The Plant Journal (2007) **50**, 995–1006 doi: 10.1111/j.1365-313X.2007.03102.x

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

Corrinne E. Grover[1], HyeRan Kim[2,†], Rod A. Wing[2], Andrew H. Paterson[3] and Jonathan F. Wendel[1,*]
[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA,*
[2]*Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, and*
[3]*Plant 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; emai 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 polyteny (Thomas, 1971), genome size evolution is now recognized as reflecting the net effects of a suite of mechanisms that sometimes work antagonistically to expand and contract the genome. Best understood are the array of mechanisms responsible for genome size expansion, most prominently polyploidy (Wendel, 2000) and transposable element amplification (Bennetzen, 2000, 2002; Kidwell, 2002; Piegu *et al.*, 2006), but also smaller scale processes such as tandem repeat expansion (Ellegren, 2002; Morgante *et al.*, 2002), gene duplication and pseudogenization (Zhang, 2003), organellar transfer to the nucleus (Shahmuradov *et al.*, 2003), and intron size expansion (Deutsch and Long, 1999; Vinogradov, 1999). Less is known about mechanisms of genome size contraction, of which unequal intrastrand homologous recombination (Shepherd *et al.*, 1984; San-Miguel *et al.*, 1996; Chen *et al.*, 1998; Vicient *et al.*, 1999; Shirasu *et al.*, 2000), double-strand break repair (Kirik *et al.*, 2000; Orel and Puchta, 2003), and illegitimate recombination

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

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

To continue to investigate the patterns and processes responsible for genome size evolution in *Gossypium*, we report 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 agronomically important *G. hirsutum*, the genomes of which are slightly less than additive with respect to their diploid progenitors (Hendrix and Stewart, 2005).

We present here an analysis of comparative sequencing of a BAC-sized region surrounding the alcohol dehydrogenase A gene (*AdhA*), from two diploid species representing the closest living relatives of the A- and D-genome species involved in the allopolyploidization event (reviewed in Wendel and Cronn, 2003), as well as from both homoeologous genomes (A_T and D_T) from the tetraploid, *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.

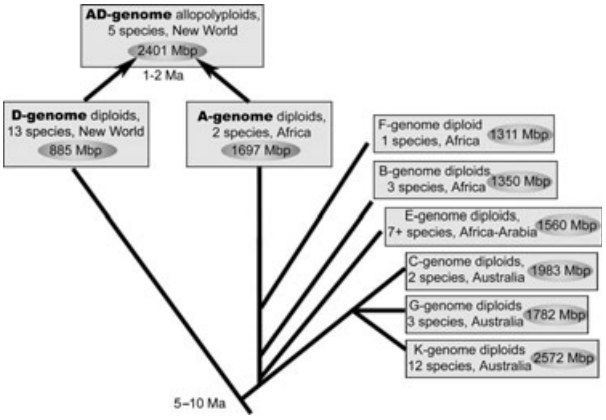

Figure 1. The evolutionary history of diploid and tetraploid *Gossypium* species groups ($n = 13$ and 26, respectively), as inferred from multiple molecular datasets (Seelanan *et al.*, 1997; Small *et al.*, 1998; Cronn *et al.*, 2002). The eight diploid genome groups, determined by interspecific meiotic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

Results

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

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

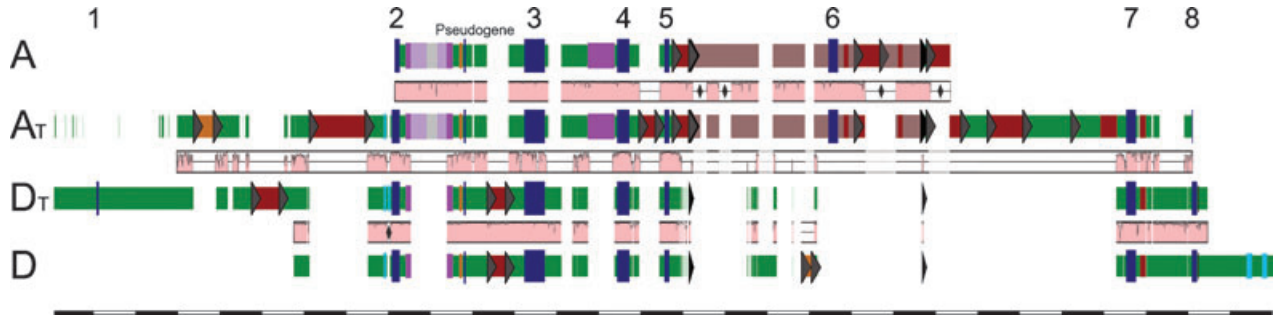

Figure 2. Multiple alignment of orthologous *AdhA* BACs from four different genomes (A, D, A_T and D_T; the latter two are co-resident in the nucleus of polyploid cottons). Numbered blue boxes are predicted genes corresponding to the list presented in Table 1; *copia* elements are in orange, *gypsy*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

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 A_T, 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 A_T sequences; these are unequally distributed as 28 gaps in the A sequence (151 bp) and 64 gaps in the A_T 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 A_T 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 A_T genome (between genes 4 and 5; Figure 2), accounting for 4799 bp, which by its exclusivity is inferred to have been inserted since the origin of the polyploids.

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

The *AdhA* BACs from the D genome diploid (101.3 kb) and the D_T genome from the allotetraploid (130.9 kb) were also shotgun sequenced and assembled. The aligned length of the two genomes was 86.7 kb, accounting for 85.7 kb in D and 80 kb in D_T, again indicating a size differential between the diploid and polyploid that 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 D_T, 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 D_T genome BACs had GC contents of approximately 33.6% and were determined to be 98.2% identical in sequence (89.38% including gaps). A total of 121 phylogenetically unpolarized gaps (i.e. gaps that were not distinguishable as insertions or deletions, see Experimental procedures) differentiate the D and D_T genomes, distributed as 57 gaps in D (943 bp) and 64 gaps in D_T (699 bp), and again reflecting indels that arose since polyploidization. When large gaps are excluded (>400 bp; Figure 2) the number and length of gaps in D reduces to 56 gaps (309 bp), whereas the number and length in D_T remains the same. A single *copia* insertion in the D genome (between genes 5 and 6; 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 A_T, 55 kb in D and 49 kb in D_T. The size differential between the A/A_T genomes and the D/D_T 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 A_T genome (Table 1; Figure 2). The pairwise comparison of A BACs with D BACs, irrespective of origin (diploid versus tetraploid), gave an average of 92% sequence identity (91.97–92.01%; 28.6–32.9% including gaps).

As previously reported for *Gossypium* (Grover *et al.*, 2004), the intergenic space was remarkably conserved

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

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

*Truncated by BAC end or does not exist because BAC ends.
'A' has internal stop in D_T, may be pseudogenized.
'B' Expressed sequence tag (EST) support for D/D_T mRNA only.

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 (χ^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 A_T BACs was approximately half of that observed in the D and D_T BACs.

Each indel was assigned a probable mechanism of origin (Table S1), regardless of whether the indel was polarized. Transposable element insertions and probable insertions account for the majority of sequence difference between the four genomes, representing over half of the alignment for the A genomes. Illegitimate recombination, a RecA independent form of recombination involving regions of microhomology (2–15 bp) flanked by short direct repeats, incorporates a variety of mechanisms, most notably double-stranded break (DSB) repair and slipstrand annealing. Both were common in the alignment, occurring mostly within transposable elements. A majority (50.42%) of indels were classified as having been generated by an 'unknown mechanism', 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 D_T genomes, determined by using A and A_T to approximate the ancestral state, was similar for each genome (six insertions and 13 deletions versus six insertions and 20 deletions in the D and D_T genomes, respectively); however, this was not the case for the A and A_T genomes (where D and D_T represent the ancestral state), where we inferred 12 insertions and four deletions in the A genome, but two insertions and 13 deletions in the A_T genome. Thus,

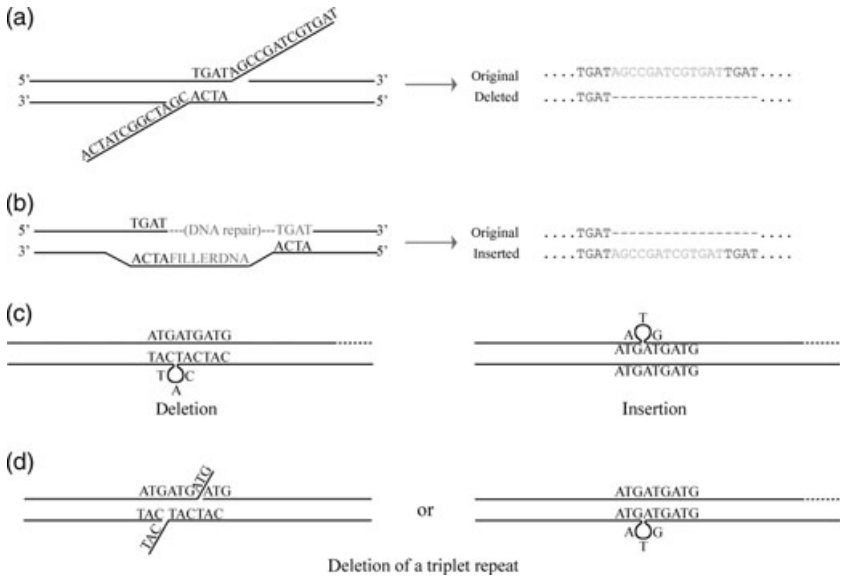

Figure 3. Illegitimate recombination represents several different mechanisms leading to the deletion of a sequence bounded by small repeats (only 1 bp of homology required), as well as one of the bounding repeats, or, less commonly, the addition of an intervening sequence. Three mechanisms are thought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

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 D_T and A_T, 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 A_T and D, and 1.8 nt for D_T), whereas the average deletion size in the A genome mirrored the average deletion size in A_T (2 and 1.9 nt, respectively), and was approximately half the average deletion size in D/D_T (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

Mechanism	Type	*G. arboreum, A*	*G. hirsutum, A_T*	*G. hirsutum, D_T*	*G. raimondii, D*
Overall	#Deletions	4	13	20	13
	nt deletions	8	29	77	91
	#Insertions	12	2 (1)	6	6 (4)
	nt insertions	34	4799 (1)	11	5971 (4)
	#Unknown gaps (excluding TEs)	159	169	154	152
	nt missing (excluding TEs)	7809	21 715	15 440	15 404
Small indels (<400 bp) per 100 kb in the *AdhA* region	#Deletions	3.57	12.78	36.36	26.53
	nt deletions	7.14	28.52	140	185.71
	#Insertions	10.71	0.99	10.91	8.16
	nt Insertions	30.28	0.99	20	8.16

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

Analysis of putative genes

Six genes and one pseudogene are predicted to occur ancestrally in the *AdhA* region (Table 1). These six genes range in size from a 1.1-kb putative integral membrane protein-encoding gene to a 4.9-kb putative FAD-dependent oxidoreductase protein-encoding gene. The structures of four of the six genes were confirmed fully by 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 D_T genome. The matching EST is derived from a D-genome library, indicating transcription at the diploid level, and extends past the point in which the D_T genome has acquired a stop codon. This pseudogenization is inferred to be relatively recent, as no acceleration in non-synonymous mutations is observed (K_a A–D = 0.0024; K_a A–D_T = 0.0024). A conserved-domain search (Marchler-Bauer and Bryant, 2004) indicated that this unknown gene bears a slight similarity (*E*-value = $2e^{-6}$) to nucleotide-sugar transporters.

A single gene duplication, involving a putative caffeic acid *O*-methyltransferase-encoding gene, differentiates the *AdhA* BACs of the A/A_T genomes from those of the D/D_T genomes. By virtue of its shared presence in the former two genomes, and its absence from the latter two genomes, we infer that the duplication event happened subsequent to the divergence of the A and D genome diploids from their common ancestor 5–10 Ma, but prior to polyploid formation 1–2 Ma. The duplicate falls within a block of several nested *gypsy* elements (full length as well as remnant) present in both the A and A_T genomes. Interestingly, the predicted intron/exon structure of the duplicate in the A genomes more closely resembles the structure found in the D genomes than its syntenic copy, primarily because of the predicted compensatory intron/exon boundary changes in the original A_T copy necessary to restore function in response to a 22-bp frame-shifting insertion (Figure 4). Alternatively, the original copy of the caffeic acid *O*-methyltransferase encoding gene may be pseudogenized by the 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/A_T genomes), a *copia*-like pol fragment (820 bp), and possible long terminal repeats (LTRs) of an ancient retroelement, representing transposable element insertions that occurred prior to or concurrent with the origin of the genus (and hence are not relevant to genome size evolution within the genus). Two TEs in the D/D_T genomes

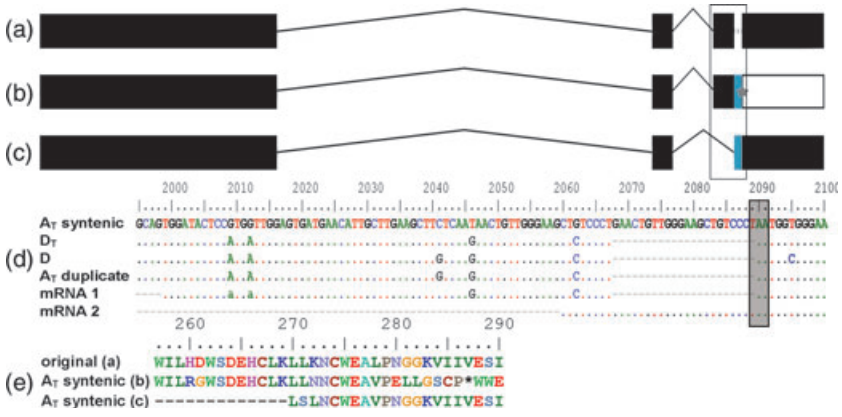

Figure 4. Possible splicing of the putative caffeic acid encoding genes in the *AdhA* region.
(a) Structure of the putative caffeic acid encoding gene in D, D_T and the duplicate (non-syntenic) copy in A_T. The location of the insertion in the syntenic A_T copy is noted by a dotted line.
(b) Structure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

Table 3 Repetitive element lengths in diploid and polyploid cotton

	A	A_T	D_T	D
Repetitive sequence				
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	6333	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
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	0	2029	2005
Unknown repetitive	0	0	2339	2334
Gypsy-like repetitive	*	0	1983	1972
Limited similarity to Calcineurin-like Phosphoesterase and Mutator-like elements	*	*	*	3009
Unknown repetitive	*	*	*	1941
Unknown repetitive	*	0	674	*
GORGE3 gypsy-like repetitive	*	0	584	*
Unknown repetitive	*	0	314	*
GORGE3 gypsy-like repetitive	*	0	1423	*
Unknown repetitive	*	0	3215	*
Unknown repetitive	*	0	219	*
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.

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

The D and D_T genomes share a *gypsy* element insertion of approximately 5 kb in length (*gypsy*D2, 4.8 kb in D and 5.1 kb in D_T; between genes 2 and 3 of Figure 2). The LTRs are approximately 98 and 97% identical (excluding gaps) in D and D_T, respectively. Of the 11 mutations in the LTRs, 10 have occurred since the D/D_T divergence (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 (*copia*D; 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 (*gypsy*A4; possibly *Gorge1* or *Gorge3*), and one full-length element in A that is represented by only a solo-LTR in A_T (*gypsy*A6; possibly *Gorge3*). Characterization of the particular family to which each element belongs was made possible by a larger survey of cotton repetitive sequences (Hawkins *et al.*, 2006). The range in full-length element size is from 8.1 to 16.7 kb, and all LTRs are approximately 95% identical (within elements). For the three discrete *gypsy* elements examined, the orthologous elements in the A_T genome were smaller than their A genome counterparts, by a minimum of 20%. In addition, the A_T genome contains a 4.8-kb unique LTR-retrotransposon of probable *gypsy* origin (*gypsy*A2). Overall, aside from being more abundant in the A genomes, intact *gypsy* elements were larger than those found in the D genomes. The largest *gypsy* represented in the D genomes was still smaller than the smallest intact *gypsy* in the A genomes, and less than half the size of the largest (7.5 kb in D versus 8.1 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

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–D_T]. The divergence of the unassigned intergenic space between the A and D genomes was nearly identical, regardless of which A and D genome were compared (0.058 for A–D, A–D_T, A_T–D_T and 0.059 for A_T–D), and these values were over double the divergence calculated from nuclear genes (0.022 A–D and 0.024 A_T–D_T; Senchina *et al.*, 2003), possibly indicating the presence of previously (and perhaps currently) rapidly evolving, severely degraded TEs that are unrecognizable.

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

Unidentified repetitive DNA Repetitive sequences not assigned to a class were uncovered through BLAST identity to repetitive whole-genome shotgun sequences of unknown origin. These did not substantially contribute to the alignment, representing approximately 5.7 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 intrastrand 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 counteracting genome size expansion in certain plants (Devos *et al.*, 2002; Ma *et al.*, 2004), and has been suggested to have influenced *Gossypium* genomes (Grover *et al.*, 2004). Although the present study was able to polarize only a small number of indels as insertions or deletions via illegitimate recombination, a substantial body of unpolarized sequence data reveals the hallmarks of illegitimate recombination, particularly in the A_T genome. The gaps represented by these events contribute, in a large part, to the total *gypsy* element length difference between A and A_T.

A bias in the formation of small indels has been implicated in genome size differences (Petrov *et al.*, 1996; Kirik *et al.*, 2000; Petrov *et al.*, 2000; Petrov, 2002; Orel and Puchta, 2003), but has not been demonstrated to date for cotton (Grover *et al.*, 2004). The limited polarized indel data available indicate a possible insertional bias, which suggests that the A genome is more prone to insertions than the other genomes and that it is the only genome where small insertions outweigh small deletions. Furthermore, the polarized deletions suggest that a deletional bias exists between A/A_T and D/D_T, with small deletions occurring more frequently and of greater average length in the D genomes. The polarized indels represent insertion and deletion events occurring since polyploid formation and, when extrapolated to the entire genome, indicate that a bias in small indels could be responsible for adding several hundred kb to the A genome and removing several hundred kb (in increasing amounts) from the A_T, D and D_T genomes in the last 1–2 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 then the D genome, representing a propensity for insertions in A, deletions in D, or, a combination of these two processes, as reflected in the polarized gap data.

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 (Soltis and Soltis, 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 (GeneMachines; 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×. The resulting sequences were base-called using the program PHRED (Ewing and Green, 1998; Ewing *et al.*, 1998), vector sequences were removed by CROSS_MATCH (Ewing and Green, 1998; Ewing *et al.*, 1998), and assembled by the program PHRAP (Green, 1999). Contigs were visualized and edited with CONSED (Gordon *et al.*, 1998). The output from three *ab initio* gene prediction programs, FGENESH (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 REPBASE (version 8.5) (Jurka, 2000) and GenBank. Each BAC was again queried in 500-bp fragments against whole-genomic shotgun (WGS) sequences representing approximately 0.1% of 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

Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4649–4654.

Adams, K.L., Percifield, R. and Wendel, J.F. (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168**, 2217–2226.

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

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

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

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

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

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269.

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

Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95**, 127–132.

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

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

Cavalier-Smith, T. (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot.* **95**, 147–175.

Chantret, N., Salse, J., Sabot, F. *et al.* (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell*, **17**, 1033–1045.

Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A. and Bennetzen, J.L. (1997) Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. *Proc. Natl Acad. Sci.* **94**, 3431–3435.

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

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

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

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

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

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

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

Endrizzi, J.D., Turcotte, E.L. and Kohel, R.J. (1985) Genetics, cytology, and evolution of *Gossypium*. *Adv. Genet.* **23**, 271–375.

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

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

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

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

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

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

Gregory, T.R. (2005) The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* **95**, 133–146.

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

Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H. and Wendel, J.F. (2004) Incongruent patterns of local and global genome size evolution in cotton. *Genome Res.* **14**, 1474–1482.

Gu, Y.Q., Salse, J., Coleman-Derr, D. *et al.* (2006) Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes. *Genetics*, **174**, 1493–1504.

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

Hendrix, B. and Stewart, J.M. (2005) Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* **95**, 789–797. .%R 10.1093/aob/mci078.

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

Jurka, J. (2000) Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.* **9**, 418–420.

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

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

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

Knight, C.A., Molinari, N.A. and Petrov, D.A. (2005) The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* **95**, 177–190.

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

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

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

Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194–200.

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

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

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

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

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

Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17**, 23–28.

Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.*, **61**, 531–544.

Petrov, D.A. and Wendel, J.F. (2006) Evolution of eukaryotic genome structure. In *Evolutionary Genetics: Concepts and Case Studies* (Fox, C.W. and Wolf, J.B. eds). Oxford University Press, USA, pp. 144–156.

Piegu, B., Guyot, R., Picault, N. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269. %R 10.1101/gr.5290206.

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

SanMiguel, P., Tikhonov, A., Jin, Y.K. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

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

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

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

Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J.K., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A. and Wendel, J.F. (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20**, 633–643.

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

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

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

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

Small, R.L., Ryburn, J.A. and Wendel, J.F. (1999) Low levels of nucleotide diversity at homoeologous Adh loci in allotetraploid cotton (*Gossypium* L.). *Mol. Biol. Evol.* **16**, 491–501.

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

Thomas, C.A. (1971) The genetic organisation of chromosomes. *Ann. Rev. Genet.* **5**, 237–256.

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

Udall, J.A., Swanson, J.M., Haller, K. *et al.* (2006) A global assembly of cotton ESTs. *Genome Res.* **16**, 441–450.

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

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

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

Vitte, C. and Panaud, O. (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* **20**, 528–540.

Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L. and Senchina, D.S. (2002) Intron size and genome size in plants. *Mol. Biol. Evol.* **19**, 2346–2352.

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

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

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

Zhang, J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298.

Accession numbers: submission in progress.

