiC1** were not significantly different. This finding does not conform to theoretical expectations; since pseudogenes are freed from
selective constraints, they should exhibit an increased nucleotide substitution rate (Li, Gojobori and Nei, 1981) detectable by one of several relative rate tests (e.g., Tajima, 1993). The absence of replacement substitutions, combined with the lack of acceleration in nucleotide substitution rate in ψPgiC2, led the authors to conclude that the locus had been recently pseudogenized (Gottlieb and Ford, 1997). Hence, patterns and rates of nucleotide divergence observed in pseudogenes may not fulfill theoretical predictions and may be highly dependent upon the age of the pseudogene.

(iii) After the duplication of an ancestral PgiC, the genus Clarkia diverged into 32 diploid species, 17 of which maintained two functional PGI genes (PgiC1 and PgiC2) and 15 of which had one functional gene and a pseudogene (PgiC1 and ψPgiC2). Thus, PgiC2 appears to be dispensable. If mutation occurred in the redundant PgiC2 gene to yield a PgiC2 allele with diminished catalytic ability, theory suggests that the mutant form of enzyme would have a selectively neutral effect (Nei and Roychoudhury, 1973; Li, 1980) since it can be compensated by the functional PgiC1 gene product. Gottlieb and Ford (1997) point out, however, that functional PGI is a dimer composed of similar (homodimers) or dissimilar (heterodimers) subunits. Incorporation of mutant (m) PgiC2 gene products into homodimers (PgiC2m: PgiC2m) or heterodimers (PgiC1: PgiC2m) may be selectively disadvantageous, since only one combination (PgiC1:PgiC1) may produce a functional PGI enzyme. Hence, silencing of one of two duplicate loci may actually be favorable in instances where mutant alleles lead to a selectively unfavorable interaction.

These simple examples, while not inclusive for all possible pseudogenization scenarios, show that theoretical predictions concerning the fate of a “silenced” gene (in this case by deleterious mutation) may be either accurate or incorrect, depending upon the pseudogene attributes under consideration. The two instances where theoretical predictions appear equivocal highlight the importance of the amount of time passed since pseudogene formation and the effect selection may have upon variant alleles. The effect of selection upon duplicate genes has been addressed a number of ways in theoretical models, yet at present there remains no “best assumption” for how selection acts to promote or prevent silencing and pseudogenization. Selection regimes have proposed that null alleles can be strictly neutral (Nei and Roychoudhury, 1973; Kimura and King, 1979), or that one of several null allele genotypes can be under selection. These include the “double recessive” fitness models, where individuals homozygous for null alleles at one duplicate locus (Takahata and Maruyama, 1979; Watterson, 1983) or null alleles at reciprocal duplicate loci (Werth and
Windham, 1991) are at a selective disadvantage. Alternatively, Allendorf (1979) proposed a selection regime whereby individuals possessing three or more functional alleles (e.g., AABb, AaBB or AABB) at duplicate loci had reduced fitness due to an excess of gene product. Walsh (1995) has argued that such selection is unlikely in diploids, since the fixation of a maladaptive gene duplication would be unfavorable (i.e., removed by purifying selection) in the first place. However, this scenario seems quite likely at duplicate loci created during polyploidization (as in the tetraploid salmonids modeled in Allendorf's paper), since the process of polyploidization can immediately fix potentially unfavorable gene combination.

**A Third Fate - Maintenance of Duplicate Gene Function**

The processes responsible for creating gene duplications play a vitally important role in determining whether a previously non-redundant gene will be able to function in its redundant state. Processes which duplicate individual genes or portions of genes (i.e., unequal-crossing over), or through RNA intermediates (i.e., transposition or other reverse transcriptase-mediated events) have a higher likelihood of making duplications which cannot be properly expressed than do processes which duplicate larger chromosomal segments (translocations, chromosomal fusions or polyploidy; Li and Grauer, 1991). In fact, small-scale duplication processes can only produce a functional duplicate locus when gene coding sequences are copied intact (not prematurely truncated), when regulatory regions upstream and/or downstream of the gene are faithfully reproduced, or if the freshly duplicated gene is fortuitously inserted in close proximity to a suitable gene control region. If duplicated genes arrive in a non-functional state, they will degenerate rapidly into unrecoverable pseudogenes unless they are immediately rescued, either by additional rounds of crossing-over or by gene conversion.

When genes are faithfully duplicated in a functional state, the forces of selection — both “positive” selection (Tautz, 1992) and “negative” or purifying selection (Ferris and Whitt, 1979) — will act upon both loci and determine whether a redundant gene remains functional, and if so, for how long. Most models suggest that if functional divergence between duplicated loci does not take place, positive selection will be minimal and redundancies will be resolved by the silencing of a locus (Walsh, 1995; see previous section for other references). Despite this prediction, isozyme surveys (e.g., Gottlieb, 1982) and molecular genetic analyses (Meagher et al., 1989; Walker et al., 1995; Popadic and Anderson, 1995) show that a vast array of duplicated genes are maintained in a functional state for long periods of evolutionary time, even in the apparent absence of functional
divergence. These observations provide evidence that duplicate genes can be maintained in part through the action of purifying selection (Hughes and Hughes, 1993).

Two remarkable cases that highlight the maintenance of gene function post-duplication are provided by the small subunit of ribulose bisphosphate carboxylase (rbcS; Meagher et al., 1989) in flowering plants and glutamine synthetase (GS; Walker et al., 1995) in Pisum. In each of these cases, members of major lineages (such as rbcS in the Solanaceae and GS in the Fabaceae) each possessed a conserved number of loci for these genes, indicating that the duplications which gave rise to redundancy occurred prior to the radiation of these groups. When sequence "genealogies" were reconstructed for rbcS, however, paralogous loci within a taxon (e.g., Nicotiana) were more similar to each other than to their orthologous counterparts in other closely related taxa (e.g., Solanum). To explain this lack of correct orthology/paralogy relationships, the authors concluded that the multiple copies of rbcS in each taxon had undergone concerted evolution, effectively "erasing" evidence of orthological relationships among these loci. A similar finding has been reported for the duplicate, unlinked GS genes of Pisum sativa (Walker et al., 1995). These genes are present at 2-3 copies per genome in papilionoid legumes, indicating a duplication event ~ 50 million years ago. Despite this long history of divergence, twin GS genes (GS3A and GS3B) from Pisum show approximately 1% nucleotide divergence in coding regions and only three amino acid replacements. In addition, these genes show conserved reading frames (no deleterious mutations) and are both expressed at the level of mRNA. The only substantial divergence between GS3A and GS3B are in introns five and ten; these introns differ considerably in size (due to indels) and in nucleotide sequence (93% similarity) between these loci, while pairwise comparisons of all other introns show conserved sized and 96 - 98% nucleotide identity. Based upon these data, the authors concluded that GS3A and GS3B have interacted via concerted evolutionary processes, creating two nearly-identical GS genes which are both functional and highly constrained by selection.

These two examples provide evidence that purifying selection can act upon duplicate genes to maintain the original function of presumably redundant genes. This type of selection works by preventing the accumulation of mutant alleles within loci, as well as by minimizing the divergence between duplicate loci. In doing so, processes that promote sequence homogeneity, such as gene interaction (i.e., concerted evolution) between non-orthologous loci, can become selectively advantageous. This process can make a pair of duplicated, non-orthologous genes from a single genome appear more similar to each other than to their orthologues in more divergent lineages (Meagher et al., 1989; Walker et al.,
The mechanism(s) responsible for maintaining sequence homogeneity certainly includes concerted evolutionary processes such as gene conversion or unequal crossing-over, but they may also include other (as of yet unknown) phenomena.

**A Final Fate - Gene Interaction through Polyploidy**

As early as 1951, S. G. Stephens speculated on the genetic novelty that may be created by combining two divergent loci into one genome during allopolyploidization.

*What happens to a genetic locus after it has become duplicated as a result of amphidiploidy? Does it retain the same potentialities which it possessed in the diploid parent?... Has anything radically new been added to the ancestral system?* (Stephens, 1951, pg. 131)

Since allopolyploid genomes combine two or more related genomes into a common nucleus, there exist at a minimum two closely-related homoeologous loci (one per subgenome) in a single nucleus. These homoeologues typically share greater sequence similarity to each other than to other duplicate (paralogous) genes in the nucleus, since they are often separated by a more recent duplication event. As a consequence, it can be anticipated that homoeologous loci could interact transiently during homology-driven scanning processes such as recombination or gene repair, even if homoeologous chromosomes themselves show little evidence of interaction. These interactions could lead to crossing-over in low-copy genes (Spiersen, 1991), and concerted-evolution between high-copy genes in both subgenomes (Wendel et al., 1995). This aspect of duplicate gene evolution is unique to polyploid genomes. Only after combining two divergent genomes — once separated and evolving in different trajectories — into a single nucleus can two loci interact to create novel combinations, which in the words of Stephens, “may be radically new... to the ancestral system.”

While nuclear RFLP surveys (Song and Osborne, 1995) and linkage mapping (Reinisch et al., 1994; Shoemaker et al., 1996) indicate that inter-genomic interaction must occur frequently during polyploid evolution, only two instances of sequence-level interactions have been proposed to date. In the first report, Sperisen et al. (1991) found that two 1,3-β-glucanase cDNA clones (GL31 and GL43) isolated from ST-genome *Nicotiana tabacum* were unique relative to most genomic clones they had isolated. Typical 1,3-β-glucanase gene sequences fall into two classes: one class called “GLA” which are similar to loci from the diploid *N. sylvestris* (the model S-genome progenitor), and a second class called “GLB” which are similar to loci from *N. tomentosiformis* (the model T-genome progenitor). The cDNA clones showed unique restriction patterns that were missing from both the S- and T-genomes,
indicating that they had arisen after polyploid formation. Sequence analysis revealed that GL31 had a higher overall sequence similarity to the \textit{N. sylvestris}-like (GLA) locus and that GL43 showed a higher sequence similarity to the \textit{N. tomentosiformis}-like (GLB) locus. However, the sequence similarity was not uniform across the length of these genes, as the central one-third of GL31 showed higher similarity to GLB, and the corresponding region of GL43 showed higher similarity to GLA. These authors suggested that the two cDNAs — GL31 and GL43 — represented products from two homoeologous genes that had interacted in a reciprocal manner (e.g., by crossing-over) between the S- and T-subgenomes. In the absence of linkage mapping information, this example must be considered as equivocal with respect to the issue of intergenomic interaction. Nevertheless, it remains a promising piece of evidence and highlights how low-copy genes from two divergent subgenomes may recombine to yield new gene combinations.

The second example of inter-genomic interaction in plants involves the 45S rDNA loci from allotetraploid members of \textit{Gossypium} (Wendel et al., 1995). \textit{In-situ} hybridization reveals that these loci exist as highly-repeated tandem arrays which localize to two major loci per diploid genome, with corresponding sets in each subgenome of the allopolyploid (Price et al., 1991; Crane et al., 1993). In diploid A- and D-genome cottons, sequencing of the internal transcribed spacer (ITS) region of the 45S locus revealed that rDNA had undergone complete intra-locus (within array) and inter-locus (between arrays) concerted evolution in these taxa, yielding homogenized sequences with low heterogeneity and no locus specificity. As a consequence of this rapid concerted evolution, the A- and D-genome sequences differed by approximately 40 substitutions across a 685 bp region. Interestingly, sequence analysis of the five AD-genome allopolyploid species of cotton did not show an "additive" (i.e., A + D) pattern; rather, sequences obtained from each of the species were homogeneous, indicating that concerted evolution had homogenized rDNA sequences both within and between subgenomes. In addition, phylogenetic analysis of the sequences and genomic southern hybridization with rDNA probes showed that four species (\textit{G. hirsutum}, \textit{G. tomentosum}, \textit{G. barbadense} and \textit{G. darwinii}) possessed D-genome-like arrays, while a single species (\textit{G. mustelinum}) possessed an A-genome-like array. Since these allopolyploids are believed to have formed only once, the resulting rDNA sequences from these AD-genome allotetraploids provide convincing evidence that concerted evolution can occur between sequences localized on different subgenomes, and that concerted evolution can occur in either one or both directions in incipient lineages.
How Frequent Each Fate?

An obvious question that can be posed concerning the fate of duplicate, homoeologous genes is "how frequently do each of these fates (new gene formation, pseudogenization or maintenance) occur"? At present, surprisingly little empirical data are available to answer this question. This absence of data largely reflect the difficulties of determining whether a duplicate locus has become a pseudogene, if a duplicate locus has acquired a novel function, and calculating a rate for pseudogene formation among duplicate gene pairs. Clearly, mutations that give rise to structural aberrations in exon sequences (such as the deletion events that created \( \psi PgrC2 \); Gottlieb and Ford, 1997) can provide convincing evidence that a duplicated gene has become pseudogenized. However, there have been an increasing number of reports which show genes to be silenced via epigenetic processes such as methylation or chromatin "repatternning" (reviewed in Flavell, 1994). These mechanisms can inhibit gene expression at the level of transcription or translation, and can be influenced by increasing gene copy number (Matzke et al., 1996) and changes in ploidy (Chen and Pikaard, 1997; Scheid et al., 1996). All of these processes effect gene expression without imposing changes upon the nucleotide sequence, making sequence-level analyses (for nucleotides or proteins) uninformative at best.

Nevertheless, if we assume that we can confidently measure the rate of silencing at a given duplicate locus, it may still provide an inaccurate measure of pseudogenization since the fate of a pseudogene may not always be sealed at the time of formation. The processes of compensatory mutation (Ohta, 1988), gene conversion (Walsh, 1995) and homologous and xenologous recombination (Matzke et al., 1996) can resurrect "dead" pseudogenes, and in the process, restore the original gene function or endow the gene with a new function (Marshall et al., 1994). Accordingly, any estimate of the rate of pseudogene formation must be corrected for the frequency of pseudogene "rescue," a value which may not reach steady-state until the passage of several millions of years (Marshall et al., 1994).

Given these caveats, the theoretical predictions of Walsh (1995) — while admittedly simple — illuminate parameters which may influence the frequency of each of the three fates for a duplicated locus, providing the closest thing to a concrete answer to the question posed at the beginning of this section. In his model, Walsh calculates the probability that a duplicate gene can escape pseudogenization to be approximately

\[
P = \left( \frac{(1-e^\alpha)}{(pS)} + 1 \right)^{-1}
\]

where \( S = 4^*Ne^*s \), or four times the product of the effective population size and the selective advantage of a duplicate locus, \( p \) is the ratio of the advantageous to null allele mutation rates,
and \( s \) is the selective advantage of a duplicate locus. Using Walsh's own example, if we assume that \( s = 0.0001 \), \( Ne = 5000 \) and \( p = 5 \times 10^{-5} \), the probability of fixing a duplicate, functional gene is \( \sim 1\% \), and the probability of pseudogene formation is 99%. While these values may be arbitrary, they are nevertheless instructive; though the probability of an advantageous gene being fixed appears low in this example (\( \sim 1\% \)), it is quite large compared to the value of \( p \) (\( = 5 \times 10^{-5} \)) which yields 1 advantageous mutations in 20,000, while the remaining 19,999 can silence the locus. Note that when the effective population size \( Ne \) is increased, the probability of fixing an advantageous gene is also increased; if \( Ne \) is increased to 50,000, \( P = 9\% \); if \( Ne = 500,000 \), \( P = 50\% \). The reason for this "magnification" is the assumption that duplicate genes can yield *selectable advantages* (through the term \( s \)); hence, even though a gene duplication may yield a minute selective advantage, populations with large effective populations may fix these duplications at a surprisingly high rate. An important final note with respect to this model is that the selection term \( s \) applies both to duplications that preserve the original function of a duplicated gene (whether unmodified or gene-converted), or to duplications that create a novel gene function. Ultimately, the long-term preservation of a duplicate, redundant gene depends upon the selective advantage it confers.

**Literature Cited**


ribosomal RNA loci in meiotic chromosomes of cotton (*Gossypium hirsutum*).  
Genome 36: 1015-1022.

Nature 265: 258-260.


Duplication of the α-globin gene mediated by repetitive L1 sequences in an early  
ancestor of simian primates.  

Flavell, R. B. 1994. Inactivation of gene expression in plants as a consequence of specific  
sequence activation.  

Ford, V. S., B. R. Thomas and L. D. Gottlieb. 1995. The same duplication accounts for the  
PgiC genes in *Clarkia xantiana* and *C. lewisii* (Onagraceae).  

revisited.  

Science 216: 373-380.

Gottlieb, L. D., and V. S. Ford. 1996. Phylogenetic relationships among the sections of  
*Clarkia* (Onagraceae) inferred from the nucleotide sequences of *PgiC*.  


Haldane, J. B. S. 1932. The causes of evolution.  
Longmans Green, London.

Hillis, D. M. and M. T. Dixon. 1991. Ribosomal DNA: Molecular evolution and  
phylogenetic inference.  

Hughes, M. K. and A. L. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal,  
*Xenopus laevis*.  


J. Biol. Chem. 265: 4944-4952.


CHAPTER THREE

POLYMORPHISM AND CONCERTED EVOLUTION IN A TANDEMLY REPEATED GENE FAMILY:
5S RIBOSOMAL DNA IN DIPLOID AND ALLOPOLYPLOID COTTONS

A paper published in the Journal of Molecular Evolution

Richard C. Cronn, Xinping Zhao, Andrew H. Paterson, and Jonathan F. Wendel

SUMMARY

5S RNA genes and their non-transcribed spacers are tandemly repeated in plant genomes at one or more chromosomal loci. To facilitate an understanding of the forces that govern 5S rDNA evolution, copy-number estimation and DNA sequencing were conducted for a phylogenetically well-characterized set of 16 diploid species of cotton (Gossypium) and 4 species representing allopolyploid derivatives of the diploids. Copy-number varies over twenty-fold in the genus, from approximately 1000 - 20,000 copies/2C genome. When superimposed on the organismal phylogeny, these data reveal examples of both array expansion and contraction. Across species, a mean of 12% of nucleotide positions are polymorphic within individual arrays, for both gene and spacer sequences. This shows, in conjunction with phylogenetic evidence for ancestral polymorphisms that survive speciation events, that intralocus concerted evolutionary forces are relatively weak and that the rate of inter-repeat homogenization is approximately equal to the rate of speciation. Evidence presented also shows that duplicated 5S rDNA arrays in allopolyploids have retained their subgenomic identity since polyploid formation, thereby indicating that interlocus concerted evolution has not been an important factor in the evolution of these arrays. A descriptive model, one which incorporates the opposing forces of mutation and homogenization within a selective framework, is outlined to account for the empirical data presented. Weak homogenizing forces allow equivalent levels of sequence polymorphism to accumulate in the 5S gene and spacer sequences, but fixation of mutations is nearly prohibited in the 5S gene. As a consequence, fixed interspecific differences are statistically under-represented for 5S

1 Reprinted with Permission from the Journal of Molecular Evolution, vol. 42: 685-705
2 Department of Botany, Iowa State University, Ames, IA 50011
3 MIPM, Michigan State University, Lansing, MI 48109-0672
genes. This result explains the apparent paradox that despite similar levels of gene and spacer diversity, phylogenetic analysis of spacer sequences yields highly resolved trees, whereas those based on 5S gene sequences do not.

**Key Words:** 5S rDNA; Concerted evolution; *Gossypium*; Polyploidy; Molecular evolution; Genome evolution; Repetitive DNA; Molecular phylogeny

**INTRODUCTION**

A universal feature of ribosomes is the presence of small 5S RNA molecules that are associated with the large subunit. These RNAs usually are encoded by genes organized into tandemly repeated arrays that occur at one or more chromosomal loci (Long and Dawid, 1980; Gerbi, 1985; Appels and Honeycutt, 1986; Sastri et al., 1992). At each locus, the 5S genes, which are approximately 120 bp in length, are separated from one another by intergenic, non-transcribed spacers, which in plants vary in length from approximately 100 to 700 bp. In plants, the total number of repeats (= gene + spacer) per genome varies by over two orders of magnitude, from less than 1,000 to over 100,000 (Schneeberger et al., 1989; Sastri et al., 1992).

Because of this ubiquity and prominence, 5S rDNA diversity and evolution have been studied in broad range of plants, with a particular focus on patterns of nucleotide conservation and divergence (Wolters and Erdmann, 1988; Campell et al., 1992; Baum and Johnson, 1994; Kellogg and Appels, 1995), and structural organization of 5S arrays (Scoles et al., 1988; Dvorák et al., 1989; Schneeberger et al., 1989; Gottlob-McHugh, 1990; Appels et al., 1992; Röder et al., 1992; Kanazin et al., 1993). This accumulating data base demonstrates that 5S RNA genes are highly conserved in the plant kingdom, both with respect to length and nucleotide sequence, whereas the intergenic spacers evolve more rapidly. An additional generalization is that individual 5S rDNA repeats do not evolve independently. As with 18S-26S rDNA and other tandemly repeated multigene families, the hundreds to thousands of repeats within 5S arrays retain a high degree of identity due to homogenizing forces collectively referred to as concerted evolution (Zimmer et al., 1980).

Most models of concerted evolution invoke one or two molecular processes, unequal crossing-over and gene conversion (Hood et al., 1975; Smith, 1976; Dover, 1982; Arnheim, 1983; Ohta, 1983, 1984, 1990; Ohta and Dover, 1983; Nagylaki, 1984a, 1984b, 1990; Basten and Ohta, 1992; Schlötterer and Tautz, 1994). Regardless of mechanism, the rate at

---

4Department of Soil and Plant Sciences, Texas A&M University, College Station, TX 77843
which variant repeat types become homogenized depends upon several factors, including the number of repeats in an array, the frequency of homogenization events relative to the formation of new variants via mutation, the intensity of natural selection, and effective population size (Smith, 1976; Ohta, 1983, 1984, 1990; Nagylaki, 1984a, 1984b, 1990; Li et al., 1985; Basten and Ohta, 1992; Linares et al., 1994). Interplay among these and other variables leads to a continuum in the degree of heterogeneity exhibited by repeats within arrays: when concerted evolution is “strong”, repeats are identical or nearly so; when weaker, sequence heterogeneity is observed. With respect to 5S rDNA arrays, sequence heterogeneity among repeats within individual arrays has been reported from several plant groups (Rafalski et al., 1982; Gottlob-McHugh et al., 1990; Cox et al., 1992; Kellogg and Appels, 1995). In each case, a moderate level of sequence variation has been detected, demonstrating that concerted evolutionary forces have not been strong enough to overcome those that generate variation.

An additional complexity arises when 5S arrays occur at more than one chromosomal location. This is the situation in the majority of plants, where polyploidy is prevalent (Masterson, 1994). In these cases, the outcome of concerted evolutionary processes depends not only on the previously listed factors, but also on the frequency of genetic exchanges between homologous sequences on homoeologous chromosomes (Ohta and Dover, 1983; Nagylaki, 1984b, 1990; Schlötterer and Tautz, 1994). When barriers prevent such exchange, chromosome specific arrays may evolve independently despite concerted evolution of repeats within arrays, as appears to be the case for 5S arrays in the grass tribe Triticeae (Scoles et al., 1992; Kellogg and Appels, 1995). The alternative scenario, homogenization of repeats from different arrays (interlocus concerted evolution; e.g., Wendel et al., 1995a), has yet to be demonstrated for 5S rDNA in plants (Sastri et al., 1992).

As part of an ongoing effort to elucidate evolutionary process in a model system involving diploid and allopolyploid members of *Gossypium* (VanderWiel et al., 1993; Reinisch et al., 1994; Wendel et al., 1995a, 1995b), we here describe patterns of polymorphism in and concerted evolution of 5S rDNA sequences. Fluorescent *in situ* hybridization analysis has revealed that *Gossypium* 5S arrays occupy only a single centromeric chromosomal location in A-genome and D-genome diploid species (R. Hanson and D. Stelly, pers. comm.) and two corresponding loci in the AD-genome allopolyploids (Crane et al., 1993). This organization contrasts with that of the major 18S-26S rDNA arrays, which occupy two relatively distal chromosomal loci in each diploid and in each subgenome of the allopolyploids. Repeats of *Gossypium* 18S-26S rDNA arrays evolve under
strong inter- and intralocus concerted evolution (Wendel et al., 1995a), leading us to ask whether other tandemly repeated sequences evolve in a similar manner. To address this issue, we cloned and sequenced multiple 5S rDNA repeats from allopolyplloid species of cotton and all lineages representing their diploid progenitor genomes. We were particularly interested in whether 5S sequences from individual arrays are homogeneous or polymorphic in diploid and allopolyplloid species, and whether 5S sequences from the two arrays of the allopolyplloid species evolve independently or in a concerted fashion.

**MATERIALS AND METHODS**

*Organismal Context - Gossypium* is a genus of approximately 50 species of perennial herbs to small trees with centers of diversity in Australia, Africa-Arabia and Mexico (Fryxell, 1979; Fryxell et al., 1992). Extensive chromosomal diversification accompanied radiation of the genus, leading to the evolution of diploid “genome groups”, designated A through G based on chromosome size differences and meiotic pairing behavior in interspecific hybrids (reviewed in Endrizzi et al., 1985). Phylogenetic analyses, based on chloroplast DNA restriction site variation (DeJoode, 1992; Wendel and Albert, 1992) reveal diploid clades that are congruent with taxonomic alignments, geographic distributions, and genome designations. In addition to the diploid species, there are five species of tetraploid (2n = 52) *Gossypium* (Brubaker and Wendel, 1993; Fryxell, 1992), all endemic to the New World. A wealth of data establishes that these tetraploid species are derived from allopolyplloidization between A-genome and D-genome progenitors (Endrizzi et al., 1985; Wendel, 1989; Reinisch et al., 1994). Chloroplast DNA sequence divergence data suggest that the two parental genomes diverged from a common ancestor 6 - 11 million years ago (mya) and that they became reunited in a common nucleus in the mid-Pleistocene (1 - 2 mya; Wendel, 1989; Wendel and Albert, 1992). Although the actual progenitor diploid taxa are most likely extinct, data suggest that the best models of the ancestral D-genome parent are *G. raimondii* (Endrizzi et al., 1985) and *G. gossypioides* (Wendel and Albert, 1992; Wendel et al., 1995b), and that the A-genome donor was most similar to present-day *G. herbaceum* (Endrizzi et al., 1985). Following polyplloidization, allopolyplloids diverged into five modern species (DeJoode and Wendel, 1992; Brubaker and Wendel, 1993; Wendel et al., 1994), including the commercially important *G. hirsutum* (“upland cotton”) and *G. barbadense* (“Pima” and “Egyptian” cotton).

*Plant Materials -* We isolated total DNA, using methods detailed in Paterson et al. (1993), from individual plants representing 20 diploid and allopolyplloid *Gossypium* species
(Table 1). Included were both Old World, A-genome diploids, all 13 New World, D-genome diploids, and four of five New World, AD-genome allopolyploids. Because Australian C-genome cottons are basal within the genus (Wendel and Albert, 1992), we selected one Australian species (G. robinsonii) for inclusion as an outgroup. For comparative purposes, we also isolated DNAs from a synthetic allopolyploid, 2(A2D1), derived via colchicine-doubling the sterile inter-genomic F1 hybrid G. arboreum x G. thurberi (synthesized by J. O. Beasley).

5S rDNA Amplification and Cloning - We amplified 5S repeats from each genomic DNA by the polymerase chain reaction (PCR), using the cycling parameters specified in Cox et al. (1992). Reactions were conducted in 50μl volumes containing 10-500 ng genomic DNA and 10 pmol of each primer. Primers 5SF (5'-GAG-AGT-AGT-[A/T]-[A/T][C/G]G-ATG-GG) and 5SR (5'-GGA-GTT-CTG-[C/T]G-GGA-TCC-GG) were designed to anneal to the 5S rRNA gene at nucleotides 69 to 88 and 28 to 49, respectively. Under the PCR conditions of Cox et al. (1992) and a variety of other amplification protocols, products consisted of a ladder of 5S repeats that ranged from single repeats (~300 bp) to multimers greater than 10 kb in length. These ladders were digested with BanII to yield 5S pools that were mostly monomeric in length, and were readily cloned into M13mp18 (Gibco-BRL). Individual clones were isolated using routine procedures (Sambrook et al., 1989) and were sequenced using an ABI automated sequencer. A total of 99 5S rDNA sequences were generated for the 20 species and the synthetic allopolyploid 2(A2D1), with an average of four clones per diploid and eight clones per polyploid species. To evaluate sequencing error and to verify the intra-individual polymorphisms observed, residual DNAs from 40 clones were sequenced manually using standard methods of dideoxy sequencing.

5S rDNA Copy-number Estimation - To quantify the number of 5S repeats in each species examined, DNAs were digested to completion with RNase A (to remove potentially contaminating RNAs) and subjected to slot-blot analysis. Previous studies (Edwards et al., 1974; Kadir, 1976; Michaelson et al., 1991) demonstrate that 2C DNA content varies little among Gossypium species within genome groups, but more between species from different genome groups. For estimating slot-blot loadings, we used published 2C estimates (Edwards et al., 1974; Kadir, 1976; Michaelson et al., 1991) that are in close agreement with newer data (J. S. Johnston and H. J. Price, pers. comm.): A genome = 3.8 pg; C genome = 5.0 pg; D genome = 2.0 pg; AD genome = 5.8 pg.

Samples of genomic DNA, quantified by microfluorometry, were transferred onto MSI MagnaGraph nylon membranes using a 72 position slot-blot apparatus (Schleicher and Schuell). For each sample, 10^4 2C genomic equivalents of DNA were denatured in 0.4M
NaOH at 65°C for 30 minutes, neutralized with an equal volume of 2M NH₄OAc (pH 7), and transferred to the slot wells. Copy-number standards for A- and D-genome 5S rDNA were generated using PCR-amplified inserts from the M13 clones *G. herbaceum* pGh328-3 (A-genome) and *G. raimondii* pGr330-1 (D-genome). Quantified amounts (ranging from 10⁶ to 10⁹ copies) were applied in duplicate to each blot. DNAs were UV-crosslinked to membranes that were dried at 65°C for two hours. Hybridization probes were generated from both clones using random-primer labeling and the hybridization and wash conditions detailed in Wendel et al. (1995a). After washing, signal intensity was quantified by phosphorimaging on a Molecular Dynamics 400 phosphorimager. Absolute 5S rDNA copy-numbers for each slot were estimated by linear regression using a standard curve of “volume above background” values (obtained from ImageQuant™ software). Copy-numbers per 2C genome were calculated by dividing absolute number of copies per slot by the number of genomic equivalents loaded in that slot. To estimate standard errors for 5S copy number, four to eight replicate experiments were conducted.

**Sequence Alignment and Analysis** - Alignment of multiple 5S sequences from diploid individuals was straightforward, as there were low levels of nucleotide substitution and length variation. Sequence alignment was more problematic, however, for sequences from different diploid genome groups and from different subgenomes of the allopolyploid species. A global alignment among all 99 5S rDNA sequences was accomplished using the PILEUP module of the Wisconsin GCG computer package, version 8.0 (Devereux et al., 1984). After visual inspection of the resulting alignment, minor changes were made in the intergenic spacer region. The alignment used is shown in Figure 1.

To quantify the diversity of 5S rDNA sequences detected at various levels of organization (e.g., within individuals, within genome groups), we calculated two descriptors: $p_a$, the proportion of nucleotide sites that are polymorphic and $\pi$, nucleotide diversity (Nei, 1987). These calculations were facilitated by using the Molecular Evolutionary Genetics Analysis software (MEGA v 1.0; Kumar et al., 1993). To test for equivalent patterns of sequence evolution in the 5S gene and spacer regions, we made 2 x 2 contingency tables (McDonald and Kreitman, 1991; Kellogg and Appels, 1995) for selected taxa, where observed fixed and polymorphic differences (columns) were tabulated for 5S gene and spacer sequences (rows). Because the number of expected fixed differences in the 5S gene was low, two-tailed Fisher exact tests (Sokal and Rohlf, 1981) were used to determine significance.

Gene trees of aligned 5S rDNA sequences were generated using maximum parsimony and distance-based methods of phylogenetic reconstruction. Maximum parsimony analysis
was performed using PAUP v. 3.1 (Swofford 1990). Several search strategies were employed in an effort to find the most parsimonious trees. In all analyses, characters and character-state transformations were weighted equally. We explored several alternatives for coding of alignment gaps, including treating all gaps as binary, presence/absence characters, coding them as missing data, and excluding gapped positions from the data set prior to analysis. Separate analyses were performed for the complete data set (5S gene + spacer) and for subsets of the data (5S only, spacer only). In all cases, at least five independent heuristic searches were conducted per data set using the random data addition option, in an effort to find shorter “islands” of trees than those recovered from initial searches (Maddison, 1991). To evaluate relative levels of support for individual clades, strict consensus trees were generated for all topologies found that were up to five steps longer than the most parsimonious trees (“decay analysis”: Bremer, 1988; Donoghue et al., 1992).

For distance-based phylogeny estimation, complete and partial data sets were analyzed using MEGA (Kumar et al., 1993). We translated the observed distances between all pairs of sequences to evolutionarily "corrected" (for superimposed substitutions) Kimura two-parameter distances (Kimura, 1980), and subjected the resulting distance matrices to neighbor-joining analysis (Saitou and Nei, 1987).

RESULTS

5S rDNA Repeats from Gossypium - We sequenced 99 5S rDNAs from 20 taxa representing diploid and allopolyploid Gossypium (Fig. 1). In all species, digestion of genomic DNA with BamHI and subsequent probing with 5S rDNA reveals the characteristic ladder of tandemly repeated genes. 5S rDNA organization is conventional in that each repeat within an array consists of 5S RNA genes separated from one another by intergenic spacers (Sastri et al., 1992). Although the 5S gene is nearly invariant in length (121 - 122 bp; nucleotides 1-122 in Fig. 1), the non-transcribed intergenic spacer is considerably more variable, ranging from 175 to 191 bp (nucleotides 123-316 of the aligned data set).

Alignment of full-length 5S gene sequences within and between the 20 taxa was simple, requiring only a single nucleotide insertion at position 38 for one sequence from G. robinsonii. Visual inspection showed that five 5S gene sequences (A2D1syn3, tomentosum2, raimondii3, raimondii6 and hirsutum7) were considerably shorter than the expected length. We feel that these truncated sequences are deletion artifacts of cloning rather than actual sequences, since M13 can delete insert DNA (Sambrook et al., 1989). In contrast to the 5S gene, alignment of the spacer region was not trivial, due to higher sequence divergence.
between taxa from different genomic groups. Indels were introduced in a region of relatively low conservation in the 5' half of the spacer region (nucleotides 135 - 170 and 182 - 205, Fig. 1), two of which are genome-specific. Indel 1 is seven bp in length (nucleotides 183-189) and is interpreted as either a deletion in the 5S rDNA of the common ancestor of A- and D-genome diploid species or an insertion in the G. robinsonii lineage. Indel 2 is eight bp in length (nucleotides 195-202) and is unique to sequences from A-genome diploids, implying that a deletion occurred in 5S rDNA of the common ancestor of these species. These diagnostic indels, combined with phylogenetic analyses (detailed below), allowed us to infer the subgenomic origin of each allopolyploid sequence. Clones from the A-subgenome were recovered from all allopolyploid species studied, whereas clones originating from the D-subgenome were detected only in G. mustelinum, G. hirsutum and the synthetic polyploid 2(A2D1). As with the 5S gene sequences, apparent M13 deletion artifacts were encountered in two spacer sequences (raimondiiS and hirsutumS).

Two of our 99 clones, hirsutum6 and hirsutum10, are unusually divergent from each other and from the remaining 5S rDNA sequences. Coding regions of these sequences show lower similarity to each other and to other Gossypium 5S sequences than exhibited by all other pairwise comparisons (discussed in the following section). In addition, based upon the number of substitutions in the 5S gene, these sequences appear to have experienced accelerated rates of sequence evolution. These features suggest that hirsutum6 and hirsutum10 are pseudogenes, although we point out the absence of explicit criteria for determining (from DNA sequence data) whether rRNAs are functional in translational machinery. Since these two cloned sequences are clearly different from all other sequences, they are considered separately (under the name G. hirsutum-Dy) from D-subgenomic sequences of G. hirsutum in subsequent analyses.

Intra-Individual Sequence Variation - The sequence data of Figure 1 demonstrate that 5S rDNA sequences are highly polymorphic in Gossypium, not only between species but also within individual plants. Even though we sampled only a small fraction of the existing repeats from any single genome (2 to 10 clones per individual), all sequences were unique. Polymorphism appears to be partitioned approximately equally between the 5S gene and the intergenic spacer, although variation is not uniformly distributed across all nucleotides.

A mean of 12% of nucleotide positions are polymorphic within individual arrays of Gossypium species, both for genes and spacers, although there is considerable variance around this mean (Table 2).
Intra-individual values of $p_\pi$ range from 0.033 (for *G. herbaceum*) to 0.223 (*G. hirsutum* Dγ) for the 5S gene, and from 0.033 (*G. armourianum*) to 0.289 (*G. gossypioides*) for the intergenic spacer (Table 2). Overall, the least polymorphic 5S sequences are from *G. herbaceum* ($p_\pi = 0.044$) and *G. turneri* ($p_\pi = 0.046$). At the other extreme are the putative pseudogenes from *G. hirsutum* ($p_\pi = 0.223$ and 0.148 for the gene and spacer, respectively; $p_\pi = 0.178$ overall) and the sequences from *G. gossypioides* ($p_\pi = 0.132$ and 0.289 for the gene and spacer, respectively; $p_\pi = 0.226$ overall). Despite the wide range of $p_\pi$ values observed among species and the overall equivalence of polymorphism in the gene and spacer regions, there does not appear to be a correlation between $p_\pi$ values for the gene and values for the spacer ($r^2 = 0.17$).

Because $p_\pi$ tracks the proportion of polymorphic positions without regard to frequency, the measure may be influenced by sampling intensity. To provide estimates of polymorphism that are less biased with respect to sample size, we calculated nucleotide diversity ($\pi;$ Nei, 1987), which is numerically equivalent to the mean number of nucleotide differences per site between all pairs of sequences. Table 2 shows $\pi$ within species for 5S genes, spacers, and entire repeats. In general, $\pi$ values parallel the patterns of sequence variation revealed by $p_\pi$. Intra-individual values for $\pi$ range from a low of 0.018 in sequences from *G. herbaceum* to a high of 0.188 for repeats from the putative *G. hirsutum* pseudogenes. Excluding these pseudogene sequences, the highest nucleotide diversity is observed in *G. gossypioides* (0.120), as was the case for $p_\pi$. For those taxa where only two sequences were sampled (*G. aridum*, *G. davidsonii*, *G. armourianum*), $\pi$ is identical to $p_\pi$.

Overall mean nucleotide diversity for 5S genes is significantly higher than for spacer sequences in the A-genome ($\pi = 0.057$ vs. 0.043; probability from one-tailed t-test = 0.02), but estimates were nearly identical for the two regions from the D-genome repeats ($\pi = 0.064$ vs. 0.067 for genes and spacers, respectively; $p = 0.41$). For entire repeats, nucleotide diversity is higher in D-genome than A-genome diploids (0.058 vs. 0.039). Within each genome group, no particular pattern of nucleotide diversity is apparent.

We estimated $p_\pi$ and $\pi$ separately for sequences from each of the two subgenomes of the allopolyploid species and the synthetic allopolyploid 2(A2D1) (Table 2). On average, both $p_\pi$ and $\pi$ values for 5S genes in natural allopolyploids are slightly higher than values obtained for the spacer sequences of the same repeats, although not significantly so. Approximately the same proportion of nucleotides are polymorphic (13.5%) in repeats from both subgenomes, although nucleotide diversity in the D-like sequences is nearly double.
(0.090) that of the A-like sequences (0.053). Nucleotide diversities in the A- and D-subgenomic sequences from the synthetic allopolyploid 2(A₂D₁) were approximately equivalent (0.041 vs. 0.035 overall).

*Interspecific Sequence Variation* - Excluding the potentially biasing *G. hirsutum-Dω* sequences, only 17.7% of the 5S rDNA nucleotides are conserved across the 20 taxa; these include 28 of 122 nucleotides (23.0%) in the 5S gene and 28 of 194 (14.4%) aligned nucleotides in the spacer region (Fig. 1). Among the unanticipated observations of this high degree of sequence polymorphism are the cases where mean sequence polymorphism and nucleotide diversity for 5S genes and spacers are higher from *single individuals* than they are in comparisons between species. Nucleotide diversity for spacer sequences from *G. thurberi* (π = 0.091), for example, is higher than that estimated from comparisons of spacer sequences of *G. thurberi* with those from *G. klotzschianum, G. trilobum, G. turneri, G. armourianum,* and *G. harknessii* (Table 3). In general, this effect appears to be related to the degree of evolutionary divergence, viz., as time since organismal divergence increases, fixed interspecific differences in 5S rDNAs increasingly overwhelm inter-repeat polymorphisms within individual arrays. This effect is most pronounced for spacer regions, where mean inter-genomic distances are always larger than intra-individual distances (cf. Tables 2 and 3). In contrast, there are a number of cases where inter-genomic distances are equivalent to or lower than intra-individual distances for 5S genes (e.g., D-subgenomic sequences from *G. mustelinum* versus nearly all A-type sequences).

**5S Copy-Number in Gossypium** - Copy-number varies over twenty-fold among the species sampled, from approximately 1,150 copies per 2C genomic equivalent in the D-genome diploid *G. gossypioides* to approximately 23,500 copies in the allopolyploid *G. barbadense* (Table 4). In the A-genome, the two closely-related species *G. arboreum* and *G. herbaceum* differ two-fold in 5S rDNA copy-number, with an average of 5,500 copies. Copy-number variation is even greater in the D-genome species, with a mean of 4,500 copies but a range that varies from 1150 in *G. gossypioides* to 10,300 in *G. davidsonii*. As with the A-genome species, closely related D-genome species often vary several-fold in the number of 5S genes. For example, copy-number varies from approximately 1,700 to 5,800 within the distinctive and closely related (Fryxell, 1979; 1992; Wendel and Albert, 1992) Mexican arborescent species that comprise the taxonomic subsection *Erioxylum* (*G. aridum, G. laxum, G. lobatum, G. schwendimanii*). In allopolyploid species, there is a two-fold range in copy-number estimates, from 11,200 (in *G. hirsutum*) to 23,500 (in *G. barbadense*).
Three of the four allopolyploid species examined (all but G. hirsutum) have approximately equal numbers (22,000 to 23,500) of 5S genes.

We expected that 5S rDNA copy-number of the synthetic allopolyploid 2(A2D1) would be approximately additive with respect to its A-genome and D-genome diploid progenitors. We found that additivity does not hold for the mean copy number, but does when 95% confidence intervals are considered, i.e. summing copy-numbers of parents A2 (5,900 - 9,000) and D1 (1,850 - 2,300) gives an expected range of 6,750 - 11,300 copies, compared to an observed 95% CI of 10,500 - 17,500.

Phylogenetic Analysis - We conducted parsimony and distance analyses on 5S rDNA data sets that were treated several ways. In one series of analyses, all 99 sequences were included; this data set was analyzed using sequences from the 5S gene alone (121 aligned characters), the intergenic spacer alone (195 characters), and the entire 5S repeat (316 characters). We explored several alternative gap codings, and found that these treatments did not significantly alter the topologies obtained. In a second series of analyses, we performed parsimony and distance-based analyses on data sets of reduced dimensionality, generated by computing “consensus” sequences for each diploid species and for sequences recovered from each subgenome of each allopolyploid species. To accomplish this, positions that were polymorphic within a species (or within a subgenome) were coded using appropriate ambiguity coding. The resulting data set contained 25 rows, consisting of one outgroup (G. robinsonii), seven “A-genome” repeat types (two from the two A-genome diploids examined, four from the allopolyploids included in the study and one from the synthetic allopolyploids 2(A2D1)), and 17 “D-genome” repeat types (13 from D-genome diploids, three from tetraploid subgenomes and one from 2(A2D1)). This 25 by 316 matrix was similarly analyzed using the same gap coding alternatives as described above.

Parsimony analysis of the 5S gene sequences for the original 99 sequences resulted in more than 3000 equally most-parsimonious trees, at which time the program was stopped and a strict consensus was computed. Each constituent tree had a length of 299 and a relatively low retention index (RI = 0.54). Little resolution was retained in the strict consensus of the shortest trees; nearly all sequences form an unresolved “rake” at the base of the tree (Fig. 2b). The exceptions are two, poorly supported clades, one uniting sequences from the phylogenetically distant D-genome species G. turneri and G. schwendimanii, and the other joining sequences from the more closely related species G. aridum and G. lobatum. Upon decay analysis, however, both of these clades collapse in trees that are only a single step longer than the most parsimonious trees. It is noteworthy that these analyses failed to
recover “clades” of sequences from individual species, even within the outgroup species *G. robinsonii*, which is arbitrarily shown as monophyletic in Figure 2b.

In contrast to a near-absence of phylogenetically useful information in the 5S gene sequences, considerable resolution is evident in parsimony trees from the intergenic spacer (Fig. 2a). As shown in the strict consensus of the 3000 shortest trees saved (each of length 576 and with a RI of 0.87), each genomic group is monophyletic and is strongly supported by decay analysis (Fig. 2a). Within the D-genome clade (Fig. 2a), spacer sequences are resolved into seven well-supported sub-clades that are generally concordant with previous phylogenetic results (DeJoode, 1992; Wendel and Albert, 1992; Wendel et al., 1995b). In contrast, A-genome spacer sequences are not resolved into taxonomically meaningful clades, although sequences from individual species often form higher level organization (e.g., clades of *tomentosum1*, *tomentosum6* and *tomentosum10*, and of *herbaceum1*, *herbaceum2*, and *herbaceum5*).

Analysis of 3000 minimal length trees derived from entire 5S rDNA repeats (gene + spacer) resulted in a strict consensus topology (not shown) that is identical to that illustrated for the intergenic spacer alone (Fig. 2a). Due to the additional and “noisy” characters from the 5S gene sequences, these trees are longer (978 steps) and show slightly more homoplasy (RI = 0.80) than trees generated from intergenic spacer sequences alone.

Our final parsimony analyses utilized “consensus” sequences for each diploid species and sets of sequences from individual subgenomes of the allopolyploids. The topology of strict consensus trees generated from spacer sequences alone (3,000 trees saved, each of length 121 and a RI of 0.95) and from entire 5S repeats (5S gene + spacer; 3,000 trees saved, each of length 159 and a RI of 0.94) were identical to each other and had identical decay values for most clades. The consensus tree from the spacer sequence data set is shown in Figure 3, which shows that even when the data are reduced to consensus sequences, genomic identity is retained, with the formation of two strongly supported clades corresponding to the A and D genomes. In general, trees generated from consensus sequences exhibited slightly less resolution than cladograms derived from individual sequences.

Distance-based methods of phylogenetic analysis produced trees that were concordant with those derived from parsimony analysis, in that all clades observed in strict consensus trees generated by parsimony methods are supported by neighbor-joining trees (Fig. 4). In general, analysis of spacer sequences alone (as opposed to the entire 5S repeats) provides the greatest degree of resolution, in the sense that more sequences form species-specific clades. It also appears that the high level of homoplasious polymorphism contained within 5S gene
sequences acts to reduce inter-sequence distances, which shortens internode lengths, thereby reducing confidence in the resolution obtained. In the tree shown (Fig. 4), groups within the D-genome clade appear to be more clearly resolved (with longer internode lengths) than sequences within the A-genome, as was the case with parsimony analysis. Among the more intriguing aspects of the neighbor-joining results are relationships revealed between individuals of closely related species-pairs, such as G. thurberi/G. trilobum or G. davidsonii/G. klotzschianum. Both of these species-pairs form well-defined clusters that are distinct from other taxa, yet within each cluster species-specific groups are not formed. Included in the alternative explanations for this pattern are high mutability of particular nucleotide positions (homoplasies that mimic synapomorphies at this level), and retention of 5S polymorphisms that are older than the speciation event that separated the species.

**DISCUSSION**

*Gossypium 5S rDNA Structure and Organization* - The 5S rDNA arrays in all *Gossypium* species examined exhibit a conventional organization of tandemly repeated 5S genes and intergenic spacers. Fluorescent *in situ* hybridization work demonstrates that these arrays occupy a single centromeric location in A-genome and D-genome diploid species (R. Hanson and D. Stelly, pers. comm.) and two corresponding loci in the AD-genome allopolyploids (Crane et al., 1993). *Gossypium* 5S genes and spacers range from 121-122 and 175-191 nucleotides in length, respectively. Totaling 296 to 311 bp, these repeats are among the shortest known in plants, with only three reports of smaller repeat units (*Datisca glomerata* = 219 bp; *Gleditsia triacanthos* = 278 bp; *Gymnocladus dioicus* = 215 bp; Gottlob-McHugh et al., 1990). Most variation in the length of the *Gossypium* 5S rDNA repeats is attributable to the seven-bp (distinguishing A-genome and D-genome from C-genome taxa) and eight-bp (A-genome) genome-specific indels, although minor intra-individual length variation is also evident; the latter appears to result from contraction/expansion of the T-rich region downstream of the 5S gene coding region. Three highly conserved hallmarks characterize *Gossypium* 5S genes and spacers: (i) the BamHI site used for cloning (nucleotides 30-35, Fig. 1), which is a conserved feature of land plants (Sastri, 1992); (ii) a pentanucleotide “TATRA” motif 22-26 bp upstream of the 5S gene (nucleotides 291-295), which is commonly observed in plants (*Acacia*, Playford et al., 1992; *Arabidopsis*, Campell et al., 1992; *Bromus*, Sastri et al., 1992; *Glycine*, Gottlob-McHugh et al., 1988; *Lupinus*, Rafalski et al., 1982; *Sinapis*, Capesius, 1991; *Vigna*, Hemleben and Werts, 1988; *Zea*, Sastri et al., 1992; the Triticeae, Cox et al., 1992, Kellogg and Appels,
and has been implicated in transcription initiation of 5S (Tyler, 1987; Sharp and Garcia; 1988) and other class III genes (White, 1994): (iii) a “TTTTATAT” motif immediately downstream of the gene at nucleotides 126-133, which is thought to facilitate efficient transcription termination (Korn, 1982).

**5S Copy-number is Evolutionarily Labile** - The number of 5S rDNA repeats per genome varies over twenty-fold among species (Table 4), thereby demonstrating that arrays have expanded and contracted since the origin of the genus. Specific examples of array expansion or contraction are evident, despite the sizable errors associated with most estimates. These inferences are dependent on correct diagnoses of ancestral conditions, which are in turn dependent on the organismal phylogeny (DeJoode, 1992; Wendel and Albert, 1992; Wendel et al., 1994, 1995b). For example, if the common ancestor of the D-genome species had a similar number of 5S rDNA copies as both the A-genome diploids (5,500 copies/2C) and the C-genome outgroup species *G. robinsonii* (3,000 copies/2C genome), then: (1) the species-pair *G. trilobum* and *G. thurberi* experienced a lineage-specific decrease in copy-number to an average of 1,700 copies/2C genome; and (2) the species-pair of *G. davidsonii* and *G. klotzschianum* experienced a lineage-specific increase to approximately 8,600 copies per 2C genome equivalent. Similarly, in the monophyletic subsection *Erioxylum*, mean copy-number is 3,700 copies/2C genome, which is lower than for two members of the subsection (*G. schwendimanii* = 4,600; *G. lobatum* = 5,800), but higher than for the other two included species (*G. laxum* = 2,800; *G. aridum* = 1,700). While considerable variation exists in our copy number estimates, comparisons of 95% confidence intervals (Table 4) and the results of t-tests (not shown) show that many of these differences are statistically significant. This shows that both array expansion and contraction can occur within a relatively brief evolutionary time frame.

In contrast to the additivity observed for the synthetic allopolyploid, 5S copy-numbers from the putative progenitors of natural allopolyploids (*G. herbaceum* = 3,400, *G. raimondii* = 4,750; Σ = 8,150) add up to less than half of the average copy-number for the AD-genome species. These numbers range from 11,200/2C genome in *G. hirsutum* to over 22,000 copies in the other polyploids. Even when we consider experimental error and interspecific variation (Table 4), 5S rDNA copy number clearly is not additive in the allopolyploids. This suggests, but does not prove, that 5S rDNA arrays have expanded in the allopolyploids since their formation.
A final comment with respect to 5S copy-number is stimulated by the observation of relatively low copy-numbers in *G. gossypioides* and *G. aridum*. Of the diploid species included in this study, only these two are known to have evolutionary histories that include episodes of interspecific hybridization and introgression (DeJoode, 1992; Wendel and Albert, 1992; Wendel et al., 1995b). In the case of *G. aridum*, the cytoplasmic parent was similar to present-day *G. klotzschianum* (with 6,950 5S rDNA copies/2C genome) and the paternal parent was similar to members of the present-day subsection *Erioxylum* with a mean of 3,700 copies/2C genome. Although copy-number estimates for *G. aridum* have a large error (Table 4), the mean of 1,700 copies/2C genome is much lower than that of either parental lineage. Similarly, the values obtained for the inter-genomic derivative *G. gossypioides* (1,150 copies/2C genome) are substantially lower than for all other species examined. In this respect it is noteworthy that Zimmer et al. (1988, p. 1134) reported that in hybrids between *Zea mays* and *Z. diploperennis*, “teosinte-specific genes (rDNA and 5S DNA) are underrepresented in the F1 hybrids analyzed”. Given our small sample of two species, the association between reticulation and 5S rDNA copy-number reduction may be coincidental. On the other hand, selection may operate to reduce copy-number in hybrids as a means of rapidly eliminating excess sequence variation within an array, perhaps as a necessary component of optimizing 5S rDNA expression and ribosomal composition or function. If this hypothesis is true, our observation of 5S rDNA copy-number reduction may be a significant aspect of the stabilization of hybrid evolutionary products. Clearly, other natural and synthetic interspecific hybrids and later-generation segregates need to be screened for 5S (and perhaps other repetitive DNA) copy-number to evaluate whether contraction of repeated sequence arrays is a common consequence of interspecific hybridization.

5S Sequence Evolution and Phylogeny Reconstruction - Several authors have examined the value of 5S rRNA genes (Wheeler and Honeycutt, 1988; Steele et al., 1991; Halanych, 1991; Vawter and Brown, 1993) and spacer sequences (Scoles et al., 1988; Baimi and Appels, 1992; Kellogg and Appels, 1995) for phylogeny reconstruction. Although the 5S gene may provide useful information, it has been applied mostly to relatively ancient divergences, such as between major clades of prokaryotes (Woese, 1987) and between the major eukaryotic phyla (Wheeler and Honeycutt, 1988; Steele et al., 1991). In addition, its length (ca. 120 bp) places practical limitations on its ability to record evolutionary history. The spacer region, although also reasonably short (100-700 bp in plants; Sastri et al., 1992), has a much higher rate of sequence substitution, and hence it is more likely to provide phylogenetically useful information at lower taxonomic ranks. As with many spacer
sequences, however, alignment difficulties are likely to arise as more divergent taxa are included in an analysis, due to the characteristic occurrence of simple repeats (Kanazin et al., 1993) and indels (Cox et al., 1992; this paper).

We evaluated the phylogenetic utility of each region in the *Gossypium* 5S rDNA repeats by using both character-based and distance-based approaches to phylogeny estimation. Analysis of the gene sequences alone resulted in a large number of minimal length trees (> 3000) with low retention indices (RI = 0.54) and a strict consensus with virtually no resolution (Fig. 2b). A notable feature of this tree is that in addition to the absence of cladistic structure among species, there is a complete lack of resolution of sequences from individual species. This demonstrates that the 5S gene in *Gossypium*, despite exhibiting reasonably high intraspecific polymorphism (mean $p_n = 0.12$; $\pi = 0.06$; Table 2) and interspecific divergence (28 of 121 nucleotides conserved across taxa; also Table 3), is not phylogenetically useful within *Gossypium*. The lack of phylogenetic information is not due to an absence of variation, which might have been the *a priori* expectation given the presumed slow rate of sequence evolution in 5S genes. Instead, there is an abundance of sequence variation, but it is highly homoplasious.

A different pattern emerged when the intergenic spacer sequences were analyzed. Although numerous (> 3000) minimal length trees were still found, they each had a high retention index (0.87). In addition, considerable resolution is retained in the strict consensus tree, and clades consisting of sequences from closely related species are often recovered. Also, as shown in Fig. 2a, each diploid genomic group is monophyletic (excluding allopolyploids) and is strongly supported by decay analysis.

There is considerable congruence between the 5S intergenic spacer “gene tree” (Fig 2a) and previous phylogenetic results (Dejoode, 1992; Wendel and Albert, 1992; Wendel et al., 1995b). Within the D-genome clade, seven well-supported sub-clades were recovered. The two most basal of these are comprised solely of sequences from the Mexican species *G. gossypioides*, the evolutionary history of which has recently been reviewed (Wendel et al., 1995b). In brief, all sources of evidence, including comparative analysis of chloroplast DNA restriction site data and interspecific fertility relationships, indicate that the sister-species of *G. gossypioides* is *G. raimondii*. The sole previous exception to this unanimity consisted of DNA sequence data from the internal transcribed spacer (ITS) region of the 18S - 5.8S - 26S array, which allied *G. gossypioides* most closely to the A-genome cottons, albeit in a phylogenetically basal position. Wendel et al. (1995b) argued that the incongruence between all other sources of data and the ITS data reveals an ancient hybridization and introgression
event between the antecedent of modern *G. gossypioides* and an A-genome ITS. The present results for the 5S rDNA are similar, in that *G. gossypioides* again occupies a phylogenetically basal position, albeit in the D-genome clade rather than the A-genome clade. As hybrid taxa are expected to occupy phylogenetically basal positions in cladistic analyses (McDade, 1990, 1992), we interpret these results as additional support for the hybridization and introgression hypothesis advanced by Wendel et al. (1995b).

Resolution of the remaining D-genome sequences is into a polytomy that unites five major clades: (1) the arborescent species comprising the Mexican subsection *Erioxylum* (*G. laxum, G. schwendimanii, G. lobatum, G. aridum*); (2) the Baja/Galapagos Islands species-pair (Wendel and Percival, 1990) that comprises subsection *Integifolia* (*G. klotzschianum, G. davidsonii*); (3) the two members of subsection *Houzingenia* (*G. thurberi, G. trilobum*); (4) the members of subsection *Caducibracteolata* from Baja California (*G. tumeri, G. harknessii, G. armourianum*); and (5) the clade uniting the D-type repeats from the allopolyploids with all sequences from *G. raimondii*. This last clade provides additional evidence in support of the traditional hypothesis that *G. raimondii* is the best living model of the original D-genome donor to the allopolyploids (Endrizzi et al., 1985; Wendel, 1989; but see Wendel et al., 1995b).

**Duplicated 5S Arrays Evolve Independently in Allopolyploids** - Among the more important results is that different sequences from single allopolyploid species often occur in both the A-genome and D-genome clades (Figs 1-4). Using sample sizes of 9 to 10 clones per taxon, we were able to isolate two distinct classes of 5S rDNA sequences from the natural allopolyploids *G. hirsutum* and *G. mustelinum* and from the synthetic allopolyploid 2(A2D1). These two classes of sequences evidently originated from the two different allopolyploid subgenomes (A and D), as each class shares a high degree of sequence similarity to 5S repeats from the putative subgenome donors. Moreover, in each species both repeat types were detected in nearly equal proportions. This demonstrates that for these species, orthology-paralogy relationships have been retained for since allopolyploid formation (1-2 mybp; Wendel, 1989; Wendel and Albert, 1992).

The significance of this observation is that it constitutes compelling evidence that intralocus concerted evolution has predominated over interlocus interactions. This conclusion seems firm, although it is not without precedent; in fact, we are aware of no case where interlocus concerted evolution of 5S rDNA arrays has been demonstrated in plants (Cox et al., 1992; Sastri et al., 1992; Kellogg and Appels, 1995). Available information,
therefore, suggests that the predominant homogenizing forces acting on 5S ribosomal genes and spacers operate at the level of the individual array.

From previous analyses of the 45S arrays in the same allopolyploid *Gossypium* species (Wendel et al., 1995a), we know that interlocus concerted evolution has homogenized, to near-identity, sequences located on homoeologous chromosomes. The differences in the evolutionary behavior of 5S and 45S arrays indicate that relative to 45S arrays, interlocus interactions among 5S arrays are prohibited or are too infrequent to be detected (cf. Dover, 1994; Schlötterer and Tautz, 1994). While 5S rRNA and 45S rRNA genes exist as highly repetitive, tandemly arranged arrays, differences exist in both the number and organization of those arrays. Specifically, *Gossypium* allopolyploids have inherited one 5S locus but two major 45S loci from each parent (Crane et al., 1993). Moreover, 5S rDNA loci are located near the centromere whereas 45S rDNA loci occupy telomeric or subtelomeric locations. Under the assumptions that (1) unequal crossing over is the operative mechanism of interlocus homogenization in *Gossypium*, and (2) unequal crossing over in centromerically located arrays would lead to unbalanced and presumably inviable gametes, Wendel et al. (1995a) suggested that long-term maintenance of interlocus polymorphism following allopolyploidization is more likely for sequences that are centromeric rather than telomeric in distribution. The 5S results presented here appear to meet that prediction.

Although both A- and D-subgenomic homoeologues were recovered from *G. hirsutum*, *G. mustelinum* and the synthetic 2(A2D1), only the A-subgenome repeat was detected from *G. barbadense* (N = 7) and *G. tomentosum* (N = 6). There are at least three possible explanations for this observation: (1) that the 5S rDNA locus has been lost from the D-subgenome of these two species; (2) that D-type 5S arrays exist but were missed due to sampling or experimental bias; and (3) that interlocus concerted evolution has converted D-type 5S arrays to A-type only in *G. barbadense* and *G. tomentosum*.

The first alternative, reduction or loss of a 5S array, has been documented for hexaploid wheat (Dvorák, 1990), but is contraindicated for *Gossypium* by fluorescent *in situ* hybridization results that reveal two arrays - one on each of the homoeologous chromosomes - for *G. barbadense*, *G. hirsutum* and *G. mustelinum* (R. Hanson and D. Stelly, pers. comm.). By extrapolation, these results suggest that loss of the 5S locus is also unlikely in *G. tomentosum*. To discriminate between the alternatives (2) and (3) - sampling versus interlocus concerted evolution - we used the sequence data (Fig. 1) to develop a 24 bp PCR primer ("gapR" = 5'-TCA-AAT-TAT-TTA-TTT-CAC-AAA-ACG) that hybridizes
specifically to D-subgenomic sequences in the region of indel 2 (nucleotides 204 - 227). This primer, when paired with 5SF, was expected to amplify a 159 bp fragment from D-genome but not from A-genome repeats. Using M13 clones of known genomic origin as templates, this expectation was met, since only clones from D-genome diploids or the D-subgenome of allotetraploids showed the expected PCR fragment (data not shown). When genomic DNAs were used as templates, D-genome diploids and all allotetraploids showed the expected amplification product, whereas A-genome diploids yielded no PCR product. These results constitute strong evidence that D-subgenomic repeats (and by inference, D-subgenomic arrays) are present in the genomes of *G. barbadense* and *G. tomentosum*, and that these sequences were not detected in our M13 clones due to sampling or experimental bias. We conclude, therefore, that there is no evidence of interlocus concerted evolution of 5S arrays in *Gossypium* allotetraploids.

*Evolution of 5S rDNA: A Balance of Mutational, Homogenizing and Selective Forces*

- In *Gossypium*, mean intraindividual nucleotide diversity for 5S genes is nearly identical to the mean diversity found in spacer sequences (0.061 vs. 0.060; Table 2). This level of intraindividual polymorphism is approximately equal to that observed in other plant species. A survey of 28 diploid species from the wheat tribe *Triticeae* revealed nucleotide diversity values of 0.00 - 0.06 for the 5S gene and 0.00 - 0.11 for spacer sequences (Kellogg and Appels, 1995), while lower values were obtained for nine 5S rDNA sequences from *Glycine max* (p = 0.01 for both gene and spacer; Gottlob-McHugh, 1990). Despite these data demonstrating intraindividual 5S rDNA polymorphism, concerted evolutionary processes are clearly homogenizing 5S rDNA sequences. In *Gossypium*, this is evidenced by the presence of highly conserved repeat length in diploid and polyploid genomic groups (A = A' = 295-298 bp; D = D' = 301-304 bp; C = 310-314 bp) and the phylogenetic conclusion (Figs. 2 - 4) that different sequences from single diploid species are, for the most part, more similar to each other than they are to sequences from other species. Neither of these observations is consistent with a mode of evolution in which each individual repeat evolves independently. Similar results have been obtained from phylogenetic analysis of long and short 5S repeats from members of the *Triticeae* (Sastri et al., 1992; Kellogg and Appels, 1995). Taken as a whole, these results underscore the apparent contradiction that although 5S genes are subjected to concerted evolutionary forces, they display considerable intraindividual polymorphism. The heterogeneity observed, therefore, must reflect the net effect of opposing and complementary forces that operate on 5S arrays; these include mutation, homogenization and selection.
The rate at which repeated sequences interact is an important factor in determining the degree of polymorphism maintained in an array. Although rates of concerted evolution can be predicted by models (Nagylaki and Petes, 1982; Ohta, 1983, 1984, 1990; Ohta and Dover, 1983; Nagylaki, 1984a, 1984b, 1990; Basten and Ohta, 1992), the application of these models to empirical observations can be problematic. At the simplest level, a survey of the amount of polymorphism that is retained across speciation events allows inferences to be made regarding the frequency with which polymorphism is removed from arrays. If concerted evolutionary processes homogenize 5S repeats at rates greater than the rate of speciation, novel mutations are expected to become fixed or removed and sequence polymorphism is expected to be low within species, with the absolute level determined by the severity of the homogenizing forces. Alternatively, if concerted evolutionary events homogenize 5S rDNA at rates equivalent to or slower than the rate of speciation, one expects greater levels of polymorphism within arrays. In addition, since polymorphism can survive through one or more speciation events, a corollary expectation is that closely related species will share 5S rDNA polymorphisms. In our data (Fig. 1), polymorphisms that are shared between closely related species are evident in both the 5S gene and spacer (e.g., nucleotide positions 13, 182, 253). The most parsimonious interpretation of these shared polymorphisms is that they reflect a single mutation in the common ancestor that survived through a speciation event and has escaped homogenization in both daughter species. That these shared polymorphisms are restricted to closely related species leads to the qualitative generalization that fixation rates are approximately equal to rates of speciation in *Gossypium*.

Although this interpretation accounts for the majority of the shared polymorphism in the sequence data, there are notable exceptions at positions 49 and 68 of the 5S gene (Fig. 1). At these two positions, polymorphisms occur throughout the genus, reflecting, in both cases, transitional substitutions (purines at position 49, pyrimidines at 68). Position 49 of the 5S gene, for example, is polymorphic in nearly every species examined, with guanine and adenine represented in approximately equal ratios in all species where polymorphism was detected. Considering the amount of time since divergence of C-genome cottons (20 - 30 million years; Wendel and Albert, 1992) from the remainder of the genus, as well as since separation of A- and D-genome lineages from each other (5 - 10 million years), it is unlikely that this polymorphism reflects the retention of a single ancestral polymorphism that has yet to become fixed or lost. A more likely scenario is that these sites are evolutionary labile, allowing transitional mutations to occur repeatedly during radiation of the genus. An
explanation for the absence of transversions may be that these nucleotides occur within the
5S gene; transversions at these two positions, once propagated and homogenized across the
5S array, may alter ribosome function and reduce relative fitness.

This example may be revealing with respect to the forces that govern 5S sequence
evolution. Trees based on 5S genes yield unresolved “rakes,” whereas those based on spacer
sequences contain considerable resolution (Fig. 2). The absence of resolution in trees based
on the coding sequences clearly does not reflect a paucity of variation, as diversity and
polymorphism are approximately equivalent for genes and spacers. It also seems unlikely
that concerted evolutionary processes discriminate between 5S gene and spacer sequences,
although no empirical evidence directly eliminates this as a formal possibility. Instead, the
impressive difference in resolution between the trees based on 5S genes and those based on
spacer sequences reflects a fundamental contrast in the nature of 5S gene and spacer sequence
evolution. Specifically, although nucleotide substitutions appear to accumulate in roughly
equal proportions in genes and spacers, fixation of these differences among repeats in an
array is limited to spacer regions.

This conclusion, which emerged from inspection of the sequence data (Fig 1) and
from the list of character-state changes in parsimony-based trees (Fig 2), was evaluated
statistically (after Kellogg and Appels, 1995). To do this, we made pairwise comparisons
among a subset of taxa for which five or more sequences were generated; in each comparison,
we compiled 2 x 2 contingency tables of fixed versus polymorphic differences (columns) in
the 5S gene and spacer (rows), and tested for independence of row and column categories by
using the two-tailed Fisher exact test (Sokal and Rohlf, 1981). The results show that for all
comparisons of taxa from separate, well-supported clades, probabilities of independence are
less than 0.05 (Table 5). In all cases, it is a deficiency of fixed differences in the 5S gene and
a surplus of fixed differences in the spacer that are responsible for the statistical significance.
This is most evident in comparisons involving species from the A- vs. D-genomes: despite an
accumulation of from 32 to 42 fixed differences in the spacer region, not a single fixed
difference has evolved in the 5S gene. A consequence of this phenomenon is that there is no
phylogenetic content in the 5S genic data: the ample polymorphism that exists resolves only
as autapomorphy and homoplasy (Fig. 2).

To extend the analysis beyond these select taxa we computed consensus sequences for
all species and for the genus as a whole. This confirmed that not a single mutation has
become fixed in the 5S gene during the 20 - 30 million year history of the genus (Wendel and
Albert, 1992). Moreover, the consensus for Gossypium is identical to that computed for 152
sequences from the Triticeae (Kellogg and Appels, 1995), implying that selection has preserved the 5S gene consensus sequence since the most recent common ancestor of the Malvaceae and Poaceae, or at least 120 million years.

Figure 5 presents a descriptive model of 5S rDNA evolution that incorporates the differential ability of gene and spacer sequences to fix novel variants (see also Schlötterer and Tautz, 1994; Kellogg and Appels, 1995). In this model concerted evolutionary forces eliminate variants or spread them throughout individual arrays, but not between different arrays, in keeping with our observations on allopolyploid species. Most nucleotide positions in the spacer region are presumed to be free to vary because variants are selectively neutral or near-neutral. Consequently, variant nucleotides can become either fixed or lost, thereby causing fixed interspecific differences to accumulate. In contrast, most mutations in the 5S gene are presumed to be selectively neutral or near-neutral only when they occur in a subcritical proportion of repeats in an array. Because 5S genes are present in high copy number, departures from the consensus sequence (perhaps even leading to non-functional 5S rRNAs) in a small proportion of genes are expected to have little overall effect on the fitness of an organism, due to the buffering effect of functional 5S genes. Polymorphisms are therefore expected at "moderate" frequency. As variants move toward fixation by concerted evolution and stochastic factors, however, the number of functional 5S genes is reduced. Once this threshold frequency is reached (Fig. 5B), variant nucleotides become disadvantageous due to their effects on 5S transcription or 5S RNA function, the relative fitness of the entire array is thereby reduced (Williams, 1990). As a consequence, fixed interspecific differences in 5S genes fail to accumulate, despite the fact that polymorphic nucleotide positions are as common as those in the spacer. In this respect, our results are consistent with those previously observed in 5S rDNA from diploid Triticeae (Kellogg and Appels, 1995) and ITS sequences from Drosophila (Schlötterer and Tautz, 1994), suggesting that this generalized model may have broad applicability.

The degree of 5S rDNA polymorphism observed in Gossypium and in other groups (Kellogg and Appels, 1995) raises important questions concerning the biological consequences of heterogeneity in 5S genes. Because of low sequence similarity between plant 5S rDNA and homologues from model organisms such as Xenopus, Drosophila and Neurospora, it is difficult to evaluate the effect of substitutions within previously-defined transcription signals and control regions on rRNA transcription and/or function. At present, only two criteria, gross mutation/rearrangements in putative regulatory regions and an unexpectedly high accumulation of substitutions (as in putative pseudogenes such as
hirsutum6 and hirsutum10), may be used to assess the likelihood of 5S genes as being either functional or non-functional. If the 5S rDNA sequences reported here are representative of the genes that are transcribed, then the 5S rRNA pool within each species is heterogeneous. At present, no data address whether 5S rRNAs are as polymorphic as the genes that encode them. Further, it has yet to be determined whether the 5S rRNA pool is a random sample of the 5S rDNA genes, or whether there are transcriptional consequences of genomic and spacer polymorphisms. Finally, if only a subset of 5S repeats lead to functional 5S RNAs, what mechanisms promote selective transcription and/or filter out less-than-optimal rRNAs? Answers to these functional and mechanistic questions are essential to achieving a more complete understanding of 5S rDNA evolution.

ACKNOWLEDGMENTS

We thank C. Brubaker, E. Kellogg and A. Schnabel for comments on the manuscript, E. Kellogg for sharing unpublished data, Xiaoling Ding for assistance in automated sequencing, the USDA Plant Genome Research Program and Texas Higher Education Coordinating Board (to AHP), and the National Science Foundation for financial support (to JFW).

LITERATURE CITED


Fryxell PA (1979) *The natural history of the cotton tribe*. Texas A&M Univ Press, College Station, TX


Kumar S, Koichir T, Nei M (1993) MEGA, molecular evolutionary genetics analysis, v 1.0. *Penn State Univ, University Park, PA*


Ohta T, Dover GA (1983) Population genetics of multigene families that are dispersed into two or more chromosomes. Proc Natl Acad Sci USA 80:4079-4083


White RJ (1994) RNA polymerase III transcription. R. G. Landes, Austin, TX


Table 1. Description of *Gossypium* accessions used. Taxa are arranged according to the classification of Fryxell (1992). Cytogenetic (“genome”) designations follow the conventions of Endrizzi et al. (1985). Accession names are those used in the National Collection of *Gossypium* Germplasm (Percival, 1987) or by our own laboratory. Geographic origin refers to site of accession collection rather than the aggregate range of the species. Between 2 and 9 clones were isolated per species.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Genome</th>
<th>Accession</th>
<th>Geographic Origin</th>
<th>Clones Isolated</th>
<th>GenBank Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgenus Sturtia (R. Brown) Todaro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section Sturtia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. robinsonii F. von Mueller</td>
<td>C_1</td>
<td>AZ-50</td>
<td>Western Australia</td>
<td>3</td>
<td>U31852 - U31854</td>
</tr>
<tr>
<td>Subgenus Gossypium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section Gossypium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. herbaceum L.</td>
<td>A_1</td>
<td>A1-73</td>
<td>Botswana</td>
<td>5</td>
<td>U32006 - U32010</td>
</tr>
<tr>
<td>G. arboreum L.</td>
<td>A_1</td>
<td>A2-74</td>
<td>China</td>
<td>5</td>
<td>U31855 - U31856, U31999 - U32001</td>
</tr>
<tr>
<td>Subgenus Houzingenia Fryxell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section Houzingenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. trifolium (Mocino &amp; Sesse ex DC) Skovsted</td>
<td>D_k</td>
<td>D8-1</td>
<td>western Mex.</td>
<td>3</td>
<td>U32056 - U32058</td>
</tr>
<tr>
<td>G. thurberi Todaro</td>
<td>D_1</td>
<td>T17</td>
<td>Arizona, USA</td>
<td>3</td>
<td>U32059 - U32061</td>
</tr>
<tr>
<td>Subsection Integrifolia (Todaro) Todaro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. davidsonii Kellogg</td>
<td>D_k,D_1</td>
<td>D3d-32a</td>
<td>Baja California, Mex.</td>
<td>2</td>
<td>U32054 - U32055</td>
</tr>
<tr>
<td>G. klotochianum Andersen</td>
<td>D_k,D_1</td>
<td>D3k-3</td>
<td>Galapagos Islands</td>
<td>4</td>
<td>U32050 - U32053</td>
</tr>
<tr>
<td>Subsection Caducibracteolata Mauer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. armourianum Kearney</td>
<td>D_1</td>
<td>D2-1</td>
<td>Baja California, Mex.</td>
<td>2</td>
<td>U32072 - U32073</td>
</tr>
<tr>
<td>G. harknessii Brandegee</td>
<td>D_1</td>
<td>D2-2</td>
<td>Baja California, Mex.</td>
<td>4</td>
<td>U32068 - U32071</td>
</tr>
<tr>
<td>G. turneri Fryxell</td>
<td>D_1</td>
<td>D10-3</td>
<td>Sonora, Mex.</td>
<td>3</td>
<td>U32066 - U32067, U39496</td>
</tr>
<tr>
<td>Section Erioxylum (Rose &amp; Standley) Prokhanov</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subsection Erioxylum (Rose &amp; Standley) Fryxell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. tridentum (Rose &amp; Standley ex Rose) Skovsted</td>
<td>D_1</td>
<td>D4-12</td>
<td>Colima, Mex.</td>
<td>2</td>
<td>U32040 - U32041</td>
</tr>
<tr>
<td>G. lobatum H. Gentry</td>
<td>D_1</td>
<td>D7</td>
<td>Michoacan, Mex.</td>
<td>5</td>
<td>U32042 - U32046</td>
</tr>
<tr>
<td>G. laxum Phillips</td>
<td>D_1</td>
<td>LP</td>
<td>Guerrero, Mex.</td>
<td>3</td>
<td>U32047 - U32049</td>
</tr>
<tr>
<td>G. schwendimanii Fryxell &amp; Koch</td>
<td>-</td>
<td>JMS</td>
<td>Michoacan, Mex.</td>
<td>4</td>
<td>U32036 - U32039</td>
</tr>
<tr>
<td>Subsection Selera (Ulbrich) Standley</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. gossypioides (Ulbrich) Standley</td>
<td>D_1</td>
<td>D6-5</td>
<td>Oaxaca, Mex.</td>
<td>4</td>
<td>U32032 - U32035</td>
</tr>
<tr>
<td>Subsection Austroamericana Fryxell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. raimondii Ulbrich</td>
<td>D_1</td>
<td>D5-37</td>
<td>Peru</td>
<td>6</td>
<td>U32074 - U32077, U39497 - U39498</td>
</tr>
<tr>
<td>Subgenus Karpas Rafinesque</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. hirsutum L.</td>
<td>(AD)_1</td>
<td>Tx2094</td>
<td>Yucatán, Mex.</td>
<td>9</td>
<td>U32027 - U32031, U32083 - U32085, U39499</td>
</tr>
<tr>
<td>G. barbadense L.</td>
<td>(AD)_1</td>
<td>K101</td>
<td>Bolivia</td>
<td>7</td>
<td>U32011 - U32017</td>
</tr>
<tr>
<td>G. tomentosum Nuttall ex Scannan</td>
<td>(AD)_1</td>
<td>WT936</td>
<td>Hawaii, USA</td>
<td>6</td>
<td>U32018 - U32021, U39494 - U39495</td>
</tr>
<tr>
<td>G. mustelianum Miers ex Watt</td>
<td>(AD)_1</td>
<td>JL</td>
<td>Ceará, Brazil</td>
<td>10</td>
<td>U32022 - U32026, U32078 - U32082</td>
</tr>
<tr>
<td>2X(G. arboreum x G. thurberi)</td>
<td>2(A_1,D_1)</td>
<td>Besamley</td>
<td>Synthetic allopolyploid</td>
<td>9</td>
<td>U32002 - U32005, U32062 - U32065, U39491</td>
</tr>
</tbody>
</table>
Table 2. Polymorphism ($p_n$) and diversity ($\pi$) values for *Gossypium* 5S rDNA sequences. Nucleotide variability measures are partitioned into values for 5S genes, spacers, and entire 5S rDNA repeats. Polymorphism values are expressed as the proportion of nucleotide positions that are variable. Nucleotide diversity values represent the mean proportion of nucleotide differences among all sequences in a single array from an individual. $N =$ the number of clones sequenced per taxon.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>$N$</th>
<th>5S Gene</th>
<th></th>
<th>Spacer</th>
<th></th>
<th>Entire repeat</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$p_n$</td>
<td>$\pi$</td>
<td>$p_n$</td>
<td>$\pi$</td>
<td>$p_n$</td>
<td>$\pi$</td>
</tr>
<tr>
<td><strong>C-genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. robinsonii</em></td>
<td>4</td>
<td>0.123</td>
<td>0.078</td>
<td>0.073</td>
<td>0.052</td>
<td>0.093</td>
<td>0.062</td>
</tr>
<tr>
<td><strong>A-genome/subgenome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. arboreum</em></td>
<td>5</td>
<td>0.141</td>
<td>0.060</td>
<td>0.096</td>
<td>0.057</td>
<td>0.107</td>
<td>0.059</td>
</tr>
<tr>
<td><em>G. herbaceum</em></td>
<td>5</td>
<td>0.033</td>
<td>0.013</td>
<td>0.051</td>
<td>0.020</td>
<td>0.044</td>
<td>0.018</td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>5</td>
<td>0.157</td>
<td>0.069</td>
<td>0.147</td>
<td>0.059</td>
<td>0.151</td>
<td>0.063</td>
</tr>
<tr>
<td><em>G. tomentosum</em></td>
<td>6</td>
<td>0.157</td>
<td>0.069</td>
<td>0.125</td>
<td>0.044</td>
<td>0.138</td>
<td>0.054</td>
</tr>
<tr>
<td><em>G. barbadense</em></td>
<td>7</td>
<td>0.149</td>
<td>0.053</td>
<td>0.119</td>
<td>0.037</td>
<td>0.131</td>
<td>0.044</td>
</tr>
<tr>
<td><em>G. mustelinum</em></td>
<td>5</td>
<td>0.157</td>
<td>0.081</td>
<td>0.096</td>
<td>0.048</td>
<td>0.121</td>
<td>0.061</td>
</tr>
<tr>
<td><em>2(2A2D)</em></td>
<td>5</td>
<td>0.124</td>
<td>0.056</td>
<td>0.096</td>
<td>0.037</td>
<td>0.104</td>
<td>0.041</td>
</tr>
<tr>
<td>Overall mean</td>
<td></td>
<td>0.131</td>
<td>0.057</td>
<td>0.109</td>
<td>0.043</td>
<td>0.118</td>
<td>0.049</td>
</tr>
<tr>
<td>A-genome diploids</td>
<td></td>
<td>0.087</td>
<td>0.037</td>
<td>0.090</td>
<td>0.038</td>
<td>0.089</td>
<td>0.039</td>
</tr>
<tr>
<td>Allopolyploid A-subgenomes</td>
<td></td>
<td>0.155</td>
<td>0.066</td>
<td>0.122</td>
<td>0.045</td>
<td>0.135</td>
<td>0.053</td>
</tr>
<tr>
<td><strong>D-genome/subgenome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>6</td>
<td>0.074</td>
<td>0.027</td>
<td>0.214</td>
<td>0.077</td>
<td>0.158</td>
<td>0.059</td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>2</td>
<td>0.076</td>
<td>0.066</td>
<td>0.069</td>
<td>0.065</td>
<td>0.072</td>
<td>0.065</td>
</tr>
<tr>
<td><em>G. hirsutum (Dy)</em></td>
<td>2</td>
<td>0.223</td>
<td>0.248</td>
<td>0.148</td>
<td>0.148</td>
<td>0.178</td>
<td>0.188</td>
</tr>
<tr>
<td><em>G. mustelinum</em></td>
<td>5</td>
<td>0.157</td>
<td>0.072</td>
<td>0.159</td>
<td>0.069</td>
<td>0.158</td>
<td>0.070</td>
</tr>
<tr>
<td><em>2(2A2D)</em></td>
<td>4</td>
<td>0.074</td>
<td>0.040</td>
<td>0.060</td>
<td>0.031</td>
<td>0.066</td>
<td>0.035</td>
</tr>
<tr>
<td><em>G. gossypioides</em></td>
<td>4</td>
<td>0.132</td>
<td>0.071</td>
<td>0.289</td>
<td>0.154</td>
<td>0.226</td>
<td>0.120</td>
</tr>
<tr>
<td><em>G. laxum</em></td>
<td>3</td>
<td>0.025</td>
<td>0.022</td>
<td>0.093</td>
<td>0.055</td>
<td>0.066</td>
<td>0.042</td>
</tr>
<tr>
<td><em>G. lobatum</em></td>
<td>5</td>
<td>0.116</td>
<td>0.064</td>
<td>0.071</td>
<td>0.029</td>
<td>0.089</td>
<td>0.038</td>
</tr>
<tr>
<td><em>G. aridum</em></td>
<td>2</td>
<td>0.066</td>
<td>0.066</td>
<td>0.094</td>
<td>0.094</td>
<td>0.083</td>
<td>0.083</td>
</tr>
<tr>
<td><em>G. schwendimanii</em></td>
<td>4</td>
<td>0.083</td>
<td>0.041</td>
<td>0.115</td>
<td>0.066</td>
<td>0.102</td>
<td>0.056</td>
</tr>
<tr>
<td><em>G. klotzschianum</em></td>
<td>4</td>
<td>0.066</td>
<td>0.039</td>
<td>0.072</td>
<td>0.038</td>
<td>0.070</td>
<td>0.038</td>
</tr>
<tr>
<td><em>G. davidsonii</em></td>
<td>2</td>
<td>0.058</td>
<td>0.058</td>
<td>0.067</td>
<td>0.067</td>
<td>0.063</td>
<td>0.063</td>
</tr>
<tr>
<td><em>G. thurberi</em></td>
<td>3</td>
<td>0.099</td>
<td>0.064</td>
<td>0.132</td>
<td>0.091</td>
<td>0.119</td>
<td>0.080</td>
</tr>
<tr>
<td><em>G. trilobum</em></td>
<td>3</td>
<td>0.050</td>
<td>0.039</td>
<td>0.077</td>
<td>0.053</td>
<td>0.066</td>
<td>0.047</td>
</tr>
<tr>
<td><em>G. turneri</em></td>
<td>3</td>
<td>0.050</td>
<td>0.033</td>
<td>0.044</td>
<td>0.032</td>
<td>0.046</td>
<td>0.032</td>
</tr>
<tr>
<td><em>G. harknessii</em></td>
<td>4</td>
<td>0.107</td>
<td>0.059</td>
<td>0.071</td>
<td>0.039</td>
<td>0.086</td>
<td>0.049</td>
</tr>
<tr>
<td><em>G. armourianum</em></td>
<td>2</td>
<td>0.083</td>
<td>0.083</td>
<td>0.033</td>
<td>0.033</td>
<td>0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>Overall mean</td>
<td></td>
<td>0.103</td>
<td>0.064</td>
<td>0.127</td>
<td>0.067</td>
<td>0.117</td>
<td>0.066</td>
</tr>
<tr>
<td>D-genome diploids</td>
<td></td>
<td>0.078</td>
<td>0.051</td>
<td>0.094</td>
<td>0.063</td>
<td>0.088</td>
<td>0.058</td>
</tr>
<tr>
<td>Allopolyploid D-subgenomes</td>
<td></td>
<td>0.152</td>
<td>0.101</td>
<td>0.126</td>
<td>0.078</td>
<td>0.136</td>
<td>0.090</td>
</tr>
</tbody>
</table>
Table 3. Mean differences among 5S gene (above diagonal) and spacer (below diagonal) sequences for diploid and allotetraploid *Gossypium* species. Each entry represents the mean proportion of nucleotide differences between all sequences involved in the comparison. Two lines and columns are listed for allopolyploid species for which sequences were recovered from both the A- and D-subgenomes.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <em>G. robinsonii</em></td>
<td>-</td>
<td>0.077</td>
<td>0.052</td>
<td>0.072</td>
<td>0.076</td>
<td>0.077</td>
<td>0.086</td>
<td>0.071</td>
<td>0.081</td>
<td>0.058</td>
<td>0.071</td>
<td>0.067</td>
</tr>
<tr>
<td>2. <em>G. arboreum</em></td>
<td>0.277</td>
<td>-</td>
<td>0.038</td>
<td>0.063</td>
<td>0.070</td>
<td>0.064</td>
<td>0.076</td>
<td>0.053</td>
<td>0.067</td>
<td>0.045</td>
<td>0.059</td>
<td>0.054</td>
</tr>
<tr>
<td>3. <em>G. herbaceum</em></td>
<td>0.278</td>
<td>0.061</td>
<td>-</td>
<td>0.035</td>
<td>0.045</td>
<td>0.044</td>
<td>0.054</td>
<td>0.035</td>
<td>0.042</td>
<td>0.019</td>
<td>0.035</td>
<td>0.027</td>
</tr>
<tr>
<td>4. <em>G. barbadense</em></td>
<td>0.267</td>
<td>0.057</td>
<td>0.038</td>
<td>-</td>
<td>0.061</td>
<td>0.061</td>
<td>0.071</td>
<td>0.053</td>
<td>0.064</td>
<td>0.042</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>5. <em>G. tomentosum</em></td>
<td>0.260</td>
<td>0.062</td>
<td>0.044</td>
<td>0.043</td>
<td>-</td>
<td>0.067</td>
<td>0.081</td>
<td>0.065</td>
<td>0.074</td>
<td>0.051</td>
<td>0.061</td>
<td>0.060</td>
</tr>
<tr>
<td>6. <em>G. hirsutum (A)</em></td>
<td>0.273</td>
<td>0.071</td>
<td>0.049</td>
<td>0.054</td>
<td>0.057</td>
<td>-</td>
<td>0.078</td>
<td>0.065</td>
<td>0.068</td>
<td>0.048</td>
<td>0.055</td>
<td>0.056</td>
</tr>
<tr>
<td>7. <em>G. mustelinum (A)</em></td>
<td>0.259</td>
<td>0.067</td>
<td>0.043</td>
<td>0.048</td>
<td>0.053</td>
<td>0.060</td>
<td>-</td>
<td>0.073</td>
<td>0.085</td>
<td>0.060</td>
<td>0.067</td>
<td>0.070</td>
</tr>
<tr>
<td>8. 2*(A_D _D)* (A)</td>
<td>0.269</td>
<td>0.049</td>
<td>0.048</td>
<td>0.045</td>
<td>0.052</td>
<td>0.062</td>
<td>0.052</td>
<td>-</td>
<td>0.068</td>
<td>0.038</td>
<td>0.047</td>
<td>0.053</td>
</tr>
<tr>
<td>9. <em>G. gossypioides</em></td>
<td>0.281</td>
<td>0.281</td>
<td>0.272</td>
<td>0.276</td>
<td>0.274</td>
<td>0.281</td>
<td>0.271</td>
<td>0.271</td>
<td>-</td>
<td>0.051</td>
<td>0.062</td>
<td>0.057</td>
</tr>
<tr>
<td>10. <em>G. raimondii</em></td>
<td>0.277</td>
<td>0.298</td>
<td>0.282</td>
<td>0.290</td>
<td>0.282</td>
<td>0.297</td>
<td>0.283</td>
<td>0.285</td>
<td>0.191</td>
<td>-</td>
<td>0.045</td>
<td>0.038</td>
</tr>
<tr>
<td>11. <em>G. hirsutum (D)</em></td>
<td>0.271</td>
<td>0.287</td>
<td>0.268</td>
<td>0.281</td>
<td>0.271</td>
<td>0.284</td>
<td>0.269</td>
<td>0.273</td>
<td>0.191</td>
<td>0.087</td>
<td>-</td>
<td>0.063</td>
</tr>
<tr>
<td>12. <em>G. mustelinum (D)</em></td>
<td>0.264</td>
<td>0.285</td>
<td>0.270</td>
<td>0.279</td>
<td>0.269</td>
<td>0.282</td>
<td>0.266</td>
<td>0.270</td>
<td>0.197</td>
<td>0.091</td>
<td>0.070</td>
<td>-</td>
</tr>
<tr>
<td>13. <em>G. schwendimanii</em></td>
<td>0.250</td>
<td>0.271</td>
<td>0.260</td>
<td>0.266</td>
<td>0.262</td>
<td>0.273</td>
<td>0.261</td>
<td>0.261</td>
<td>0.192</td>
<td>0.130</td>
<td>0.127</td>
<td>0.130</td>
</tr>
<tr>
<td>14. <em>G. aridum</em></td>
<td>0.257</td>
<td>0.280</td>
<td>0.269</td>
<td>0.275</td>
<td>0.269</td>
<td>0.275</td>
<td>0.270</td>
<td>0.270</td>
<td>0.199</td>
<td>0.144</td>
<td>0.138</td>
<td>0.144</td>
</tr>
<tr>
<td>15. <em>G. lobatum</em></td>
<td>0.252</td>
<td>0.276</td>
<td>0.264</td>
<td>0.271</td>
<td>0.266</td>
<td>0.273</td>
<td>0.267</td>
<td>0.266</td>
<td>0.187</td>
<td>0.133</td>
<td>0.130</td>
<td>0.135</td>
</tr>
<tr>
<td>16. <em>G. laxum</em></td>
<td>0.266</td>
<td>0.291</td>
<td>0.279</td>
<td>0.285</td>
<td>0.281</td>
<td>0.282</td>
<td>0.279</td>
<td>0.279</td>
<td>0.206</td>
<td>0.150</td>
<td>0.148</td>
<td>0.150</td>
</tr>
<tr>
<td>17. <em>G. klotzschianum</em></td>
<td>0.245</td>
<td>0.254</td>
<td>0.242</td>
<td>0.249</td>
<td>0.243</td>
<td>0.252</td>
<td>0.243</td>
<td>0.241</td>
<td>0.167</td>
<td>0.090</td>
<td>0.089</td>
<td>0.093</td>
</tr>
<tr>
<td>18. <em>G. davidsonii</em></td>
<td>0.264</td>
<td>0.273</td>
<td>0.262</td>
<td>0.268</td>
<td>0.264</td>
<td>0.271</td>
<td>0.263</td>
<td>0.261</td>
<td>0.186</td>
<td>0.108</td>
<td>0.100</td>
<td>0.109</td>
</tr>
<tr>
<td>19. <em>G. trilobum</em></td>
<td>0.263</td>
<td>0.268</td>
<td>0.254</td>
<td>0.262</td>
<td>0.255</td>
<td>0.268</td>
<td>0.257</td>
<td>0.255</td>
<td>0.180</td>
<td>0.100</td>
<td>0.097</td>
<td>0.103</td>
</tr>
<tr>
<td>20. <em>G. thurberi</em></td>
<td>0.272</td>
<td>0.269</td>
<td>0.258</td>
<td>0.259</td>
<td>0.255</td>
<td>0.266</td>
<td>0.256</td>
<td>0.256</td>
<td>0.188</td>
<td>0.110</td>
<td>0.107</td>
<td>0.113</td>
</tr>
<tr>
<td>21. 2*(A_D _D)* (D)</td>
<td>0.256</td>
<td>0.252</td>
<td>0.239</td>
<td>0.247</td>
<td>0.242</td>
<td>0.252</td>
<td>0.241</td>
<td>0.239</td>
<td>0.170</td>
<td>0.090</td>
<td>0.084</td>
<td>0.090</td>
</tr>
<tr>
<td>22. <em>G. turneri</em></td>
<td>0.259</td>
<td>0.262</td>
<td>0.247</td>
<td>0.254</td>
<td>0.248</td>
<td>0.260</td>
<td>0.247</td>
<td>0.248</td>
<td>0.172</td>
<td>0.080</td>
<td>0.079</td>
<td>0.083</td>
</tr>
<tr>
<td>23. <em>G. harknessii</em></td>
<td>0.257</td>
<td>0.260</td>
<td>0.249</td>
<td>0.256</td>
<td>0.251</td>
<td>0.259</td>
<td>0.249</td>
<td>0.248</td>
<td>0.173</td>
<td>0.084</td>
<td>0.079</td>
<td>0.093</td>
</tr>
<tr>
<td>24. <em>G. armourianum</em></td>
<td>0.256</td>
<td>0.263</td>
<td>0.250</td>
<td>0.258</td>
<td>0.253</td>
<td>0.261</td>
<td>0.276</td>
<td>0.250</td>
<td>0.167</td>
<td>0.074</td>
<td>0.073</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1.</td>
<td>$G. \text{robinsonii}$</td>
<td>0.084</td>
<td>0.095</td>
<td>0.071</td>
<td>0.056</td>
<td>0.065</td>
<td>0.075</td>
<td>0.064</td>
<td>0.080</td>
<td>0.066</td>
<td>0.058</td>
<td>0.074</td>
</tr>
<tr>
<td>2.</td>
<td>$G. \text{arboreum}$</td>
<td>0.072</td>
<td>0.034</td>
<td>0.056</td>
<td>0.040</td>
<td>0.049</td>
<td>0.060</td>
<td>0.046</td>
<td>0.066</td>
<td>0.053</td>
<td>0.045</td>
<td>0.059</td>
</tr>
<tr>
<td>3.</td>
<td>$G. \text{herbaceum}$</td>
<td>0.044</td>
<td>0.024</td>
<td>0.032</td>
<td>0.018</td>
<td>0.027</td>
<td>0.037</td>
<td>0.024</td>
<td>0.040</td>
<td>0.031</td>
<td>0.020</td>
<td>0.039</td>
</tr>
<tr>
<td>4.</td>
<td>$G. \text{barbadense}$</td>
<td>0.068</td>
<td>0.037</td>
<td>0.055</td>
<td>0.037</td>
<td>0.048</td>
<td>0.058</td>
<td>0.047</td>
<td>0.063</td>
<td>0.048</td>
<td>0.043</td>
<td>0.060</td>
</tr>
<tr>
<td>5.</td>
<td>$G. \text{tomentosum}$</td>
<td>0.076</td>
<td>0.060</td>
<td>0.065</td>
<td>0.049</td>
<td>0.057</td>
<td>0.064</td>
<td>0.057</td>
<td>0.070</td>
<td>0.058</td>
<td>0.051</td>
<td>0.071</td>
</tr>
<tr>
<td>6.</td>
<td>$G. \text{hirsutum (A)}$</td>
<td>0.076</td>
<td>0.024</td>
<td>0.061</td>
<td>0.044</td>
<td>0.053</td>
<td>0.058</td>
<td>0.053</td>
<td>0.064</td>
<td>0.055</td>
<td>0.047</td>
<td>0.066</td>
</tr>
<tr>
<td>7.</td>
<td>$G. \text{mustelinum (A)}$</td>
<td>0.088</td>
<td>0.100</td>
<td>0.068</td>
<td>0.052</td>
<td>0.064</td>
<td>0.070</td>
<td>0.064</td>
<td>0.079</td>
<td>0.065</td>
<td>0.060</td>
<td>0.072</td>
</tr>
<tr>
<td>8.</td>
<td>2(A2D1) (A)</td>
<td>0.084</td>
<td>0.034</td>
<td>0.054</td>
<td>0.036</td>
<td>0.046</td>
<td>0.059</td>
<td>0.046</td>
<td>0.070</td>
<td>0.047</td>
<td>0.043</td>
<td>0.054</td>
</tr>
<tr>
<td>9.</td>
<td>$G. \text{gossypiioides}$</td>
<td>0.075</td>
<td>0.072</td>
<td>0.061</td>
<td>0.048</td>
<td>0.057</td>
<td>0.062</td>
<td>0.052</td>
<td>0.067</td>
<td>0.061</td>
<td>0.048</td>
<td>0.069</td>
</tr>
<tr>
<td>10.</td>
<td>$G. \text{raimondii}$</td>
<td>0.051</td>
<td>0.032</td>
<td>0.039</td>
<td>0.024</td>
<td>0.035</td>
<td>0.044</td>
<td>0.032</td>
<td>0.051</td>
<td>0.039</td>
<td>0.028</td>
<td>0.044</td>
</tr>
<tr>
<td>11.</td>
<td>$G. \text{hirsutum (D)}$</td>
<td>0.052</td>
<td>0.063</td>
<td>0.052</td>
<td>0.037</td>
<td>0.046</td>
<td>0.037</td>
<td>0.043</td>
<td>0.064</td>
<td>0.044</td>
<td>0.039</td>
<td>0.057</td>
</tr>
<tr>
<td>12.</td>
<td>$G. \text{mustelinum (D)}$</td>
<td>0.066</td>
<td>0.082</td>
<td>0.068</td>
<td>0.050</td>
<td>0.060</td>
<td>0.065</td>
<td>0.058</td>
<td>0.072</td>
<td>0.055</td>
<td>0.053</td>
<td>0.068</td>
</tr>
<tr>
<td>13.</td>
<td>$G. \text{schwendimanii}$</td>
<td>0.053</td>
<td>0.045</td>
<td>0.033</td>
<td>0.043</td>
<td>0.054</td>
<td>0.040</td>
<td>0.053</td>
<td>0.050</td>
<td>0.035</td>
<td>0.055</td>
<td>0.068</td>
</tr>
<tr>
<td>14.</td>
<td>$G. \text{aridum}$</td>
<td>0.080</td>
<td>0.059</td>
<td>0.054</td>
<td>0.062</td>
<td>0.071</td>
<td>0.057</td>
<td>0.071</td>
<td>0.068</td>
<td>0.054</td>
<td>0.072</td>
<td>0.079</td>
</tr>
<tr>
<td>15.</td>
<td>$G. \text{lobatum}$</td>
<td>0.080</td>
<td>0.056</td>
<td>0.034</td>
<td>0.046</td>
<td>0.056</td>
<td>0.042</td>
<td>0.058</td>
<td>0.050</td>
<td>0.040</td>
<td>0.054</td>
<td>0.072</td>
</tr>
<tr>
<td>16.</td>
<td>$G. \text{laxum}$</td>
<td>0.105</td>
<td>0.114</td>
<td>0.100</td>
<td>-</td>
<td>0.028</td>
<td>0.042</td>
<td>0.029</td>
<td>0.043</td>
<td>0.030</td>
<td>0.024</td>
<td>0.040</td>
</tr>
<tr>
<td>17.</td>
<td>$G. \text{klotzschianum}$</td>
<td>0.109</td>
<td>0.126</td>
<td>0.114</td>
<td>0.127</td>
<td>-</td>
<td>0.046</td>
<td>0.039</td>
<td>0.054</td>
<td>0.041</td>
<td>0.036</td>
<td>0.049</td>
</tr>
<tr>
<td>18.</td>
<td>$G. \text{davidsonii}$</td>
<td>0.121</td>
<td>0.138</td>
<td>0.130</td>
<td>0.144</td>
<td>0.054</td>
<td>-</td>
<td>0.047</td>
<td>0.064</td>
<td>0.048</td>
<td>0.044</td>
<td>0.062</td>
</tr>
<tr>
<td>19.</td>
<td>$G. \text{trilobum}$</td>
<td>0.118</td>
<td>0.135</td>
<td>0.118</td>
<td>0.132</td>
<td>0.072</td>
<td>0.090</td>
<td>-</td>
<td>0.050</td>
<td>0.040</td>
<td>0.032</td>
<td>0.048</td>
</tr>
<tr>
<td>20.</td>
<td>$G. \text{thurberi}$</td>
<td>0.126</td>
<td>0.141</td>
<td>0.131</td>
<td>0.133</td>
<td>0.083</td>
<td>0.100</td>
<td>0.071</td>
<td>-</td>
<td>0.055</td>
<td>0.049</td>
<td>0.067</td>
</tr>
<tr>
<td>21.</td>
<td>2(A2D1) (D)</td>
<td>0.107</td>
<td>0.125</td>
<td>0.112</td>
<td>0.120</td>
<td>0.063</td>
<td>0.082</td>
<td>0.044</td>
<td>0.059</td>
<td>-</td>
<td>0.037</td>
<td>0.052</td>
</tr>
<tr>
<td>22.</td>
<td>$G. \text{turneri}$</td>
<td>0.089</td>
<td>0.112</td>
<td>0.101</td>
<td>0.112</td>
<td>0.054</td>
<td>0.077</td>
<td>0.066</td>
<td>0.079</td>
<td>0.055</td>
<td>-</td>
<td>0.046</td>
</tr>
<tr>
<td>23.</td>
<td>$G. \text{harknessii}$</td>
<td>0.092</td>
<td>0.115</td>
<td>0.104</td>
<td>0.121</td>
<td>0.059</td>
<td>0.080</td>
<td>0.071</td>
<td>0.082</td>
<td>0.061</td>
<td>0.037</td>
<td>-</td>
</tr>
<tr>
<td>24.</td>
<td>$G. \text{armourianum}$</td>
<td>0.094</td>
<td>0.112</td>
<td>0.102</td>
<td>0.119</td>
<td>0.052</td>
<td>0.071</td>
<td>0.065</td>
<td>0.079</td>
<td>0.055</td>
<td>0.033</td>
<td>0.037</td>
</tr>
</tbody>
</table>
Table 4. Mean 5S rDNA copy-number estimates in *Gossypium*, based on slot-blot hybridization and phosphorimaging. N = the number of replicate experiments; 95% CI = 95% confidence interval; CV = coefficient of variation.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>N</th>
<th>Copy Number</th>
<th>95% CI</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C-genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. robinsonii</em></td>
<td>4</td>
<td>2940</td>
<td>2121 - 3755</td>
<td>17.5%</td>
</tr>
<tr>
<td><strong>A-genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. herbaceum</em></td>
<td>8</td>
<td>3415</td>
<td>2609 - 4222</td>
<td>28.2%</td>
</tr>
<tr>
<td><em>G. arboreum</em></td>
<td>9</td>
<td>7550</td>
<td>5900 - 9037</td>
<td>21.0%</td>
</tr>
<tr>
<td><strong>D-genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. gossypioides</em></td>
<td>8</td>
<td>1145</td>
<td>812 - 1477</td>
<td>34.8%</td>
</tr>
<tr>
<td><em>G. trilobum</em></td>
<td>8</td>
<td>1420</td>
<td>1101 - 1737</td>
<td>26.8%</td>
</tr>
<tr>
<td><em>G. aridum</em></td>
<td>8</td>
<td>1685</td>
<td>492 - 2880</td>
<td>84.7%</td>
</tr>
<tr>
<td><em>G. thurberi</em></td>
<td>8</td>
<td>2070</td>
<td>1862 - 2280</td>
<td>12.1%</td>
</tr>
<tr>
<td><em>G. laxum</em></td>
<td>8</td>
<td>2760</td>
<td>2289 - 3230</td>
<td>20.4%</td>
</tr>
<tr>
<td><em>G. armourianum</em></td>
<td>8</td>
<td>4205</td>
<td>3301 - 5105</td>
<td>25.7%</td>
</tr>
<tr>
<td><em>G. harknessii</em></td>
<td>8</td>
<td>4385</td>
<td>3015 - 5750</td>
<td>37.3%</td>
</tr>
<tr>
<td><em>G. schwendimanii</em></td>
<td>8</td>
<td>4615</td>
<td>3964 - 5266</td>
<td>16.9%</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>8</td>
<td>4730</td>
<td>3838 - 5623</td>
<td>22.6%</td>
</tr>
<tr>
<td><em>G. lobatum</em></td>
<td>8</td>
<td>5830</td>
<td>4048 - 7614</td>
<td>36.6%</td>
</tr>
<tr>
<td><em>G. klotzschianum</em></td>
<td>8</td>
<td>6930</td>
<td>5509 - 8346</td>
<td>24.5%</td>
</tr>
<tr>
<td><em>G. turneri</em></td>
<td>7</td>
<td>8495</td>
<td>6934 - 10054</td>
<td>19.9%</td>
</tr>
<tr>
<td><em>G. davidsonii</em></td>
<td>8</td>
<td>10280</td>
<td>8518 - 12039</td>
<td>20.5%</td>
</tr>
<tr>
<td><strong>AD-genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. hirsutum</em></td>
<td>8</td>
<td>11190</td>
<td>8154 - 14228</td>
<td>32.5%</td>
</tr>
<tr>
<td>2(A2D1)</td>
<td>8</td>
<td>14015</td>
<td>10503 - 17522</td>
<td>30.0%</td>
</tr>
<tr>
<td><em>G. mustelinum</em></td>
<td>8</td>
<td>21845</td>
<td>15007 - 28684</td>
<td>37.4%</td>
</tr>
<tr>
<td><em>G. tomentosum</em></td>
<td>8</td>
<td>22290</td>
<td>18822 - 25765</td>
<td>18.6%</td>
</tr>
<tr>
<td><em>G. barbadense</em></td>
<td>8</td>
<td>23515</td>
<td>20100 - 26929</td>
<td>17.4%</td>
</tr>
</tbody>
</table>
Table 5. Tests for equivalent patterns of sequence evolution in 5S genes and spacers. Two-
by-two contingency tables were constructed for the taxa shown, where the observed numbers
of fixed and polymorphic differences (columns) were tabulated for 5S gene and spacer
sequences (rows).

<table>
<thead>
<tr>
<th>Pairwise Comparison</th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within genomes/subgenomes:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. herbaceum</em> vs. <em>G. arboreum</em></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><em>G. herbaceum</em> vs. <em>G. mustelinum (A)</em></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>G. raimondii</em> vs. <em>G. lobatum</em></td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td><em>G. raimondii</em> vs. <em>G. mustelinum (D)</em></td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Between genomes:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. herbaceum</em> vs. <em>G. lobatum</em></td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td><em>G. herbaceum</em> vs. <em>G. raimondii</em></td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td><em>G. arboreum</em> vs. <em>G. raimondii</em></td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td><em>G. mustelinum (A)</em> vs. <em>G. raimondii</em></td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td><em>G. mustelinum (A)</em> vs. <em>mustelinum (D)</em></td>
<td>0</td>
<td>32</td>
</tr>
</tbody>
</table>
Figure 1: Aligned nucleotide sequences of 99 cloned 5S rDNA repeats from *Gossypium*. Periods represent residues identical to those of the reference sequence from *G. robinsonii*; dashes indicate alignment gaps; and question marks denote missing information. The first 122 nucleotides correspond to the 5S gene; the intergenic spacer encompasses nucleotides 123-316.

|   | robinsonii11 | robinsonii18 | robinsonii10 | A2Dlsyn3 | A2Dlsyn5 | A2Dlsyn6 | A2Dlsyn15 | A2Dlsyn16 | arboreum1 | Arboreum2 | Arboreum4 | Arboreum5 | Arboreum6 | Herbaceum1 | Herbaceum2 | Herbaceum3 | Herbaceum4 | Herbaceum5 | Barbadense1 | barbadense2 | barbadense3 | barbadense6 | barbadense7 | barbadense8 | barbadense10 | tomentosum1 | tomentosum2 | tomentosum6 | tomentosum7 | tomentosum10 | tomentosum11 | mustelinum9 |
|---|--------------|--------------|--------------|---------|---------|---------|---------|---------|---------|---------|----------|---------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 10| GGGTCGATC  | TGATGCGACG  | TGATGCGACG  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  | TGGCGAC  |
| 20|             |             |             | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   | AAAAA   |
| 30| C           | A           | A           | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       | C       |
| 40|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 50|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 60|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 70|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 80|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 90|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 100|            |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
| 110|             |             |             |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
Figure 1, Continued

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>hirsutum9</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>gossypioide1</td>
<td>A..</td>
<td>.. T..</td>
<td>.. A..</td>
<td>.. A..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>schwendemannii5</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>aridum4</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>lobatum4</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>lobatum1</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>lobatum5</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>lobatum2</td>
<td>T..</td>
<td>.. C</td>
<td>.. G..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>laxum4</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>laxum5</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>laxum6</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>klotzschianum4</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>klotzschianum5</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>klotzschianum6</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>klotzschianum8</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>davidsonii5</td>
<td>T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>davidsonii6</td>
<td>A..</td>
<td>.. T..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>trilobum1</td>
<td>A..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>trilobum10</td>
<td>C..</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
<tr>
<td>trilobum11</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
<td>.. C</td>
</tr>
</tbody>
</table>
Figure 1, Continued

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>thurber13</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
</tr>
<tr>
<td>thurber15</td>
<td>...</td>
<td>A</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>AC</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
</tr>
<tr>
<td>thurber17</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A2Dlayn11</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
</tr>
<tr>
<td>A2Dlayn13</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
</tr>
<tr>
<td>A2Dlayn14</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A2Dlayn17</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>turner11</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>turner12</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>turner13</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>harkness11</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>G</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>AC</td>
<td></td>
</tr>
<tr>
<td>harkness12</td>
<td>...</td>
<td>...</td>
<td>TC</td>
<td>...</td>
<td>...</td>
<td>G</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>harkness13</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>TG</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
</tr>
<tr>
<td>harkness15</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>armourianum1</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>T</td>
<td>...</td>
</tr>
<tr>
<td>armourianum10</td>
<td>...</td>
<td>...</td>
<td>CC</td>
<td>...</td>
<td>...</td>
<td>AC</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>A</td>
<td>...</td>
</tr>
<tr>
<td>raimondi11</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>raimondi13</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>raimondi14</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>raimondi15</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>raimondi16</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>C</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>raimondi17</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mustelinum6</td>
<td>...</td>
<td>...</td>
<td>A</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mustelinum7</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mustelinum13</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mustelinum14</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>mustelinum17</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>hirsutum6</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>hirsutum7</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>hirsutum10</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>hirsutum12</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Figure 1, Continued
Figure 1, Continued

<table>
<thead>
<tr>
<th>230</th>
<th>240</th>
<th>250</th>
<th>260</th>
<th>270</th>
<th>280</th>
<th>290</th>
<th>300</th>
<th>310</th>
<th>316</th>
</tr>
</thead>
</table>
| robbinsou111 | AAATGTGACTGTTATTTGTCTAAATTACAATGAAATTCGGGCGGAGGTGGCCTAAGGGGGACAGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
Figure 1, Continued

<table>
<thead>
<tr>
<th></th>
<th>230</th>
<th>240</th>
<th>250</th>
<th>260</th>
<th>270</th>
<th>280</th>
<th>290</th>
<th>300</th>
<th>310</th>
<th>316</th>
</tr>
</thead>
<tbody>
<tr>
<td>mustellinus16</td>
<td>C.C........G...GT......A......C........TC.A.C........C........T...........A........A...G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mustellinus18</td>
<td>C.C........A......G...GT......A......C........TC.A.C........C........T...........A........A...G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mustellinus19</td>
<td>C.C........A......G...GT......A......C........TC.A.C........C........T...........A........A...G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hirsutum1</td>
<td>C.CAC........GT...GT......A......C......C..GT.....A........C........A........GG...A........A...G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hirsutum8</td>
<td>C.C........G...GT......A......C........TC.A.C........C........T...........G........A........A...A...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hirsutum9</td>
<td>C.C........G...GT......A......C........TC.A.C........C........T...........G........A...T........A...G.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hirsutum11</td>
<td>C.C........G...GT......A......C........TC.A.C........C........T...........G........A...A......A...G.T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gossypioides1</td>
<td>C.T.C........C.C......G......G......C........C........T........C.AA........G........A...G...A.GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gossypioides4</td>
<td>C.T.C........A......C......G......C........C........T........C.AA........C........C........C........G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gossypioides6</td>
<td>C.T.C........TC.C......C......G......GT......C........C.AA........C........C........C........C........G...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1, Continued

<table>
<thead>
<tr>
<th></th>
<th>230</th>
<th>240</th>
<th>250</th>
<th>260</th>
<th>270</th>
<th>280</th>
<th>290</th>
<th>300</th>
<th>310</th>
<th>316</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>T..</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A...</td>
<td>C.A...</td>
<td>.C...</td>
<td>...</td>
<td>G...</td>
</tr>
<tr>
<td>h1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>i1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>j1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>k1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>m1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>n1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>o1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>q1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>r1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>s1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>t1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>u1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>w1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>x1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>y1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
<tr>
<td>z1</td>
<td>T...</td>
<td>A...</td>
<td>G..</td>
<td>G..</td>
<td>.C...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
<td>A.A...</td>
</tr>
</tbody>
</table>
Figure 2: Consensus gene trees resulting from maximum parsimony analysis of 99 cloned 5S rDNA sequences from *Gossypium*. Topologies shown are the strict consensus of 3,000 trees recovered from heuristic searches. Numbers above branch segments indicate the number of additional steps that are required for each resolved clade to collapse ("decay values"). For example, in trees 2 steps longer than the most parsimonious, the clade consisting of *schwendimanii*5 and *schwendimanii*8 is no longer supported. (A) Consensus tree based on sequence data from the intergenic spacers alone (length of each constituent tree = 576 steps; consistency index = 0.55; retention index = 0.87). (B) Consensus tree based on sequence data from the 5S genes alone (treelength = 299 steps; CI = 0.60; RI = 0.54). Both trees are rooted with sequences from the outgroup taxon *G. robinsonii*, which resolved as monophyletic in the tree based on 5S spacer sequences (Fig. 2A), but was constrained to be monophyletic in the tree based on 5S gene sequences (Fig. 2B).
Figure 3: Consensus tree resulting from maximum parsimony analysis of 23 5S rDNA spacer sequences from *Gossypium*. For each diploid species, a consensus sequence was computed prior to phylogenetic analysis, as described in the text. Similarly, consensus sequences were computed separately for sequences from each subgenome of the allopolyploids. The topology shown, rooted with the outgroup taxon *G. robinsonii*, is the strict consensus of 3,000 minimum length trees recovered from heuristic searches. Numbers above branch segments indicate the number of additional steps that are required for each clade to collapse. Each constituent tree had a length of 121 steps, a consistency index of 0.87, and a retention index of 0.95.
Figure 4: Neighbor-joining tree based on Kimura 2-parameter distances between 99 cloned 5S rDNA spacer sequences from *Gossypium*. The tree is rooted with sequences from the outgroup taxon *G. robinsonii*. Branch lengths are drawn to scale (shown at bottom).
Figure 5: Model illustrating the major features of 5S rDNA evolution in *Gossypium*. Panel A: Nucleotide substitutions in individual repeats of an array generate polymorphisms. Concerted evolutionary forces may eliminate variants or spread them throughout individual arrays (intralocus), but not between different arrays (interlocus). Most nucleotide positions in the spacer region are presumed to be free to vary, i.e., variants are selectively neutral or near-neutral. Consequently, variant nucleotides can become either fixed or lost. In contrast, most mutations in the 5S gene are presumed to be selectively neutral or near-neutral only when they occur in a sub-critical proportion of repeats in an array. Once this threshold frequency is reached, the variant nucleotide becomes disadvantageous due to its effects on 5S transcription or 5S RNA function; thus, the array acquires reduced fitness and is selectively removed from the population. As a consequence, fixed interspecific differences in 5S genes fail to accumulate, despite the fact that polymorphic nucleotide positions are common. Panel B: Diagrammatic representation of fitness curves as a function of the proportion of functional 5S genes in an array. Several possibilities are illustrated, although neither the actual shapes of these curves nor the level at which a selective threshold is reached are known. When the number of functional copies drops below a threshold level, selection can operate to remove arrays.
CHAPTER FOUR

SIMPLE METHODS FOR ISOLATING HOMOELOGOUS LOCI
FROM ALLOPOLYPLOID GENOMES

A paper submitted to the journal Genome

Richard C. Cronn and Jonathan F. Wendel

ABSTRACT
During allopolyploid formation two divergent sets of chromosomes are combined in a common nucleus. Accordingly, loci that are single-copy in progenitor diploid genomes become duplicated, homoeologous loci in allopolyploids. Although many homoeologous loci have been identified using classical genetic and linkage mapping approaches, there are few examples of the isolation and molecular characterization of homoeologous loci from allopolyploids. Here we describe two methods for the isolation of strictly orthologous duplicated loci and demonstrate the feasibility of the techniques using allotetraploid cotton. The methods utilize restriction-digested, size-fractionated genomic DNA as a template for either plasmid cloning or PCR amplification. This fractionation procedure, when combined with Southern hybridization and mapping information, effectively separates homoeologues from each other and from more divergent cross-hybridizing sequences (paralogues). Each homoeologue is thereby confidently isolated from different pools of size-fractionated genomic DNA. While these methods were specifically designed for isolating orthologous duplicated loci from allotetraploids, they should be applicable to a broad spectrum of diploid and polyploid plants and be useful for studying both ancient and recent gene duplications. In addition, the procedures yield strictly orthologous sequences, which are required for phylogenetic analysis. Thus, a general approach is detailed for the isolation of nuclear genes for phylogeny reconstruction.

Key words: Polyploidy, homoeology, orthology, phylogenetic analysis, Gossypium
INTRODUCTION

Gene duplication is an important process in the evolution of eukaryotic genomes because it creates genetic redundancy, a condition required for the functional diversification of gene families (Ohno 1970; Ohta 1994). Of the numerous mechanisms known to create gene duplications, polyploidy is unique in that entire genomes become duplicated and combined within a common nucleus, hence yielding genome equivalents of redundant genetic information. While polyploidy is a prevalent process in plant evolution (Masterson 1994), little is known regarding the fate of genes that become duplicated via polyploidy. These duplicate loci, termed “homoeologues”, are thought to play a central role in the evolutionary success and adaptive radiation of polyploid taxa. Suggested benefits include “fixed heterozygosity”, which can mask deleterious alleles (Spofford 1969; Werth and Windham 1991) or furnish a presumptive advantage of hybrid vigor (Grant 1981), and the acquisition of new function due to relaxed selection. Relaxation of purifying selection has been suggested to facilitate divergence of one of the two homoeologues, possibly leading to novel gene function (Ohno 1970; Li 1983). In addition to the possibility of functional divergence, duplicated loci in polyploids may maintain their original function (Hughes 1994), or one copy may be silenced or degenerate into a non-functional pseudogene (Soltis and Soltis 1993; Clark 1994; Walsh 1995).

While the importance of gene duplication in polyploidy has been recognized for almost half a century (Stephens 1951), molecular genetic analysis of homoeologous loci has been hampered by technical challenges. Perhaps most vexing has been the problem of distinguishing loci whose close relationship reflects a relatively recent polyploidization event from loci related by other duplication processes or more ancient cycles of whole-genome duplication. In this respect, it is noteworthy that most diploid plant genomes contain a high degree of genetic redundancy (Cavalier-Smith 1985; Tanksley and Pichersky 1988; Pickett and Meeks-Wagner 1995). This is commonly evidenced in Southern hybridization analyses and RFLP-mapping studies, with many or even most cDNA and genomic clones detecting multiple DNA fragments when used as hybridization probes against genomic DNAs, even under stringent hybridization conditions (e.g., Bonierbale et al. 1988; Jena et al. 1994; Brubaker et al. 1998). Even the highly streamlined genome of Arabidopsis thaliana contains appreciable genetic redundancy, with more than 15% of probes revealing duplicated genomic regions (Kowalski et al. 1994). Accordingly, many if not most genes in plant genomes occur as members (i.e., paralogous copies, or paralogues) of small multi-gene families whose intra-family relationships reflect a complex evolutionary history of gene duplication and
divergence. This history poses a stumbling block for those wishing to isolate a specific member of a gene family, or the "same" member of a gene family (i.e., strictly orthologous copies, or orthologues) from two or more taxa.

This technical difficulty is compounded in polyploid plants, since for each gene family, the entire set of paralogues becomes duplicated during polyploidization (Fig. 1). To examine the fate of specific loci that became duplicated by virtue of allopolyploidization (homoeologues), it becomes necessary to isolate each locus not just from its corresponding copy in the other subgenome, but also from the suite of divergent paralogues that presumably exist in both subgenomes. It is this task — confidently deciphering whether a pair of sequences isolated from an allopolyploid are related by homoeology (duplication via polyploidization) rather than paralogy (duplication via other mechanisms or more ancient cycles of polyploidy) — that remains the crux of homoeologue analysis.

With the long-term goal of assessing the evolutionary fate of duplicated loci in allopolyploid cotton (*Gossypium hirsutum* L.), we devised two simple approaches for isolating homoeologous genes and genomic regions. Using mapped loci revealed by anonymous *PstI* genomic DNA fragments, we modified standard cloning and sequence-tagged-site (STS) PCR methods to create pools of DNA enriched for homoeologous target loci while minimizing contamination from the numerous paralogues that are present. Both methods utilize restriction enzyme-digested genomic DNA which has been size-fractionated by agarose gel electrophoresis. This simple fractionation step frequently separates two homoeologous loci into different DNA size pools, thereby affording a means for isolation and characterization of each duplicated copy. Two such means are described here, one entailing the construction of plasmid libraries and the other using STS-PCR. In conjunction with Southern hybridization and RFLP-mapping data, the methods allow the strict homoeology and subgenomic origin of each particular homoeologue to be confidently assessed. To demonstrate the feasibility of the methods, we show how they were applied to the isolation of a pair of homoeologous P-glycoprotein genes from allotetraploid cotton. In addition, we isolated the orthologous copies of these genes from the two diploid species that represent the ancestral genome donors to the allopolyploids. While these methods were specifically designed for isolating homoeologous loci from allotetraploid cotton, the methods appear to be general, and thus they should be applicable to a broad spectrum of diploid and polyploid plants, and be useful for studying both ancient and recent gene duplications.
MATERIALS AND METHODS

Organismal Context
Available evidence suggests that allotetraploid species of cotton (AD-genome, $2n = 4X = 52$) formed approximately 1 - 2 million years ago, in the New World, after hybridization between two divergent diploid parents, one similar to the present-day African species *G. herbaceum* L. (A-genome; $2n = 2x = 26$) and the other to the Peruvian species *G. raimondii* Ulbrich (D-genome; $2n = 2x = 26$) (Wendel 1989; Wendel and Albert 1992; Endrizzi et al. 1985). After polyploidization, the AD-genome group diverged into three distinct lineages (Brubaker and Wendel 1994; Small et al. 1998), presently represented by five modern species, including the economically important cottons *G. hirsutum* L. and *G. barbadense*. The interrelationships between diploid and allopolyploid cottons and their respective genomes are depicted in Fig. 1.

Plant Materials
Genetic linkage maps for allotetraploid cotton (Reinisch et al. 1994) and the diploid A- and D-genome groups (Brubaker et al. 1998) have been constructed. Comparisons among the maps demonstrate that, in general, a high degree of synteny is conserved between the two diploid genomes, and that this synteny is further retained within the allopolyploids, notwithstanding a small number of structural mutations that have arisen at both the diploid and allopolyploid levels. Thus, homoeologous linkage groups have been demonstrated for allopolyploid *Gossypium*, with clear counterparts in diploid cottons. Considerable genomic redundancy is indicated for both diploids and allotetraploids: using stringent hybridization conditions, anonymous "low-copy" probes, for example, hybridize to an average of 2.96 and 2.71 genomic EcoRI fragments in A-genome and D-genome diploids, respectively, and to 4.12 genomic EcoRI fragments in the allopolyploids (Reinisch et al. 1994).

For the present study, we chose one representative from each of the populations used for constructing the RFLP maps described in Reinisch et al. (1994) and Brubaker et al. (1998). *Gossypium hirsutum* race palmeri was used to represent allopolyploid, AD-genome cottons, while *G. herbaceum* and *G. raimondii* were selected for A-genome and D-genome diploids, respectively. DNAs were extracted using methods detailed in Paterson et al. (1993).

Locus Selection
Homoeologous linkage groups and their constituent loci were revealed using clones from several libraries, including those constructed using anonymous cDNA and genomic clones (Reinisch et al. 1994; Brubaker et al. 1998). We are presently investigating sequence evolution for a number of mapped sets of orthologous and homoeologous loci in the diploids...
and allopolyploids. Here, we focus on the gene revealed by the anonymous PstI probe G1262, because this gene appears to exist in only one copy per diploid genome and two copies in the allopolyploids. This 1,018 bp clone, originally isolated from PstI-fractionated G. raimondii DNA (Brubaker and Wendel 1994), shows high sequence similarity to the transcribed P-glycoprotein gene pgpl from A. thaliana (Genbank accession number 99741; Dudler and Hertig 1992). When hybridized with probe G1262 under conditions of high stringency, HindIII-digested genomic DNAs from the A-genome and D-genome diploids show single fragments of approximately 1.9 kb and 4.5 kb, respectively (Fig. 2A). These fragments are additive in allotetraploid cottons, providing indirect evidence of homoeology. Comparative linkage mapping of this locus in interspecific A-genome, D-genome, and AD-genome populations confirms that the hybridizing fragments from diploids are orthologous loci, and that the two fragments in the allotetraploid are homoeologous loci (Fig. 2B). Specifically, marker loci revealed by G1262 fall onto a single homoeologous assemblage of linkage groups corresponding to diploid linkage groups A3 and D10 and tetraploid chromosomes 6A and 25D (Brubaker et al. 1998).

Size-Fractionation of Genomic DNA

Approximately 10 μg of genomic DNA from allotetraploid and diploid cottons was digested to completion using the restriction enzyme HindIII, and the products were separated on a 1.0% low-melting point agarose gel (SeaPlaque GTG; FMC BioProducts) in a low-EDTA, TAE buffer (0.1 mM EDTA, 40 mM Tris-Acetate). After staining with ethidium bromide, DNAs were visualized under long-wavelength UV light. Using a 1-kb DNA ladder for reference, each lane of the gel was carefully excised using 1-kb intervals (e.g., 1-2 kb, 2-3 kb, 3-4 kb, etc., up to 10 kb), and the DNA was isolated from each agarose slice using either Gene-Clean™ (Bio-101; used for DNA larger than 4 kb) or electrophoresis onto DEAE-cellulose membrane (Schleicher & Schuell; used for DNA smaller than 4 kb).

Isolation of Homoeologues - Assumptions of the Approach

Two methods were designed to isolate homoeologues, one employing selective cloning and the other STS-PCR. For either approach to succeed, two important criteria must be satisfied:

(1) Restriction enzyme and PCR primer sites must be conserved in the allopolyploid subgenomes. Both the PstI-selective cloning and STS-PCR approaches require sequence conservation of homoeologous loci in regions corresponding to the 5' and 3' ends of the probe of interest. In addition, due to the methylation sensitivity of PstI, the homoeologous loci of interest must be unmethylated. Restriction site conservation and methylation status
can be verified experimentally by checking the susceptibility of subgenomic sequences to restriction enzyme digestion prior to cloning. In the case of PstI-probe G1262, for example, digestion of allotetraploid *G. hirsutum* genomic DNA with PstI yields a homogeneous digestion product of ~1.0 kb (Fig. 2A). This band is identical in size to the original PstI probe, confirming that the PstI sites that flank locus G1262 are conserved (with respect to nucleotide sequence and methylation status) in both the A- and D-subgenomes of the allotetraploid. It is more difficult to evaluate conservation of PCR priming sites *a priori*, although in some cases other evidence may be used to infer the degree of divergence in homoeologues. In the present example, the high degree of nucleotide conservation between P-glycoprotein-like genes of plants and animals (Dudler and Hertig, 1992) suggested that the primer sites had a similarly high likelihood of conservation.

(2) Mapping probes must be fully contained within the RFLP bands they reveal. Since the PstI-selective cloning and STS-PCR approaches require conservation of the regions corresponding to the termini of the cloned probe, both PstI sites and PCR primer sites must be located internal to the enzyme sites used to generate RFLPs. The probability of an RFLP probe being entirely contained within an RFLP band depends both upon the size of the probe (X) and the RFLP band (Y). This relationship can be expressed mathematically as \( P = (Y-X)/(X+Y) \), where the numerator describes the number of possible “alignment states” for containment of the probe within the RFLP band, and the denominator describes the total number of alignment states possible for the probe and the RFLP. Using the 1,018 bp probe G1262 as an example, the probability that both PstI sites lie within the HindIII digestion products are 0.30 for the 1.9 kb *G. herbaceum* band and 0.63 for the 4.5 kb *G. raimondii* band. While these probabilities appear low, sequencing of probe G1262 revealed no internal HindIII sites, providing evidence that the probe is fully contained within the RFLP band from its source genome (*G. raimondii*) and suggesting that it may be similarly internal in the other relevant genomic HindIII fragments.

**PstI-Selective Cloning of Homoeologues**

To isolate homoeologues of locus G1262 by selective cloning, we digested ~500 ng of the 1-2 kb and 4-5 kb HindIII size-fractionated DNA pools from *G. hirsutum* with the restriction enzyme PstI. Size-fractionated DNAs from *G. herbaceum* (from the 1-2 kb HindIII pool) and *G. raimondii* (4-5 kb HindIII pool) were treated similarly for isolation of orthologous diploid loci. Following digestion, products were cleaned with Gene-Clean and ligated into the PstI site of the plasmid vector pZero (Invitrogen) as per manufacturers instructions. After ligation, mixtures were heated to 65°C, diluted 1/10 in water and aliquots
(1 μl) were electroporated into TOP10F+ E. coli cells. Cells were incubated overnight at 37°C on LB-agar plates containing Zeocin (0.1 mM) and IPTG (1 mM). Once colonies had grown to ~0.5 - 1.0 mm in diameter, a replica was made of the original transformants onto a second LB-agar/Zeocin/IPTG plate, and these petri dishes were incubated for 8 hours at 37°C. After incubation, cells from the original transformation plate were used to make a permanent mini-library by eluting the cells in a minimal volume (1-2 ml) of LB medium containing Zeocin (0.1 mM) and 35% glycerol. These cells were frozen in dry ice/ethanol and stored at -70°C. The second replica plate was used to make a colony lift onto nylon membrane (Nytran™, Schleicher & Schuell) as directed by the manufacturer. Colony lifts were later probed using methods described in Ausubel (1994). To evaluate the efficiency of cloning, colony lift membranes were screened with three probes: (1) the G1262 probe insert, to reveal positive clones; (2) labeled, size-fractionated DNA from the original size fraction, to reveal clones containing highly-repetitive DNA; and (3) labeled pZero vector, to provide an estimate of the total number of clones on the membrane.

PCR-amplification of Homoeologues

To amplify the homoeologues and diploid orthologues of locus G1262, clone G1262 was sequenced using universal sequencing primers (M13U and T7) of the vector and primers were designed. The priming sites for “G1262F” (5’-GGCGGCGAGGAAGCAGCCTY-3’) and “G1262R” (5’-CGGAGGGTCAAGCAGCTTCT-3’) were selected to minimize intermolecular primer dimers and to maintain an annealing temperature ≥54°C. Primer selection was facilitated using the computer program OLIGO 4.0 (Rychlik and Rhoades 1989). Amplification cycling parameters were similar to those described in Don et al. (1991): a two minute hotstart (80°C) was followed by 10 cycles of 1 minute at 94°C for denaturation, 1 minute at 60°C (with a decrease of 0.6°C per cycle) for primer annealing, and a 1 minute extension at 72°C. An additional 25 cycles were run once the annealing temperature reached 54°C. Each reaction mixture (25 μl) contained 10 picomoles of each primer, a final MgCl₂ concentration of 2.5 mM, and approximately 10 ng of HindIII size-fractionated DNA.

DNA Sequencing

Isolated clones and PCR products were sequenced manually using the Amersham cycle sequencing kit with [³²P]ddNTP terminators. Sequencing reactions typically used 0.25 picomoles of template and 2 picomoles of either the G1262F or G1262R primer. Cycle sequencing parameters were as follows: 45 cycles of 95°C for 30 seconds, 54°C for 30 seconds
and 72°C for 45 seconds. Sequencing products were resolved on 5% acrylamide gels (Long Ranger, FMC) and visualized by autoradiography.

**RESULTS**

**Selective Cloning**

As shown in Fig. 2, Southern hybridization of HindIII-digested DNAs with probe G1262 yields a simple pattern where the A-genome diploid *G. herbaceum* shows a single band of 1.9 kb, the D-genome diploid *G. raimondii* shows a single band of 4.5 kb, and the allotetraploid *G. hirsutum* shows an additive pattern. In addition, hybridization of probe G1262 against PstI-digested *G. hirsutum* DNA yields a single band with a size identical to the original probe. These data indicate that the loci corresponding to the PstI clone G1262 are conserved in all four genomes (A-, D-, and both subgenomes of the allopolyploid) with respect to PstI sites, suggesting that it should be possible to clone all four loci using a selection strategy employing PstI and size-fractionated genomic DNAs.

Using the PstI selective cloning approach described above, each electroporation yielded between 75 and several thousand primary recombinants for each of the size-fractionated DNA libraries. Screens for recombinants using colony hybridization indicated that as many as one in 75 to as few as approximately one in 2000 colonies contained inserts homologous to probe G1262. A representative colony lift showing hybridization of probe G1262 to ~ 800 colonies isolated from the 4-5 kb HindIII DNA pool of *G. hirsutum* is illustrated in Fig. 3. This hybridization reveals a maximum of six potentially positive clones, of which three were isolated and shown to contain inserts similar to G1262. To verify the subgenomic origin of these clones, plasmids were isolated and inserts were digested with the restriction enzyme SpeI. This enzyme is diagnostic for D-subgenomic sequences at nucleotide 427, as a single nucleotide substitution (from 5'-ACTAGT-3' to 5'-ACGAGT-3') makes the A-subgenome orthologue refractory to SpeI digestion. As expected, the three clones isolated from this pool were susceptible to digestion with SpeI, confirming they had originated from the D-subgenome. These results indicate that HindIII size-fractionation had effectively separated the A- and D-subgenome G1262 homoeologues into different size pools, making it possible to obtain each subgenomic locus in the absence of its homoeologue.

In addition to screening colony lifts with probe G1262, filters containing positive G1262 clones were probed with labeled DNA from the size-fraction used in constructing the plasmid library. This screen was performed to estimate the proportion of clones present in each library which contained highly-repetitive DNA sequences. After high stringency washes
and long (1 week) exposures, all library filters showed significant signal. As shown in Figure 3B, the 4-5 kb HindIII DNA pool from *G. hirsutum* contained a minimum of 80 clones with inserts that could be considered highly-repetitive, suggesting that clones containing repetitive DNA can account for a substantial portion (>10%) of the library.

**STS-PCR Approach**

As shown in Fig. 2, Southern hybridization of HindIII-digested DNAs with probe G1262 yielded a simple pattern where the diploids *G. herbaceum* and *G. raimondii* showed a single band (1.9 and 4.5 kb, respectively) and *G. hirsutum* showed an additive pattern with two bands. Hybridization of G1262 to PstI-digested *G. hirsutum* DNA also yielded a single band with a size identical to the original probe. These data, combined with the nucleotide sequence of clone G1262, suggested that PCR primers designed internally to the *Pst*I sites would have a high likelihood of amplifying PCR products from all genomes under consideration.

Using genomic DNA as the template for PCR reactions with primers G1262F and G1262R, amplification products of uniform size (~1000 bp) were readily obtained from the diploids *G. herbaceum*, *G. raimondii* and from the allotetraploid *G. hirsutum*. Direct DNA sequencing showed the PCR products from the diploids to be homogeneous, while the sequence from the allotetraploid was polymorphic. Nearly all polymorphic sites were additive relative to the A- and D-genome diploid sequences, indicating that both the A- and D-subgenomic loci had been amplified. When HindIII size-fractionated DNA was used as a template in PCR reactions, amplification products were again obtained from the diploids *G. herbaceum* (from the 2-3 kb DNA pool) and *G. raimondii* (from the 4-5 kb DNA pool), and from the allotetraploid *G. hirsutum* (from these two pools and from adjacent size fractions; Fig. 4). Sequences of PCR products from the diploids were identical to those obtained when genomic DNA was used in amplification, indicating the locus-specific nature of the primers. However, in contrast to the polymorphic sequence obtained when using genomic *G. hirsutum* DNA as a PCR template, amplification products from the 1-2 kb (A-subgenome) and 4-5 kb (D-subgenome) HindIII-digested size fractions of *G. hirsutum* showed no sequence polymorphism and showed near-identity to the sequences from their counterparts in the respective diploids (Fig. 4D). These results indicate that HindIII size-fractionation had effectively separated the A- and D-subgenome G1262 homoeologues into different size pools, making it possible to amplify each homoeologue separately and to thereby generate uncontaminated sequence.
DISCUSSION

We have described two methods for the isolation of homoeologous loci from an allotetraploid and for their orthologous counterparts from progenitor diploids. While the basic principles used to isolate homoeologues (selective cloning and STS-PCR amplification) are not novel, the application of these approaches to size-fractionated genomic DNA facilitates the isolation of specific loci. Rather than attempting to amplify or clone homoeologues from a genomic DNA pool which contains one homoeologue per subgenome and numerous unwanted paralogues, the size fractions are considerably less complex and are highly enriched for the locus of interest. When combined with Southern hybridization analysis and mapping information, these procedures represent a powerful tool for the analysis of orthologous and homoeologous loci. Their use increases confidence that a specific target orthologue has been isolated from the many other related sequences in a genome, and that its direct duplicate in an allopolyploid or counterparts in other diploids have been similarly isolated. The example provided here, involving the suite of loci corresponding to probe G1262, was selected because Southern hybridization and RFLP mapping data showed that this was a simple system without interfering paralogous copies, at least under high stringency hybridization conditions. We have had success, however, applying these same techniques to other, more complex gene families (unpublished data), indicating that the methods are generally useful.

A number of authors have used sequence-tagged primers to amplify mapped loci from allopolyploid plant genomes (Zhu et al. 1994; Chee et al. 1995; Talbert et al. 1995; Liu et al. 1996). Using similar approaches, we find that the majority of PCR amplification products from sequence-tagged sites (including locus G1262) are polymorphic when allotetraploid cotton genomic DNA is used as a template. This indicates that most STS primers are not subgenome-specific in *Gossypium*, perhaps as a consequence of the limited sequence divergence that has arisen between subgenomes of allotetraploid cotton (1.8 % for locus G1262; similar values for other homoeologous pairs being studied [data not shown]). This outcome is not unique to polyploid cotton, as a similar absence of genome-specificity was observed in STS-PCR amplification from allotetraploid *Stylosanthes* (Liu et al. 1996). In addition, nearly all other STS primer pairs examined in both diploid and allopolyploid cottons lead to heterogeneous amplification pools. Considering the inherent redundancy of allopolyploid and most diploid genomes, lack of genome and locus specificity may prove to be the rule rather than the exception. In these cases, size-fractionation of restriction-
digested template DNA combined with Southern hybridization offers a convenient route to enhanced locus selectivity.

While the use of size-fractionated DNA in cloning and STS-PCR has several advantages over the use of total genomic DNA, there are some disadvantages to each of the approaches described. For the isolation of homeologues by the STS-PCR method, abundant, clean, un-sheared genomic DNA is paramount for the fractionation procedure. Sheared genomic DNA and/or partially digested genomic DNA will lead to the presence of unwanted paralogues in the desired size pool as well as anomalous migration of fragments that include the desired orthologue or homoeologue. These artifacts arise because shearing yields fragments that migrate into smaller than expected size fractions and partially-digested products remain in higher than predicted size fractions (data not shown). Due of the inherently high sensitivity of PCR, these artifacts may even appear in DNA preparations that appear to be reasonably free from shearing (e.g., Fig. 4C). A second factor which influences the success of amplifying a target from a discrete size fraction involves the accuracy with which genomic digests are size-fractionated. If a band of interest migrates close to or across a zone of fractionation (e.g., the A-genome/sub-genome RFLPs of ~1.9 kb), inexact gel slicing and/or anomalous migration of DNA can place a target sequence into either one of the adjacent pools (e.g., the 2-3 kb pool of A1; Figure 4A) or possibly both pools (Figure 4C). For this reason, gel-fractionation should yield the best results when bands of interest are separated by a large physical distance on the gel (i.e., two or more size fractions).

The selective cloning approach appears to be less sensitive to the influence of sheared and partially-digested genomic DNA, as “contamination” in incorrect size-fractionated libraries has not been observed. This appears to be due, at least in part, to the low cloning efficiency of inserts from the heterogeneous size-fractionated DNA mixtures. Recalling that two different restriction enzymes are employed to generate the size-fractionated libraries (one to size-fractionate genomic DNA, the other to generate PstI inserts), three populations of products are expected: (1) clonable products with two PstI ends (like locus G1262); (2) products with a single PstI end and a second terminus created by the enzyme used for size-fractionation (e.g., HindIII); and (3) products with no PstI ends. While the first population may be ligated into the vector and sampled after transformation, the second population titrates away vector by ligating the single compatible PstI end to linearized vector. These opposing reactions yield an overall cloning efficiency that is low. The method does, however, provide an additional selective filter for sequences present in high
abundance, making low frequency homoeologue "contaminants" rare. A practical
consideration dictated by the low cloning efficiency is that the cloning vector chosen should,
ideally, provide presence/absence selection through the use of a cell death gene (such as *ccdB*
in the vector pZero), as transformants containing re-circularized vector fail to grow.

The presence of repetitive DNA in size-fractionated *PstI* libraries is a second and
unavoidable weakness of the cloning approach. Preliminary surveys show that different size
fractions contain proportions of clonable (i.e., with two *PstI* ends) repetitive DNA that spans
the full range from a minority to the majority of the DNA in a size fraction (data not
shown).

The loci corresponding to G1262 account for less than a millionth of the DNA
present in allotetraploid cotton (~2 kb out of 2.9 x 10^6 kb total; Michaelson et al. 1991;
Gomez et al. 1993; H. J. Price, pers. comm.). Isolation of both homoeologues from the
allotetraploid and the corresponding orthologues from the parental diploids, via a two-
enzyme, size-fractionation method, represents a significant enrichment relative to cloning
methods that sample at random. For example, to have a 90% probability of obtaining both
G1262 homoeologues in YAC or cosmid libraries (assuming insert sizes of 150 kb and 35 kb,
respectively), approximately 58,000 independent YAC or 250,000 cosmid clones would need
to be surveyed (Ausubel 1994). Despite the disadvantages of the selective cloning method,
we have successfully used the approach to clone one or both of the allotetraploid cotton
homoeologues for 10 different mapped *PstI* loci (unpublished data).

The methods described in this paper were expressly designed to isolate homoeologous
loci from allopolyploid genomes, although they clearly are applicable to the isolation of
target orthologues from diploids as well. Thus, the methods provide a general tool for
isolating a specific nuclear sequence from a suite of diploid and polyploid taxa, as well as a
means for accessing divergent members of a given gene family from a single taxon. In the
arena of genome mapping, an important assumption of STS-PCR analyses is that PCR
amplification products are orthologous to the locus originally mapped (Olson et al. 1989).
While this assumption may be warranted for some loci, especially in diploids with small
genome sizes (Inoue et al. 1994), it appears to be less reliable in diploids with larger genomes.
In a survey of 99 barley-specific STS-primers, for example, Blake et al. (1996) found that
amplification products from 36 primer pairs did not correspond to the loci originally
identified in the RFLP map. This illustrates how frequently un-mapped paralogues of
sequence-tagged sites may be amplified if total genomic DNA is used as the PCR template.
As demonstrated here, the use of size-fractionated DNA obviates this problem in many cases,
thereby increasing the accuracy of STS-amplifications.

Finally, we note that the ability to target specific nuclear sequences may find utility in evolutionary and phylogenetic analyses. The use of nuclear sequences for phylogenetic purposes has frequently been hampered by intractable problems associated with paralogy/orthology conflation. Isolation of strict orthologues is an essential prerequisite to the use of nuclear sequences in phylogeny reconstruction; the methods detailed here should prove useful in this regard.

ACKNOWLEDGEMENTS

The authors thank C. L. Brubaker for sharing comparative mapping results prior to publication, A. Paterson and X. Zhao for assistance in probe selection and J. Ryburn for technical assistance. This research was supported by the National Science Foundation.

REFERENCES


Figure 1: Organismal and genome relationships of diploid and allopolyploid taxa in the genus *Gossypium*. *Left:* Diploid progenitors of the extant allotetraploids diverged approximately 6-11 million years before present (mybp), giving rise to two lineages represented by the A-genome diploids in Africa and D-genome diploids in the New World. Allopolyploid cotton formed 1-2 million years ago in the New World, when A- and D-genomes were combined in a common nucleus (Endrizzi et al., 1985; Wendel, 1989). *Right:* Types of homology among related sequences within and between diploids and their derived allopolyploids. For simplicity, only a single chromosome having three “similar” sequences (clear, grey, and black fill, respectively) is shown for the diploids; both homoeologues are illustrated in the allopolyploid. At the diploid level, loci with the same fill pattern are orthologues, reflecting the speciation event that created the opportunity for genic divergence. Genic relationships reflecting past duplication events rather than speciation events are paralogous, as exemplified by all comparisons among loci having different fill colors, be they within or between genomes. When orthologues become united in a common nucleus as a consequence of allopolyploid formation, homoeologous relationships are established, as illustrated for three sets of orthologues (“clear”, “grey”, “black”) in this example.
Figure 2: Evidence for homoeology and orthology relationships of loci revealed by probe G1262. Panel A: Southern hybridization of HindIII and PstI-digested genomic DNA from diploid and allotetraploid cottons using probe G1262. Hybridization reveals a single band in the A- and D-genome diploids while two bands are evident in the AD-allotetraploid, one from each progenitor genome. Digestion of AD-allotetraploid DNA with PstI generates a single band of 1 kb, equal in size to the original mapping probe G1262. Abbreviations of taxa: G. raimondii = D5; G. hirstutum = AD1; G. herbaceum = A1. Size marker (M) is lambda DNA digested with Accl and HindIII. Panel B: Comparative linkage mapping supports orthology of locus G1262 in the two diploid genomes (linkage groups D10 and A3 in the D-genome and A-genome, respectively) and homoeology of the corresponding loci on two syntenic chromosomes (25D and 6A) in the allotetraploid genomes (adapted from Brubaker et al. 1998).
Figure 3: Colony hybridization of a single mini-library from the 4-5 kb HindIII pool from *G. hirsutum*. Panel A: Hybridization with probe G1262 reveals at least three positive clones that contain the desired insert. Panel B: Hybridization with labeled 4-5 kb HindIII-digested DNA from *G. hirsutum* reveals approximately 80 clones that contain highly-repetitive DNA. Panel C: Hybridization with labeled pZero vector reveals approximately 800 colonies.
Figure 4: Agarose gel analysis of G1262 PCR-amplification products using HindIII size-fractionated DNA as a template for amplification. Panel A: Amplification products from the 1-8 kb HindIII size-fractions (lanes 1 through 7) of the A-genome diploid G. herbaceum; product is only amplified from the 2-3 kb pool (lane 2). Panel B: Amplification products from the 1-8 kb HindIII size-fractions (lanes 1 through 7) of the D-genome diploid G. raimondii; amplification is apparent only in the 4-5 kb pool (lane 4). Panel C: Amplification products from the 1-8 kb HindIII size-fractions (lanes 1 through 7) of allotetraploid G. hirsutum; products are apparent in lanes corresponding to 1-2, 2-3, 4-5, 5-6, and 6-7 kb pools. Panel D: Homogeneous DNA sequences obtained from G1262 amplification products isolated from the G. hirsutum 1-2 kb (left) and 4-5 kb (right) HindIII size-fractionated DNA pools (from panel C). The arrow indicates a region of locus G1262 where the sequences from the A- and D-subgenomes have diverged, showing the homogeneity of amplification products that result from size-fractionation.
CHAPTER FIVE

MODE AND TEMPO OF SEQUENCE EVOLUTION AT HOMOELOGOUS LOCI FROM THE COTTON GENOME

A paper prepared for submission to Molecular Biology and Evolution

Richard C. Cronn and Jonathan F. Wendel

ABSTRACT

Duplicated genes created during polyploid formation (homoeologues) may experience a variety of fates depending upon the evolutionary forces operating on these loci. Homoeologue divergence may be limited if selection operates to maintain duplicate gene function, or divergence may be permitted if selective pressure on a functionally redundant locus is relaxed. In an attempt to determine the fate of duplicate loci in a polyploid genome, we isolated 14 sets of mapped, homoeologous loci from AD-genome allotetraploid cotton (Gossypium hirsutum L.) and the corresponding orthologues from diploid species representing A- and D-genome donors. In combination, these results demonstrate that low-copy homoeologous loci in the allopolyploid genome evolve independently of one another. In addition, relaxation of selective pressure due to genetic redundancy (as indicated by an increase in the substitution rate of one homoeologue) at these duplicate loci subsequent to polyploidization appears minimal across the majority of loci sampled. Ten loci (which correspond to mapped anonymous PstI-genomic probes) show rate equivalency in pairwise comparisons between polyploid subgenomes and between subgenomes and their progenitor diploid genomes, whereas four loci show significant rate differences in at least one comparison. Among the four loci showing rate inequivalence are two cellulose synthase genes (CelAl and CelA2) which show significantly elevated substitution rates in the D- and A-subgenome lineages, respectively. In particular, CelA2 from the A-subgenome of G. hirsutum has experienced a marked rate acceleration since polyploidization, and has accumulated a greater than expected number of non-synonymous substitutions without exhibiting the hallmarks of pseudogenization. These results indicate that cellulose synthase A2 may be responding to directional or diversifying selection, perhaps as a consequence of human-mediated selection upon fiber quality attributes.
INTRODUCTION

Polyploid evolution is a prominent process in plants, accounting for perhaps 70% of all angiosperms (Masterson, 1994). During allopolyploid speciation, two diverging diploid genomes are united in a single nucleus, generating genetic “redundancy” on a genome-wide scale. This redundancy is central to polyploid speciation theory in that duplication may lead to a relaxation of selection on one gene copy, allowing divergence and the acquisition of novel function (Ohno, 1970; Kimura and Ohta, 1974). This theory, while reasonable, remains largely untested and unsupported by empirical evidence. Indeed, acquisition of new function is one of several possible evolutionary fates for duplicated genes (Li, 1983; Li et al., 1985), others being loss of function via mutation (Li, 1980, 1982; Ferris and Whitt, 1977; Wilson et al., 1983; McGrath et al., 1994) and maintenance of function, either through natural selection (Hughes and Hughes, 1993; Hughes, 1994) or concerted evolution (Arnheim et al., 1980; Zimmer et al., 1980; Dover, 1982; Meagher et al., 1989; Drouin and Dover, 1990; Wendel et al., 1995).

A long-standing problem in determining the molecular evolutionary fate of loci duplicated via polyploidy has been how to demonstrate that two sequences are related by polyploidization (homoeology) rather than a more ancient duplication event (paralogy; Fitch, 1970; Doyle, 1991, 1992; VanderWiel et al., 1993). This consideration is complicated by the fact that most plant nuclear genomes are extraordinarily redundant (Cavalier-Smith, 1985; Tanksley and Pichersky, 1988). Numerous RFLP studies have demonstrated that most gene and genomic clones, even those pre-selected to be “single-copy”, detect more than a single DNA fragment when they are used as hybridization probes against genomic DNAs (e.g., Helentjaris et al., 1988; Brubaker and Wendel, 1994; Reinisch et al., 1994). Since plant genomes can experience a complicated history of sequence duplication and sequence loss in the absence of ploidy-level changes (via genomic and/or chromosomal deletion or duplication mechanisms), sequence similarity per se may be an unreliable indicator of orthology.

Although orthologous relationships within individual gene systems have occasionally been demonstrated using other criteria, such as tissue specificity (Doyle, 1991), comparative RFLP mapping provides the most convincing evidence for orthological relationships, namely, localization of putative homoeologues to duplicated, syntenic linkage groups in one or more taxa (Bonierbale et al., 1988; Tanksley and Pichersky, 1988; McGrath and Quiros, 1991; Kowalski et al., 1994; Reinisch et al., 1994; Brubaker et al., 1998). The combination of sequence similarity with shared genomic location constitutes compelling evidence that hybridizing sequences are actually homoeologous, as opposed to having a more distant, paralogous relationship.
In an effort to gain insight into processes of divergence of homoeologous loci in allopolyploids, we have initiated studies in a model system involving one allotetraploid species of AD-genome cotton (*Gossypium hirsutum*) and species that represent A-genome (*G. herbaceum*) and D-genome (*G. raimondii*) diploid progenitors. These taxa are well-suited for addressing issues of homoeologous sequence divergence since the phylogenetic relationships among these genomic groups have been well-characterized (Wendel, 1989; Wendel and Albert, 1992; Seelanan et al., 1997; Figure 1). In addition, a wealth of information exists on cotton cytogenetics and chromosomal morphology (reviewed in Endrizzi et al., 1985). Both A- and D-genome diploids have a gametic complement of 13 chromosomes, but the latter posses only two-thirds of the DNA content per cell as the former. Allopolyploid species are additive in this respect (Arumunganathan and Earle, 1991; Geever et al., 1989; Michaelson et al., 1991), and exhibit true disomic inheritance. Finally, detailed linkage maps have been constructed for all three (A, D, AD) genomic groups (Reinisch et al., 1994; Brubaker et al., 1998). This map information, derived from F2 segregation analysis and Southern hybridization data for several hundred low-copy number cDNA and genomic PstI probes, facilitates the isolation and sequencing of orthologous sequences from the A- and D-genome diploids and their counterparts in the allopolyploids.

In this paper, we describe the evolution of 14 homoeologous low-copy number nuclear loci from AD-genome allotetraploid cotton and their orthologues in A- and D-genomes from model progenitor taxa. Specifically, genomic regions corresponding to 12 anonymous, mapped PstI-genomic probes and two known genes that encode cellulose synthase (*CelAl* and *CelA2*) have been isolated and sequenced from both subgenomes (A, and D,) of the allopolyploid *G. hirsutum*. For each of these loci, orthologues have also been isolated and sequenced from the A-genome (*G. herbaceum*) and D-genome (*G. raimondii*) diploid progenitors, and a single C-genome outgroup taxon (*G. robinsonii*). Thus, for each locus, sequences have been isolated from five different genomes; three from diploid genomes (A, D, and the outgroup C) and two from the homoeologous subgenomes (A, and D,) of the allopolyploid. Sequence data from each of these loci provide an independent test of two null hypotheses concerning allopolyploid genome evolution: (1) Homoeologous genes have evolved independently following allopolyploid formation; (2) Rates of molecular evolution are equivalent in diploids and allopolyploids.

**MATERIALS AND METHODS**

**Plant materials** - The taxa selected for this study include the New World allotetraploid *Gossypium hirsutum* race Palmeri (genome designation AD), the African diploid *G. herbaceum* (A), the Peruvian diploid *G. raimondii* (D), and a diploid Australian taxon (*G.
robinsonii; C_{2} included as an outgroup. DNAs from each of these plants were isolated using the method of Paterson et al. (1993). RFLP mapping and identification of homoeologous linkage groups are described in detail in Reinisch et al. (1994) and Brubaker et al. (1998).

**Locus Selection, Isolation and Sequencing** - Twelve low-copy loci revealed by PstI genomic probes (Brubaker and Wendel, 1994) and two cellulose synthase genes (CelA; Pear et al., 1997) were selected for inclusion in this study (Table 1). Loci were selected on the basis of RFLP probe size (with preference given to those under 1.5 kb), and RFLP pattern: loci were typically included only if hybridizing bands from the subgenomes of the allotetraploid (At and Dt) differed by approximately 1-2 kb in size, and if the RFLP bands were larger than the probes used to reveal them. Genomic DNA from diploids and the allotetraploid were then digested by EcoRI or HindIII, fractionated on agarose, and then used as PCR templates or sources for cloning (detailed in Cronn and Wendel, 1998). This approach separates homoeologues (and divergent paralogues) from the allotetraploid into different pools, allowing them to be isolated in the absence of contaminating loci. Genes encoding cellulose synthase (CelA1 and CelA2; Pear et al., 1997) were isolated by PCR amplification using genomic DNA from each of the four taxa as a template. These PCR products were sequenced directly from diploid taxa, and from cloned PCR products (using the vector pGEM-T; Promega) in the allopolyploid subgenomes. All PCR primers were designed from mapping probe sequences (in the case of loci revealed by PstI-clones) or cDNA sequences (in the case of CelA1 and CelA2); primer sequences are summarized in Table 1. Orthological relationships of a subset of these PCR products (G1262, A1751, A1713, A1341, CelA1 and CelA2) were re-confirmed by linkage mapping using DNAs isolated from the same diploid F2 individuals that were used in constructing comparative RFLP linkage maps (Reinisch et al., 1994; Brubaker et al., 1998).

**Data analysis** - The analytical framework used in this study is shown in Fig. 1. Two null hypotheses were tested: (1) Homoeologous loci have evolved independently of each other subsequent to allopolyploidization; and (2) Rates of sequence evolution are equivalent in diploids and polyploids. Once sequences were obtained, they were aligned by eye and subjected to phylogenetic analysis using a parsimony framework and PAUP v. 3.1.1 (Swofford, 1993). If sequences evolve independently (satisfying null hypothesis 1) and at equivalent rates among all lineages (satisfying null hypothesis 2), a topology identical to the true organismal/genome phylogeny should be obtained (Fig. 1, scenario 1). If sequences evolve at significantly different rates in one (or more) lineages, a topology similar to that shown in scenario 2 of Fig. 1 will be revealed. Non-independent evolution of homoeologues
should lead to deviation from the expected topology, as is shown in scenario 3 of Fig. 1, and a rejection of the first null hypothesis.

To evaluate rate equivalency between subgenomes and diploid progenitors, we used the chi-square method of Tajima (1993) to test for differences in nucleotide substitution rates between pairs of taxa relative to an outgroup. In most cases, *G. robinsonii* was used as the outgroup sequence, although sequences similar to two loci could either not be isolated (A1834) or confidently aligned (A1623) from this taxon. In these instances, the alternative diploid genome was used as an outgroup to compare allotetraploid and diploid orthologue divergences. This $\chi^2$ test was performed on full length sequences individually for all loci, and on combined sequences from all loci in a single data set. Cellulose synthase genes were additionally partitioned into synonymous ($K_s$) and non-synonymous ($K_a$) nucleotide positions as indicated by the computer program SITES (Hey, 1996), and rates were evaluated on these subsets as well.

**RESULTS**

*Relative sequence divergence between diploid genomes of cotton* - Pairwise sequence comparisons at 14 nuclear loci from the A-genome diploid *G. herbaceum* and the D-genome diploid *G. raimondii* yielded overall Jukes-Cantor distances ($K_e$) spanning a 3.6-fold range, with a low value of 0.9% observed at locus A1520 and a high value of 3.2% at A1341 (Table 2). Of the 14 loci examined, three loci (A1520, A1751 and G1134) exhibited rates of nucleotide divergence at 1.2% or less, indicating that they evolve at rates comparable to that observed in the chloroplast gene *ndhF* ($K_e = 1.2\%$ divergence between $A_1$ and $D_1$; Seelanant et al., 1997). Of the remaining eleven loci, six (A1286, A1623, A1713, G1121, G1262 and *CelA1*) showed rates of divergence between 1.4% and 2.0%, a range comparable to that reported for nuclear *AdhC* from these same taxa ($K_e = 1.7\%$ between $A_1$ and $D_1$; Small et al., 1998). The five remaining loci (A1341, A1550, A1625, A1834 and *CelA2*) show divergences greater than 2.0%, which is approximately twice that observed at *ndhF*.

In order to estimate an average overall rate of divergence, we aligned sequences from the 14 loci obtained from $A_1$ and $D_1$ and calculated $K_e$ from all 10,130 nucleotides in the data set. The overall rate of nucleotide divergence between these diploid taxa is calculated to be 2.0%, a value similar to that obtained for known nuclear genes such as *AdhC* (Small et al., 1998). Interestingly, this value is intermediate to the non-synonymous ($K_a = 1.0\%$) and synonymous ($K_s = 3.0\%$) rates for *CelA1*, and is approximately equal to the non-synonymous rate for *CelA2* ($K_a = 1.8\%$). These comparisons indicate that, while the majority of loci evaluated in this study are anonymous, the relatively low-level of nucleotide divergences between diploid orthologues are indicative of "gene-like" rates of divergence, where
"genes" are composed of intron and exon nucleotides. These values provide a dramatic contrast to other types of "orthologous" nuclear sequences which have been described from these same taxa, particularly those sequences which encode the spacers of ribosomal RNA genes. Sequences from the internal transcribed spacer of the 45S rRNA genes (Wendel et al., 1995) showed a distance of ~ 8.4% between A1 and D1, while rates of divergence in 5S rDNA non-transcribed spacers were dramatically higher at 28% for these same taxa (Cronn et al., 1996).

**Absolute rates of divergence between diploid orthologues** - Exact estimates for the date of diploid cotton radiation have been debated since *Gossypium* (and other members of the Malvaceae) is poorly represented in the fossil record and because the genus has representatives native to all continents except Europe and Antarctica (Fryxell, 1979). Using molecular clock assumptions and cpDNA sequence divergence values based upon restriction-site surveys (Wendel 1989; Wendel and Albert, 1992) or direct nucleotide sequencing (Seelanan et al., 1997; Small et al., 1998), several estimates place the divergence of the A- and D-genome diploid lineages between 5 - 11 million years ago (MYA). By using this range of divergence times as lower and upper boundaries, estimates of the minimum and maximum rates of absolute divergence per nucleotide can be calculated. Using an assumed divergence time of 11 MYA (which yields the lowest rate) or 5 MYA (which yields the highest rate), overall rates of divergence between A- and D-genome cottons range between $0.4 \times 10^4$ to $0.9 \times 10^4$ substitutions/site/year at locus A1520, to a high of $1.4 \times 10^4$ to $3.2 \times 10^4$ substitutions/site/year at locus A1341 (Table 3). For the two known genes included in this study, nonsynonymous substitution rates ranged from $0.5 \times 10^4$ to $1.0 \times 10^4$ substitutions/site/year at locus *CelA1* and from $0.8 \times 10^4$ to $1.8 \times 10^4$ substitutions/site/year at locus *CelA2*. Synonymous substitution rates (which include intron nucleotides) at these loci were approximately twice as high, ranging from $0.9 \times 10^4$ to $1.9 \times 10^4$ substitutions/site/year at locus *CelA1* and from $2.0 \times 10^4$ to $4.3 \times 10^4$ substitutions/site/year at locus *CelA2*.

**Sequence divergence between tetraploid loci and their diploid orthologues** - Across all 14 loci, sequences from the allotetraploid subgenomes of *G. hirsutum* showed greater similarity to their putative diploid progenitor genome than they did to one another, e.g., A1 was more similar to A2 than D1, and D1 was more similar to D2 than A1 (Table 2). Jukes-Cantor distances ($K_2$) between diploid - tetraploid orthologues ranged from a low of 0.0% for locus G1134 (A1 and A2 were identical) and locus A1751 (D1 and D2 were identical), to a high of 2.0% for locus A1623 (between D1 and D2). In general, orthologues residing in the A1 and A2 genomes showed lower levels of divergence than did the D1 and D2 orthologues. This trend is most
easily illustrated by aligning sequences from all 14 loci for these taxa and evaluating nucleotide divergence across 10,130 bases; overall divergence between the diploid A-genome and allotetraploid A-subgenome is 0.5%, while divergence between the diploid D-genome and the allotetraploid D-subgenome is 0.7% (Table 2). Since the A, and D, genomes were combined in a common genome at a discrete point in time, differences between these values indicate that there has either been an acceleration in divergence rates between D, and D, (relative to A, - A, divergence) at these loci subsequent to polyploidization, or that the D-genome donor for the tetraploid was slightly more divergent from present-day G. raimondii than the A-genome donor was to present-day G. herbaceum.

**Patterns of nucleotide change between diploid orthologues and tetraploid homoeologues** - Figure 2 shows the topologies obtained by cladistic parsimony analysis of 14 mapped sets of loci obtained from the diploid progenitor genomes (e.g., A,, D,) and their corresponding allotetraploid subgenomes (A,, D,). While branch lengths vary among lineages and loci, the topologies are identical to the known organismal phylogeny (Fig. 1, scenario 1). Only one locus, A1520, shows a tendency towards deviation from the expected topology since the A-subgenome sequence can be placed sister to either A, or the D-genome/D-subgenome lineage without a change in tree length. This equivocal interrelationship appears to be due to inherently low level of character evolution in the A-genome (A, and A,) lineage at locus A1520. This example highlights an important consideration; that character evolution may be insufficient to reveal underlying trends of inter-subgenomic interaction in the allotetraploid sequences. To investigate the likelihood that subtle trends were going undetected, we analyzed the 10,130 nucleotides from all 14 loci of these genomes (A,, A,, D,, D, and C,) using cladistic parsimony. Results of this analysis (shown as "all loci" in Fig. 2) mirror those observed across all 14 loci, as tetraploid subgenome sequences are sister to their respective diploid genome donors (Fig. 2). Based upon the topologies obtained from these 14 loci individually and combined (all of which fail to refute the null topology), interaction between homoeologues which yields a change in cladistic topology is not evident.

**Relative rate tests between diploid orthologue and tetraploid homoeologue divergence** - As illustrated in Table 2 and Fig. 2, a number of loci show the potential for rate inequality between diploid and polyploid genomes. Differences in rate of sequence divergence between diploid genomes and allopolyploid subgenomes can arise from: (i) changes in the rate of divergence between diploid progenitor lineages prior to polyploid formation; (ii) changes in the rate of divergence among one or more lineages after allopolyploidization; and (iii) some combination of both (i) and (ii). To evaluate the potential of rate inequality at each one of
these steps, we used the $\chi^2$ rate test of Tajima (1993) to evaluate all pairwise comparisons between 14 loci isolated from the A$_1$ and D$_3$ diploids, and the A$_1$ and D$_3$ subgenomes of the allotetraploid relative to the outgroup taxon C$_3$ (Table 2). While a number of these comparisons show apparent differences in branch lengths upon cladistic parsimony analysis (Fig. 2), only seven of the 76 pairwise comparisons between these sequences proved statistically significant ($P < 0.05$). Rate inequivalence was observed in two instances at the diploid level, as D$_3$ diploid sequences from loci A1550 and CelA1 evolved more rapidly than the A$_1$ diploid orthologue. Lack of rate equivalence between model diploid progenitors and their corresponding subgenome was also detected for the locus encoding A1713, as sequences from the allotetraploid D$_3$ genome had evolved more rapidly than sequences from the D$_3$ diploid. Finally, comparisons between the A$_1$ and D$_3$ subgenomes revealed that the loci encoding CelA1 had evolved more rapidly in D$_3$ than in A$_1$, and that CelA2 was evolving more rapidly in A$_1$ than in D$_3$.

In the parsimony tree derived from the combined sequence data set (“all loci” in Fig. 2), we noticed that the terminal branches leading to the tetraploid subgenomes (A$_1$ and D$_3$) appeared approximately one-third longer than those leading to diploid genomes (A$_1$ and D$_3$). We used the Tajima $\chi^2$ test to evaluate rate equivalence between the combined sequences (10,130 bp in length) of the diploid and polyploid genome lineages using C-genome outgroup sequences for comparison. Results of this test indicate that while allotetraploid subgenome branches appear longer (Fig. 2), rate differences are insignificant between the A$_1$ and A$_3$ lineages ($\chi^2 = 1.52; P = 0.22$ for 1 d.f.) and the D$_3$ and D$_3$ lineages ($\chi^2 = 1.03; P = 0.31$ for 1 d.f.) (Table 2). In order to evaluate whether the rate of change in polyploid lineages was greater than that observed in diploid lineages, we made a final comparison of all “Diploid” and “Tetraploid” sequences by combining the two diploid sequences into a single sequence ($A_1 + D_3 = \text{"diploid"}$) and by combining and the two tetraploid sequences into a single sequence ($A_1 + D_3 = \text{"tetraploid"}$). For outgroup comparisons, we doubled the C-genome outgroup sequence to yield an aligned sequence of comparable length. These pooled sequences had an aligned length of 20,260 bp, and were used to test for rate inequivalence between diploid and tetraploid genomes. Results from this test also indicate that while rates of divergence appear elevated in allotetraploid subgenomes relative to diploid genomes, the magnitude of this increase is insignificant as indicated by the test of Tajima ($\chi^2 = 2.46; P = 0.12$ for 1 d.f.).

**DISCUSSION**

The focus of this paper has been to describe nucleotide change at 14 nuclear-encoded, homoeologous loci from allopolyploid cotton (*Gossypium hirsutum*). By
incorporating two model genome donors (G. herbaceum [A-genome] and G. raimondii [D-genome]) and an additional outgroup taxon (G. robinsonii) in this study, nucleotide-level changes between A- and D-subgenome homoeologues can be evaluated in two ways. First, homoeologous sequences can be evaluated with respect to pattern of nucleotide divergence, since a genealogy reconstructed from any locus should be expected to mirror the known phylogeny of these taxa. Deviation from the expected pattern may be indicative of subgenome interaction, either through gene-conversion or recombination-mediated events. Second, homoeologous sequences can be evaluated with respect to rate of nucleotide divergence, not only between the homoeologous pair but also between the subgenomic sequences and their orthologues from the appropriate diploid genome donor. If all sequences diverge at equivalent rates, pairwise distances between the homoeologues (e.g., A, and D,) should be equivalent to pairwise distances between their corresponding diploid orthologues (A, and D,); similarly, pairwise comparisons between tetraploid and diploid orthologues (e.g., sequences from A, and A,) should be equivalent to the rates of divergence at orthologous loci in the alternative genome (e.g., D, and D,). Rate inequality in one or more lineages can arise at random, although certain patterns of substitution (e.g., an increase in non-synonymous substitutions in exon sequence without the introduction of frameshift mutations) may be indicative of a shift in the selective forces acting upon a given locus.

* Rates and patterns of nucleotide divergence at low-copy, homoeologous loci in polyploid cotton - One important conclusion of this work is that all low-copy, homoeologous loci evaluated in this study appear to evolve independently of each other, despite being held in the intimate confines of a single nucleus (Figure 2). This conclusion suggests that interaction between homoeologous loci residing on the different subgenomes of G. hirsutum has been infrequent over the 0.5 - 2.0 million years since polyploidization occurred (Wendel, 1989; Seelanam et al., 1987). These conclusions provide a dramatic contrast to results previously obtained from highly-repeated genes in these same taxa, as the internal transcribed spacer (ITS) region from 45S rDNA loci shows evidence of both inter-subgenomic homogenization and lineage-specific rate differences in allotetraploid cottons (Wendel et al., 1995). Despite this apparent contradiction, each of these results are robust. Evidence supporting independent evolution at low-copy, homoeologous sequences is provided by 14 independent tests (= loci) and over 10,000 nucleotides of information; conversely, evidence supporting interaction between high-copy homoeologous sequences is bolstered by the fact that all five allopolyploid species of Gossypium show evidence of complete concerted evolution at ITS regions. This discrepancy is most likely due to intrinsic differences in the behavior of low-copy or unique sequences and highly repetitive sequences, as the latter display a greater
propensity towards evolving in a "concerted" manner (Arnheim et al., 1980; Zimmer et al., 1980; Dover, 1982).

A second conclusion of this work is that nearly all low-copy, homoeologous loci evaluated in this study evolve at statistically equivalent rates to each other and to their diploid orthologues. Of the 76 rate tests summarized in Table 2, only seven tests from five sequence sets (A1550, A1713, CelA1, CelA2 and "all loci" combined) showed evidence of rate inequality. This result could be interpreted as evidence that loci duplicated via polyploidization are maintained by selective forces that operate to preserve the original function of the locus. It should be noted, however, that the 0.5 - 2.0 million years of divergence time which separates these polyploid subgenomes from their diploid progenitors may be insufficient for a significant number of substitutions to accumulate. If we simply assume that (i) polyploidization occurred 1 MYA and (ii) that these low-copy loci evolve at rates similar to those published for other plant nuclear genes (Wolfe et al., 1987; Gaut et al., 1996), we can expect approximately 4 to 60 substitutions per 1 kb of synonymous sites (e.g., non-coding sequence) and 0.6 to 2 substitutions per 1 kb of non-synonymous sites (e.g., exon sequence). Deviations in divergence rates between such small values will likely go undetected by most statistical tests, particularly if the sequences contain a large proportion of exon-encoding nucleotides.

*Rates and patterns of nucleotide divergence at cellulose synthase genes in diploid and polyploid cotton* - Considering the limited sequence divergence that should be expected between the relatively "young" subgenomes of *G. hirsutum*, it seems all the more remarkable that significant rate differences were observed at both of the two cellulose synthase loci, CelA1 and CelA2. At each of these loci, sequences from one tetraploid subgenome showed evidence of rate acceleration relative to the other sequences examined; in the case of CelA1, the D-subgenome sequence showed significant rate enhancement, while at CelA2 the A-subgenome sequence showed rate enhancement (Table 2). At both of these loci, synonymous changes only slightly outnumbered nonsynonymous changes, as ratios averaged 2:1 for CelA1 (1.9% versus 0.8% for the A, - D, comparison) to nearly 1.5:1 for CelA2 (4.3% versus 2.8% for the same comparison; Table 2). Hence, at these two cellulose synthase loci, synonymous (silent) substitutions appeared only 1.5 to 2 times more frequent than amino acid replacement substitutions. While this value is marginally lower than those reported for nuclear genes in murine rodents (~ 3:1; Hughes and Yeager, 1997), this pattern of substitution is considerably different than that observed in plant nuclear genes reported to date. For example, nuclear Adh loci from palms and grasses show synonymous to nonsynonymous replacement values in
excess of 6:1 (Gaut et al., 1996), while other plant nuclear genes show values ranging from ~8:1 (Pgi; Terauchi et al., 1997) to \( \geq 10:1 \) (gapC and phytochrome; Wolfe et al., 1987).

While these two cellulose synthase loci show similar patterns of synonymous-to-replacement substitutions, it is the A-subgenome sequence from CelA2 that appears to be undergoing the most dramatic change. Rates of A\(_{i}\) and D\(_{i}\) homoeologue divergence are twice as high in CelA2 as in CelA1, whether measured as overall rates (3.7% vs. 1.6%), synonymous rates (4.3% vs. 1.9%) or non-synonymous rates (2.8% vs. 0.8%)(Table 2). Despite the acceleration in the rate of divergence at CelA2 locus in the A\(_{i}\) genome, the inferred coding sequence is free of insertions and deletions, providing an indication that the correct reading frame has been maintained despite the large number of nucleotide changes. Similarly, intron/exon splice junctions are maintained in CelA2 sequence from the A\(_{i}\) genome. These data hint at a possible role for diversifying and/or directional selection in CelA2 divergence, since changes in cellulose synthase activity may be responsible for variation in the quantity or quality of seed coat fiber and lint. In this context, it is interesting to note that A-genome diploid cottons have a long history of cultivation and are the only diploid species in the genus which possess spinnable seed fibers (Wendel et al., 1989). It is likely that this trait - development of seed lint coat - evolved once in an early ancestor of the A-genome lineage, and that it has been enhanced in AD-genome allotetraploids like Upland Cotton through selection (Saunders, 1961). Such a process – gain of unique function, combined with intense agronomic selection – could yield distorted synonymous:nonsynonymous substitution patterns and elevated rates of nucleotide divergence like that observed at CelA2, although these data are far too preliminary to adequately evaluate these hypotheses. Clearly, however, a greater effort should be given to characterizing the rates and patterns of sequence divergence of these “fiber-specific” genes, particularly within a phylogenetic context as has been used in this study.

Rates of polyploid subgenome-diploid genome divergence and the time of allopolyploid cotton formation - Estimates for the date of polyploid cotton formation vary from the mid-Pleistocene (0.5 - 2 million years ago [MYA]; Wendel, 1989; Wendel and Albert, 1992; Seelanan et al., 1998) to as distantly as \(- 10 - 20 \) MYA (Endrizzi et al., 1989). Similarly, estimates of A- and D-genome diploid divergence range from 6 - 11 MYA (Wendel et al., 1989), to an ancient value of 100 MYA (Endrizzi et al., 1989). The large discrepancies in these estimates are largely due to the fact that *Gossypium* (like other members in the Malvaceae) are poorly represented in the fossil record, and that the diploid genomes (A, B, C, D, E, F, G and K) show dramatic allopatry, with representatives native to all continents except Europe and Antarctica (Fryxell, 1979). Consequently, estimates of divergence tend to reflect
how various authors treat biographical disjunctions in light of other data, such as nucleotide divergence estimates from nuclear (Endrizzi et al., 1989) or chloroplast (Wendel, 1989; Seelanan et al., 1997) sequences.

Estimates defining the upper bounds (e.g., ~10 - 20 MYA for polyploid formation and 100 MYA for diploid divergence) were provided by Endrizzi et al. (1989) and are based upon the assumption that A- and D-genome diploids were separated by continental drift. The separation of the African and South American continents (at ~ 100 MYA) were then used to "calibrate" nuclear sequence divergence data, which were then used to estimate the date of polyploid formation. Using thermal stability (ΔTm) measurements of nucleotide divergence at low copy DNA, Geever et al. (1989) showed the diploids *G. herbaceum* and *G. raimondii* to be approximately 6% divergent, yielding a rate of change of ~ 0.022 to 0.046% divergence per genome per million years. Since low-copy divergences for *G. herbaceum* and the A-subgenome of *G. hirsutum* were measured at ≤ 0.25%, the approximate time of A-genome/A-subgenome divergence was placed at 5.4 - 11.4 MYA. Similarly, pairwise divergences for *G. raimondii* and the D-subgenome of *G. hirsutum* were measured at 0.4%, yielding divergence dates of 8.7 - 18.2 MYA.

By contrast, Wendel (1989), Wendel and Albert (1992) and Seelanan et al. (1998) used cpDNA sequence divergence to estimate times of divergences for these genomes based upon molecular clock assumptions. Estimates of nucleotide divergences based upon restriction site differences (Wendel, 1989; Wendel and Albert, 1992) or direct nucleotide sequencing (Seelanan et al., 1998) yielded comparable values ranging from 1.1% for the A-D diploid genome divergence and 0.2% for the A-genome/A-subgenome divergence. Accordingly, divergence times were nearly identical in these three studies, with A- and D-genome diploid divergence estimated at 5 - 11 MYA and polyploid formation estimated at 0.5 - 2 MYA.

Results from the present study can be used to shed additional light on the relative time of formation for allopolyploid cotton. Based upon 10,130 nucleotides from 14 nuclear loci, we find the A- and D-genome diploids to differ at 184 nucleotide positions, giving an overall pairwise divergence of 1.95%. Comparisons of sequences from diploid genomes and their corresponding polyploid subgenome yield 41 substitutions (0.45% divergence) for the A, - A, comparison, and 60 substitutions (0.65% divergence) for the D, - D, comparison. This indicates that the relative ratio between diploid genome divergence and diploid genome/ allopolyploid subgenome divergence ranges from 4.3:1 to 3.0:1, with an average of ~ 3.6 to 1. Thus, if we assume that sequences within polyploid genomes evolve at identical rates as diploid progenitor genomes (an assumption supported by this work; Table 2), the time of divergence for the A- and D-genome lineages must be 3.6 times more distant that the time of
diploid/polyploid genome divergence. This value lies most closely to the value of 5:1 obtained by Wendel (1989) than the value of 9:1 obtained by Endrizzi et al. (1989).

While it is tempting to estimate absolute divergence times using this data set, such an exercise requires a number of assumptions regarding relative numbers of synonymous and nonsynonymous sites in the data set and the average evolutionary rates of these 14 loci. It may be most prudent to consider that the values we obtained for diploid genome/allopolyploid subgenome divergence are nearly identical to those obtained by Endrizzi et al. (1989). It is interesting to note that while Endrizzi et al. (1989) showed confidence in their calculations for rates of diploid divergence (at 0.03% per $10^8$ years), they noted that this rate was "markedly slower than that reported for other organisms." By comparison, measured rates of divergence from other organisms (Drosophila, Xenopus, birds, primates, rodents and ferns) yield values approximately 10 times higher (from 0.17% to 0.4% per $10^8$ years). These authors acknowledged this discrepancy, and then investigated times of divergence using the average rate of sequence divergence from humans (~0.2% per $10^8$ years). Using this molecular clock hypothesis, their estimate of diploid genome/allopolyploid genome divergence was reduced to 1.1 MYA, a value which is in perfect agreement with Wendel (1989) and Seelanan et al. (1997). If this value is accurate, our data set would place diploid genome divergence at 3.6 times this value, or ~ 4 million years ago.

ACKNOWLEDGMENTS
The authors wish to thank Randy Small for sharing his results on AdhC prior to publication, and Leigh Rasmussen and Julie Ryburn for competent technical assistance. This work was supported by the National Science Foundation.

LITERATURE CITED


Hey, J. 1996. SITES. Rutgers University, Piscataway, NJ.


Table 1: Low-copy loci isolated from tetraploid and diploid cottons and amplification primers used for isolation.

<table>
<thead>
<tr>
<th>Probe Name</th>
<th>Map Location</th>
<th>Aligned Size (bp)</th>
<th>Putative Identification</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1286</td>
<td>HA 10</td>
<td>291</td>
<td>transketolase precursor</td>
<td>GTACTCGCGTGAACATGCATGAAAC</td>
<td>ACTCTCTCTTTAGCATCAACAGCTTCACCA</td>
</tr>
<tr>
<td>A1341</td>
<td>HA 7A</td>
<td>765</td>
<td>unknown</td>
<td>GACTGCCTGAAATGACAGAAACGCTY</td>
<td>CACCTCAAAAAGTTATTGCGCGATGY</td>
</tr>
<tr>
<td>A1520</td>
<td>HA 6</td>
<td>1032</td>
<td>unknown</td>
<td>GGCCTGAAAACCCCCTAGTTATTATY</td>
<td>CAAAACGAGCAAAGCACTCCAAGA</td>
</tr>
<tr>
<td>A1550</td>
<td>HA 9</td>
<td>912</td>
<td>aldehyde dehydrogenase</td>
<td>CCAAGTCAAGGAAATGTCATTACAG</td>
<td>GGGATCAATGAGTGGATCACATT</td>
</tr>
<tr>
<td>A1623</td>
<td>HA 5</td>
<td>908</td>
<td>unknown</td>
<td>GTATGGAATGACACATGTCGAAT</td>
<td>CTTTATTACTGTTGTATAGGCTAG</td>
</tr>
<tr>
<td>A1625</td>
<td>HA 1</td>
<td>1118</td>
<td>unknown</td>
<td>GAGAGGAAAGTTTGTACAACATG</td>
<td>CATGGACACTGCTTACAAAGGTA</td>
</tr>
<tr>
<td>A1713</td>
<td>HA 6</td>
<td>702</td>
<td>unknown</td>
<td>GAGAGGAAAGTTTGTACAACATG</td>
<td>CATGGACACTGCTTACAAAGGTA</td>
</tr>
<tr>
<td>A1751</td>
<td>HA 10</td>
<td>841</td>
<td>subtilisin-like protease</td>
<td>GCGGAATTGCGTTATAGTAGA</td>
<td>GATGAGCTCTGCTCACCACAAAGG</td>
</tr>
<tr>
<td>A1834</td>
<td>HA 8b</td>
<td>930</td>
<td>a-mannosidase precursor</td>
<td>TCACTAAAGAACCCCATGGAAT</td>
<td>AGGUGACTCAAATGGCAACCA</td>
</tr>
<tr>
<td>G1121</td>
<td>HA 3</td>
<td>786</td>
<td>unknown</td>
<td>GTCGCTAGCTCATACATGAGACG</td>
<td>TCAACCTAGTGGAGACACTCTY</td>
</tr>
<tr>
<td>G1134</td>
<td>HA 8b</td>
<td>586</td>
<td>unknown</td>
<td>CAAGCAGGAAATGACATGGCTTC</td>
<td>GACACTGACAGAAAGCAACC</td>
</tr>
<tr>
<td>G1262</td>
<td>HA 9</td>
<td>974</td>
<td>P-glycoprotein</td>
<td>GGGGAGCAGGCAAAGCACCTCY</td>
<td>CCGAGGCTCATACTTTCAAGGTA</td>
</tr>
<tr>
<td>CelA1</td>
<td>(--)³</td>
<td>1051</td>
<td>cellulose synthase A1</td>
<td>GATGGAAATCGGAGGGTTTCCGTTGTCG</td>
<td>GGGAACGATCCAACACCCAGGA</td>
</tr>
<tr>
<td>CelA2</td>
<td>(--)³</td>
<td>1079</td>
<td>cellulose synthase A2</td>
<td>GATGGAAATCGGAGGGTTTCCGTTGTCG</td>
<td>GGGAACGATCCAACACCCAGGA</td>
</tr>
</tbody>
</table>

¹ PstI probe construction and nomenclature are described in Brubaker and Wendel (1994); CelA loci are described in Pear et al. (1996).
² From Brubaker et al., 1998.
³ Mapping of CelA1 and CelA2 from these genomes is still in progress.
Table 2: Rates of divergence between homoeologs (A, vs D), homoeologs and their corresponding loci in diploid progenitors (A, vs. A; D, vs. D), and between orthologs from diploid genomes (A, vs. D). The outgroup for all comparisons was the homologue from C, when available; when this sequences could not be isolated, the alternative diploid genome was used as the outgroup.

<table>
<thead>
<tr>
<th>Locus Name</th>
<th>Sites Evaluated</th>
<th>Aligned Length (bp)</th>
<th>Distance between sequence pairs and results of rate tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A, vs. A, D, vs. D, A, vs. D, A, vs. D,</td>
</tr>
<tr>
<td>A1286</td>
<td>Entire</td>
<td>290</td>
<td>0.003 (-) 0.007 (-) 0.014 (-) 0.014 (-)</td>
</tr>
<tr>
<td>A1341</td>
<td>Entire</td>
<td>602</td>
<td>0.002 (-) 0.008 (-) 0.032 (-) 0.028 (-)</td>
</tr>
<tr>
<td>A1520</td>
<td>Entire</td>
<td>831</td>
<td>0.001 (-) 0.007 (-) 0.009 (-) 0.011 (-)</td>
</tr>
<tr>
<td>A1550</td>
<td>Entire</td>
<td>912</td>
<td>0.005 (-) 0.014 (-) 0.024 (*) 0.023 (-)</td>
</tr>
<tr>
<td>A1623</td>
<td>Entire</td>
<td>827</td>
<td>0.005 (-) 0.020 (-) 0.017 (NA) 0.024 (NA)</td>
</tr>
<tr>
<td>A1625</td>
<td>Entire</td>
<td>800</td>
<td>0.005 (-) 0.011 (-) 0.025 (-) 0.032 (-)</td>
</tr>
<tr>
<td>A1713</td>
<td>Entire</td>
<td>615</td>
<td>0.006 (-) 0.006 (*) 0.019 (-) 0.016 (-)</td>
</tr>
<tr>
<td>A1751</td>
<td>Entire</td>
<td>809</td>
<td>0.002 (-) 0.000 (-) 0.012 (-) 0.015 (-)</td>
</tr>
<tr>
<td>A1834</td>
<td>Entire</td>
<td>942</td>
<td>0.003 (-) 0.005 (-) 0.022 (NA) 0.023 (NA)</td>
</tr>
<tr>
<td>G1121</td>
<td>Entire</td>
<td>628</td>
<td>0.008 (-) 0.004 (-) 0.019 (-) 0.016 (-)</td>
</tr>
<tr>
<td>G1134</td>
<td>Entire</td>
<td>510</td>
<td>0.000 (-) 0.002 (-) 0.011 (-) 0.013 (-)</td>
</tr>
<tr>
<td>G1262</td>
<td>Entire</td>
<td>895</td>
<td>0.003 (-) 0.004 (-) 0.019 (-) 0.020 (-)</td>
</tr>
<tr>
<td>CelA1</td>
<td>Entire</td>
<td>1019</td>
<td>0.003 (-) 0.011 (-) 0.017 (*) 0.016 (**)</td>
</tr>
<tr>
<td></td>
<td>synon</td>
<td>722</td>
<td>0.001 (-) 0.013 (-) 0.019 (-) 0.019 (**)</td>
</tr>
<tr>
<td></td>
<td>nonsynon</td>
<td>298</td>
<td>0.008 (-) 0.007 (-) 0.010 (-) 0.008 (-)</td>
</tr>
<tr>
<td>CelA2</td>
<td>Entire</td>
<td>946</td>
<td>0.012 (-) 0.004 (-) 0.030 (-) 0.037 (*)</td>
</tr>
<tr>
<td></td>
<td>synon</td>
<td>612</td>
<td>0.013 (-) 0.005 (-) 0.037 (-) 0.043 (-)</td>
</tr>
<tr>
<td></td>
<td>nonsynon</td>
<td>334</td>
<td>0.009 (-) 0.003 (-) 0.018 (-) 0.028 (-)</td>
</tr>
<tr>
<td>All Loci</td>
<td>Entire</td>
<td>10130</td>
<td>0.005 (-) 0.007 (-) 0.020 (-) 0.021 (*)</td>
</tr>
</tbody>
</table>

1 Pairwise distances were calculated using the method of Jukes and Cantor (1969).
2 Results of Tajima ID rate tests are indicated by the following symbols; "-", not significant; "*, p ≤ 0.05; "**, p ≤ 0.01.
3 These comparisons could not be evaluated since C-genome outgroup sequences were unavailable.
4 Loci lacking C-genome outgroup sequences (A1623 and A1834) were excluded from this summary.
Table 3: Estimated rates of nucleotide divergence per site per 10^9 years at 14 low-copy nuclear loci from *Gossypium herbaceum* and *G. raimondii*.

<table>
<thead>
<tr>
<th>Locus</th>
<th>$K_s$</th>
<th>Low Rate</th>
<th>High Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1286</td>
<td>0.014</td>
<td>0.6</td>
<td>1.4</td>
</tr>
<tr>
<td>A1341</td>
<td>0.032</td>
<td>1.4</td>
<td>3.2</td>
</tr>
<tr>
<td>A1520</td>
<td>0.009</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>A1550</td>
<td>0.024</td>
<td>1.1</td>
<td>2.4</td>
</tr>
<tr>
<td>A1623</td>
<td>0.017</td>
<td>0.8</td>
<td>1.7</td>
</tr>
<tr>
<td>A1625</td>
<td>0.025</td>
<td>1.1</td>
<td>2.5</td>
</tr>
<tr>
<td>A1713</td>
<td>0.019</td>
<td>0.9</td>
<td>1.9</td>
</tr>
<tr>
<td>A1751</td>
<td>0.012</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>A1834</td>
<td>0.022</td>
<td>1.0</td>
<td>2.2</td>
</tr>
<tr>
<td>G1121</td>
<td>0.019</td>
<td>0.9</td>
<td>1.9</td>
</tr>
<tr>
<td>G1134</td>
<td>0.011</td>
<td>0.5</td>
<td>1.1</td>
</tr>
<tr>
<td>G1262</td>
<td>0.019</td>
<td>0.9</td>
<td>1.9</td>
</tr>
<tr>
<td>CelA1</td>
<td>0.017</td>
<td>0.8</td>
<td>1.7</td>
</tr>
<tr>
<td>CelA2</td>
<td>0.030</td>
<td>1.4</td>
<td>3.0</td>
</tr>
</tbody>
</table>

* Overall rate = ($K_s \times 10^9$)/2T, where $T$ (divergence time) is 11 million years for calculating the lowest rate of divergence, and 5 million years for calculating the highest rate.
Figure 1. Null hypothesis for sequence evolution in allopolyploids. **Scenario 1:** Phylogenetic expectations of independence and equal rates of sequence evolution following allopolyploid formation. Shown are topological relationships between sequences from diploid progenitor genomes (A and D) and their orthologues (At and Dt) in derived allopolyploids. C-genome diploids serve as outgroups for testing both rate equivalence and independence. **Scenario 2:** An accelerated rate of sequence evolution in allopolyploids will generate longer branches leading to At and/or Dt than to A and D. **Scenario 3:** Concerted evolutionary forces may lead to non-independent sequence evolution following allopolyploidization. Illustrated is conversion of an A-subgenome orthologue to a D-subgenomic form.
Figure 2. Trees obtained from parsimony analysis of 14 low-copy loci in *Gossypium* genomes. Numerical values along branches indicate branch lengths, while values in boxes indicate the overall tree length ($L$) and the consistency index ($CI$).
CHAPTER SIX

GENERAL CONCLUSIONS

Investigations into the nature of polyploid subgenome evolution may entail a variety of approaches. First, it is possible to characterize subgenome divergence at the cytological level by observing the frequency and effectiveness of meiotic pairing between homoeologous chromosomes in natural allopolyploids, or between homoeologous chromosomes and related diploid chromosomes which have been combined into a common nucleus via interspecific hybridization. Second, a more detailed picture of subgenome divergence via chromosomal organization/re-organization can be gained through linkage mapping using molecular probes and, in some cases, in-situ hybridization. A third approach for evaluating homoeologue divergence entails the use of nucleotide sequences from homoeologous loci to directly measure the degree of subgenome sequence divergence. In combination, these three approaches may provide an accurate and relatively complete picture of the degree and pattern of subgenome change in allopolyploids.

The focus of this research has been to characterize nucleotide change at a number of nuclear-encoded, homoeologous loci from allopolyploid cotton (*Gossypium*). By including two model genome donors (*G. herbaceum* [A-genome] and *G. raimondii* [D-genome]) and an additional outgroup taxon (*G. robinsonii*) in this study, nucleotide-level changes between A- and D-subgenome homoeologues can be characterized in two important ways. First, divergence between a homoeologous pair of loci can be measured by their pattern of sequence divergence, since a genealogy reconstructed from any locus should be expected to mirror the known phylogeny of these taxa. Deviation from the expected pattern may be indicative of subgenome interaction, either through gene-conversion or recombination-mediated events. In addition, homoeologous loci can also be evaluated for their rates of nucleotide divergence, not only between the homoeologous pair but also between the homoeologous sequences and their orthologues from the appropriate diploid genome donor. Rate inequality in one or more lineages can arise as a consequence of random mutation, although certain patterns of substitution (e.g., an increase in non-synonymous substitutions in exon sequence without the introduction of frameshift mutations) may be indicative of a shift in the selective forces acting upon a given locus.

Results from this study indicate that interaction between the A- and D-subgenomes of *G. hirsutum* is rare, since no deviation was observed between the genealogies obtained from 5S rDNA or low-copy mapped loci, and the known phylogenetic relationships of the taxa.
examined. This general result contrasts with the report by Wendel et al. (1995) which demonstrated inter-subgenomic homogenization of ITS sequences in the highly-repeated 45S rDNA loci. The reasons for such dramatically different behavior are not immediately obvious from the data alone, although there are two observations that may be revealing in this regard. First, 5S rDNA arrays show considerable heterogeneity, indicating that concerted evolutionary forces act only "weakly" in removing variant repeats as they arise; conversely, Wendel et al. reported that there was little-to-no heterogeneity in ITS regions from diploid and allopolyploid cottons. Secondly, Crane et al. (1993) have shown that 45S arrays occupy four major subtelomeric locations in the tetraploids (two per subgenome), while 5S arrays occupy only two centromeric locations in the tetraploids. Assuming (i) that unequal crossing-over plays an important role in interlocus homogenization, and (ii) that unequal crossing-over near the centromere may lead to unbalanced and presumably inviable gametes, Wendel et al. (1995) suggested that maintenance of subgenomic identity in repetitive arrays may be more likely in centromeric, rather than telomeric, tandem arrays. This prediction may explain the differences observed in the relative "strength" of concerted evolution between the 5S and 45S rDNA loci, as well as their proclivity for intergenomic interaction.

Rate equivalency, on the other hand, was not observed at four of the fourteen low-copy loci evaluated in this study. Since a small number of significant differences should arise due to chance alone (e.g., 5 out of 100 using the $X^2$ test of Tajima), it would be prudent to consider these tests "significant" only when meaningful information is available concerning the identity (known or putative) of a locus so that rate changes can be considered along with their potential impact on either gene structure or a change in the gene product. In this light, it is interesting to consider the rate deviations observed in the two pairs of homoeologous cellulose synthase genes, $CelA1$ and $CelA2$. In both of these cases, one diploid genome lineage appears to have experienced a significant increase in the rate of sequence divergence: in the case of $CelA1$, the D-genome diverged more rapidly than the A-genome, and in $CelA2$ the A-genome diverged more rapidly than the D-genome. While these rate deviations were due primarily to an increase in the number of silent substitutions, the A-genome and A-subgenome lineages of $CelA2$ are characterized by an unusually large number of non-synonymous substitutions that lead to amino acid replacements. These data hint at a possible role for diversifying and/or directional selection in $CelA2$ divergence, since changes in cellulose synthase activity may be responsible for the greater fiber and lint yields observable in the A-genome diploid relative to the D-genome diploid. Clearly, a greater effort should be given to characterizing the rates and patterns of "fiber-specific" gene evolution, particularly within a phylogenetic framework as has been used in this study.
Perhaps the single most important conclusion reached during the course of this study is that duplicated, homoeologous loci in the allopolyploid genome of Upland Cotton continue to evolve independently and at rates equivalent to each other and to their respective diploid genome donors. If the anonymous loci sampled in this study were assumed to contain a high proportion of coding sequence (perhaps a valid assumption for the majority of these loci, considering their observed rates of divergence), there would be three general conclusions that could explain the results obtained:

(1) That the majority of loci duplicated via polyploidy are conserved at the sequence level (at least initially) by selective forces which operate to preserve the original function of the locus. This seems to be a likely scenario, since the majority of the anonymous loci used in this study have "gene-like" rates of divergence comparable to known genes. Clear exceptions to this trend are the two cellulose synthase genes, both of which may be experiencing accelerated rates of divergence as a consequence of a shift in selective forces.

(2) That polyploid Upland Cotton is too young (at 1-2 million years; Wendel, 1989) for rate differences to have accumulated between redundant loci, and that a significant number of loci may be currently diverging at different rates as a consequence of allopolyploidization. This is also a likely scenario, particularly when consideration is given to the published rates of nucleotide sequence divergence in known plant nuclear genes (such as Adh; Gaut et al., 1996). These range from $-2 \times 10^{-9}$ substitutions/site/year for synonymous sites (introns and third codon positions), to $0.3 - 0.9 \times 10^{-8}$ substitutions/site/year for non-synonymous sites. If we assume that polyploidization occurred 1 million years ago, a 1 kb sequence from either subgenome of the tetraploid will differ from its diploid progenitor by $-4$ to 14 nucleotide substitutions in synonymous sites, and $-0.6$ to 2 bp in nonsynonymous sites. Deviations in divergence rates between such values may go undetected by statistical rate tests, particularly if the sequences contain a high proportion of coding sequence.

(3) That the sample of loci used in this study is either insufficient in overall number, or sufficiently biased with respect to sampling that true trends of subgenome divergence in Upland Cotton cannot be detected. This conclusion is more difficult to evaluate, particularly as one considers the alternatives: How many more loci would yield an accurate (true) answer? What sampling method(s) would eliminate the bias? Answers to these questions need to be reinforced with the necessary assurances of homoeology (i.e., everything would need to be mapped), as well as considerations for the complexity of the cotton genome (i.e., what works in Arabidopsis may not work in Gossypium). At present, only one observation can be offered to address this potential concern: a brief survey of the 200 low-copy mapping probes used to map diploid and cotton genomes show a predominance of RFLP patterns similar to those...
revealed by the loci included in this study. More specifically, allotetraploid genomes typically contain two alloalleles that are identical in size to either one or both diploid progenitor genomes ("additivity", as was demonstrated for probe G1262 in chapter four). Only 3% (6 out of 200) of these survey autorads showed patterns which do not exhibit some degree of additivity; unfortunately, none were mapped to more than one linkage group in the comparative map, so homoeology relationships cannot be convincingly established. Nevertheless, rearrangements and/or intergenome interactions which might lead to changes in RFLP patterns appear rare, suggesting that the low-copy loci selected for study are not unusual at the RFLP level.

In order to verify the generality of the conclusions reached in this paper, it would be instructive to carry out similar studies as those described herein using a small number of known mapped genes from cotton, such as those encoding isozymes. This would have the advantage of allowing variation to be partitioned into suitable pools (exon versus intron, synonymous versus non-synonymous sites, etc.) so that topological and/or rate changes could be evaluated in an evolutionarily meaningful context. Similarly, the generality of these conclusions could also be evaluated in other allopolyploid taxa, particularly those that may have a more ancient date of origin. This would obviate issues concerning "recency of formation" and add greater statistical power to rate tests. It should be noted that as the age of allopolyploidization increases, the potential for finding suitable diploid progenitor taxa will likely decrease. This may prevent an accurate determination of "changes in topology" or "intergenome interaction," since there is no genome progenitor available for confirmation. However, for evaluating rates of subgenome change, the lack of diploid genome progenitors would not weaken arguments as long as evidence of homoeology was strong.

In an attempt to reveal how subgenomic sequences evolve within the confines of a single polyploid plant genome, I seem to have uncovered a small number of results which remain dwarfed by the remaining, nagging questions of polyploid genome evolution. In fact, one of the main products of this work seems to be additional questions, largely restatements and refinements of previously existing questions concerning polyploidy and polyploid evolution. Given the complexity, the staggering frequency, and the central importance of this process to flowering plant evolution, the contribution of these results (and their attending questions) somehow seems appropriate.
ACKNOWLEDGMENTS

As a new arrival in the field of cotton evolutionary genetics, I began this work by literally "standing on the shoulders of giants." The genetic and cytogenetic contributions of early investigators such as J. Beasley, J. Hutchinson, S. Harland and S. Stephens set a solid foundation for the past half-century of cotton geneticists to build upon. More recent geneticists and taxonomists such as C. Brabaker, M. Endrizzi, P. Fryxell, A. Paterson and J. Wendel have helped make cotton an accessible and well-characterized allopolyploid model. Both indirectly and directly, these pioneers have made this body of work possible, and they deserve both my highest praise and thanks.

A number of people have contributed to the more immediate success of this project. First and foremost on this list is Jonathan Wendel (Iowa State University), who had the both the vision and the grant writing skills to make this rather unwieldy project both tractable and attractive. My successes in the lab and my development as a scientist have been largely due to Jonathan, and I owe him thanks far beyond what can be repayed. I am also indebted to Curt Brubaker for his tireless and dogged determination in developing the PstI clone libraries and making the cotton comparative maps: Curt, you are a saint among men. Thanks are also in order for Andy Paterson and Xinping Zhao for providing the tetraploid maps and sharing cDNA clones in the early phase of this project. I also should thank my committee members for providing their assistance as needed and for helping me stay focused; included are Jim Colbert, Randy Shoemaker, Dan Voytas and Rob Wallace.

Engaging discussion makes for better science and a more interesting life. The following cast of friends and acquaintances helped in numerous ways, sometimes technically, most times socially, and always enjoyably: Peter Bretting, Curt Brubaker, Hugo Cota, Steve Dickie, Phil Dykema, Kristin Eibauer, Jennifer Lee Leonard, Donald Pratt, Leigh Rasmussen, Julie Ryburn, Andy Schnabel, Tosak Seelanan, Randy Small and Paul Wetzel. You all know who you are and what you've done, and I thank you for it.

Finally, there's Adriene who supported me through it all. The average life is too short to repay so much love and kindness.