Heredity (2010) 105, 463–472

www.nature.com/hdy

ORIGINAL ARTICLE

Gene copy number evolution during tetraploid cotton radiation

J Rong[1,2], FA Feltus[1,3], L Liu[1], L Lin[1] and AH Paterson[1]
[1]*Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA;* [2]*School of Agriculture and Food Science, Zhejiang Forestry University, Lin'an, Hangzhou, PR China and* [3]*Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA*

After polyploid formation, retention or loss of duplicated genes is not random. Genes with some functional domains are convergently restored to 'singleton' state after many independent genome duplications, and have been referred to as 'duplication-resistant' (DR) genes. To further explore the timeframe for their restoration to the singleton state, 27 cotton homologs of genes found to be 'DR' in *Arabidopsis* were selected based on diagnostic Pfam domains. Their copy numbers were studied using southern hybridization and sequence analysis in five tetraploid species and their ancestral A and D genome diploids. DR genes had significantly lower copy number than gene families hybridizing to randomly selected cotton ESTs. Three DR genes showed complete loss of D genome-derived homoeologs in some or all tetraploid species. Prior analysis has shown gene loss in polyploid cotton to be rare, and herein only one randomly selected gene showed loss of a homoeolog in only one of the five tetraploid species (*Gossypium mustelinum*). BAC sequencing confirmed two cases of gene loss in tetraploid cotton. Divergence among 5′ sequences of DR genes amplified from *G. arboreum*, *G. raimondii*, and *Gossypioides kirkii* was correlated with gene copy number. These results show that genes containing Pfam domains associated with duplication resistance in *Arabidopsis* have also been preferentially restored to low copy number after a more recent polyploidization event in cotton. In tetraploid cotton, genes from the progenitor D genome seem to experience more gene copy number divergence than genes from the A genome. Together with D subgenome-biased alterations in gene expression, perhaps gene loss may contribute to the relatively larger portion of quantitative trait variation attributable to D than A subgenome chromosomes of tetraploid cotton.
Heredity (2010) **105,** 463–472; doi:10.1038/hdy.2009.192; published online 17 February 2010

Keywords: genome duplication; duplication-resistant gene; evolution; polyploidy; cotton

Introduction

Whole genome duplication occurred during the evolution of a wide range of organisms. The discovery that one polyploidization event may predate the monocot-eudicot divergence (Bowers *et al*., 2003) suggests that all angiosperms are ancient polyploids. Whole genome duplication or large segmental duplication is followed by gene loss, gene functional divergence, gene movement, and chromosome structural changes such as translocation and inversion, collectively forming a process known as diploidization. This process presumably contributes to eventual restoration of polyploids to a diploid-like genome structure after a long period of evolution. Diploidization-associated changes obscure our ability to detect syntenic regions and trace the fates of duplicated genes with increasing evolutionary time. New computational methods improve our ability to deduce relationships among paleopolyploid genomes and take early steps toward inferring the genome composition and organization of their ancestors (Tang *et al*., 2008a, b).

The ability to 'synthesize' newly polyploid plants by artificial crosses and chromosomal manipulation using colchicine has revealed striking immediate reactions of genomes to duplication. These reactions include loss and restructuring of low-copy DNA sequences (Song *et al*., 1995; Feldman *et al*., 1997; Ozkan *et al*., 2001, 2002; Shaked *et al*., 2001; Kashkush *et al*., 2002), activation of genes and retrotransposons (O'Neill *et al*., 2002; Kashkush *et al*., 2003), gene silencing (Chen and Pikaard, 1997a, b; Comai *et al*., 2000; Lee and Chen, 2001), and subfunctionalization of gene expression patterns (Adams *et al*., 2003, 2004; Samuel Yang *et al*., 2006; Hovav *et al*., 2008). These responses are closely paralleled in animals by inactivation (Lee and Jaenisch, 1997; Avner and Heard, 2001) and differential regulation of X-chromosome gene expression (Disteche *et al*., 2002; Parisi *et al*., 2003) and retroelement activation (O'Neill *et al*., 2002).

Retention/loss of duplicated gene copies in lineages that survive gene duplication is not random. In angiosperms, we find three 'fates' of individual gene pairs after duplication:

1. *Most gene functional groups show post-duplication gene preservation/loss rates that are indistinguishable from the genome-wide average*. Such 'neutral' loss of duplicated

Correspondence: Dr AH Paterson, Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Road, Rm 228, Athens, GA 30602, USA.
E-mail paterson@uga.edu
Received 29 August 2009; revised 24 November 2009; accepted 30 November 2009; published online 17 February 2010

genes presumably involves inactivating mutations opposed by very weak selection (Haldane, 1933). Population genetic models predict that this may happen over a few million years (Lynch and Conery, 2000), with a prediction supported by recent comparisons of genomes that diverged millions of years after a shared duplication (Paterson *et al*., 2009).

2. *Genes in some specific functional categories duplicate and reduplicate.* Genes duplicated by gene duplication survive much longer than those duplicated individually (Lynch and Conery, 2000), and some gene functional groups are preferentially preserved in duplicate (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere *et al*., 2005; Chapman *et al*., 2006; Paterson *et al*., 2006; Tang *et al*., 2008b). Coding regions of genes preserved in duplicate tend to be functionally complex (Chapman *et al*., 2006), under purifying selection (Brunet *et al*., 2006; Chapman *et al*., 2006), and may evolve in concert (Gao and Innan, 2004; Wang *et al*., 2007). However, regulatory divergence between members of preserved gene pairs may contribute much to morphological complexity (Freeling and Thomas, 2006), perhaps offering important benefits to polyploidized lineages (Comai, 2005). Classical ideas about one gene copy diverging to new function (neofunctionalization—Stephens, 1951; Ohno, 1970) have now been tempered by findings that many duplicated genes may subdivide ancestral functions (subfunctionalization—Lynch and Force, 2000). Subfunctionalization may be a stepping stone to neofunctionalization (He and Zhang, 2005).
3. *Some specific genes and gene functional groups show more extensive loss of duplicate copies than the genome-wide average, and this loss is often convergent after independent duplications separated by hundreds of millions of years*. Some gene functional groups are preserved in duplicate significantly *less* frequently than the genome-wide average (Paterson *et al*., 2006). This observation alone might be viewed as noise—among thousands of functional groups, some must incur more gene loss than others because of random factors alone. However, across thousands of gene functional groups, frequencies of loss of duplicated gene copies are closely correlated after independent gene duplication events that occurred $\sim$60 and $\sim$70 MYA in the lineages of Arabidopsis (α) (Bowers *et al*., 2003) and Oryza (Paterson *et al*., 2004), at statistical probabilities that essentially rule out false positives. The fates of thousands of duplicated genes are correlated ($r = 0.6$) in the two species (Paterson *et al*., 2006). Post-duplication convergence of gene copy number is also found in divergent yeasts (Scannell *et al*., 2006), and genes from the same metabolic pathway show similar retention/loss trends in Paramecium (Aury *et al*., 2006). Repeated restoration of certain genes to singleton status at a greater-than random frequency suggests that an underlying set of principles of molecular evolution may contribute to the fates of gene and genome duplications (Paterson *et al*., 2006).

Population genetic models predict that the majority of gene loss after whole genome duplication should be relatively rapid, with duplicated genes having a half-life of about 4 MYA (Lynch and Conery, 2000). If restoration of duplication-resistant (DR) genes to singleton status is important to the success of polyploid lineages, then losses (or loss of function alleles) of duplicate copies of these genes should reach fixation relatively rapidly. Even so, one could envision this loss happening on various time scales. Duplicated copies of most DR genes presumably impose only a modest genetic load individually (otherwise the collective load of the hundreds of such genes thought to exist would be so high that new polyploids would not survive). However, a few might cause fundamental functional problems such that only those lineages that very rapidly silence one of the duplicated copies can survive. Recent empirical studies of neopolyploid *Tragopogon* spp. formed about 80 years ago revealed that 3.2% of homoeologs had been lost and a further 3.4% had been silenced. The homoeolog losses and silencing events found were not fixed within natural populations, and did not form a predictable pattern between populations. This suggests that genome evolution after polyploid formation is highly dynamic, leading to a low rate of haphazard homoeolog loss that is far from complete at this early stage (80 years) after polyploidization (Buggs *et al*., 2009).

The well-understood phylogenetic relationships among five tetraploid Gossypium (cotton) species, and the availability of many diverse wild accessions for each of these species, provide an attractive system in which to clarify the tempo of DR gene loss and permits one to distinguish ancient evolutionary events (monomorphic in population samples) from recent ones (polymorphic in population samples) in each lineage. Indeed, one of the first to appreciate the importance of genome duplication in the evolution was a cotton geneticist (Stephens, 1951). *Gossypium* (Malvaceae) includes 45 diploid and five allopolyploid species, the latter including *G. hirsutum*, the leading commercial cotton. The five allopolyploids each contain two divergent subgenomes, A and D, which presently exist in diploid species (both $2n = 26$) in different hemispheres. A genome diploids are African in origin, whereas D genome diploids are primarily Mexican. A and D genome groups are estimated to have diverged from a common ancestor 5–10 MYA (Senchina *et al*., 2003), then been reunited through polyploidization in an A genome cytoplasm (Wendel, 1989; Small and Wendel, 1999) about 1–2 MYA (Wendel and Cronn, 2003) after trans-oceanic dispersal to the New World of an A genome propagule closely resembling the extant species *G. herbaceum*. After hybridization with a native D genome diploid resembling *G. raimondii* and chromosome doubling, the polyploid spread throughout the American tropics and subtropics, radiating into different lineages, now represented by three clades and five species.

Southern blot analysis of thousands of randomly sampled cDNAs (Reinisch *et al*., 1994; Rong *et al*., 2004) and detailed comparison of diploid and tetraploid versions of hundreds of genes (Senchina *et al*., 2003; Udall *et al*., 2006) show that loss of random gene copies in tetraploid cotton is exceedingly rare. Loss of even a few DR gene candidates would stand out as significant.

We now know that the lineage of *Arabidopsis* has experienced two genome duplications as divergence from a common ancestor shared with cotton (Ming *et al*., 2008; Tang *et al*., 2008a), accounting for the majority of *Arabidopsis* gene duplications (Bowers *et al*., 2003). In this same time period, cotton has independently experienced a paleoduplication (Rong *et al*., 2005) and the

well-known polyploid formation. Nonetheless, cotton and *Arabidopsis* are relatively closely related in the angiosperm phylogeny (Bowers *et al*., 2003; Rong *et al*., 2005). Correlated patterns of retention/loss of duplicated genes in the *Arabidopsis* and cotton genomes would lend further support to the notion that an underlying set of principles of molecular evolution may contribute to the fates of gene and genome duplications (Paterson *et al*., 2006). More importantly, careful scrutiny of specific gene loss events in the context of the well-defined tetraploid cotton phylogeny may shed light on the timing of this dimension of adaptation by a genome to the duplicated state.

In this research, we investigated copy number evolution of cotton homologs of *Arabidopsis* DR genes. After the study of the copy number variation of candidate DR genes among different diploid A and D genome cotton species as well as five tetraploid species, we explored the evolutionary fate of these genes during speciation and radiation of tetraploid cottons.

Materials and methods

Plant materials

Cotton species used in this study are listed in Supplementary Table 1 including diverse representatives from all five tetraploid cotton species: four from *G. mustelinum* (Gm), six *G. darwinii* (Gd), six *G. barbadense* (Gb), six *G. tomentosum* (Gt), six *G. hirsutum* (Gh), and from their putative diploid ancestor species, five each from *G. arboreum* (Ga, A2) and *G. herbaceum* (Gh, A1), six from *G. raimondii* (Gd, D5) and two from *G. gossypioides* (Gg, D6). Seeds were generously provided by the USDA-ARS collection (College Station, TX, USA), J Wendel, and G Mergeai.

Determination of cotton homologs of *Arabidopsis* DR genes and polymerize chain reaction primer design

Ninety five *Arabidopsis* DR genes (Paterson *et al*., 2006) were used to search for homologs in cotton ESTs downloaded from NCBI, including *G. arboreum* fiber cDNA, and *G. raimondii* seedling and floral cDNA (Udall *et al*., 2006). Those cotton ESTs matching the *Arabidopsis* genes with alignment length $\geqslant$100 bp and e-value $<10^{-20}$ were checked for diagnostic Pfam domains (Paterson *et al*., 2006). The resulting cotton ESTs were used to design polymerize chain reaction (PCR) primers for which amplicons would include the Pfam domain and surrounding exon regions. The selected cotton DR candidate genes were compared with corresponding *Arabidopsis* genomic DNA sequences to find possible locations of introns, with the aim of designing exonic primers that amplify across introns (Feltus *et al*., 2006).

DNA extraction, probe preparation, and southern hybridization

Young leaves were collected from plants grown in the greenhouse. DNA extraction, restriction digestion, gel electrophoresis, southern blotting, probe labeling, southern hybridization, and autoradiography followed Rong *et al*. (2004). To determine the fragment number and polymorphism between the putative diploid ancestral species, DNAs from the two A genome (*G. arboreum* and *G. herbaceum*) and one D genome species (*G. raimondii*) were digested with four restriction enzymes, *Eco*RI, *Eco*RV, *Hin*dIII, and *Xba*I, and arranged in one 'survey' blot. The DNAs from all races of diploid and tetraploid species digested with the respective enzymes were arranged in a 'garden blot'.

For preparation of probes, genomic DNA from *Gossyploides kirkii*, an outgroup species, was used as template for touchdown PCR as follows: denaturation at 95 °C for 3 min, then dropped to 94 °C for 20 s, annealing at 60 °C for 20 s, extension at 72 °C for 30 s, then denaturation at 94 °C. This cycle was repeated four times, reducing by 1 °C per cycle from 60 °C down to 56 °C, followed by regular PCR for 29 cycles at 94 °C denaturation for 20 s, annealing at 56 °C for 20 s and extension at 72 °C for 30 s. In the final cycle, extension prolonged at 72 °C for 20 min. The amplified PCR products were run in 1% agarose gel and the bands were cut and cleaned with a Qiagen kit. A subset (39) of cotton unigenes from *G. arboreum* immature fiber (Arpat *et al*., 2004) was selected randomly as a control for comparison with the candidate DR genes.

Amplified fragments were first hybridized to survey blots to determine the fragment number of the diploid progenitor species and identify polymorphism. A garden blot of additional diploid and tetraploid cottons listed above (Supplementary Table 1) digested with the enzyme that distinguished between diploid progenitor genomes was hybridized using the same probes.

Probes that show evidence of allele elimination in tetraploid species were hybridized to BAC libraries made from *G. arboreum* (GAMBO), *G. raimondii* (GR), *G. barbadense* (GAD), and *G. hirsutum* (MAXXA). The hybridizing BACs were digested with *Hin*dIII, blotted, and again hybridized with candidate DR probes. From each BAC library, one or more BACs were selected from each group with the same RFLP band pattern and used as template to produce amplicons for sequencing.

DNA sequencing

DNA amplified from BACs mentioned above or genomic DNA of *G. arboreum* (A2), *G. raimondii* (D5), and *G. kirkii* (K) using the indicated primers was sequenced as reported (Rong *et al*., 2004).

Results

Cotton homologs of *Arabidopsis* DR genes and their amplification

When 95 *Arabidopsis* DR genes were blasted against the cotton EST database, 49 corresponded to 229 cotton ESTs at alignment length $\geqslant$100 bp and e-value $<10^{-20}$. A total of 29 cotton ESTs, each a best match for a different *Arabidopsis* DR gene, contained the diagnostic Pfam domain and were used in this study. *Arabidopsis* genes, corresponding cotton ESTs, primer sequences, Pfam domain names, and their start and end points in *Arabidopsis* genes are summarized in Supplementary Table 2.

We estimated copy number variation of the candidate DR genes from southern blots, based on differences in restriction fragment number. Probes used were amplified from gDNA of *G. kirkii* with primers (Supplementary Table 2) designed from the 29 cotton homologs of *Arabidopsis* (DR) genes identified above. Among these,

23 were also expected to include the introns based on the comparison of cotton EST sequences to *Arabidopsis* genomic DNA. We obtained PCR products from 27 primers, excluding DRs 9 and 24.

Gene copy number in diploid cotton species

Most PCR amplicons from *G. kirkii* produced clear distinguishable bands in survey southern blots, except for DR18. Band numbers produced with four enzymes in each of the two A genome species (*G. arboreum* and *G. herbaceum*) and one D genome species (*G. raimondii*) were listed in Supplementary Tables 3 and 4, and summarized in Tables 1 and 2, for both candidate DR genes and randomly selected unigenes.

DR genes consistently have fewer bands than randomly selected genes across different cotton genotypes and restriction enzymes (Table 1). About half of the candidate DR genes had 1–2 bands in all genotypes and enzymes, and about 40% had 2–3 bands (Table 2). Among 27 amplified DR genes, 9 (33%) had only one band in at least one enzyme digest. In contrast, only 64.1% of randomly selected genes had band numbers <3 versus 92.6% of DR genes. A *t*-test showed that average band number was significantly lower in candidate DR genes than randomly selected genes ($P=0.0045$).

In contrast, no significant difference was observed between the two A genome species (*t*-test: $P=0.368$ and 0.392) in both DR genes and unigenes. *G. raimondii* (D5 genome) had more bands than the two A genome species for most genes, especially in the randomly selected group in which it had significantly more bands than *G. arboreum* (*t*-test: 0.033) and was close to significantly more bands than *G. herbaceum* (*t*-test: 0.064).

Preferential elimination of some candidate DR genes during tetraploid cotton speciation and radiation

To compare the fates of genes from A and D progenitors during the speciation and radiation of the tetraploid cottons, the garden blots were hybridized to specific DR gene probes that showed polymorphism between the two ancestral genomes (A and D) and the fewest bands of the four enzymes surveyed. In these garden blots, two *G. gossypioides* accessions (D6-1 and D6-2) were also included, but D6-1 was later excluded from analysis because of poor enzyme digestion. D6-2 showed slightly higher gene copy number (about 0.2 more bands on average) than the two A genome species for randomly selected genes, and the same average band numbers as A genome species in DR genes. All races of both *G. arboreum* and *G. raimondii* showed virtually the same band numbers in both randomly selected and DR genes. The *G. herbaceum* races have slightly different band number between the two groups of genes.

In the tetraploid species, 4.09–4.59 restriction fragments were found on average for randomly selected genes and 3.06–3.23 for DR genes (Table 3). Fragment number in tetraploid cotton is not additive of that in the two ancestral genomes, which had about 3.4 in D and 2.3 in A for randomly selected genes and 2.4 in D and 1.9 in A for DR genes. Randomly selected genes had significantly more fragments than DR genes for all races of all five tetraploid species (Supplementary Table 5). Among the five tetraploid species, *G. mustelinum* possesses about 0.5 fewer randomly selected gene bands than the other four. In DR genes, *G. mustelinum* has the same number or even more bands. As expected, polymorphism among different races within species was less than that among different species (Supplementary Table 5).

To further explore the hypothesis that genes containing 'DR' Pfam domains may be returned to singleton status soon after whole genome duplication, DR genes exhibiting only one band in surveys of band number and polymorphic between diploid species are of particular interest. A total of nine cotton DR genes (33.3%) had only one band detected in all diploid species in at least one enzyme digestion. Two of these genes (DRs 8 and 21) showed no polymorphism among diploid or tetraploid

Table 1 Average band number of cotton homologs detected by genes showing random versus duplication-resistant retention patterns in *Arabidopsis*, with each of four restriction enzymes

	Eco*RI*			Eco*RV*			Hin*dIII*			Xba*I*		
	Gh(A1)	*Ga(A2)*	*GR(D5)*	*Gh(A1)*	*Ga(A2)*	*GR(D5)*	*Gh(A1)*	*Ga(A2)*	*GR(D5)*	*Gh(A1)*	*Ga(A2)*	*GR(D5)*
Unig	2.59	2.42	3.10	2.09	2.06	2.37	2.50	2.70	2.95	2.06	1.94	2.37
DR	2.15	2.08	2.37	1.84	1.79	2.13	2.15	2.00	2.31	1.62	1.73	1.96
Unig-DR	0.44	0.34	0.73	0.25	0.27	0.24	0.35	0.70	0.64	0.44	0.21	0.41

Abbreviation: DR, duplication resistant.

Table 2 Homolog copy number distributions in cotton of genes showing random versus duplication-resistant retention patterns in *Arabidopsis*

Genes	*Total genes*	*One band*[a]		*1-1.99*[b]		*2-2.99*		*3-3.99*		*>4*	
		Gene	*%*	*Gene*	*%*	*Gene*	*%*	*Gene*	*%*	*Gene*	*%*
Unig	39	9	23.08	12	30.77	13	33.33	7	17.95	7	17.95
DR	27	9	33.33	14	51.85	11	40.74	2	7.41	0	0

Abbreviation: DR, duplication resistant.
[a]One band in at least one enzyme.
[b]Average band number.

species, providing no evidence about gene copy number evolution. Three genes (DRs 14, 16, and 20) showed the presence of bands from both A and D genomes in all races of five tetraploids, except for *G. hirsutum* accession TX45 for DR16 (Figure 1a), suggesting no gene loss after tetraploidization. In TX45, the band from D was lost, but a new band detected when DR16 was hybridized to the garden blot digested with *Xba*I, representing an intraspecific RFLP. The other four DR genes showed evidence of elimination/divergence of genes from one ancestor, mainly from the D genome. Among them, three genes (DRs 4, 5, and 25) showed complete loss of the band from one genome (Figures 1b, c, and d). Figure 1b shows the hybridization pattern of cotton DR4 digested with *Eco*RV. One band of about 7000 bp could be detected in the two A genome species. In the D genome species *G. raimondii*, one larger band (~8500 bp) is detected. All races of the five tetraploid species had a band of the same size as in the A genome diploids, whereas the *G. raimondii* band was found only in three tetraploid species (*G.mustelinum*, *G. tomentosum*, and *G. hirsutum*) and is absent from *G. darwinii* and *G. barbadense*. This pattern of absence is congruent with the widely accepted phylogenetic relationship among these species (3) and is suggestive that the D genome gene might have been eliminated after divergence of the *G. darwinii-barbadense* clade from the other tetraploids. DR5 has only the A genome band in all five tetraploids, lacking the D genome band in all cases (Figure 1c) when DNA was digested with *Eco*RV. This result may either suggest that the copy from D was lost in the common ancestor of tetraploid species before its radiation into multiple lineages, or that there was no polymorphism between A and D when they formed the tetraploid and a later mutation became fixed in the diploid D genome. A further study by sequencing the corresponding BACs has shown the latter scenario to be more likely, detailed below. DR25 retained the A genome band in all tetraploid species and lacked the D genome band in all except *G. hirsutum*, which had one accession showing loss/gain of the D genome band (Figure 1d). One gene (DR3) showed the loss/gain genotype in five tetraploids.

Among the 12 DR genes displaying two or more bands at diploid level, as mentioned above, six showed evidence of loss/divergence in some or all five tetraploids, with five of the six cases affecting the D genome-

Table 3 Average band number in two A, two D, and five tetraploid *Gossypium* species detected with genes showing random versus duplication-resistant retention patterns in *Arabidopsis*

Species (Genome)	*Unig*	*DR*	*Unig-DR*
G. herbaceum (A1)	2.32	1.98	0.33
G. arboreum (A2)	2.31	1.92	0.39
G. mustelinum $(AD)_4$	4.09	3.22	0.86
G. darwinii $(AD)_5$	4.41	3.10	1.32
G. barbadense (AD)	4.40	3.06	1.34
G. tomentosum $(AD)_3$	4.59	3.17	1.42
G. hirsutum (AD)	4.42	3.23	1.19
G. raimondii (D5)	3.47	2.38	1.08
G. gossypioides (D6)	2.56	1.96	0.60

Abbreviation: DR, duplication resistant.

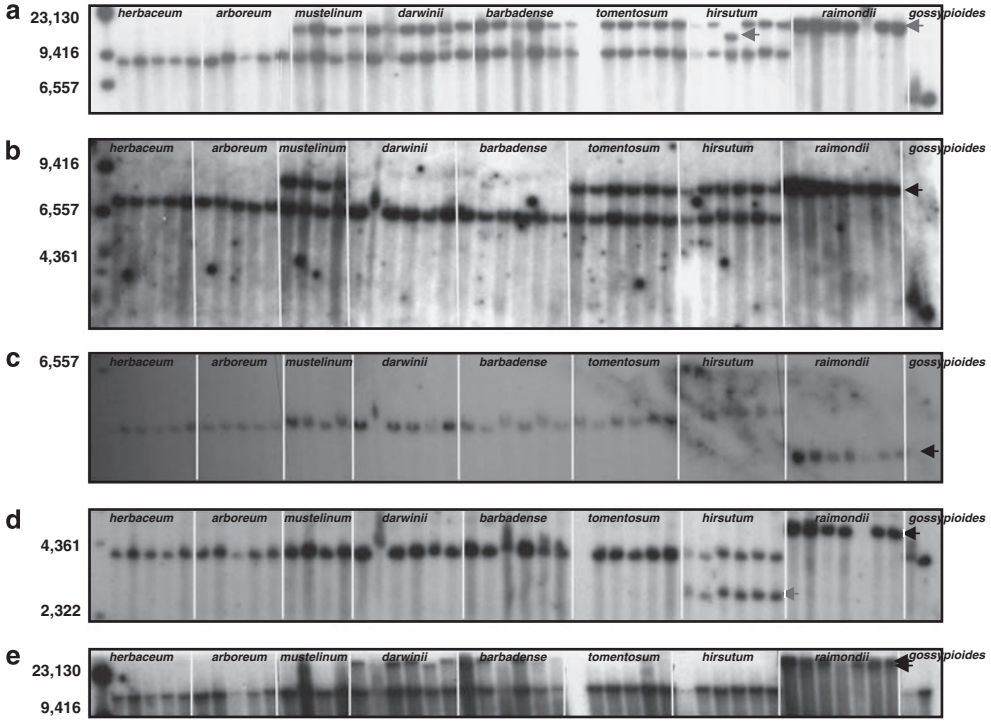

Figure 1 Autoradiograph from southern hybridization with DR candidate genes showing gene loss/divergence. Lane 1: lambda *Hin*dIII, lanes 2–48: cotton species listed in Supplementary Table 1. Black arrows indicated gene loss and red ones sequence divergence of restriction site. (**a**) DR16, showing presence of alloalleles from both A and D genome species except for a loss/gain in a *G. hirsutum* accession (TX45). (**b**) DR4, showing loss of the alloallele from D genome species in *G. darwinii* and *G. barbadense*. (**c**) DR5, showing loss of the alloallele from D genome species in all tetraploid species. (**d**) DR25, showing loss of the alloallele from D genome species in all five tetraploid species and a new band in *G. hirsutum*. (**e**) DR28, showing loss of two alloalleles from D genome species in *G. mustelinum*, *G. tomentosum*, and *G. hirsutum*. A full colour version of this figure is available at the *Heredity* journal online.

derived allele. DR28 is a typical example showing apparent absence of the D genome allele in some tetraploids (Figure 1e). The autoradiograph shows two bands having similar sizes near 23 kb in *G. raimondii*. No such bands were found in *G. mustelinum*, *G. tomentosum*, or *G. hirsutum*, and there are one or two bands around 23 kb in some races of *G. darwinii* and *G. barbadense* (Figure 1e), suggesting the loss of this allele in some cotton species.

The 39 randomly selected genes produced interpretable bands in most genotypes. Among them, nine (23.1%) displayed only one band in at least one enzyme, one (unig6B02) showing loss of the copy from *G. raimondii* in *G. mustelinum*. Three genes (unig6D09, 22D10, and 24E01) showed no polymorphism between diploid A and D genomes. One gene (unig66F01) had two bands with the same size as in diploid A and D genomes in all tetraploid species. Three genes had the loss/gain pattern in one or more tetraploids. For example, Unig22B04 and 23B11 showed the loss/gain of the D genome homeolog in *G. darwinii* and also in TX45 for 22B04. Contrary to these two genes, unig23H09 had loss/gain from the A genome in four tetraploids excluding *G. darwinii*.

Copy number of candidate DR genes evaluated by sequencing of hybridizing BACs

To further evaluate copy number variation of the DR genes in diploid and tetraploid cottons, a subset of BACs hybridized by three DR genes (DRs 4, 5, and 28), which showed apparent loss of the D genome allele in some or all tetraploids, were used as templates for sequencing. DRs 4 and 5 showed single restriction fragments in at least one enzyme in survey in both A and D diploids and loss of the one from the D genome in two and all tetraploid cottons, respectively (Figures 1b and c). DR28 showed two restriction fragments in each diploid, and loss of one or both D genome copies in all tetraploid cottons (Figure 1e). DR4 alleles were sequenced from a total of 12 BACs, 3 from *G. arboreum* (A diploid), 3 from *G. barbadense* (tetraploid), and 6 from *G. hirsutum* (tetraploid), producing a consensus sequence of 332 bp when both 5′and 3′sequences were aligned. These sequences can be classified into two groups. Three *G. hirsutum* BACs in one group that differ in seven polymorphic sites (six SNPs and one indel) from the other nine BACs including the other three *G. hirsutum* BACs and all three *G, barbadense* BACs, suggesting that there are two copies of this gene in *G. hirsutum* and one copy in *G. barbadense*. The three *G. arboreum* BACs closely resemble the second group, differing only in one site by an SNP of G/A. This sequence result matches the hybridization pattern of genomic DNA, that is two bands in *G. hirsutum* and one band each in *G. arboreum* and *G. barbadense*, confirming that the *G. barbadense* clade has lost the D genome allele of this gene.

For DR5, sequences from 10 BACs (3 *G. arboreum*, 3 *G. raimondii*, 3 *G, barbadense,* and 1 *G. hirsutum*) could be divided into two main groups. One *G, barbadense* and three *G. raimondii* BACs differed from the other 6 BACs in 14 sites of a 650 bp consensus sequence, including 12 SNPs and 2 indels (one two bp and another 20 bp deletion). The three *G. raimondii* sequences differ from the *G, barbadense* sequence in only three SNPs. Among the other group of six BACs, the three *G. arboreum* BACs are differentiated by only two SNPs from two *G, barbadense* and one *G. hirsutum* BAC(s). In partial summary, these findings suggest that the respective A and D genome versions of DR5 genes are clearly differentiated, and that each corresponds to one of the two copies in *G. barbadense*. As two groups of BACs were detected in *G. barbadense* corresponding to the band from *G. arboreum* and *G. raimondii*, respectively, it is inferred that there are two copies of this gene in at least *G. barbadense* and that the band number difference detected in southern hybridization most probably reflects a nucleotide substitution. One possible explanation is that DNA sequence at the restriction site in diploid A and D was the same when they were joined in the tetraploid. As a result, there seems to be only one band in the tetraploid. The different fragment size in the modern D genome than the A genome may be the result of the mutation in the D genome since polyploid formation.

DR28 sequences were obtained from 17 BACs (3 *G. arboreum*, 4 *G. raimondii*, 3 *G. barbadense*, and 7 *G. hirsutum*). There seem to be two copies of DR28 in each diploid and tetraploid species, respectively, accounting for 1 and 2 *G. arboreum*, 3 and 1 *G. raimondii*, 1 and 2 *G. barbadense*, and 4 and 3 *G. hirsutum* BACs, with copy 1 very closely matching the original cotton EST (gi 48878853) used for primer design (Figure 2). The polymorphic regions of the two copies spanned 53 bp in copy 1 and 142 bp in copy 2. Copy 1 sequences of three *G. raimondii* BACs all differed in one SNP (C/T) from *G. arboreum* and tetraploid cotton sequences (Figure 2). This result, that is that all tetraploid BACs had the same sequences as the *G. arboreum* BAC, indicates that copy 1 from the diploid D genome has been lost in tetraploid cotton. In the polymorphic region of copy 2, 121 of 122 bp (indicated in gray; Figure 2) matched another *G. hirsutum* EST (gi:73860087) . For copy 2, the single *G. raimondii* sequence differed from those of *G. arboreum* (2), *G. barbadense* (2), and *G. hirsutum* (3) in four SNPs, suggesting that only the A genome copy had been retained in the tetraploids. In summary, these findings support

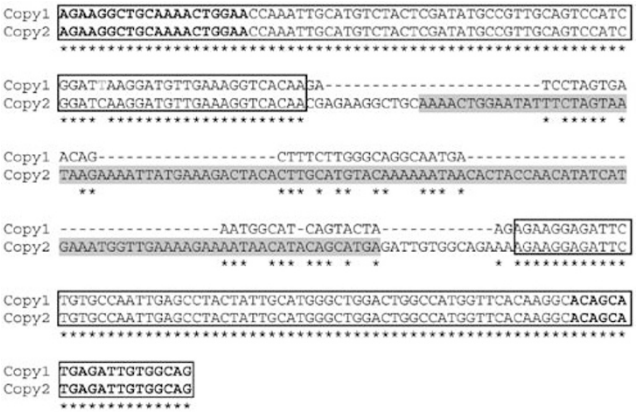

Figure 2 Sequence alignment of two DR28 copies amplified from four BAC libraries (GAMBO, GR, GAD, and MAXXA) showing their divergence. Boxed sequences are exons matching the original cotton EST (gi:48878853) used to design the primers (bold bps) for amplification of targeted gene. The bp highlighted with red in copy 1 is an SNP between A genome (in both diploid and tetraploid) and *G. raimondii* as well copy 2 The bps highlighted with gray in copy 2 match another *G. hirsutum* EST (gi:73860087). A full colour version of this figure is available at the *Heredity* journal online.

Table 4 Correlation coefficient between sequence identities and average band (homolog) number in cotton of genes showing duplication-resistant retention patterns in *Arabidopsis*

Identity	*Sequence*	*A1 band#*	*A2 band#*	*D5 band#*
A/K	5′	−0.399	−0.465*	−0.503*
	3′	0.014	0.071	0.118
D/K	5′	−0.347	−0.410	−0.419
	3′	0.041	0.072	0.019
A/D	5′	−0.517*	−0.484*	−0.382
	3′	0.137	0.138	0.29

*5% significance.

the inference from southern hybridization that each of two copies of DR28 from the D genome was lost in at least *G. hirsutum*.

Correlation between gene copy number and DNA sequence divergence

Amplicons of 18 cotton candidate DR genes (Supplementary Table 3) from genomic DNA of *G. arboreum*, *G. raimondii*, and *G. kirkii* were sequenced. The three sequences of each DR gene were compared using blastn and pairwise identities were calculated. The average identity between A and D for 18 sequenced DR genes is 97% (ranging from 85 to 100%), with 5′ sequences of DRs 1, 7, and 17 each having <90% identity. The identities between A or D and K is about 95%, ranging from 90 to 100% except for the 5′ sequence of DR17, which has 87.6% identity between D and K.

Pairwise sequence identities at the 5′ end of the gene(s) among the A, D, and K genome species have a strong negative correlation with band number in the A and D genomes, in several cases reaching statistical significance (Table 4). Corresponding analysis using 3′ sequences yielded correlation coefficients close to zero (Table 4). The identities of best-matched sequences between randomly selected *G. arboreum* and *G. raimondii* ESTs also showed no correlation with band number. In partial summary, high 5′ sequence identity between *Gossypium* diploids for a DR gene is associated with low copy number of the gene.

Discussion

Preferential loss of cotton candidate DR genes

Retention and loss of duplicated genes, an important source of gene copy number variation, are not random and are related to the proteome in a range of ways (Korbel *et al*., 2008). In angiosperms, most gene functional groups show post-duplication gene preservation/loss rates that are indistinguishable from the genome-wide average, whereas genes in some specific functional categories duplicate and reduplicate, and genes in other functional categories show more extensive loss of duplicate copies than the genome-wide average.

Gossypium (cotton) homologs of genes suggested to be 'DR' in *Arabidopsis* (Paterson *et al*., 2006) show significantly lower gene copy number at both the diploid and tetraploid levels than randomly selected cotton genes. Cotton candidate DR genes also show stronger evidence than randomly selected genes of recent gene loss in tetraploids, with preferential loss from the D subgenome. These results indicate that at least some DR genes have followed a distinctly different evolutionary path than most genes, after the formation of tetraploid cotton. Even 'diploid' cotton has experienced at least one genome duplication as its divergence from a common ancestor shared with *Arabidopsis* (Rong *et al*., 2005), a period during which *Arabidopsis* has experienced two genome duplications. DR genes have lower copy number than randomly selected genes in 'diploid' cotton, indicating that the ancestors of modern diploid cotton may have experienced gene loss similar to that being observed now in tetraploid cotton. In partial summary, these findings support the hypothesis that fundamental principles of molecular evolution favor the loss of duplicated copies of certain genes and functional groups across many lineages (Paterson *et al*., 2006). This hypothesis also gains support from the notion that transcriptome dominance of genes from one genome partly determines patterns of gene loss (Wu *et al*., 2008).

Although rapid loss of DNA fragments can occur just after the synthesis of allopolyploid wheat and *Brassica* (Song *et al*., 1995; Feldman *et al*., 1997), it seems that duplicated gene loss in cotton is a slow and long-term process. Liu *et al*. (2001) did not find any loss in nine sets of newly synthesized allotetraploid and allohexaploid cotton plants, their parents, and the self-crossed progeny from colchicine-doubled synthetics among 22 000 genomic loci analyzed using AFLP or five retrotransposons using southern hybridization. The few clear cases of gene loss we report are striking exceptions to the otherwise high level of DNA sequence preservation in polyploid cotton. Their assignment to different 'branches' of the well-established *Gossypium* phylogeny illustrates that these gene losses have been distributed over a considerable period of time, rather than (for example) occurring shortly after formation of a polyploid ancestor (Figure 3).

Although it is widely accepted that gene loss happens after whole genome or segmental duplication in angiosperms in a manner that is related to gene function (Paterson *et al*., 2006) and transcriptome dominance (Wu *et al*., 2008), very little is known about the mechanism causing the gene loss. In resynthesized *Brassica napus* allopolyploids, a considerable amount of gene loss resulted from homoeologous nonreciprocal transpositions (Udall *et al*., 2005; Gaeta *et al*., 2007; Nicolas *et al*., 2009). Homoeologous nonreciprocal transposition occurred nonrandomly across the genome and are related to the degree of divergence between homoeologous chromosomes. The least divergent homoeologs in the A and C genomes (N1–N11) of *B. napus* had the most homoeologous exchange events (Udall *et al*., 2005). Similar gene loss has been documented in the natural *Tragopogon miscellus* (Buggs *et al*., 2009). In our research, 5′ sequence similarity between homologous genes and copy number are negatively correlated for the cotton DR candidate genes, but not for randomly selected genes. In other words, DR genes with high sequence similarity between homoeologous A, D, and K genomes of *Gossypium* normally have low copy number. Collectively, evidence from *Gossypium*, *Brassica*, and *Tragopogon* all indicate that high sequence similarity may promote pairing and further recombination between homoeologous pairs, which thus may present a potential mechanism for sequence elimination through nonreciprocal crossover events.

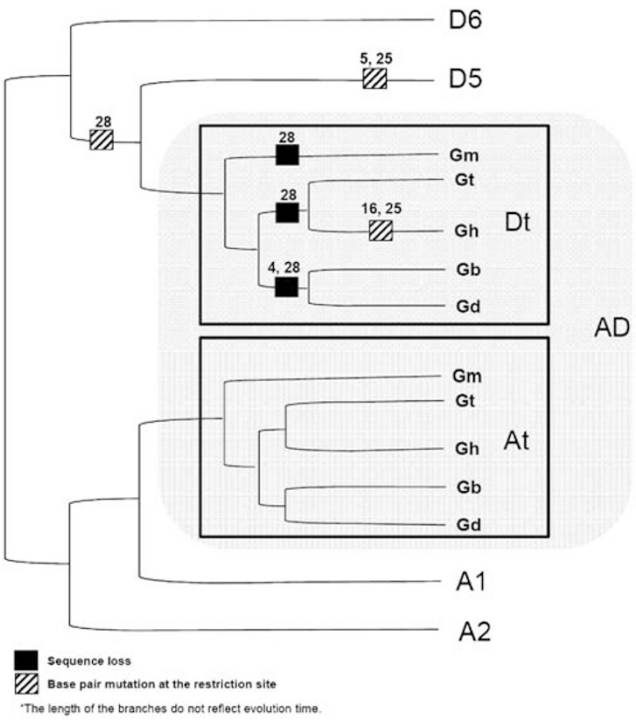

Figure 3 Inferred gene loss/divergence in relation to the Gossypium phylogeny. Numbers represent the DR genes. Branch lengths are symbolic, and may not directly reflect the evolution time.

D genome seems to have incurred more gene loss than the A genome during radiation of tetraploid cotton

The genome size of A genome species is almost twice that of D genome species because of the enrichment of repeated sequences (mainly transposable elements) in the A genome (Zhao *et al*., 1998; Hawkins *et al*., 2006), but both species have very similar numbers of genes. Earlier surveys of fragment number in diploid A and D genome species indicated only small differences (Reinisch *et al*., 1994), with slightly more fragments detected in the A genome by *Pst*I-digested genomic probes, and slightly more bands in D genome detected by cDNA probes. In the present research, which focused entirely on expressed genes, D genome *G. raimondii* usually showed equal or greater numbers of restriction fragments than the two diploid A genome species for both DR genes and randomly chosen genes. Other investigators also found fiber-related cDNA/genes to show more copies in *G. raimondii* than in diploid A genome species (Orford *et al*., 1999; Orford and Timmis, 2000). The possibility that *G. raimondii* may have slightly higher gene copy numbers than A genome species might contribute to recurring reports of a slightly higher level of DNA polymorphism in the D genome than that can be detected in A genome (Reinisch *et al*., 1994; Jiang *et al*., 1998; Small and Wendel, 2002; Grover *et al*., 2007).

All DR gene copies lost from tetraploid species in this study were from the D genome (DRs 4 and 28; Figures 1b, e, and 3). Further, for both DR genes and randomly selected genes, most homoeologous restriction site mutations occurred in the D genome (Figures 1 and 3), in accordance with the discovery of more DNA-level variations in the D than the A genome (Reinisch *et al*., 1994; Jiang *et al*., 1998; Small and Wendel, 2002; Grover *et al*., 2007). More DNA-level variation found in the D genome than the A genome may be symptomatic of mechanisms that also contribute to the greater abundance of QTLs responsible for phenotypic variation, including fiber-related QTLs detected on this genome from a nonfiber-producing ancestor (Jiang *et al*., 1998; Rong *et al*., 2007). Collectively, these results point to a pivotal function of the D genome in tetraploid cotton evolution.

DR genes experienced divergent fates in different tetraploid cottons

Southern hybridization of genomic DNA by two single copy DR genes (DRs 4 and 5) and one two-copy gene (DR28) showed the loss of the copy from one diploid progenitor in two or more tetraploid cottons, displaying the diploid A genome haplotype in these tetraploid species. Sequencing from BAC templates revealed one case (DR5) to be explicable by comigration in tetraploid cotton of two bands from the respective diploids, suggesting that the restriction site of the gene in the diploid D genome has mutated since tetraploid formation (Figures 1c and 3).

Southern hybridization and sequence analysis confirm that the copy of the DR4 gene from the diploid D genome was lost in the *G. darwinii–G. barbadense* clade (Figures 1b and 3). As the absence of the allele is fixed in both species, it seems probable that this loss occurred soon after the divergence of this clade from the other two tetraploid clades (Figure 3).

For DR28, southern hybridization and sequence analysis confirm that each of the two copies of the gene from the diploid D were lost in *G. hirsutum*, *G. tomentosum*, and *G. mustelinum*, with these tetraploids retaining each of the two A genome copies. Different *G. darwinii* and *barbadense* accessions have bands corresponding to different, single, diploid D genome alleles (Figure 1e), suggesting the loss of one D genome allele from each of these accessions (Figures 1d and 3).

Although there are only a small number of informative genes that have been identified to date, evidence of DR gene loss is consistent in all cases with the well-documented phylogeny of polyploid *Gossypium* (Figure 3). Evidence from additional loci is needed (and is being sought), but the present data does support the notion that return to singleton status of genes containing some specific protein domains (Paterson *et al*., 2006) may be an important dimension of adaptation by a genome to the duplicated state.

Conflict of interest

The authors declare no conflict of interest.

References

Adams KL, Cronn R, Percifield R, Wendel JF (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci USA* **100**: 4649–4654.

Adams KL, Percifield R, Wendel JF (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168**: 2217–2226.

Arpat A, Waugh M, Sullivan J, Gonzales M, Frisch D, Main D *et al*. (2004). Functional genomics of cell elongation in developing cotton fibers. *Plant Mol Biol* **54**: 911–929.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM *et al*. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**: 171–178.

Avner P, Heard E (2001). X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet* **2**: 59–67.

Blanc G, Wolfe KH (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**: 1679–1691.

Bowers JE, Chapman BA, Rong JK, Paterson AH (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O *et al*. (2006). Gene loss and evolutionary rates following whole genome duplication in teleost fishes. *Mol Biol Evol* **23**: 1808–1816.

Buggs RJA, Doust AN, Tate JA, Koh J, Soltis K, Feltus FA *et al*. (2009). Gene loss and silencing in Tragopogon miscellus (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity* **103**: 73–81.

Chapman BA, Bowers JE, Feltus FA, Paterson AH (2006). Buffering crucial functions by paleologous duplicated genes may impart cyclicality to angiosperm genome duplication. *Proc Natl Acad Sci USA* **103**: 2730–2735.

Chen ZJ, Pikaard CS (1997a). Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes Dev* **11**: 2124–2136.

Chen ZJ, Pikaard CS (1997b). Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica. *Proc Natl Acad Sci USA* **94**: 3442–3447.

Comai L (2005). The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**: 836–846.

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y *et al*. (2000). Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. *Plant Cell* **12**: 1551–1567.

Disteche CM, Filippova GN, Tsuchiya KD (2002). Escape from X inactivation. *Cytogenet Genome Res* **99**: 36–43.

Feldman M, Liu B, Segal G, Abbo S, Levy AA, Vega JM (1997). Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**: 1381–1387.

Feltus FA, Singh HP, Lohithaswa HC, Schulze SR, Silva T, Paterson AH (2006). Conserved intron scanning primers: targeted sampling of orthologous DNA sequence diversity in orphan crops. *Plant Physiol* **140**: 1183–1191.

Freeling M, Thomas BC (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* **16**: 805–814.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007). Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell* **19**: 3403–3417.

Gao LZ, Innan H (2004). Very low gene duplication rate in the yeast genome. *Science* **306**: 1367–1370.

Grover C, Kim H, Wing R, Paterson AH, Wendel JF (2007). Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton. *Plant J* **50**: 995–1006.

Haldane JBS (1933). The part played by recurrent mutation in evolution. *Am Nat* **67**: 5–19.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res* **16**: 1252–1261.

He XL, Zhang JZ (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.

Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF (2008). Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. *Genetics* **179**: 1725–1733.

Jiang CX, Wright RJ, El-Zik KM, Paterson AH (1998). Polyploid formation created unique avenues for response to selection in Gossypium (cotton). *Proc Natl Acad Sci USA* **95**: 4419–4424.

Kashkush K, Feldman M, Levy AA (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**: 1651–1659.

Kashkush K, Feldman M, Levy AA (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33**: 102–106.

Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M *et al*. (2008). The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol* **18**: 366–374.

Lee HS, Chen ZJ (2001). Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *ProcNatl Acad Sci USA* **98**: 6753–6758.

Lee JT, Jaenisch R (1997). The (epi)genetic control of mammalian X-chromosome inactivation. *Curr Opin Genet Dev* **7**: 274–280.

Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF (2001). Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome* **44**: 321–330.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Force A (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459–473.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M *et al*. (2005). Modeling gene and genome duplications in eukaryotes. *ProcNatl Acad Sci USA* **102**: 5454–5459.

Ming R, Hou S, Feng Y, Yu QY, Dionne-Laporte A, Saw J *et al*. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**: 991–997.

Nicolas SD, Leflon M, Monod H, Eber F, Coriton O, Huteau V *et al*. (2009). Genetic regulation of meiotic cross-overs between related genomes in Brassica napus haploids and hybrids. *Plant Cell* **21**: 373–385.

O'Neill RJW, O'Neill MJ, Graves JAM (2002). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid (vol 393, pg 68, 1998). *Nature* **420**: 106–106.

Ohno S (1970). *Evolution by Gene Duplication*. Springer: Berlin.

Orford SJ, Carney TJ, Olesnicky NS, Timmis JN (1999). Characterisation of a cotton gene expressed late in fibre cell elongation. *Theor Appl Genet* **98**: 757–764.

Orford SJ, Timmis JN (2000). Expression of a lipid transfer protein gene family during cotton fibre development. *Biochimica et Biophysica Acta (BBA)—Mol Cell Biol Lipids* **1483**: 275–284.

Ozkan H, Levy AA, Feldman M (2001). Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. *Plant Cell* **13**: 1735–1747.

Ozkan H, Levy AA, Feldman M (2002). Rapid differentiation of homeologous chromosomes in newly-formed allopolyploid wheat. *Isr J Plant Sci* **50**: S65–S76.

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J *et al*. (2003). Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science* **299**: 697–700.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H *et al*. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551–556.

Paterson AH, Bowers JE, Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* **101**: 9903–9908.

Paterson AH, Chapman BA, Kissinger J, Bowers JE, Feltus FA, Estill J *et al*. (2006). Convergent retention or loss of gene/domain families following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces, and Tetraodon. *Trends Genet* **22**: 597–602.

Reinisch A, Dong J-M, Brubaker C, Stelly D, Wendel J, Paterson A (1994). A detailed RFLP map of cotton (Gossypium hirsutum × G. barbadense): chromosome organization and evolution in a disomic polyploid genome. *Genetics* **138**: 829–847.

Rong J, Bowers JE, Schulze SR, Waghmare VN, Rogers CJ, Pierce GJ *et al.* (2005). Comparative genomics of Gossypium and Arabidopsis: unraveling the consequences of both ancient and recent polyploidy. *Genome Res* **15**: 1198–1210.

Rong J, Feltus EA, Waghmare VN, Pierce GJ, Chee PW, Draye X *et al.* (2007). Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics* **176**: 2577–2588.

Rong JK, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW *et al.* (2004). A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). *Genetics* **166**: 389–417.

Samuel Yang S, Cheung F, Lee JJ, Ha M, Wei NE, Sze SH *et al.* (2006). Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *Plant J* **47**: 761–775.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341–345.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD *et al.* (2003). Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* **20**: 633–643.

Seoighe C, Gehring C (2004). Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet* **20**: 461–464.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**: 1749–1759.

Small RL, Wendel JF (1999). The mitochondrial genome of allotetraploid cotton (Gossypium L.). *J Hered* **90**: 251–253.

Small RL, Wendel JF (2002). Differential evolutionary dynamics of duplicated paralogous Adh loci in allotetraploid cotton (Gossypium). *Mol Biol Evol* **19**: 597–607.

Song KM, Lu P, Tang KL, Osborn TC (1995). Rapid genome change in synthetic polyploids of brassica and its implications for polyploid evolution. *Proc Natl Acad Sci USA* **92**: 7719–7723.

Stephens S (1951). Possible significance of duplications in evolution. *Adv Genet* **4**: 247–265.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008a). Synteny and colinearity in plant genomes. *Science* **320**: 486–488.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008b). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**: 1944–1954.

Udall JA, Quijada PA, Osborn TC (2005). Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of Brassica napus L. *Genetics* **169**: 967–979.

Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J *et al.* (2006). A global assembly of cotton ESTs. *Genome Res* **16**: 441–450.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH (2007). Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**: 1753–1763.

Wendel JF (1989). New world tetraploid cottons contain old-world cytoplasm. *Proc Natl Acad Sci USA* **86**: 4132–4136.

Wendel JF, Cronn RC (2003). Polyploidy and the evolutionary history of cotton. *Adv Agronomy* **78**: 139–186.

Wu Y, Zhu Z, Ma L, Chen M (2008). The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. *Mol Biol Evol* **25**: 1003–1006.

Zhao X, Si Y, Hanson R, Crane C, Price H, Stelly D *et al.* (1998). Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**: 479–492.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)