Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium)

JONATHAN F. WENDEL*, ANDREW SCHNABEL, AND TOSAK SEELANAN

Department of Botany, Iowa State University, Ames, IA 50011

Communicated by Major M. Goodman, North Carolina State University, Raleigh, NC, October 3, 1994 (received for review July 12, 1994)

ABSTRACT Polyplody is a prominent process in plant evolution; yet few data address the question of whether homoeologous sequences evolve independently subsequent to polyploidization. We report on ribosomal DNA (rDNA) evolution in five allopolyploid (4D genome) species of cotton (Gossypium) and species representing their diploid progenitors (A genome, D genome). Sequence data from the internal transcribed spacer regions (ITS1 and ITS2) and the 5.8S gene indicate that rDNA arrays are homogeneous, or nearly so, in all diploids and allopolyploids examined. Because these arrays occur at four chromosomal loci in allopolyploid cotton, two in each genome, repeats from different arrays must have become homogenized by interlocus concerted evolution. Southern hybridization analysis combined with copy-number estimation demonstrate that this process has gone to completion in the diploids and to completion or near-completion in all allopolyploid species and that it most likely involves the entire rDNA repeat. Phylogenetic analysis demonstrates that interlocus concerted evolution has been bidirectional in allopolyploid species—i.e., rDNA from four polyploid lineages has been homogenized to a D genome repeat type, whereas sequences from Gossypium mustellimum have been concerted to an A genome repeat type. Although little is known regarding the functional significance of interlocus concerted evolution of homoeologous sequences, this study demonstrates that the process occurs for tandemly repeated sequences in diploid and polyploid plants. That interlocus concerted evolution can occur bidirectionally subsequent to hybridization and polyploidization has significant implications for phylogeny reconstruction, especially when based on rDNA sequences.

Approximately 70% of angiosperms are thought to have experienced one or more episodes of polyploidy at some point in the past (1). Despite the prevalence of polyploidy in plants, surprisingly little is known regarding its genetic and genomic consequences. There are few data, for example, on the mode and tempo of homoeologous sequence evolution in polyploid genomes. These sequences, which evolved independently while isolated in diverging diploid genomes, are united in a common nucleus at the time of polyploid formation, thereby providing the possibility of nonindependent molecular evolution. We have initiated studies of repeated sequence evolution (2) in a model system involving five allopolyploid species of cotton (Gossypium) and species that represent their two diploid progenitors (Fig. 1). The allopolyploid ("4D genome"); (2n = 4 x = 52) species originated following hybridization of an African or Asian diploid ("A genome"); (2n = 26) species with a diploid American ("D genome"); (2n = 26) species (4, 5). Data suggest that the best models of the ancestral D genome parent are G. raimondii (4) and G. gossypioides (3) and that the A genome donor was similar to present-day G. herbaceum and G. arboreum (4). Polyploidization is suggested to have occurred in the mid-Pleistocene, following trans-oceanic dispersal of an A

Genome taxon to the New World, which served as the female (seed) parent in the initial hybridization (3, 5). Following polyploidization, the allopolyploids diverged into several lineages represented by five modern species, including the commercially important species G. hirsutum ("upland cotton") and G. barbadense ("pima" and "Egyptian" cotton).

We chose to study rDNA as an example of a highly repeated, tandemly arranged sequence for two reasons. (i) Higher plant rDNA is organized into arrays at one or more chromosomal locations (6, 7); each array contains hundreds to thousands of identical to near-identical repeats (Fig. 2), the repeats having become homogenized by evolutionary forces [e.g., unequal crossing-over (9, 10), gene conversion (11–13)] that are collectively referred to as concerted evolution (14, 15). Although concerted evolutionary forces have long been known to homogenize repeats within individual arrays (8, 15), there are few robust demonstrations of concerted evolution among repeats from different arrays (9, 13, 15). (ii) In situ hybridization experiments have shown that four rDNA loci exist in allopolyploid G. hirsutum, two in each of the A and D genomes (16), thereby providing the opportunity for interlocus concerted evolution at the diploid and polyploid levels. The ITS region, containing the 5.8S RNA gene flanked by two spacer sequences (Fig. 2), was selected for study due to its relatively rapid rate of sequence evolution (7, 8). Describing the evolutionary dynamics of multilocus ITS sequences is additionally significant in that sequence variation in this region recently has become widely used for phylogeny reconstruction in plants (17–22).

MATERIALS AND METHODS

Genomic DNAs were isolated (23) from individual plants of both extant A genome diploids (G. arboreum and G. herbaceum), two representative D genome diploids (G. thurberi and G. raimondii), and all five existing allopolyploids (Table 1). For comparative purposes, we included an allopolyploid synthesized by colchicine-doubling the sterile intergenomic hybrid G. arboreum X G. thurberi. Because the C genome diploids are basal within the genus (3), we selected the C genome species G. robinsonii for purposes of rooting phylogenetic trees.

To generate single-stranded ITS DNA for sequencing, we conducted two-stage PCR amplifications (24). We performed symmetric amplifications using primers a and b (Fig. 2), followed by asymmetric amplifications using aliquots of the double-stranded PCR products as template and only a single primer (a or b). Single-stranded DNAs were purified by ultrafiltration prior to sequencing. Because of problems presumably introduced by secondary structure, we sequenced both DNA strands with and without dGTP analogs, using the two amplification primers and

Abbreviations: ITS, internal transcribed spacer; rDNA, ribosomal DNA.
*To whom reprint requests should be addressed.
†The sequences discussed in this paper have been deposited in the GenBank data base (accession nos. U12710–U12719).

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.
two internal sequencing primers (c and d; Fig. 2). Sequences were generated by standard methods of dyeceq sequencing.

We aligned the sequences using the computer program MALIGN (25) and made minor adjustments manually to the resulting alignments. The phylogeny of these sequences was explored by using character-based and distance-based approaches to phylogeny reconstruction. Maximum parsimony analysis was conducted with the aid of the computer program PAUP version 3.0s (26). In this analysis, nucleotide transformations were weighted equally, and gaps were treated as additional character states, with the exception of a single 15-bp gap that was treated as a binary, presence/absence character. For distance-based phylogeny estimation, we translated the observed distances between all pairs of sequences to Jukes-Cantor or Kimura-2 “corrected” distances and subjected the resulting distance matrices to neighbor-joining analysis. These analyses were facilitated by the computer program MEGA (27).

![Fig. 1. Organisal context for the evaluation of concerted evolution of ribosomal DNA (rDNA) in the genus *Gossypium*. Phylogenetic relationships (solid lines) of three groups of diploid cotton are illustrated (3). Genomes of these groups vary 2-1.2-fold in DNA content and have been assigned genomic designations (boxes) based on meiotic pairing behavior in interspecies hybrids (4). A genome species are African–Asian in origin while D genome species are endemic to the New World tropics, primarily Mexico. The Australian, C genome diploids are basal within the genus (3), thereby providing outgroup taxa for phylogenetic analysis. Hybridization between A genome and D genome representatives led to the evolution of the New World allopolyploid cottons (4 AD genome). Chloroplast DNA sequence divergence data suggest that the A genome and D genome groups diverged from a common ancestor 5-10 million years ago and that the two diverged genomes became reunited in a common nucleus, via allopolyploidization, in the mid-Pleistocene (3, 5).](image1)

![Fig. 2. Nuclear ribosomal genes in higher plants are organized into one or more arrays containing hundreds to thousands of tandemly arranged repeats (6, 8). Each repeat, which is ∼9.4 kb in length in *Gossypium*, consists of a nontranscribed spacer region and an external transcribed spacer (solid lines, upper) and a coding region (boxes, upper) that contains genes for the large and small subunit RNAs and the 5.8S RNA (163-164 bp in length). The two internal transcribed spacer sequences (ITS1 and ITS2, lower) are cleaved from precursor transcripts during formation of the mature RNAs. Half-arrows denote PCR primers used for amplification and sequencing: a = 5'-GGAAAGTAAAATCGATAACAGG-3'; b = 5'-TCCTCCTCCGTCTATTGATAATGC-3'; c = 5'-GCATCGATGAAGAAAAGAGCCG-3'; d = 5'-CTCGTGTTCTTCATCGATGC-3'.](image2)

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Genome</th>
<th>Accession</th>
<th>Geographic origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G. herbaceum</em></td>
<td>A1</td>
<td>A1-73</td>
<td>Botswana</td>
</tr>
<tr>
<td><em>G. arboreum</em></td>
<td>A2</td>
<td>A2-74</td>
<td>China</td>
</tr>
<tr>
<td><em>G. thurberi</em></td>
<td>D1</td>
<td>T no. 71</td>
<td>Arizona, USA</td>
</tr>
<tr>
<td><em>G. raimondii</em></td>
<td>D2</td>
<td>D2-37</td>
<td>Peru</td>
</tr>
<tr>
<td><em>G. hirsutum</em> (AD1)</td>
<td>TX2009</td>
<td>Yucatán, Mexico</td>
<td></td>
</tr>
<tr>
<td><em>G. hirsutum</em> (AD2)</td>
<td>Palmeri</td>
<td>Oaxaca, Mexico</td>
<td></td>
</tr>
<tr>
<td><em>G. barbadense</em> (AD2)</td>
<td>K101</td>
<td>Bolivia</td>
<td></td>
</tr>
<tr>
<td><em>G. tomentosum</em> (AD3)</td>
<td>W7936</td>
<td>Hawaii, USA</td>
<td></td>
</tr>
<tr>
<td><em>G. mustelinum</em> (AD4)</td>
<td>WL600</td>
<td>Bahia, Brazil</td>
<td></td>
</tr>
<tr>
<td><em>G. barbadense</em> (AD4)</td>
<td>JL</td>
<td>Ceará, Brazil</td>
<td></td>
</tr>
<tr>
<td><em>G. darwinii</em> (AD5)</td>
<td>WL125</td>
<td>Galapagos Islands</td>
<td></td>
</tr>
<tr>
<td>Doubled [ <em>G. arboreum</em> (2AD1)]</td>
<td>Beasley</td>
<td>Synthetic</td>
<td></td>
</tr>
<tr>
<td><em>G. robbinsii</em></td>
<td>C2</td>
<td>AZ-50</td>
<td>Western Australia</td>
</tr>
</tbody>
</table>

Genomic designations follow ref. 4.

rDNA copy number was estimated for each taxon by Southern blot analysis. Aliquots of genomic DNAs that had been quantified by fluorometry were digested with restriction enzymes (Acc I, Mbo I, Xho I, Xmn I) that were predicted, based on the DNA sequence data, to lead to genome-specific ITS cleavage patterns. Resulting fragments were separated by agarose gel electrophoresis (3.5% NuSieve 3:1, FMC) and transferred to nylon membranes using standard procedures. Previous quantitation of genome sizes for *Gossypium* genomes and allopolyploids (28, 29) facilitated the loading of specific numbers of 1C genomic equivalents per lane, these ranging from 5 × 10^3 to 5 × 10^5. For internal standards on each blot, PCR-amplified ITS fragments were quantitated fluorometrically and digested with the appropriate restriction enzymes; amounts ranging from 5 × 10^3 to 5 × 10^5 ITS copies were loaded per lane. Hybridization probes were 32P-labeled, PCR-amplified ITS sequences from *G. mustelinum* or *G. hirsutum*. Hybridization and wash conditions were as described (30).

Restriction site variation in the rDNA nontranscribed spacer, external transcribed spacer, and the 18S and 26S coding regions (Fig. 2) was assessed by Southern blot analysis. Fifteen restriction enzymes were used (listed in ref. 31). Membranes were probed with a labeled clone, derived from *G. hirsutum* (Ghr1), containing a full-length (9.4 kb) rDNA repeat (31). Restriction site maps of the ITS sequence data allowed us to determine whether the resulting polymorphisms were due to variation in the ITS region or elsewhere within the repeat.

RESULTS

ITS sequences were deposited in GenBank under the following accession numbers: *G. robbinsii*, U12710; *G. arboreum*,
U12712; *G. herbaeaeum*, U12713; *G. mustelinum*, U12714; *G. barbadense*, U12715; *G. darwinii*, U12716; *G. hirsutum*, U12719; *G. tomentosum*, U12717; *G. thurberi*, U12711; *G. raimondii*, U12718. Comparisons of these data with published sequences (17, 32–34) allowed us to infer the boundaries of the three components of the ITS region (ITS1, ITS2, and the 5.8S gene). Because all *Gossypium* sequences are nearly identical in length, sequence alignment necessitated the insertion of only five gaps. Four of these (three in ITS1, two in ITS2) were a single nucleotide in length, while the fifth was a 15-bp gap in positions 670–684 of the sequences from *G. arboreum*, *G. herbaeaeum*, and *G. mustelinum*. ITS1 is longer in *Gossypium* (aligned length = 295 bp; absolute length = 293–294 bp) than in other angiosperms studied to date (ITS1 is usually 200–260 bp in length; refs. 17–22), but it has a typical GC content (58%). For ITS2, GC content (61%) and length (210–226 bp) are typical of other angiosperms. As expected, the 164-bp 5.8S gene is strongly conserved, with no variable sites detected. Excluding polymorphisms introduced by indels, we observed over twice as many variable nucleotide positions in ITS1 (59) than in ITS2 (25).

Although the DNA sequences were generated from a potentially heterogeneous PCR-amplified pool of ITS fragments, monomorphic sequencing ladders were observed in all diploids and allopolyploids examined. Sequences from a synthetic allopolyploid (doubled *G. arboreum* × *G. thurberi*) and from diploid hybrids, however, displayed all polymorphisms predicted from the parental sequences. These observations suggest that the PCR pools were relatively homogeneous—i.e., that sequence heterogeneity among repeats is minimal in diploid and allopolyploid *Gossypium*. When combined with the *in situ* hybridization data (16) that demonstrate four rDNA loci in allopolyploid *G. hirsutum* (two in each of its two subgenomes), the observation of ITS homogeneity leads us to conclude that there must have been complete or nearly complete interlocus concerted evolution of the ITS region in diploid and polyploid cotton species.

To evaluate whether the phenomenon of interlocus concerted evolution has extended beyond the ITS region into other components of the 9.4-kb rDNA repeat, we analyzed genomic DNAs, digested with 15 restriction enzymes, by Southern hybridization, using labeled rDNA as a probe. Particularly informative were restriction enzymes (e.g., *Apa* I, *Bam* HI, *Eco* RI, *Sac* I) that have no recognition sequences in the ITS region but reveal *A* genome and *D* genome specific hybridization patterns. In all such cases, we observed only one of the two patterns in allopolyploid cotton (not shown). Although we did not map the restriction site variants observed, they must have arisen from polymorphisms located in either the 18S gene, the 26S gene, or the large intergenic spacer (= nontranscribed spacer plus external transcribed spacer; Fig. 2). The absence of heterogeneity among repeats within individuals for these polymorphisms is congruent with the results for the ITS region, suggesting that the entire repeat has undergone interlocus concerted evolution.

Following an exhaustive search, phylogenetic analysis of the ITS sequences produced a single most-parsimonious tree (Fig. 3), with two primary, strongly supported branches (an *A* genome clade and a *D* genome clade). Topologically congruent trees were obtained from distance-based (neighbor-joining) phylogeny reconstruction methods and from analyses based on nucleotide variation for only ITS1 or ITS2 (not shown). The robustness of this phylogenetic hypothesis is indicated by (i) strong character support for the two divergent branches (22 and 10 character-state changes, respectively); (ii) high inter-character congruence, as demonstrated by the fact that only two nucleotide characters were homoplasious; and (iii) topological stability with respect to distance-based phylogenetic methods.

The most striking feature of the ITS phylogeny is the incongruence between the well-established organinal tree (Fig. 1) and the “gene tree” (Fig. 3). In the absence of concerted evolutionary forces, and given complete sampling, sequences from all allopolyploids should be evident on both major clades (*A* genome and *D* genome), because each allopolyploid is expected to contain two sets of paralogous loci. Our results, however, show only one sequence for each allopolyploid species, and, moreover, these sequences are not monophyletic: four of the five allopolyploids and both *D* genome diploids constitute one of the two primary branches,

![Fig. 3. Phylogeny of rDNA ITS sequences in Gossypium. Taxon names are followed by their alphabetical genome designations. This single most-parsimonious cladogram (length = 90), obtained from an exhaustive search, is topologically identical to trees obtained using neighbor-joining analysis. The phylogeny is rooted with the Australian, outgroup taxon G. robinsonii. Numbers indicate character support for each branch segment (overall consistency index = 0.98; retention index = 0.98). ITS sequences from the allopolyploid species (AD genome) occur on both primary branches, demonstrating that different polyploid lineages have experienced homogenization to alternative rDNA repeat types.](image-url)
while the other clade consists of the two $A$ genome diploids and the sole remaining allopolyploid, *G. mustelinum*. To confirm this result and check for PCR contamination, we analyzed sequences from an additional accession of one allopolyploid species on each branch—i.e., *G. hirsutum* and *G. mustelinum* (Table 1). We selected these accessions according to expectations that they would be maximally divergent from the initial accessions studied, based on previous evidence (35–37). In both cases, the newly generated ITS sequence was identical to that generated from the first accession studied. When combined with the observation of relatively homogeneous PCR pools and the *in situ* hybridization data, the phylogenetic result that the $AD$ genome allopolyploid *Gossypium* species occur on two divergent clades leads to the conclusion that not only has there been interlocus concerted evolution among rDNA repeats on four different chromosomes but also that this phenomenon has occurred bidirectionally—i.e., rDNA of *G. mustelinum* has “concerted” to an $A$ genome repeat type, whereas rDNA from the remaining polyploids has become homogenized to a $D$ genome repeat type (the *in situ* hybridization observations exclude the formal possibility of differential loss of entire rDNA arrays in different polyploid lineages).

We used Southern blot analysis to address the question of whether interlocus concerted evolution has reciprocally homogenized all rDNA repeats or simply most rDNA repeats. As shown in Fig. 4, the synthetic allopolyploid exhibits both repeat types in approximately equal amount, demonstrating that the phenomenon of interlocus concerted evolution requires a longer time to become evident than the few generations that have elapsed since synthesis of the laboratory allopolyploid. In contrast to the synthetic allopolyploid, only one repeat type is visible in the naturally occurring allopolyploids (Fig. 4). Prolonged exposure of the autoradiogram confirmed that homogenization to an $A$ genome repeat type has been complete in the *G. mustelinum* lineage and that in *G. hirsutum*, where concerted evolution has operated in the direction of the $D$ genome repeat type, a small amount (two or three orders of magnitude less) of $A$ genome rDNA remains.

To quantify the extent of interlocus homogenization, we estimated the number of rDNA repeat units in the diploids and the allopolyploids. These experiments (data not shown) demonstrated that the $A$ genome and $D$ genome diploids contain $\approx 3800$ rDNA repeats per haploid genome and that this number is approximately additive in the allopolyploids.

We suggest, accordingly, that since polyploid formation $\approx 3800$ $D$ genome repeats have become homogenized to an $A$ genome repeat type in the *G. mustelinum* lineage, while a nearly equivalent number of repeats have been homogenized in the opposite direction in the *G. hirsutum* lineage.

**DISCUSSION**

The sequence data and *in situ* hybridization results, combined with phylogenetic reconstruction and Southern blot analyses, lead to the conclusion that *Gossypium* rDNA has experienced interlocus as well as intralocus concerted evolution. Although several studies have documented the retention of more than one rDNA repeat type in various plant groups following hybridization (Fig. 4 and refs. 38–42) or polyploid formation (43–45), there appear to be few, if any, demonstrations of interlocus rDNA homogenization in plants (46, 47). This raises the question of whether our results will prove to be exceptional as data accumulate for other plant groups. The paucity of documented cases of rDNA interlocus concerted evolution may reflect, in part, the severity of the criteria required to demonstrate the process: a well-established phylogenetic framework; justified inferences of allopolyploidy (as opposed to autopolyploidy); knowledge of rDNA locus number in descendant polyploids, which is critical for ruling out rDNA locus loss (48) as an alternative explanation for sequence homogeneity. This combination of requirements has been rarely met [e.g., *Gossypium*, *Brassica* (43), *Triticum* (46)]. Concerted evolution among rDNA loci may be more common than previously recognized, however, given accumulating evidence of ITS sequence homogeneity in “diploid” plant groups suspected of being paleopolyploid (refs. 17 and 49; this study).

Although we have no direct evidence bearing on the mechanism of interlocus homogenization in *Gossypium*, perhaps the most plausible scenario is mitotic or meiotic unequal crossing-over between repeats located on different chromosomes (9). In this respect, it may be more than coincidental that *Gossypium* rDNA arrays occupy telomeric or subtelomeric chromosomal loci (16), because this would allow ephemeral, perhaps sporadic, pairing and exchange to occur without deleterious cytogenetic consequences. In this respect, the pattern of allopolyploid *Gossypium* rDNA evolution is remarkably similar to that in humans, where repeats at five telomeric locations appear to have experienced interlocus homogenization (9, 15). If interlocus unequal crossing-over is the operative mechanism, we would expect that following polyploidization, long-term maintenance of more than a single repeat type, as in the flowering plant family Winteraceae (21), is more likely in organisms where rDNA loci are not telomeric in distribution.

At present, we do not know whether interlocus concerted evolution has homogenized other tandemly repeated sequences or whether there is functional or selective significance to the process. In this vacuum of empirical evidence, we can only speculate about the possibility that functionally or selectively unequal repeated families can become united in a common nucleus as a consequence of polyploidization (or diploid hybridization) and thereby provide the opportunity for differential selection.

Much clearer is the significance of our results for phylogeny reconstruction based on ITS (and other rDNA) sequencing, which has become the tool of choice for inferring organismal relationships based on nuclear sequences in plants (17–22, 49). As demonstrated in allopolyploid *Gossypium*, bidirectional interlocus concerted evolution is a possibility whenever two or more divergent repeat types are maintained through at least one cladogenic (speciation) event. When followed by interlocus homogenization to alternative repeat types, strongly supported, but positively misleading, inferences of organismal relationships may obtain. A necessary (but insufficient) con-
dation for this process to operate is that two (or more) rDNA arrays diverge in a single nucleus or that two divergent arrays become united in a common nucleus via polyploidization or hybridization at the diploid level. The known high frequency of these latter two phenomena in plants (1, 50) underscores the potential phylogenetic significance of bidirectional, interlocus concerted evolution.

Finally, we note that even under conditions of complete and bidirectional homogenization to alternative repeat types, where misleading phylogenetic relationships are expected, rDNA data may still reinforce or extend our understanding of evolutionary history. This is most likely to be true when other independent lines of phylogenetic evidence are available, as in *Gossypium*. In such cases, ITS and other rDNA data provide the possibility of corroborating or providing insight into the genominc composition of allopolyploids or diploids suspected of having an introgressive history.

We thank B. Baldwin for initial advice regarding sequencing of the ITS region and B. Baldwin, J. Doyle, and an anonymous reviewer for comments on the manuscript. Financial support (to J.F.W.) was provided by the National Science Foundation.