Duplicated genes evolve independently after polyploid formation in cotton

Richard C. Cronn*, Randall L. Small†, and Jonathan F. Wendel*‡

*Department of Botany, Bessey Hall 353, Iowa State University, Ames, IA 50011; and †Department of Botany, the University of Tennessee, Knoxville, TN 37996-1100

Communicated by Major M. Goodman, North Carolina State University, Raleigh, NC, October 15, 1999 (received for review July 23, 1999)

Of the many processes that generate gene duplications, polyploidy is unique in that entire genomes are duplicated. This process has been important in the evolution of many eukaryotic groups, and it occurs with high frequency in plants. Recent evidence suggests that polyploidization may be accompanied by rapid genomic changes, but the evolutionary fate of discrete loci recently doubled by polyploidy (homoeologues) has not been studied. Here we use locus-specific isolation techniques with comparative mapping to characterize the evolution of homoeologous loci in allopolyploid cotton (*Gossypium hirsutum*) and in species representing its diploid progenitors. We isolated and sequenced 16 loci from both genomes of the allopolyploid, from both progenitor diploid genomes and appropriate outgroups. Phylogenetic analysis of the resulting 73.5 kb of sequence data demonstrated that for all 16 loci (14.7 kb/genome), the topology expected from organismal history was recovered. In contrast to observations involving repetitive DNAs in cotton, there was no evidence of interaction among duplicated genes in the allopolyploid. Polyploidy was not accompanied by an obvious increase in mutations indicative of pseudogene formation. Additionally, differences in rates of divergence among homoeologues in polyploids and orthologues in diploids were indistinguishable across loci, with significant rate deviation restricted to two putative pseudogenes. Our results indicate that most duplicated genes in allopolyploid cotton evolve independently of each other and at the same rate as those of their diploid progenitors. These indications of genic stasis accompanying polyploidization provide a sharp contrast to recent examples of rapid genomic evolution in allopolyploids.

Gene duplication is recognized as an important requirement for the diversification of gene function. The resulting genetic redundancy is thought to permit novel mutations to accumulate because of relaxation of selective constraints on one redundant copy (1–4). Of the many processes known to create gene duplications, polyploidy is unique in that *entire genomes* become duplicated. This process of creating "genome equivalents" of genetic redundancy may be one of the most important features of polyploidy, because no other known mechanism can provide a comparable increase of genetic material on which selection may act. Indeed, the present genomic and biochemical complexity in higher plants, animals, and fungi may result in part from the gain of new genes and gene functions after genome duplications (1, 3–14).

Although the creation of new gene function is one important result of gene duplication, the frequency of this particular outcome is not known. Other possibilities include silencing of one of the duplicated ("homoeologous") copies (15–19), molecular interaction mediated by concerted evolutionary processes (20, 21), and long-term evolutionary maintenance of both duplicated copies (14, 22, 23). Growing evidence indicates that gene silencing rates are surprisingly low (4, 23), and that maintenance of gene function is a common fate for homoeologous genes. This conclusion is supported mostly by studies of genes duplicated by rounds of polyploidy in the relatively distant past, e.g., in *Xenopus* [≈30 million years ago (mya); (22)], catostomid and salmonid fishes [≈50–100 mya; (15, 18)], and tetrapod vertebrates [≈250 mya; (14)].

Although the foregoing studies describe the *ultimate* fate of genes duplicated via polyploidization, the tempo and pattern of duplicate gene and genome divergence subsequent to polyploid formation remains incompletely understood. For example, there is an apparent contradiction between long-term gene maintenance in older polyploids (as described above) and a growing body of evidence indicating that in younger allopolyploids, there may be rapid sequence conversion of homoeologous loci [as shown for rDNA (20, 21, 24)], homoeologue-specific sequence elimination (25, 26), and extensive genomic rearrangements (27). The latter studies reveal phenomena that may be important in stabilization of newly formed polyploid genomes, all of which violate null expectations (Fig. 1) of additivity, independence, and evolutionary rate equivalence for duplicated factors. Evidence of genic interactions accompanying the early stages of polyploid stabilization may become obscured in older polyploids by subsequent evolutionary change. In addition, with increasing time since polyploidization, inferences become more tenuous that any pair of similar genes *actually* are derived from a genome-doubling event (i.e., are homoeologous) rather than from some other gene duplication process. This problem is exacerbated by the increased probability of extinction of progenitor diploids, which provide the comparative organismal context necessary for inferences of polyploid ancestry and homeology. Accordingly, younger polyploids may be more appropriate for empirical studies of rate equivalence and independence of homoeologous genes.

To better understand patterns and processes of homoeologue divergence in a young allopolyploid genome, we have used locus-specific isolation methods (28) and comparative linkage mapping (29, 30) to study homoeologous gene evolution in allopolyploid cotton, *Gossypium hirsutum* ($2n = 4x = 52$; AD-genome). We used a phylogenetic perspective by including orthologous copies from representatives of each of the progenitor diploid ($2n = 26$) genomes (Fig. 1*A*). *G. hirsutum* is one of five extant allopolyploid species that originated from a single polyploidization event involving an A-genome maternal parent and a D-genome paternal parent approximately 0.5–2.0 mya (31–33). Because phylogenetic relationships among these genome groups have been characterized (32–34), they provide a framework for evaluating sequence evolution of homoeologous and orthologous loci in a young polyploid (Fig. 1 *B* and *D*). Homoeology in allopolyploid cotton has been established by genetically mapping its constituent genomes [A- and D-subgenomes of *G. hirsutum* (A_T and D_T)] (29) and by comparing these maps to those generated for A- and D-genome diploids (30). These studies identified parallel linkage groups among the diploids and allopolyploids, thereby providing evidence of orthologous (A vs. D, A vs. A_T, D vs. D_T) and homoeologous (A_T vs. D_T) relationships. Using methods designed for isolation of strictly homoeologous sequences (28), we isolated 16 pairs of loci from allopolyploid *G. hirsutum* and the corresponding orthologues from model A- and D-genome diploid parents. Phylogenetic analysis by using an ap-

Abbreviations: A_T, A-subgenome of *G. hirsutum*; D_T, D-subgenome of *G. hirsutum*; A_1, *Gossypium herbaceum*; D_5, *Gossypium raimondii*; mya, million years ago.

‡To whom reprint requests should be addressed. E-mail: jfw@iastate.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

propriate outgroup permitted detailed evaluation of the fate of duplicated loci after polyploid formation.

Materials and Methods

Locus Isolation and Characterization. We isolated pairs of homoeologous genes from allopolyploid (AD-genome) *G. hirsutum* L. race Palmeri and their orthologous counterparts from A-genome [*Gossypium herbaceum* L. (A_1); *Gossypium arboreum* L. (A_2)] and D-genome [*Gossypium raimondii* Ulbrich (D_5)] diploids (Fig. 1). For purposes of rooting phylogenetic trees and evaluating evolutionary rate equivalence, the African taxon *Gossypioides kirkii* (Masters) J. B. Hutchinson was selected as an outgroup (32).

Of the sixteen loci studied, twelve correspond to a subset of the anonymous *Pst*I probes that were used to generate the comparative linkage maps (29, 30). These loci were selected for analysis because of their low copy number (one to two copies in diploids, additive in allopolyploids), the existence of comparative mapping information that substantiates inferences of orthology and homoeology, and because size separation of homoeologous sequences in genomic Southerns facilitated isolation of duplicated sequences. Putative identities for six of these twelve loci have been inferred from database searches (Table 1). The four remaining loci comprise partial sequences of alcohol dehydrogenase (*Adh*)*A* (exons 2–6) (35), *AdhC* (exons 2–8) (33), and two genes encoding cellulose synthase (CelA) (exons 1–5 of *CelA1* and *CelA2*) (36). Each of the 16 loci was isolated and sequenced from the two diploid species representing the progenitors of allopolyploid cotton (A, D), and from both subgenomes (A_T, D_T) of *G. hirsutum*. Homologous sequences for 15 of the 16 loci (all but *A1834*, which could not be amplified) were similarly characterized from the outgroup *G. kirkii*, although for two loci only partial sequences were obtained (*A1520*: 502 bp; *A1625*: 495 bp).

Anonymous *Pst*I-loci were isolated by using previously described methods (28); *Adh* and *CelA* gene fragments were isolated by PCR amplification from genomic DNAs. PCR products either were sequenced directly (from diploids) by using standard methods or were cloned into pGEM-T (Promega) before sequencing, so that duplicated copies in the allopolyploid could be isolated from each other. PCR primers (Table 1) were designed from *Pst*I-mapping probes or from published cDNA sequences.

Homology Relationships. A primary criterion for selecting loci in this study was confidence that sequence similarity would reflect orthologous (between A and D; A and A_T; D and D_T) and homoeologous (A_T and D_T) relationships. These inferences were based both on sequence comparisons and on comparative mapping studies (29, 30), which showed that the genes of interest mapped to comparable linkage groups in the different taxa used. Map locations for *AdhA, AdhC, CelA1,* and *CelA2* were determined by RFLP analysis of PCR products or genomic Southerns in segregating progenies of A-genome diploid (*G. arboreum* × *G. herbaceum*), D-genome diploid (*G. raimondii* × *G. trilobum*) and AD-genome allotetraploid (*G. barbadense* × *G. hirsutum*) interspecific crosses (data not shown). Segregation data were used to place these genes onto existing maps (29, 30). Map positions for other previously mapped loci were confirmed by using the same approach.

Tests of Independence and Rate Equivalence. For all loci, sequences were readily aligned by eye because of the low levels of nucleotide divergence (data sets available at www.public.iastate.edu/~botany/wendel.html). PAUP (37) and MEGA (38) were used to perform parsimony and distance analyses on individual loci, as well as on combined sequences from all loci (aligned length: 14,705 bp). The interpretive framework for this study is based on the organismal history shown in Fig. 1*A*. For each locus, the null hypothesis of independent evolution in the allopolyploid is expected to lead to a gene tree that is identical to the organismal phylogeny (Fig. 1*B*). If sequences evolve at equivalent rates among all lineages, branches

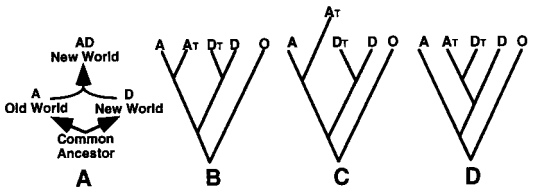

Fig. 1. Null hypothesis for sequence evolution in allopolyploids. (*A*) Phylogenetic history of diploid (A- and D-genome) and allopolyploid cotton species, as inferred from multiple lines of evidence (32–34). Allopolyploid cottons formed 0.5–2 mya from hybridization between A-genome and D-genome diploids, which diverged from each other *ca*. 5–10 mya. (*B*) Phylogenetic expectations of independence and equal rates of sequence evolution following allopolyploid formation. Shown are phylogenetic relationships between sequences from diploid progenitor ge-nomes (A and D) and their orthologous counterparts (A_T and D_T) in derived allopolyploids. *G. kirkii* serves as the outgroup (32) for testing both rate equivalence and independence. (*C*) An accelerated rate of sequence evolution in allopolyploids will generate longer branches leading to A_T and/or D_T than to A and D. (*D*) Concerted evolutionary forces may lead to nonindependent sequence evolution after allopolyploidization. Illustrated is conversion of an A-subgenome homoeologue to a D-subgenomic form, as has been demonstrated for ribosomal genes in allotetraploid *Gossypium* (20).

leading to A- and D-genome sequences will have the same length; rate deviation will yield branch length inequality, as shown in Fig. 1*C*. Nonindependent evolution of homoeologues should lead to deviation from the expected topology, as modeled in Fig. 1*D*. The possibility of recombination between duplicated locus pairs in allopolyploid cotton was evaluated by using methods described by Hudson and Kaplan (39) and Grassly and Holmes (40), facilitated by the computer programs DNASP 3.0 (41) and PLATO (40), respectively. We used the χ^2 method of Tajima (42) to test for substitution rate heterogeneity between all pairs of taxa, by using sequences from *G. kirkii* as the reference taxon (except for *A1834*, where a homologous sequence was not isolated). The χ^2 test was performed on individual loci and on the combined data set for all nucleotides and for synonymous and nonsynonymous partitions.

GenBank Accession Numbers. *A1286*, AF136808–AF136811, AF201876; *A1341*, AF136813–AF136816, AF201877; *A1520*, AF13818–AF13821, AF201878; *A1550*, AF201889–AF201893; *A1623*, AF139474–AF139477, AF201879; *A1625*, AF139417–AF13920, AF201880; *A1713*, AF139422–AF139425, AF201881; *A1751*, AF139437–AF139440, AF201883; *A1834*, AF139452–AF139459; *G1121*, AF139432–AF139435, AF201884; *G1134*, AF139427–AF139430, AF201882; *G1262*, AF061085–AF061088, AF201885; *AdhA*, AF136458, AF136459, AF085064, AF090146, AF201888; *AdhC*, AF036568, AF036569, AF036574, AF036575, AF169254; *CelA1*, AF139442–AF139445, AF201886; *CelA2*, AF139447–AF13945, AF201887.

Results

Characteristics of Included Loci. Loci included in this study ranged in aligned length from 294 bp (locus *A1286*) to 1,680 bp (*AdhC*), with an average length of 919 bp. Of the 16 loci evaluated, 4 were chosen because they were known genes (*AdhA, AdhC, CelA1, CelA2*). Five of the *Pst*I loci were identified as genes by virtue of their high sequence similarity (BLAST E values $\leq 2 \times 10^{-6}$) to genes in sequence databases. These loci include an aldehyde dehydrogenase gene (*A1550*), a subtilisin-like protease (*A1751*), an α-mannosidase precursor (*A1834*), a gene with high similarity to a human brain cDNA (*G1121*), and a P-glycoprotein gene (*G1262*). The seven remaining loci (*A1286, A1341, A1520, A1623, A1625, A1713, G1134*) showed no significant similarity to sequences in databases, nor did they display shared (across taxa) ORFs of appreciable size (>100 bp) in any frame.

EVOLUTION

Table 1. Sixteen sets of orthologous and homoeologous sequences isolated from diploid and allopolyploid *Gossypium*

Locus	Linkage group*	Known gene or database match†	*E* value‡	L§	L_A¶	L_S‖	Forward primer sequence/ reverse primer sequence
A1286	HA 10	Unknown	—	294	—	—	GTACTGCGGTAGACATGCATGAAAC/ ACTCTTCTTTAGCATCAGCTTCACCA
A1341	HA 7A	Unknown	—	624	—	—	GCATGCTGAATTGACAGAACCAGCY/ CACTCACAAAGTTATGCCGGATGY
A1520	HA 6	Unknown	—	957	—	—	GGCTGCAAAACCCCTAGGATTAGTY/ CAAGCCAGGCAAGCACTCCAAGA
A1550	HA 9	*Arabidopsis thaliana* aldehyde dehydrogenase (2961384)	8.0E-16	1,448	139	1,070	CCACCTCAGGCAAGGTTATCAY/ GGGATCAATGTGGCCATGTR
A1623	HA 5	Unknown	—	840	—	—	GATATTGAATGATCAACATGTGCGAT/ CTTTTACCTGCTGTGATGCCCAGT
A1625	HA 1	Unknown	—	1,061	—	—	CGATTCCCTACCAAATCGAGGT/ CATGGGACCCTGAAAGTTGAA
A1713	HA 6	Unknown	—	702	—	—	GAGGAGGAAGTTTGATCAACCACTG/ GGGTGCTTATGGTTATACAGGTGC
A1751	HA 10	*Lycopersicon esculentum* subtilisin protease (1771162)	1.0E-70	807	587	220	GCTGGAATGCTGGTTGTTATGAC/ GATGAGCTGCCTTCAACAAAGC
A1834	HA 8b	*A. thaliana* α-mannosidase precursor (1888357)	1.2E-23	882	162	654	TCACTAAAGAACCCATGTGGAT/ AGGGATCAATTTGCCAACCAA
G1121	HA 3	*Homo sapiens* putative brain protein, (3043668)	2.0E-06	749	670	216	CTGGATCAGCCATATGATGACAGGY/ TCAACCTAGTGGGGAGTGCTY
G1134	HA 8b	Unknown	—	546	—	—	CAGCTGGAGGATGGTTAGCTTCTCY/ GACTTGCACGTAAAGCACGAACC
G1262	HA 9	*Hordeum vulgare* P-glycoprotein (2292907)	1.0E-103	888	670	216	GGCGGCAGGCTAAGCACTTCY/ CGGAGGTCATACTTCCAGCTTY
AdhA	HA 8C	*AdhA* (similar to 1840425 from *Vitis vinifera*)	1.0E-85	951	464	468	CTTCACTGCTTTATGTCACACT/ GGACGCTCCCTGTACTCC
AdhC	HA 7B	*AdhC* (similar to 2570182 from *Arabis gemmifera*)	4.0E-62	1,680	873	765	CTGCKGTKGCATGGGARGCAGGGAAGCC/ GCACAGCCACACCCCAACCCTG
CelA1	HA 10	*CelA1* (similar to 1706956 from *G. hirsutum*)	8.0E-31	1,096	440	604	GATGGAATCTGGGGTTCCTGTTTGC/ GGGAACTGATCCAACACCCAGGA
CelA2	HA 1	*CelA2* (similar to 1706956 from *G. hirsutum*)	1.0E-22	1,180	460	682	GATGGAATCTGGGGTTCCTGTTTGC/ GGGAACTGATCCAACACCCAGGA

*Homoeologous assemblages (HA) from Brubaker *et al*. (30).
†Identities are based upon TBLASTX searches of GenBank NR database; protein sequence identification number (PID) in parentheses.
‡Smallest sum probability scores from TBLASTX search.
§Aligned sequence length in base pairs.
¶Number of synonymous sites, excluding gapped sites.
‖Number of nonsynonymous sites, excluding gapped sites.

Of the nine genes studied, two showed evidence of lesions that may interfere with expression in one or more of the included taxa. In the two A-genome diploids, *AdhC* has been either partially deleted (in *G. arboreum*; exon 6 and portions of both neighboring introns) or removed entirely (in *G. herbaceum*; as verified by Southern hybridization results). In addition, a point mutation in *AdhC* from *G. arboreum* results in a premature stop codon in exon 2. The D-subgenome sequence from *G. hirsutum* may also be a pseudogene, because it shows nonconsensus dinucleotides (AC... AG rather than GC... AG) at the splice junction of intron 6. Additionally, the D-genome diploid *G. raimondii* shows nonconsensus dinucleotides at the splice junction of intron 3 (GT... GG, position 507). Because these mutations are unique to each homoeologue and orthologue pair, pseudogene formation has likely occurred independently at both the diploid (A- and D-genomes) and polyploid (D-subgenome) levels.

CelA2 was the second locus that exhibits hallmarks of pseudogenization. In both extant A-genome diploids (*G. herbaceum, G. arboreum*) and the A-subgenome of *G. hirsutum*, identical point substitutions are observed that alter the consensus dinucleotides at the intron 1 splice site (GC... AG rather than GT... AG). In addition, a premature stop codon is found in exon 4 at amino acid position 103. Because of the conservation of these mutations, pseudogenization of *CelA2* most likely occurred in the common diploid ancestor of these taxa.

Gene Trees for Homoeologous Sequences in Polyploids and Their Progenitors. Fig. 2 shows the topologies obtained by parsimony analysis of 16 mapped sets of loci obtained from the diploid progenitor genomes (A_1, D_5) and their corresponding subgenomes (A_T, D_T) in allotetraploid cotton. In all cases, the gene trees recovered reproduced the organismal history of diploid divergence followed by allotetraploid formation. Thus, for all 16 pairs of homoeologous loci, the null hypothesis of independent evolution of duplicated genes (Fig. 1) cannot be refuted. The sole possible exception to this unanimity is for the locus *A1520*, where the A-subgenome sequence can be placed sister to either A_1 or the D_5/D_T lineage without a change in tree length. This equivocal phylogenetic placement is because of the exceptionally low amount of nucleotide divergence at *A1520* in general and in the A-genome (A_1 and A_T) lineage in particular (see Fig. 2).

Not surprisingly, when data were combined across all 16 loci (= 14,705 nt/taxon), results mirrored those observed with individual loci. Most importantly, sequences from each allotetraploid subgenome were sister to those from their respective diploid progenitors (Fig. 2). On the basis of the topologies obtained from the 16 loci,

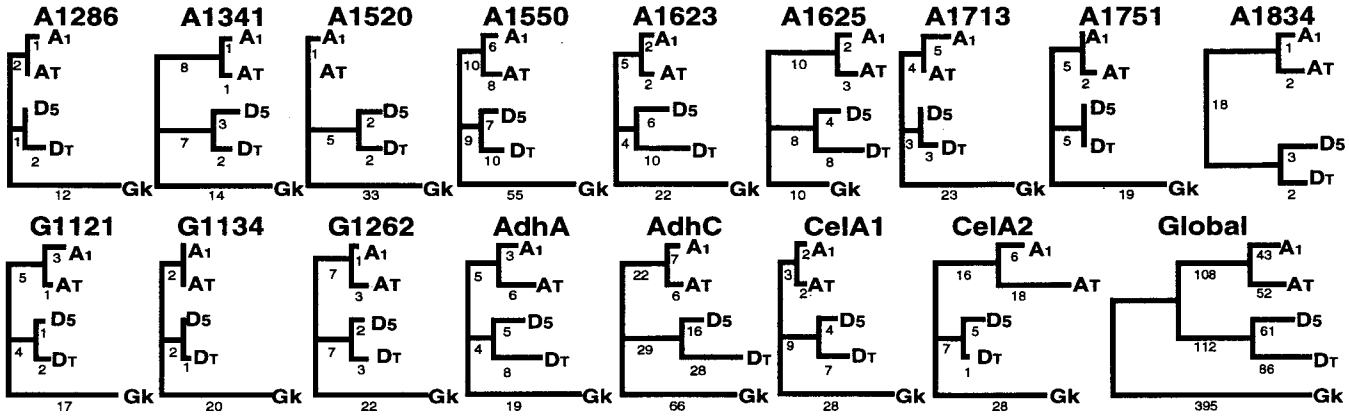

Fig. 2. Most-parsimonious trees obtained for the evolution of 16 low-copy loci in diploid (A_1, D_5) and allopolyploid (A_T, D_T) *Gossypium* genomes. The outgroup taxon, *G. kirkii,* is designated by the abbreviation Gk. Branch lengths (number of inferred changes) are indicated. (*Lower Right*) Merger of the 16 data sets leads to the global analysis shown.

individually and combined, we infer that interactions between homoeologous loci, such as interlocus gene conversion or reciprocal recombination, either have not occurred since the time that the two genomes became united via polyploidization or have been insufficient in magnitude to be detected by phylogenetic analysis.

Tests for Recombination Among Duplicated Sequences in Allotetraploid Cotton. To test for possible recombination between the A_T and D_T copies of each of the 16 sequenced loci, we subjected all locus pairs to recombination analysis (39, 40). Potential recombination events were suggested for five genes (*A1286*, *A1550, AdhC*, *CelA1*, *G1262*) by either or both methods of analysis; however, in all cases, evidence for conversion tracts was based on single anomalous nucleotides, which were flanked by sequence patterns contradictory to an interpretation of gene conversion (data not shown).

Sequence Divergence and Rate Equivalence of Homoeologous Loci in Allotetraploid Cotton. To estimate the amount of divergence for homoeologous pairs of loci within the genome of *G. hirsutum*, we calculated Jukes–Cantor distances between A_T and D_T for all 16 loci. The resulting estimates spanned a 4.5-fold range, from a low of 0.0077 (at *A1520*) to a high of 0.0494 (at *AdhC*). Combined across all loci, 343 nucleotide differences were observed (358 changes inferred by parsimony analysis; Fig. 2) among the 14,705 homologous positions sequenced in A_T and D_T, translating into a Jukes–Cantor distance of 0.0249 (Table 2).

To evaluate whether pairs of homoeologous loci were evolving at different rates, we used sequences from *G. kirkii* as a reference taxon and used relative rate tests (42). Comparisons across 15 loci (all except A1834) revealed 3 instances of rate inequality when all nucleotide sites are considered: (*i*) at *AdhC* and *CelA1*, the sequence from the D-subgenome of *G. hirsutum* has accumulated mutations at a significantly faster rate than the A-subgenome sequence ($P = 0.001$ and $P = 0.012$, respectively); and (*ii*) at *CelA2*, the sequence from the A-subgenome has evolved more rapidly than the D-subgenome sequence ($P = 0.001$). We note that two of the three cases of significant rate differences involve putative pseudogenes at *AdhC* and *CelA2* (as discussed above). *CelA2* also showed rate enhancement at nonsynonymous sites in two comparisons involving putative pseudogenes (A vs. D: $P = 0.034$; A_T vs. D_T: $P = 0.001$). The only other example of significantly elevated nonsynonymous rates involved *CelA1* (D_T vs. A_T; $P = 0.025$). Apart from these examples, there does not appear to be a bias in rate of sequence evolution for homoeologues in the two subgenomes of allotetraploid cotton (*cf.* Fig. 1*C*). This conclusion is supported by divergence data between reference sequences from *G. kirkii* (Gk) and the corresponding homoeologous locus pairs in *G. hirsutum* (Table 2). In all cases other than the two putative pseudogenes and *CelA1*, as discussed above, distances between Gk and A_T were similar to those between Gk and D_T, as expected under the hypothesis of equivalent rates of sequence evolution in the two subgenomes.

Sequence Divergence and Evolutionary Rate Equivalence in Diploid Cottons. Divergence amounts for orthologous sequences in A- and D-genome cottons at the 16 loci ranged from a low of 0.0087 for *A1520* to a high of 0.0478 for *AdhC*. When all data are considered together, 308 of 14,705 nucleotides differed between *G. arboreum* and *G. raimondii*, corresponding to a Jukes–Cantor distance of 0.0224. Comparisons to previously published sequence data from the nuclear genomes of A- and D-genome diploid cottons indicate that this mean divergence estimate is less than half the amount observed for nuclear ribosomal sequences from the internal transcribed spacer region (ITS) (ITS + 5.8S rDNA) (20) and approximately one-tenth of the divergence reported for 5S rDNA genes and spacers (43). Except for a single instance involving a putative *AdhC* pseudogene, all relative rate tests show that orthologous loci in A- and D-genome cottons have evolved at statistically equivalent rates (Table 2).

Comparisons of Evolutionary Rates Between Allotetraploid and Diploid Cotton. As shown in Table 2, rates of sequence evolution at individual loci were statistically homogeneous for all 32 comparisons between sequences from diploid cotton and their counterparts in allotetraploid cotton. When all data are combined and all nucleotide sites are considered, however, the D_T subgenome is shown by relative rate tests to evolve more rapidly than the D-genome of diploid cotton ($P = 0.02$). This effect, which is insignificant when partitions of synonymous ($P = 0.08$) and nonsynonymous ($P = 0.11$) sites are considered separately, appears primarily to be caused by the inclusion of the *AdhC* pseudogene, which has an extraordinarily long branch leading to D_T (Fig. 2). Excluding *AdhC* yields an inference of rate homogeneity ($P = 0.07$; $P = 0.49$ and 0.73 for synonymous and nonsynonymous sites, respectively).

To further evaluate whether genome doubling and the attendant genic redundancy lead to enhanced rates of sequence evolution in allotetraploid cotton, we computed genetic distances for all orthologous locus pairs in A- and D-genome diploids and compared these values to those calculated for A_T and D_T homoeologues. Because the latter are phylogenetically shown here to have evolved

Table 2. Pairwise divergences and relative rate tests among duplicated genes in allopolyploid (A_T, D_T) cotton and their orthologous counterparts (A, D) from progenitor diploid genomes

Locus	Length, nt	Distance† between sequences and relative rate test results‡					
		A vs. A_T	D vs. D_T	A vs. D	A_T vs. D_T	A_T v. Gk	D_T vs. Gk
A1286	294	0.0034	0.0069	0.0139	0.0174	0.0544	0.0583
A1341	624	0.0033	0.0082	0.0318	0.0267	0.0384	0.0364
A1520	957	0.0011	0.0044	0.0087	0.0077	0.0703	0.0749
A1550	1448	0.0104	0.0131	0.0240	0.0235	0.0552	0.0545
A1623	840	0.0048	0.0195	0.0171	0.0246	0.0362	0.0438
A1625	1061	0.0051	0.0144	0.0276	0.0325	0.0327	0.0364
A1713	702	0.0080	0.0047	0.0176	0.0158	0.0475	0.0512
A1751	807	0.0025	0.0000	0.0125	0.0150	0.0329	0.0304
A1834	882	0.0037	0.0061	0.0249§	0.0261§	—§	—§
G1121	749	0.0054	0.0040	0.0176	0.0148	0.0325	0.0312
G1134	546	0.0000	0.0018	0.0074	0.0093	0.0414	0.0402
G1262	888	0.0045	0.0057	0.0194	0.0206	0.0372	0.0372
AdhA	951	0.0096	0.0140	0.0184	0.0249	0.0325	0.0315
AdhC	1680	0.0097	0.0273	0.0478*(D)	0.0494**(D_T)	0.0510	0.0794
CelA1	1096	0.0037	0.0102	0.0168	0.0178*+(D_T)	0.0320	0.0429
CelA2	1180	0.0206	0.0051	0.0286†(A)	0.0367**†(A_T)	0.0529	0.0332
Total	14,705	0.0068	0.0105*(D_T)	0.0224	0.0249	0.0442	0.0464

† Jukes–Cantor transformations using all nucleotide sites.

‡ Significance levels using the Tajima 1D rate test (42) and *G. kirkii* (Gk) as the reference taxon are indicated by single ($P \leq 0.05$) and double ($P \leq 0.01$) asterisks (for all nucleotide sites) and plus symbols (for nonsynonymous sites); the taxon with the faster rate is shown parenthetically.

§ Relative rate tests and pairwise comparisons could not be conducted, because outgroup sequences were unavailable.

independently, and because the two allotetraploid subgenomes share the same most recent common ancestor with the diploids, genetic divergence between A_T and D_T is expected to be equal to that between A and D if evolutionary rates in diploids and polyploids are equivalent. As shown in Table 2, Jukes–Cantor distances between A_T and D_T are nearly identical to those between A and D at individual loci. Similarly, the mean divergence between A_T and D_T (0.0227) is not significantly different (paired *t* test; $n = 16$; $P = 0.07$) from the mean divergence between A and D (0.0209). Moreover, when the two pseudogenes at *CelA2* and *AdhC* are excluded, the small difference in diploid–polyploid divergences is further reduced (A vs. D = 0.0184; A_T vs. D_T = 0.0197; $P = 0.17$). These analyses indicate that duplicated sequences in allotetraploid cotton diverge from one another at rates nearly identical to those exhibited by orthologues from the diploid progenitors.

Discussion

Polyploidy is a prominent process in plants and has been significant in the evolutionary history of vertebrates and other eukaryotes (5, 10–12, 44–47). Once united in a common nucleus, genes duplicated by polyploidy may retain their independence and continue to evolve at equivalent rates, as if they were localized in different diploid nuclei. Alternatively, duplicated sequences could diverge at different rates, because of either functional divergence or pseudogenization of a redundant locus. Finally, duplicated genes may interact through concerted evolutionary mechanisms such as interlocus recombination or gene conversion. Here we explored these possibilities by isolating homoeologous loci in allotetraploid cotton and orthologous loci from model diploid progenitors. This permitted a detailed evaluation of the pattern and tempo of divergence among 16 strictly comparable genes in two diploids and their allotetraploid derivative. Results from this multilocus comparison provide insights into the early stages of homoeologue evolution in an allotetraploid plant genome.

Allopolyploidization in *Gossypium* Has not Been Accompanied with Extensive Pseudogene Formation. A central tenet of polyploid speciation theory is that resulting genetic redundancy may lead to a relaxation of purifying selection on one redundant gene copy (1, 2, 18, 19, 48). This process may be evidenced by either pseudogene formation or an enhanced rate of sequence substitution (49) in one allopolyploid subgenome relative to its diploid progenitor (Fig. 1*C*). Our survey of 10 genes revealed two potential pseudogene sequences in *G. hirsutum*, one residing in the D-subgenome (*AdhC*) and the other in the A-subgenome (*CelA2*). The hallmarks of pseudogenization for the A-subgenome *CelA2* sequence are shared with the two extant A-genome diploid taxa (*G. herbaceum* and *G. arboreum*), indicating that these mutations occurred *before* polyploid formation in a common diploid ancestor. As such, this pseudogene was not created in response to the redundancy provided by recent allotetraploid formation in *Gossypium*. In contrast, the presumptive *AdhC* pseudogene present in the D-subgenome of *G. hirsutum* does not share similar hallmarks of pseudogenization as the model D-genome diploid progenitor (*G. raimondii*), providing evidence that this locus may be undergoing pseudogenization subsequent to polyploid formation. It is worth noting that *AdhC* is also a pseudogene in the A-genome diploids, existing either in a highly degenerate form (missing exon 6 in *G. arboreum*) or missing entirely (*G. herbaceum*). These data suggest that *AdhC* may be functionally redundant with other *Adh* loci, perhaps increasing its evolutionary lability in both polyploid and diploid genomes.

In summary, the data presented here suggest that polyploidization in *G. hirsutum* has not been accompanied by an increase in deleterious mutations in homoeologous gene pairs. Accordingly, gene maintenance—*not gene silencing*—may prove to be the most common fate for genes that have been recently duplicated by polyploidization (4). As emphasized in several recent reviews (3, 4, 7, 14, 23), it is not uncommon for duplicated genes to retain functionality even after long periods of evolutionary time.

Homoeologous Low-Copy Loci Evolve Independently in Polyploid Cotton. If duplicated genes in an allopolyploid evolve independently after polyploidization, they are expected to be phylogenetically sister to their orthologous counterparts from donor diploids rather than to each other. Under this scenario (Fig. 1*B*), rooted gene trees would depict two pairs of sister sequences, A with A_T and D with D_T. Alternatively, if homoeologous sequences interact through nonhomologous crossing over or gene conversion, duplicated loci in the allopolyploid may become phylogenetically sister (Fig. 1*D*). An example of the latter was provided by the 45S rDNA arrays in allopolyploid *Gossypium* (20). *In situ* hybridization reveals that the polyploids have multiple rDNA arrays localized on several different homoeologous chromosomes (50), an organization that is additive

relative to the diploid progenitors. Nevertheless, these taxa possess only a single predominant ribosomal sequence, indicating that concerted evolutionary forces have eliminated one of the two diploid donor sequence types. In the present study, each of the 15 rooted trees shows that homoeologous sequences in *G. hirsutum* are sister to the sequences from their respective model genome donors rather than to each other (Fig. 2). Additionally, for the single unrooted tree (*A1834*), each of the duplicated copies in the polyploid is most similar to its predicted diploid orthologue.

Although phylogenetic analysis failed to reveal interaction among duplicated genes in allopolyploid cotton, the possibility remains that there has been gene conversion among particular gene segments, but that these conversion events have remained cryptic because of overwhelming phylogenetic signal in unconverted tracts. To test this possibility, we subjected all homoeologous locus pairs to recombination analysis (39, 40). For all 16 genes, these analyses failed to provide compelling evidence for gene conversion. Potentially recombinant tracts were restricted to five loci; in each case, these involved single nucleotides flanked by sequence patterns contradictory to an interpretation of recombination. Our preferred explanation of these anomalous shared nucleotides is that they reflect homoplasy rather than intergenomic interaction.

Both the phylogenetic and recombination analyses provide evidence that duplicated low-copy genes do not interact after polyploid formation but continue to evolve independently despite residing in a common nucleus. Perhaps copy number itself is an important factor underlying the evolutionary behavior of sequences duplicated via polyploidization. If this is the case, we might predict that independence will prove to be the most prevalent outcome for single-copy genes, whereas highly repetitive DNAs will be more likely to experience intersubgenomic interactions.

Evolutionary rates are equivalent in allotetraploid and diploid cotton. We evaluated the possibility of rate variation for all loci. Each of the 16 comparisons indicates that divergence amounts between homoeologues in the allotetraploid are essentially identical to those between their diploid orthologues. Relative rate tests confirm that nucleotide substitution rates are not elevated in allotetraploid *Gossypium* (Table 2), except in one case (*AdhC* in D_T), where presumptive pseudogene formation has led to an unusually long branch. Similarly, there is minimal evidence for elevated rates of nonsynonymous substitutions, apart from comparisons involving pseudogenes where such rate enhancements might be expected. Accordingly, there is little evidence for accelerated genic evolution after gene doubling via polyploidization.

Conclusions

Results presented here demonstrate that orthologous genes, recently united into a common nucleus via polyploidization, evolve independently of one another and at rates that are indistinguishable from those of their diploid progenitors. Clearly, we do not know the extent to which these results apply to the remaining thousands of homoeologous genes in allotetraploid cotton. Similarly, it is unknown whether the conclusions of rate equivalence and independence will be found generally to be true in other polyploid angiosperms, because there are no similar studies. Reciprocal recombination, gene conversion, and other forms of nonindependence among homoeologues remain evolutionary possibilities (20, 24, 51); however, the relative frequency of these outcomes and the sequences subject to each mechanism remain unknown. In this respect, it is noteworthy that nonindependence has already been demonstrated for highly repeated homoeologous sequences such as ribosomal arrays, in *Gossypium* as well as in other polyploids (20, 52–55). The present results exemplify genic stasis accompanying polyploidization, providing a sharp contrast to the several recent examples of rapid genomic evolution in allopolyploids (25–27, 56). Similar studies from other model polyploid systems (e.g., *Brassica*, *Triticum*) should facilitate an understanding of the generality of the trends described here.

We thank T. Haselkorn, L. Rasmussen, and J. Ryburn for technical assistance. This research was supported by the National Science Foundation.

1. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, New York).
2. Ohta, T. (1994) *Genetics* **138,** 1331–1337.
3. Pickett, F. B. & Meeks-Wagner, D. R. (1995) *Plant Cell* **7,** 1347–1356.
4. Wagner, A. (1998) *BioEssays* **20,** 785–788.
5. Wolfe, K. H. & Shields, D. C. (1997) *Nature (London)* **387,** 708–713.
6. Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. (1997) *Trends Genet.* **13,** 360–364.
7. Gibson, T. J. & Spring, J. (1998) *Trends Genet.* **14,** 46–49.
8. Larhammer, D. & Risinger, C. (1994) *Trends Genet.* **10,** 418–419.
9. Levin, D. A. (1983) *Am. Nat.* **122,** 1–25.
10. Pébusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. (1998) *Mol. Biol. Evol.* **15,** 1145–1159. **0**
11. Postlethwait, J. H., Yan, Y. L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., *et al.* (1998) *Nat. Genet.* **18,** 345–349.
12. Spring, J. (1997) *FEBS Lett.* **400,** 2–8.
13. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. (1997) *Nature (London)* **388,** 167–171.
14. Nadeau, J. H. & Sankoff, D. (1997) *Genetics* **147,** 1259–1266.
15. Allendorf, F. W. (1979) *Heredity* **43,** 247–258.
16. Li, W.-H. (1980) *Genetics* **95,** 237–258.
17. Wilson, H. D., Barber, S. C. & Walters, T. (1983) *Biochem. Syst. Ecol.* **11,** 7–13.
18. Ferris, S. D. (1984) in *Evolutionary Genetics of Fishes*, ed. Turner, B. (Plenum, New York), pp. 55–93.
19. Walsh, J. B. (1995) *Genetics* **139,** 421–428.
20. Wendel, J. F., Schnabel, A. & Seelanan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 280–284.
21. Wendel, J. F. (2000) *Plant Mol. Biol.,* in press.
22. Hughes, M. K. & Hughes, A. L. (1993) *Mol. Biol. Evol.* **10,** 1360–1369.
23. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-L. & Postlethwait, J. (1999) *Genetics* **151,** 1531–1545.
24. Elder, J. F. & Turner, B. J. (1995) *Q. Rev. Biol.* **70,** 297–320.
25. Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A. A. & Vega, J. M. (1997) *Genetics* **147,** 1381–1387.
26. Liu, B., Vega, J. M. & Feldman, M. (1998) *Genome* **41,** 535–542.
27. Song, K., Lu, P., Tang, K. & Osborn, T. C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7719–7723.
28. Cronn, R. C. & Wendel, J. F. (1998) *Genome* **41,** 756–762.
29. Reinisch, A. J., Dong, J., Brubaker, C. L., Stelly, D. M., Wendel, J. F. & Paterson, A. H. (1994) *Genetics* **138,** 829–847.
30. Brubaker, C. L., Paterson, A. H. & Wendel, J. F. (1999) *Genome* **42,** 184–203.
31. Wendel, J. F. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4132–4136.
32. Seelanan, T., Schnabel, A. & Wendel, J. F. (1997) *Syst. Bot.* **22,** 259–290.
33. Small, R. L., Ryburn, J. A., Cronn, R. C., Seelanan, T. & Wendel, J. F. (1998) *Am. J. Bot.* **85,** 1301–1315.
34. Wendel, J. F. & Albert, V. A. (1992) *Syst. Bot.* **17,** 115–143.
35. Small, R. L., Ryburn, J. A. & Wendel, J. F. (1999) *Mol. Biol. Evol.* **16,** 491–501.
36. Pear, J. R., Kawagoe, Y., Schreckengost, W. E., Delmer, D. P. & Stalker, D. M. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 12637–12642.
37. Swofford, D. L. (1993) *PAUP, Phylogenetic Analysis Using Parsimony, Ver. 3.1.1* (Smithsonian, Washington, DC).
38. Kumar, S., Tamura, K. & Nei, M. (1993) *MEGA: Molecular Evolutionary Genetics Analysis, Ver. 1.01* (Pennsylvania State University, University Park, PA).
39. Hudson, R. R. & Kaplan, N. L. (1985) *Genetics* **111,** 147–164.
40. Grassly, N. C. & Holmes, E. C. (1997) *Mol. Biol. Evol.* **14,** 239–247.
41. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15,** 174–175.
42. Tajima, F. (1993) *Genetics* **135,** 599–607.
43. Cronn, R. C., Zhao, X., Paterson, A. H. & Wendel, J. F. (1996) *J. Mol. Evol.* **42,** 685–705.
44. Sidow, A. (1996) *Curr. Opin. Genet. Dev.* **6,** 715–722.
45. Masterson, J. (1994) *Science* **264,** 421–424.
46. Leitch, I. J. & Bennett, M. D. (1997) *Trends Plant Sci.* **2,** 470–476.
47. Soltis, D. E. & Soltis, P. S. (1993) *Crit. Rev. Plant Sci.* **12,** 243–273.
48. Clark, A. G. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2950–2954.
49. Li, W.-H. (1985) in *Population Genetics and Molecular Evolution*, eds. Ohta, T. & Aoki, K. (Springer, Berlin), pp. 333–353.
50. Hanson, R. E., Islam-Faridi, M. N., Percival, E. A., Crane, C. F., Ji, Y., McKnight, T. D., Stelly, D. M. & Price, H. J. (1996) *Chromosoma* **105,** 55–61.
51. Arnheim, N. D. (1983) in *Evolution of Genes and Proteins*, eds. Nei, M. & Koehn, R. (Sinauer, Sunderland, MA), pp. 38–61.
52. Sang, T., Crawford, D. J. & Stuessy, T. F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6813–6817.
53. Brochmann, C., Nilsson, T. & Gabrielsen, T. M. (1996) *Symb. Bot. Ups.* **31,** 75–89.
54. Roelofs, D., Van Velzen, J., Kuperus, P. & Bachmann, K. (1997) *Mol. Ecol.* **6,** 641–649.
55. Zhang, D. & Sang, T. (1999) *Am. J. Bot.* **86,** 735–740.
56. Liu, B., Vega, J. M., Segal, G., Abbo, S., Rodova, M. & Feldman, M. (1998) *Genome* **41,** 272–277.