Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (Gossypium)

Corinne E. Grover1, HyeRan Kim2,†, Rod A. Wing2, Andrew H. Paterson3 and Jonathan F. Wendel1,*
1Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA,
2Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, and
3Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA

Received 16 October 2006; revised 26 January 2007; accepted 20 February 2007.
*For correspondence (fax 515 294 1337; email jfw@iastate.edu).
†Present address: Department of Horticulture, University of Wisconsin, Madison, WI 53706, USA

Summary
Genome sizes vary by several orders of magnitude, driven by mechanisms such as illegitimate recombination and transposable element proliferation. Prior analysis of the CesA region in two cotton genomes that diverged 5–10 million years ago (Ma), and acquired a twofold difference in genome size, revealed extensive local conservation of genic and intergenic regions, with no evidence of the global genome size difference. The present study extends the comparison to include BAC sequences surrounding the gene encoding alcohol dehydrogenase A (AdhA) from four cotton genomes: the two co-resident genomes (A_T and D_T) of the allotetraploid, Gossypium hirsutum, as well as the model diploid progenitors, Gossypium arboreum (A) and Gossypium raimondii (D). In contrast to earlier work, evolution in the AdhA region reflects, in a microcosm, the overall difference in genome size, with a nearly twofold difference in aligned sequence length. Most size differences may be attributed to differential accumulation of retroelements during divergence of the genome diploids from their common ancestor, but in addition there has been a biased accumulation of small deletions, such that those in the smaller D genome are on average twice as large as those in the larger A genome. The data also provide evidence for the global phenomenon of ‘genomic downsizing’ in polyploids shortly after formation. This in part reflects a higher frequency of small deletions post-polyploidization, and increased illegitimate recombination. In conjunction with previous work, the data here confirm the conclusion that genome size evolution reflects many forces that collectively operate heterogeneously among genomic regions.

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

Introduction
The observation that genome sizes vary tremendously among eukaryotes, and are largely uncorrelated with organismal complexity, has generated substantial interest over the last half-century. This interest has stimulated numerous genome size surveys for diverse organisms (Bennett and Leitch, 2005a; Gregory, 2006), as well as discussion of the modes and mechanisms responsible for the observed variation (Flavell et al., 1974; Bennetzen, 2000; Gregory, 2001; Petrov, 2001; Bennetzen, 2002; Gregory, 2005). Once thought to result mostly from polyploidy or polypeny (Thomas, 1971), genome size evolution is now recognized as reflecting the net effects of a suite of mechanisms that sometimes work antagonistically to expand and contract the genome. Best understood are the array of mechanisms responsible for genome size expansion, most prominently polyploidy (Wendel, 2000) and transposable element amplification (Bennetzen, 2000, 2002; Kidwell, 2002; Piegut et al., 2006), but also smaller scale processes such as tandem repeat expansion (Ellegren, 2002; Morgante et al., 2002), gene duplication and pseudogenization (Zhang, 2003), organellar transfer to the nucleus (Shahmuradov et al., 2003), and intron size expansion (Deutsch and Long, 1999; Vinogradov, 1999). Less is known about mechanisms of genome size contraction, of which unequal intrastrand homologous recombination (Shepherd et al., 1984; San-Miguel et al., 1996; Chen et al., 1998; Vicent et al., 1999; Shirasu et al., 2000), double-strand break repair (Kirik et al., 2000; Orel and Puchta, 2003), and illegitimate recombination...
(Wicker et al., 2001; Devos et al., 2002; Ma et al., 2004; Bennetzen et al., 2005) are thought to be important. Processes such as replication error and recombination in regions of tandem repeats may further contribute to genome size contraction through biases favoring small deletions over insertions (Petrov, 1997; Petrov, 2002). Superimposed on these ‘internal’ molecular and genetic mechanisms that contribute to genome size differences, are myriad ‘external’ biological and ecological factors that may potentially influence, or be influenced by, genome size (Bennett et al., 1998; Vinogradov, 2003; Cavalier-Smith, 2005; Knight et al., 2005; Petrov and Wendel, 2006), although in most cases these relationship remain unclear.

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

To continue to investigate the patterns and processes responsible for genome size evolution in *Gossypium*, we report on further comparative sequencing using both diploid and allopolyploid cotton species. *Gossypium* is an approximately 5–10 million year old genus, whose members have genomes that range 3-fold in size, from the D-genome diploids in the New World to the Australian K-genome diploids (Hendrix and Stewart, 2005). Approximately 5–10 Ma, two diploid groups, designated A-genome and D-genome, diverged and subsequently acquired genomes that differ by approximately twofold in size. Allopolyploidization reunited these two genomes approximately 1–2 Ma (Figure 1), generating five species, including the agriculturally important *G. hirsutum*. In contrast to the previously sequenced CesA region, the sequence composition of the AdhA region mirrors the overall pattern of genome size evolution in the diploid genomes. While still retaining a high level of intergenic sequence conservation, the AdhA region in the A and A_T genomes is disrupted by the presence of many gypsy elements, which serve to expand the region in a manner that reinforces the conclusions reached following analysis of sequences from whole-genome shotgun libraries (Hawkins et al., 2006). In addition to describing this phenomenon, the data presented here reveal details of ‘genomic downsizing’ in polyploids shortly after their formation, suggest an indel bias leading to frequent and larger deletions in smaller genomes, and provide evidence that increased illegitimate recombination may lead to genome size contraction.

**Results**

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

The AdhA BACs from the A genome diploid (112.3 kb) and the A_T genome from the allotetraploid (195.3 kb) were shotgun sequenced and assembled. The aligned length of the two BACs was 117.3 kb, accounting for the full 112.3 kb in A and 101.7 kb in A_T, with the elongated alignment...
reflecting gaps between the diploid and polyploid sequences (Figure 2).

Database searches led to the inference of five shared genes and one shared pseudogene (Table 1), giving gene densities of one gene per 22 kb and one gene per 20 kb for A and AT, respectively (19 and 17 kb if the pseudogene is included). Collectively, the five genes account for approximately 12.8 kb of sequence in each BAC, or approximately 10–12% of each BAC. Both BACs have a GC content of 34% and were determined to be 98.5% identical in sequence (81.28% including gaps). A total of 122 gaps appear in the alignment of the A and AT sequences; these are unequally distributed as 28 gaps in the A sequence (151 bp) and 64 gaps in the AT sequence (15 548 bp). When large indels (>400 bp) are removed, the number and length of gaps in A remains the same, but diminishes in AT to 60 gaps (449 bp). As these gaps are inferred to have evolved subsequent to the origin of the polyploids, about 1–2 Ma, the foregoing numbers reflect the differential accumulation of indels subsequent to polyploid formation. Also distinguishing the two genomes is a single retrotransposon insertion in the AT genome (between genes 4 and 5; Figure 2), accounting for 4799 bp, which by its exclusivity is inferred to have been inserted since the origin of the polyploids.

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

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

Sequence comparison between BACs from all diploid and polyploid genomes

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

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

Gap polarization and analysis

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

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

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

When the amount of sequence is considered and the gap data are normalized (e.g. per 100 kb; Table 2), the disparity in insertion rates among genomes largely disappears, whereas the disparity in the number of deletions increases. In addition, the average insertion size in the A genome, excluding TEs and the single large D genome insertion, was slightly larger than in the other genomes (2.8 nt in A versus 1 nt for both AT and D, and 1.8 nt for DT), whereas the average deletion size in the A genome mirrored the average deletion size in AT (2 and 1.9 nt, respectively), and was approximately half the average deletion size in D/DT (5.3 and 4.2 nt, respectively). Thus, the data of Table 2 highlight two salient features of genome size evolution in the AdhA region: (1) the higher frequency and size of deletions in the D genome compared with the A genome, consistent with their global difference in genome size; and (2) the higher rate of deletion in polyploid Gossypium compared with its diploid antecedent genomes, consistent with the phenomenon of ‘genomic downsizing’ following polyploid formation.

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

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Type</th>
<th>G. arboreum, A</th>
<th>G. hirsutum, AT</th>
<th>G. hirsutum, DT</th>
<th>G. raimondii, D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>#Deletions</td>
<td>4</td>
<td>13</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>nt deletions</td>
<td>8</td>
<td>29</td>
<td>77</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>#Insertions</td>
<td>12</td>
<td>2 (1)</td>
<td>6</td>
<td>6 (4)</td>
</tr>
<tr>
<td></td>
<td>nt insertions</td>
<td>34</td>
<td>4799 (1)</td>
<td>11</td>
<td>5971 (4)</td>
</tr>
<tr>
<td></td>
<td>#Unknown gaps (excluding TEs)</td>
<td>159</td>
<td>169</td>
<td>154</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td>nt missing (excluding TEs)</td>
<td>7809</td>
<td>21 715</td>
<td>15 440</td>
<td>15 404</td>
</tr>
<tr>
<td>Small indels (&lt;400 bp) per 100 kb in the AdhA region</td>
<td>#Deletions</td>
<td>3.57</td>
<td>12.78</td>
<td>36.36</td>
<td>26.53</td>
</tr>
<tr>
<td></td>
<td>nt deletions</td>
<td>7.14</td>
<td>28.52</td>
<td>28.52</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>#Insertions</td>
<td>10.71</td>
<td>0.99</td>
<td>10.91</td>
<td>8.16</td>
</tr>
<tr>
<td></td>
<td>nt Insertions</td>
<td>30.28</td>
<td>0.99</td>
<td>20</td>
<td>8.16</td>
</tr>
</tbody>
</table>

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

© 2007 The Authors
Analysis of putative genes

Six genes and one pseudogene are predicted to occur ancestrally in the AdhA region (Table 1). These six genes range in size from a 1.1-kb putative integral membrane protein-encoding gene to a 4.9-kb putative FAD-dependent oxidoreductase protein-encoding gene. The structures of four of the six genes were confirmed fully by expressed sequence tag (EST) evidence (Table 1), and the other two were partially confirmed by incomplete EST evidence (Udall et al., 2006) (http://www.genome.arizona.edu/genome/cotton.html).

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

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

Analysis of potential transposable elements and intergenic space

Differential accumulation of transposable elements was evaluated for the four genomes (Table 3). All four genomes share a LINE element (LINE 1, approximately 4 kb, contains a second LINE insertion in the A/AT genomes), a copia-like pol fragment (820 bp), and possible long terminal repeats (LTRs) of an ancient retroelement, representing transposable element insertions that occurred prior to or concurrent with the origin of the genus (and hence are not relevant to genome size evolution within the genus). Two TEs in the D/DT genomes...
have occurred since the D/D_T divergence (three mutations in D and seven in D_T), indicating that the element was likely to have been inserted just prior to that divergence, approximately 1–2 Ma. The D genome also has a unique copia insertion (copiaD; Tnt-94-like) of approximately 2.3 kb (between genes 5 and 6, Figure 2), the 420-bp LTRs of which are 96% identical (excluding gaps). The small size of the element indicates possible decay or internal deletions.

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

The A and A_T genomes share three discrete gypsy elements and an undetermined number of fragmented elements found in a large ‘gypsy landing pad’. The three discrete gypsy elements include one full-length element, one truncated by the end of the A genome BAC, but presumed to be full-length (gypsyA4; possibly Gorge1 or Gorge3), and one full-length element in A that is represented by only a solo-LTR in A_T (gypsyA6; possibly Gorge3). Characterization of the particular family to which each element belongs was made possible by a larger survey of cotton repetitive sequences (Hawkins et al., 2006). The range in full-length element size is from 8.1 to 16.7 kb, and all LTRs are approximately 95% identical (within elements). For the three discrete gypsy elements examined, the orthologous elements in the A_T genome were smaller than their A genome counterparts, by a minimum of 20%. In addition, the A_T genome contains a 4.8-kb unique LTR-retrotransposon of probable gypsy origin (gypsyA2). Overall, aside from being more abundant in the A genomes, intact gypsy elements were larger than those found in the D genomes. The largest gypsy represented in the D genomes was still smaller than the smallest intact gypsy in the A genomes, and less than half the size of the largest (7.5 kb in D versus 8.1 and 16.7 kb in A/A_T).

In intergenomic comparisons, the divergence between transposable elements, which were identical at the point of divergence, ranged from approximately 9% (when comparing either A genome to either D genome) to approximately 1% in the LINEs shared by the A/A_T genome. The two TEs shared between either A versus either D genome, and the three TEs shared between D and D_T showed less than 1% variation in sequence divergence between the different elements, whereas the seven shared TEs between A and A_T varied 2.5% in sequence divergence, from 1.2 to 3.6% divergence. These values were invariably

| Table 3 Repetitive element lengths in diploid and polytloid cotton |
|---------------------------------|--------------|--------------|----------------|----------------|
| **Repetitive sequence**          | A            | A_T          | D_T           | D             |
| Ancient retroelement, left LTR  | 8433         | 8395         | 1643          | 7614          |
| Ancient retroelement, right LTR | 1918         | 1914         | 302           | 302           |
| copia A                         | *            | 5479         | 0             | *             |
| copia remnant                   | 820          | 820          | 821           | 820           |
| gypsy A1                        | *            | 13 982       | 0             | 0             |
| gypsy A2                        | 0            | 4799         | 0             | 0             |
| gypsy A3                        | 16 698       | 9981         | 0             | 0             |
| gypsy A4                        | 6010*        | 8061         | 0             | 0             |
| gypsy A5                        | *            | 8987         | 0             | 0             |
| gypsy A6                        | 10 022       | 2632         | 0             | 0             |
| gypsy D1                        | *            | 0            | 7471          | *             |
| gypsy D2                        | 0            | 0            | 5103          | 4758          |
| copia D                         | 0            | 0            | 0             | 2348          |
| LINE 1                          | 3990         | 3995         | 4008          | 4011          |
| LINE 1b (internal)              | 6285         | 6264         | 0             | 0             |
| LINE 2                          | 6332         | 6301         | 0             | 0             |
| Mutator-like transposase         | *            | 634          | 0             | 0             |
| Mutator-like transposase         | *            | 509          | 506           | 507           |
| Mutator-like transposase         | *            | *            | *             | 698           |
| Mutator-like transposase         | *            | *            | *             | 1056          |
| pol 1                           | 500          | 500          | 0             | 0             |
| pol 2                           | 1600         | 1597         | 0             | 0             |
| pol 3                           | 4034         | 0            | 0             | 0             |
| pol 4                           | *            | 525          | 653           | 897           |
| solo-LTR; related to gypsy 4     | *            | 2697         | 0             | 0             |
| **Repetitive by genomic survey sequence** |
| Unknown repetitive               | *            | 471          | 486           | 482           |
| Unknown repetitive               | *            | 451          | 451           | 448           |
| Unknown repetitive               | 0            | 2029         | 2005          | 0             |
| Unknown repetitive               | 0            | 2339         | 2334          | 0             |
| Gypsy-like repetitive            | *            | 1983         | 1972          | 0             |
| Limited similarity to Calcineurin-like elements | * | * | 3069 |
| Phosphoesterase and Mutator-like elements | Unknown repetitive | * | 674 | * |
| GORGE3 gypsy-like repetitive     | *            | 584          | 0             | *             |
| Unknown repetitive               | 0            | 314          | 0             | *             |
| GORGE3 gypsy-like repetitive     | *            | 1423         | 0             | *             |
| Unknown repetitive               | 0            | 3215         | 0             | *             |
| Unknown repetitive               | 0            | 219          | 0             | *             |
| Unknown repetitive               | *            | 610          | 0             | *             |
| GORGE3 gypsy-like repetitive     | *            | 1238         | 0             | *             |
| Unknown repetitive               | 2533         | 2534         | 0             | 0             |
| Unknown repetitive               | 3171         | 3171         | 0             | 0             |
| Unknown repetitive               | *            | 5547         | 0             | 0             |
| Unknown repetitive               | *            | 2592         | 0             | 0             |

*Indicates the sequence is not present because of the end of the BAC.
larger than when comparing unassigned intergenic space between genomes, most likely because of a combination of factors, including conserved regulatory elements in the unassigned intergenic space, and the rapid mutation of TE sequences (SanMiguel and Bennetzen, 1998). The divergence of the unassigned intergenic space between the diploid and polyploid genomes closely mirrored the values obtained from 48 nuclear genes in Gossypium [0.008 intergenic versus 0.007 nuclear (Senchina et al., 2003) A–A$_T$; 0.0142 intergenic versus 0.010 nuclear (Senchina et al., 2003) D–DT]. The divergence of the unassigned intergenic space between the A and D genomes was nearly identical, regardless of which A and D genome were compared (0.058 for A–D, A–DT, A$_T$–DT and 0.059 for A$_T$–D), and these values were over double the divergence calculated from nuclear genes (0.022 A–D and 0.024 A$_T$–DT; Senchina et al., 2003), possibly indicating the presence of previously (and perhaps currently) rapidly evolving, severely degraded TEs that are unrecognizable.

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

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

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

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

Discussion

Mechanisms of genome evolution in the AdhA region

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

A primary difference between the present and earlier studies is that unlike the CesA region, the AdhA region mirrors, within the span of just over 100 kb, the twofold overall size difference that characterizes the 1697 and 885 Mbp genomes of the A and D genome lineages. In accordance with other plant systems and the repeat analysis of whole-genome shotgun libraries of the Gossypium genus (Hawkins et al., 2006), the primary force responsible for the size difference between the A and D genomes in the AdhA region was differential accumulation of gypsy transposable elements. Accumulation of gypsy elements in each genome accounts for >32.7, 25.3, 5.1 and 7.1 kb in the A, A$_T$, D$_T$ and D genomes, respectively. Thus, as expected based on studies in other groups (SanMiguel and Bennetzen, 1998; Bennetzen, 2002; Kidwell, 2002; Ramakrishna et al., 2002), differential TE accumulation appears to account for a large fraction of genome size evolution.

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

Illegitimate recombination has been demonstrated to have a profound effect countering genome size expansion in certain plants (Devos et al., 2002; Ma et al., 2004), and has been suggested to have influenced Gossypium genomes (Grover et al., 2004). Although the present study was able to polarize only a small number of indels as insertions or deletions via illegitimate recombination, a substantial body of unpolarized sequence data reveals the hallmarks of illegitimate recombination, particularly in the A$_T$ genome. The gaps represented by these events contribute, in a large part, to the total gypsy element length difference between A and A$_T$.

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

Although the polarized and unpolarized gap data suggest an indel bias exists in Gossypium, this bias cannot currently be described as acting homogeneously in all genomic regions. In particular, we note that in our previous study involving the CesA region (Grover et al., 2004), comparative sequencing of approximately 100 kb found the distribution of indels, with respect to size and frequency, to be equivalent for the A$_T$ and D$_T$ genomes. Thus, the mechanisms involved in generating the indel bias in Gossypium do not act homogeneously among genomic regions, but instead appear to be affected by regional dynamics. Certain mechanisms that have the ability to generate small indels, such as illegitimate recombination, may be modulated by locally operating genomic forces such as recombination rate or degree chromatin condensation, thus possibly explaining a locally operating indel bias.

Genome evolution in polyploid cotton

Polyploid formation is known to be accompanied by myriad genomic and genetic alterations, which have been the subject of a number of recent reviews (Adams and Wendel, 2005; Chen and Ni, 2006). Evidence suggests that polyploid genomes need not be additive with respect to parental genome sizes, but instead are often slightly less than the combined parental genome size (Spotts and Spotts, 1999; Ozkan et al., 2003; Bennett and Leitch, 2005b). To date, there is little information on the dynamics of genomic downsizing in polyploid genomes (Chantret et al., 2005; Gu et al., 2006). A conclusion of the present study is that in the AdhA region there has been genomic downsizing in the polyploid relative to its diploid progenitors. Of 121 ‘small’ gaps in the alignment, a greater number were in A$_T$ than in A (64 versus 28; $P < 0.0002$), as well as in D$_T$ than in D (64 versus 56), although in the latter comparison the difference is not statistically significant. In addition, the total amount of sequence attributable to transposable elements in the BACs from the polyploid was less than the sum from the homologous regions in the diploid progenitors. This was primarily a result of the insertion of a unique copia element in the D genome, but was counteracted in the A$_T$ genome by a unique gypsy insertion. Excluding the unique gypsy insertion, the solo-LTR, and the region of the third gypsy truncated by the end of the A genome BAC, the total length of gypsy elements in A$_T$ remains only approximately 65% the length of the A genome gypsy elements. This is largely because of several large gaps in the A$_T$ gypsy elements, many of which had the hallmarks of illegitimate recombination. This mirrors the results of several studies in wheat, which suggest that the evolution of genomic structures observed in polyploid wheats are largely the result of opposing influences of insertions caused by TE activity and deletions mediated through illegitimate recombination (Chantret et al., 2005) (Gu et al., 2006). Taken together, these studies suggest that increased illegitimate recombination may be a general consequence of polyploidization. Additional studies of Gossypium as well as other plant polyploids will be necessary to test the generality of this conclusion.
Finally, the present study provides an example of pseudogenization following polyploid formation in cotton, a rare fate for genes duplicated by polyploidy in the cotton genome (Cronn et al., 1999). A mutation in the \( D_T \) copy of the integral membrane protein-encoding gene caused a premature stop codon to arise halfway through the coding region, resulting in a truncated protein (182 versus 368 aa). Interestingly, this pseudogene was not the only one uncovered in the region. An ancient myosin pseudogene was shared between all genomes, and the original caffeic acid encoding gene in \( A_T \) (Table 1, gene 7) may also be silenced as a pseudogene (versus possessing an altered intron/exon structure for the last intron/exon junction). Nonetheless, the pseudogene discovered here adds a genomic example of gene silencing to an accumulating data set demonstrating expression-level changes and subfunctionalization of duplicated genes in \( Gossypium \) polyploids (Adams et al., 2003, 2004; Udall et al., 2006).

**Experimental procedures**

**BAC library screening and BAC selection**

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

**Shotgun sequencing, assembly and analysis**

\( Escherichia coli \) genomic DNA-free BAC plasmid DNA was sheared using a HydroShear (GenomeMachines; Genomic Solutions, http://www.genomicsolutions.com) DNA shearing device at speed code 12 with 25 cycles at 22°C. Fragmented DNA was end repaired using the ‘End-it’ DNA end repair kit (Epicentre, http://www.epibio.com), separated on an agarose gel, and size-selected for a range of 2 to 6 Kb. This prepared insert DNA was randomly cloned into a pBluescript II KS+ vector (Strategene, http://www.stratagene.com) and sequenced with the universal vector primers T7 and T3 to an average depth of \( 8 \times \). The resulting sequences were base-called using the program PHRED (Ewing and Green, 1998; Ewing et al., 1998), vector sequences were removed by CROSS_MATCH (Ewing and Green, 1998; Ewing et al., 1998), and assembled by the program PHRAP (Green, 1999). Contigs were visualized and edited with CONSED (Gordon et al., 1998). The output from these \( ab \) \( initio \) gene prediction programs, FGENESH (Softberry, http://www.softberry.com/), GENEMARK.HMM (Lukashin and Borodovsky, 1998) and GENSCAN+ (Burge and Karlin, 1997), was used as input for BLASTP (Altschul et al., 1997) searches against the non-redundant GenBank protein database. In addition, 500-bp segments of the sequence were subjected to BLASTX queries against the non-redundant GenBank protein database, and BLASTN queries against the cotton EST database (Udall et al., 2006). Repetitive element prediction was accomplished through RepeatMasker (http://www.repeatmasker.org), CENSOR (Jurka et al., 1996), and BLAST identity to known elements in REBASE (version 8.5) (Jurka, 2000) and GenBank. Each BAC was again queried in 500-bp fragments against whole-genomic shotgun (WGS) sequences representing approximately 0.1% of each of the four cotton genomes to uncover repetitive sequences of unknown origin (Hawkins et al., 2006).

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

**Gap polarization**

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

**Acknowledgements**

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

**Supplementary material**

The following supplementary material is available for this article online:

Table S1 Types and frequency of mechanisms contributing to genome size change in the AdhA region. This material is available as part of the online article from http://www.blackwell-synergy.com

**References**


© 2007 The Authors


SanMiguel, P. and Bennetzen, J.L. (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergenic retrotransposons. Ann. Bot. 82, 37–44.


An accession numbers submission in progress.