Extensive and biased inter-genomic non-reciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (cotton)


* Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA
§ Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
† Center for Genomics and Computational Biology, School of Life Sciences, and School of Sciences, Hebei United University, Tangshan, Hebei 063000, China
‡ MIPS/IBIS Inst. for Bioinformatics and System Biology, German Research Center for Environmental Health (GmbH), 85764 Neuherberg; Germany
** Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA
§§ Jamie Whitten Delta States Research Center, USDA-ARS, Stoneville MS 38776, USA
†† Department of Crop and Soil Science and Coastal Plain Experiment Station, Tifton GA 31794, USA
‡‡ Department of Crop and Soil Science; and Department of Genetics, University of Georgia, Athens, GA 30602, USA

1Corresponding author

Andrew H. Paterson

111 Riverbend Road, Rm 228

Athens, GA 30605

Copyright 2014.
Email: paterson@plantbio.uga.edu
Phone: 706-583-0162

**Running title**: Genome evolution in polyploid cotton

**Keywords**: cotton, homeologous gene conversion, polyploidy, heterochromatin
ABSTRACT

Genome duplication is thought to be central to the evolution of morphological complexity, and some polyploids enjoy a variety of capabilities that transgress those of their diploid progenitors. Comparison of genomic sequences from several tetraploid \((A, D_t) Gossypium\) species and genotypes with putative diploid A and D genome progenitor species revealed that unidirectional DNA exchanges between homeologous chromosomes were the predominant mechanism responsible for allelic differences between the \(Gossypium\) tetraploids and their diploid progenitors. HeGCE gradually subsided, declining to rates similar to random mutation during radiation of the polyploid into multiple clades and species. Despite occurring in a common nucleus, preservation of HeGCE is asymmetric in the two tetraploid subgenomes. \(A_t\) to \(D_t\) conversion is far more abundant than the reciprocal, is enriched in heterochromatin, is highly correlated with GC content and transposon distribution, and may silence abundant A-genome-derived retrotransposons. \(D_t\) to \(A_t\) conversion is abundant in euchromatin and genes, frequently reversing losses of gene function. The long-standing observation that the non-spinnable-fibered D genome contributes to the superior yield and quality of tetraploid cotton fibers may be explained by accelerated \(D_t\) to \(A_t\) conversion during cotton domestication and improvement, increasing dosage of alleles from the spinnable-fibered A genome. HeGCE may provide an alternative to (rare) reciprocal DNA exchanges between chromosomes in heterochromatin, where genes have ~5x greater abundance of \(D_t\) to \(A_t\) conversion than does adjacent intergenic DNA. Spanning exon-to-gene-sized regions, HeGCE is a natural
non-invasive means of gene transfer with the precision of transformation, potentially important in genetic improvement of many crop plants.
INTRODUCTION

Genome duplication is a potentially rich source of genes with new (Stephens 1951; Ohno 1970) or modified functions (Lynch and Conery 2000), and is thought to be central to the evolution of morphological complexity (Freeling and Thomas 2006). Genome doubling may confer advantages to a polyploid (Comai 2005), via mechanisms such as increased gene dosage, ‘intergenomic heterosis’ conferred by multiple alleles in a polyploid nucleus, or the evolution of novel gene functions [neofunctionalization -- (Stephens 1951; Ohno 1970)]. Over time, duplicated genes may evolve subdivisions of ancestral functions [subfunctionalization - (Lynch and Force 2000)] that render them interdependent. Subfunctionalization may sometimes lead to neofunctionalization (He and Zhang 2005).

Polyploids has been suggested to enjoy a variety of capabilities that transgress those of their diploid progenitors. For example, the notion that polyploids may adapt better than diploids to environmental extremes has been suggested based on both their geographic distribution (Muntzing 1936; Love and Love 1949; Stebbins 1950; Grant 1971), and on an inferred abundance of paleo-polyploidizations near the Cretaceous-Tertiary extinction (Fawcett et al. 2009). A variety of morphological, physiological, and gene expression changes have been associated with polyploidy. Experimental data is available to evaluate causality of only a few such cases in specific adaptations of polyploids, with more data needed but offering some support [as recently reviewed (Madlung 2013)].

Angiosperms (flowering plants) are an outstanding model for studying consequences of genome duplication salient to higher eukaryotes. All angiosperms are paleopolyploid
(Bowers et al. 2003; Jiao et al. 2011), and their abundance of multiple independent genome duplications (Paterson et al. 2010) provides ‘natural replicates’ for a variety of investigations. Study of the genes from three rounds of ancient whole genome duplications in *Arabidopsis* reveals a short phase of function relaxation followed by diversifying selection (Guo et al. 2013). The ability to study multiple independent genome duplications in a lineage also permits inference of the temporal orders and rates at which different duplication-associated events/mechanisms occur. Their larger genome sizes and smaller effective population sizes than microbes that have experienced genome duplication such as yeast (Gu et al. 2003; Christoffels et al. 2004; Scannell et al. 2006) and *Paramecium* (Aury et al. 2006), makes angiosperms more appropriate for studying consequences of genome duplication in higher eukaryotes (Lynch et al. 2001; Lynch 2006).

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; Shaked et al. 2001; Kashkush et al. 2002; Ozkan et al. 2002), activation of genes and retrotransposons (O'Neil et al. 2002; Kashkush et al. 2003), gene silencing (Chen and Pikaard 1997a, 1997b; Comai et al. 2000; Lee and Chen 2001) and subfunctionalization of gene expression patterns (Adams et al. 2003; Adams et al. 2004). Gene silencing in the allopolyploid hybrid between *Arabidopsis thaliana* and *Cardaminopsis arenosa* is arguably related to defense response against transposons (Comai et al. 2000). Changes of 24-nt siRNA and DNA
methylation levels in Arabidopsis hybrids are greatest at loci for which two parents differ substantially (Groszmann et al. 2011; Greaves et al. 2011). Chromosome rearrangement and reactivation of transposable elements are well known when plants are under ‘genomic stress’, which includes formation of polyploids (McClintock 1983). Instability of hybrid genomes has been attributed to bursts of transposition in both animals and plants; cytosine demethylation and deacetylation of lysine residues on histones may be responsible (Fontdevila 2005).

However, to learn whether immediate reactions of genomes to duplication provide raw material for the beginnings of adaptation or are merely symptoms of imminent extinction, it is necessary to investigate naturally-formed polyploids that have survived the test of time. The extinction hypothesis seems generally more likely, given that unreduced gametes are produced more or less continuously by organisms but only a tiny fraction result in successful lineages. For example, dramatic early-generation mutations in synthetic B. napus (Pires et al. 2004) are not paralleled in naturally-occurring forms (Rana et al. 2004).

A particularly intriguing example of possible advantages associated with polyploidy comes from the cotton genus, Gossypium, in which two diploids and two tetraploids have each been independently domesticated for production of the same product, seedborne epidermal fibers. A-genome diploids native to Africa, and Mexican D-genome diploids diverged ~5-10 mya (Senchina et al. 2003). They were reunited ~1-2 mya by trans-oceanic dispersal to the New World of a maternal A-genome propagule resembling G.
herbaceum (Wendel 1989), hybridization with a native D-genome species resembling *G. raimondii*, and chromosome doubling. The nascent A₁D₁ allopolyploid spread throughout the American tropics and subtropics, diverging into at least three subclades and five species, with two of those species (*G. hirsutum* and *G. barbadense*) being independently domesticated.

In India, where scientific improvement programs are active for both ploidies, tetraploid (‘AtDt’ genome) cottons consistently have substantially higher yield and superior fiber qualities than A-genome diploids (Anonymous 1997). Remarkably, the majority of genetic variation among tetraploid cottons has been ascribed to chromosomes from the D genome diploid progenitor that does not produce spinnable fiber, suggesting that post-polyploidy selection for superior fiber yield and quality of tetraploid cottons has preferentially operated upon the Dt genome (Jiang *et al.* 1998; (Rong *et al.* 2007).

In the present study, sequencing and careful comparison of several tetraploid *Gossypium* species and genotypes and representatives of their putative progenitor genomes reveals that homeologous gene conversion events (HeGCE) account for the vast majority of allelic differences between polyploid cottons and their diploid progenitors. High survivorship of alleles that were converted shortly after polyploid formation is suggested to reflect both a rapid rate of conversion at that time and also adaptive significance of many resulting alleles. A second cadre of converted alleles are closely associated with domestication, suggesting a mechanism by which chromosomes from the D genome diploid progenitor that does not produce spinnable fiber may have come to account for
the majority of genetic variation in fiber characteristics among tetraploid cottons. In partial summary, these data suggest that HeGCE is an early and important mechanism by which genomes adapt to the duplicated state, and may also contribute to plant domestication and crop improvement.

RESULTS

Unidirectional DNA exchanges between homeologous loci were the predominant mutational mechanism in the nascent Gossypium polyploid

DNA recombination, typically by reciprocal exchanges between homologous chromosomes (‘crossing over’), is a central element of eukaryotic transmission genetics and is also implicated in repair of highly deleterious DNA double-strand breaks (DSB). Reciprocal exchanges are often accompanied by tracts of unidirectional, local DNA exchanges known as ‘gene conversion’. Most models proposed to account for homologous DSB repair [synthesis-dependent strand annealing, classic double-strand break repair, and break-induced replication, although not single strand annealing (Hellday et al. 2007)], also predict the occurrence of tracts of unidirectional gene conversion, even when reciprocal crossing over occurs.

Building on rich evidence of concerted evolution in tandemly repeated sequences such as ribosomal RNA genes (Eickbush and Eickbush 2007), and multi-gene families such as primate olfactory receptor genes (Sharon et al. 1999), we recently showed gene conversion to have occurred in the past ~400,000 years between duplicated rice genes that diverged from a common ancestor 70 million years ago (mya) (Wang et al. 2009).
While this is an extreme case, non-random similarity between duplicated genes widely distributed across other genomes (Chapman et al. 2006; Wang et al. 2007) suggests the phenomenon to be widespread.

During the 1-2 my since its formation, unidirectional DNA exchanges between homeologous chromosomes have greatly outnumbered random mutations in A,D,t tetraploid cotton. Mapping of 38x Illumina coverage from the A-genome species G. herbaceum to the reference D-genome v2 (G. raimondii) (Paterson et al. 2012) revealed 2,145,177 single nucleotide variations (SNV) between the two, with 60% remaining unchanged in an A,D,t tetraploid, G. hirsutum cultivar ‘Acala Maxxa’. Among the 40% of changed sites, 25% now have only D-genome alleles and 10.6% have only A-genome alleles, with only ~4.4% having new mutations. SNVs inferred [by methods previously described (Paterson et al. 2012)] to confer striking changes of gene function are even more biased, with 45% of sites changed, 34.2% to D-genome alleles and 8% to A-genome alleles (Fig. 1, A).

To infer the levels and patterns of occurrence of gene conversions following polyploid cotton formation (Fig. 1, B), we used a parsimony-based method. For comparison with Acala Maxxa, we re-sequenced another G. hirsutum cultivar (GA120R1B3) from a different US production region (i.e. separated by less than 200 years), a wild G. hirsutum (race yucatanese) separated by about 4500 years, and G. mustelinum, the tetraploid cotton species most divergent from G. hirsutum separated by about 1 million years. If a converted allele is shared by two lineages, we assume that the event occurred in the
common ancestor rather than independently in each lineage. Indeed, most conversions were present in all four genomes (85.03-87.54%; Fig. 1, B), outnumbering omnipresent random mutations by ~8-fold ($\chi^2 = 23599, 87473$ for Dt, At, $P<<0$). Parsimony implies that these homeologous gene conversion events (HeGCE) occurred prior to speciation in the nascent *Gossypium* polyploid.

When the polyploid *Gossypium* lineage diverged into multiple species, HeGCE appears to have been abating (Fig. 1, B). This is inferred based on the observation that HeGCE accounts for only ~62% (4921) of polymorphisms among the divergent species *G. hirsutum* and *G. mustelinum*, ~47% between wild and cultivated forms of *G. hirsutum*, and 33% between *G. hirsutum* cultivars from different production regions (Fig. 1, B and Fig. S1, A). Consideration of the genomic distribution of HeGCE in these various taxa suggests that it abates sooner in heterochromatin than in euchromatin (Fig. S1, B).

**Asymmetric evolution of the polyploid cotton subgenomes**

Despite inhabiting a common nucleus, the evolution of the cotton A<sub>t</sub> and D<sub>t</sub> subgenomes differs in several striking ways. First, A<sub>t</sub> to D<sub>t</sub> conversion is enriched in heterochromatin (Fig. 2). Euchromatin, localized in the terminal regions of cotton chromosomes, shows largely similar rates of the two conversion types (log-likelihood ratio ~0). Occasional 1 Mb bins with D<sub>t</sub> to A<sub>t</sub> biased conversion (log-likelihood ratio<0) unanimously reside in euchromatin.
As in many other eukaryotes (Duret and Galtier 2009) cotton gene conversion is GC-biased, and this bias is closely related to the divergent evolution of the $A_t$ and $D_t$ subgenomes. The D genome has higher GC content than the A genome in the heterochromatin where $A_t$ to $D_t$ conversion is enriched, and the A genome has higher GC content than the D genome in the euchromatin where $D_t$ to $A_t$ conversion is enriched (Fig. 2). The high correlation of tetraploid cotton gene conversion with the GC ratio between the two progenitor genomes ($R^2=0.506$) (Fig. 3, A), may explain the 30% more A to G and T to C mutations in the D genome than the A genome since their divergence (Rong et al. 2012).

To investigate genomic features related to HeGCE, we compared regions with highly (HC)- or little (LC)-biased conversion. Across the genome, the log-ratio of $A_t$ to $D_t$ versus $D_t$ to $A_t$ conversions per 1 Mb bin averages 0.80 (SD=0.46). Log-ratios above 1.72 [mean+2SD] indicate largely $A_t$ to $D_t$ conversion (HC), and below -0.12 [mean-2SD] indicate $D_t$ to $A_t$ (LC). HC overwhelmingly locates in heterochromatin and LC in euchromatin. Heterochromatic $A_t$ to $D_t$ conversion is highly correlated with transposon distribution ($R^2=0.684$) (Fig. 3, B) and enriched for potential DNA methylation sites suitable for transposon silencing (Slotkin and Martienssen 2007). Long terminal repeat (LTR)-, particularly gypsy-type, retrotransposons are overwhelmingly enriched in HC (>70% of all types of transposable elements, and 10.02% of total length (bp) of HC versus 2.97% of LC) (Fig. 4, A). A, T to G, C conversion, providing potential DNA methylation sites, is more frequent in TEs than non-TE regions across the genome (Fig. 4, B). Further, in HC, A, T to G, C significantly outnumber G,C to A,T conversions for
A_{i} to D_{i} (50.04 vs. 34.69%), but not D_{i} to A_{i} (39.95 vs. 41.26%). In LC, A,T to G,C conversions are similar to G,C to A,T (43.97 versus 38.13%) for A_{i} to D_{i} but differ for D_{i} to A_{i} (47.35 versus 32.52%) (Table 1).

In heterochromatic regions where reciprocal DNA exchanges are rare, some conversions might simply be neutral or slightly-deleterious relics not yet purged due to inefficient selection, as is true of retroelement insertions (Paterson et al. 2009) and other rearrangements (Bowers et al. 2005). Longer persistence of harmful mutations in HC than LC may explain an increasing ratio (in HC) of non-synonymous to synonymous mutations across the phylogenetic tree of the four *Gossypium* species studied (Fig. 4, C). Indeed, the overall ratio of non-synonymous to synonymous mutations is larger in heterochromatin than euchromatin (Fisher’s exact test, $P=1.334e-6$), and heterochromatin conversion tracks are longer than the genome wide average (Fig. S2). All detected conversion tracks in the genome and HC region are listed in the Supplementary Table 1.

**Gene-altering conversions are of widespread importance**

Strong evidence suggests gene-altering conversions to be of widespread functional importance, perhaps serving as an alternative to reciprocal DNA exchanges to form new allele combinations in heterochromatic regions. Heterochromatic D_{i} to A_{i} conversions are ~5x more frequent in genes than adjacent intergenic DNA (Fig. 4, D; Fisher’s exact test, $P=0.0001$) and A_{i} to D_{i} conversions are also gene-enriched ($P=0.0157$). Conversions more frequently restored gene functions in cotton heterochromatin than euchromatin (Fisher’s exact test, $P=0.0008$). Among 59 HC and 206 LC alleles that experienced
striking mutations in the A genome since its divergence from the F genome [using the D genome as outgroup (16), including premature mutations (32 HC, 80 LC), splice site alterations (22, 89), translation initiation alterations (3, 19), and stop codon losses (2, 18)], A_t to D_t conversion found in Acala Maxxa restored function to 45.8% (27) of HC versus only 22.3% (46) of LC SNVs. Recent evidence of widespread gene conversion in the centromere cores of maize (Shi et al. 2010) and Arabidopsis (Yang et al. 2012) further supports the importance of this mechanism.

**DISCUSSION**

Whole genome comparison of four naturally-occurring tetraploid cottons and representatives of their progenitor genomes reveals extensive DNA exchanges accumulated during the past 1-2 myrs since polyploid formation. Remarkably, the two constituent ‘subgenomes’ of tetraploid cotton have experienced very different evolution while residing within a common nucleus, with more than twice as many conversions of A_t to D_t alleles than the reciprocal. The bias is unlikely to be caused by incomplete lineage sorting. Genetic diversity within both A and D-genomes are estimated to be ~10 times less than the diversity between the two genomes (Van Deynze et al. 2009). Comparison of At and Dt sub-genomes of tetraploid cotton to their corresponding diploid progenitors shows small and comparable sequence divergence (Cronn et al. 2002). G. raimondii is quite narrow genetically, so there would not be much lineage sorting on the D side (Cronn et al. 2002). The A genome is a little more heterogeneous when considering the two A-genome species (G. arboreum and G. herbaceum) (Cronn et al. 2002). Given these considerations and keeping in mind the 5-7 million years of
divergence between the A and D genome diploids (Senchina et al. 2003), it seems improbable that there would be segregation for many common alleles.

A key advantage of the cotton system is that polyploids have survived in the wild for 1 million years or more, effectively ruling out the hypothesis that HeGCE is a symptom of a dysfunctional genome destined for extinction. Extensive gene conversion in the centromere cores of maize and Arabidopsis has been described using mapping populations, in which only the first several generations are observed (Shi et al. 2010; Yang et al. 2012). On the contrary, some studies show that non-crossover gene conversion is relatively rare compared to crossover-associated gene conversion in Arabidopsis thaliana hybrids (Lu et al. 2012; Wijnker et al. 2013). Cytogenetic studies of recently and naturally formed polyploid species, Tragopogon miscellus and resynthesized Brassica napus reveal a prolonged phase of genomic instability coupled with chromosome rearrangement and translocation (Xiong et al. 2011; Chester et al. 2012). However, in each of these systems and other synthetic or recent polyploids, it is hard to assert whether the sorts of rapid responses to polyploidy that are observed (chromosome rearrangement and translocation) are the beginnings of adaptation, or symptoms of pending extinction.

Frequent conversion of mostly heterochromatic A$_4$ alleles by GC-rich D$_4$ DNA may have helped to silence abundant A-genome-derived retrotransposons, perhaps stabilizing the early polyploid. It seems intuitive that D genome alleles, from the progenitor native to the New World habitat of the polyploid, may confer many adaptations that are lacking from
the Old World A genome. However, this explanation of enrichment for A₁ to D₁ conversions is not consistent with the strong heterochromatic bias observed for these conversions. The ~2x larger physical size of the A than the D genome is largely due to retrotransposons, mostly in heterochromatin (Paterson et al. 2012) and some having spread to the D genome following polyploidy (Zhao et al. 1998). Bursts of retrotransposition following hybridization (McClintock 1983; Fontdevila 2005) can cause many DSBs (Gasior et al. 2006) that are fatal to cells if not repaired (Van Gent et al. 2001). More quickly than random mutations, GC-biased gene conversion may have provided the A₁ genome with (D₁-derived) targets for DNA methylation-based transposon silencing. Transcripts from A₁-derived retrotransposons may enter the RNAi pathway for RNA dependent DNA methylation (Slotkin et al. 2009; Groszmann et al. 2011). A existing example similar to this process is the Drosophila P cytotype female produce a repressor protein and piRNA to inhibit P element transposition in gamete (Brennecke et al. 2008). Correlation between DNA methylation and gene conversion implies that the two are mechanistically related (Colot et al. 1996) by an as-yet-unknown process. In primates, biased gene conversion is shown as a major force for stabilizing constitutively methylated CpG islands (Cohen et al. 2011).

Meiotic recombination tends to be concentrated in small regions on chromosome known as “recombination hotspot”. Recombination is initiated by introduction of double-strand break to the hotspot alleles. When DSB occurs at a hotspot that is heterozygous with an inactive (“cold”) hotspot allele, the hotspot alleles are replaced with “cold” alleles by gene conversion during DNA repair. The process will cause loss of recombination
hotspot rapidly in the genome. The existence of recombination hotspots is therefore considered as “hotspot-conversion paradox” (Boulton et al. 1997). The paradox predicts small number of hotspots in the region with high conversion rate. Consistent with this prediction, the high rate of conversion in the heterochromatin region identified in this study might be partially explained by the paradox.

The ability to ‘copy’ existing alleles via HeCGE may expedite the ability of polyploids to evolve new or more exaggerated phenotypes, by achieving allele dosages that exceed those of their progenitors. For example, D₁ to A₁ conversion steadily declines following polyploid formation, but accelerates during cotton domestication and improvement (Fig. 1B and Fig. S1). This may provide a mechanism to explain the long-standing irony that so many QTLs for yield and quality of tetraploid cotton map to chromosomes derived from an ancestor (D) that lacks spinnable fibers (Jiang et al. 1998). This genetic process may complement paramutations that copy epigenetic information. In Arabidopsis hybrids, for example, methylation levels of one parental allele change to match the other (Greaves et al. 2011). Our findings show that nonreciprocal exchange of both genetic and epigenetic information may be important to the integrity of hybrid genomes.

HeCGE may have practical value in crop improvement. The length of cotton conversion tracts averages 455 bp, ranging from 279 to 3650 bp (see Methods), somewhat longer than in mammals (Chen et al. 2007) and often spanning entire exons or occasional genes. Cotton and many other neo-polyploid crops are genetically depauperate and occasional crosses with exotic or synthetic polyploids for crop improvement may be prone to bursts
of transposition, DSBs, and conversion. A fascinating area for further study is whether occasional ‘successes’ in extracting valuable alleles from exotic germplasm (e.g. (Campbell et al. 2011)) might have occurred by gene conversion rather than crossing over. Introgression by gene conversion might solve the widespread challenge of extracting desirable alleles from exotic germplasm while leaving behind nearby undesirable ones, i.e. with the precision of transformation but by a natural non-invasive means.

While their early evolution involved extensive intergenomic exchange, modern tetraploid cottons show strict disomy (Kimber 1961). The gradual decline of intergenomic conversion during the course of polyploid cotton evolution (Fig. 1, B) may have been due in part to fewer DSBs as the nascent polyploid was stabilized, perhaps by silencing of abundant A-genome-derived retrotransposons. The evolutionary success of a newly-formed polyploid may require a delicate balance between genomic-stress induced novel variation (McClintock 1983), and stabilization via such quantitative factors as we suggest or qualitative factors such as the ph1 locus enforcing pairing specificity of wheat (Griffiths et al. 2006).

**METHODS**

**Read mapping and SNV detection**

Sequences of cotton D genome v2 (G. raimondii), A genome (G. herbaceum) and a tetraploid genome (G. hirsutum, Acala Maxxa) are from an accompanying paper (Paterson et al. 2012). G. herbaceum is sequenced with read depth of 32X and Acala
Maxxa with read depth of 82X. Sequences of three additional tetraploid genomes (G. hirsutum GA120R1B3 30X, G. hirsutum race yucatanese 15X; G. mustelinum 46X) are from Illumina sequencing of paired-end libraries. Reads are aligned to the reference genome using BWA (Li and Durbin 2009). SNVs between D and A genome are called with samtools/bcftools (Li et al. 2009; Li 2011) using reads with mapping quality >30 and base quality >30. Raw SNVs between progenitor genomes are further filtered by keeping those with read depth between 4 and 60. Raw SNVs for the tetraploid species are further filtered by keeping those with read depth between 7 and twice the average effective read coverage. After aligning reads from the tetraploid genome to the reference genome, alleles are assigned to each subgenome by referring to the parental alleles.

Detection of converted alleles

SNVs from the two progenitor genomes are identified by aligning reads from the A-genome to the reference D-genome. For each of the SNV sites, the orthologous alleles in the tetraploid genomes are sorted into subsets corresponding to the respective parental subgenomes (Fig. S6). A locus that differs between the diploid progenitors but for which a tetraploid shows only the allele from one progenitor (i.e. is monomorphic) is regarded as converted, if meeting the following criteria. Assuming that alleles at a locus follow a binomial distribution when sampling tetraploid DNA, read coverage of 7x is necessary to keep the false positive rate under 0.0078. To remove false positives caused by under-sampling, we filtered out monomorphic alleles with read coverage under 7. False-positive converted alleles can also be derived from deletion of a progenitor allele in a subgenome. In this event, read coverage of deleted sites would average half that of the sites that show
both parental alleles. Average mapped read coverage for the four tetraploid cotton species studied are shown in Fig. S7. Reduction of read coverage by half is not observed in either of the conversion categories, compared to sites that show both parental alleles.

**Quality control**

To rule out the possibility that inferred conversions are due to the At subgenome being more divergent than the Dt from the reference genome, the effect of relaxing the value of editing distance (-n flag in BWA) was investigated. If the sequence divergence was a major factor, the magnitude of conversion bias would be reduced after relaxing the editing distance. The default value for –n flag is 0.04. The estimated mean nucleotide divergence between A and D genomes is ~0.64%, ranging from 1.13e-5 to 0.012. Theoretically, the default value is sufficiently large to allow the mapping of At-genome reads even for most diverged regions. Fig. S8, A shows the proportion of allele changes resulting from relaxing the value of –n flag to 0.8. The magnitude of conversion bias is not decreased, but increased. At to Dt converted alleles became the most common one, presumably due to an increase of mapped reads from the Dt-subgenome after allowing more mismatches. That is, the relative frequency of At alleles declined under the SNP calling threshold because more Dt-subgenome reads aligned to the reference. To test this, we reduced the frequency threshold of calling a heterozygous SNP (Fig. S8, B). We observed an increase of unchanged alleles and decrease of both converted types. Expectedly, the log-ratio of the two conversion types also increased (Fig. S9). We also tried reads with mapping quality from 1 to 30 with increment of 5; all of the tests show similar results. We used reads with mapping quality > 20 for this study.
We further investigated possible artifacts of read alignment by assessing the sequence divergence between the A and D genomes near HeGCE sites. We divided the genome into non-overlapping 10Kb bins. For each bin, we calculated the number of At to Dt converted alleles and the nucleotide divergence between the A and D genomes. The nucleotide divergence in each bin is normalized by the proportion of sites covered by reads to remove the variation of reads coverage. The Pearson correlation coefficient of nucleotide divergence and At to Dt conversions is 0.089, which indicates only weak positive correlations (Fig. S10, A). To compare with other types of allele changes, we also looked into the correlation between Dt to At converted sites and unchanged sites. The other two types also show similar level of correlations (r=0.0613 for Dt to At converted alleles; r=0.0979 for unchanged alleles) (Fig. S10, B, C).

**Evaluation of gene function impact**

Gene function impact of allele changes are measured relative to the cotton reference gene models as described (Paterson *et al.* 2012), by the following four categories: a) Altered translation initiation site - allele changes in the first codon of a coding sequence that leads to an amino acid other than methionine; b) Altered splicing sites - allele changes disrupting the “GT-AG” conserved sites flanking introns; c) Introduced stop codons - allele changes introducing premature stop codons into the normal protein coding sequence; d) Altered stop codons - allele changes altering stop codon to encode an amino acid.
**Estimation of the size of biased-conversion tracks**

Due to the large variance of measurement of the length of conversion tracks in base pairs (bp), we use number of continuous At to Dt conversion alleles as a measure. We searched for continuous tracks of conversion alleles both in the genome and in HC. If we assume random mutation follows a Poisson distribution, the length of continuous mutations should follow an exponential distribution. Both genome wide and HC conversion tracks show longer continuous tracks than expected (Fig. S2). Conversion tracks in the HC show an excess of long ones (4-6 alleles) and a deficiency of shorter ones (<3 alleles) than the genome wide set. From the distribution, the average length of conversion tracks across the genome is 3.75 continuous alleles. Genome wide, the average distance between three continuous conversion alleles is 320.3 bp and 412.6 bp for four. Estimation of the average length of conversion tracks across the genome is (320.3/3+412.6/4)*3.75/2 = 394 bp. In HC, the average length of conversion tracks in HC is 4.32 continuous alleles, with the average length of four continuous At to Dt conversion alleles being 600.8 bp and 647.3 bp for five. Estimation of the average length of conversion tracks in HC is (600.8/4+647.3/5)*4.32/2 = 604bp.

**DATA ACCESS**

The *G. raimondii* genome sequence is in NCBI with BioProject accession PRJNA171262. Other sequences are available at the NCBI short read archive for *G. herbaceum* (accession F1-1, SRA061660), *G. hirsutum* (accession GA120R1B3, SRA068148), *G. hirsutum* (race yucatanese, SRA068479) and *G. mustelinum* (SRA068485).
ACKNOWLEDGMENTS

The authors appreciate financial support from the U. S. National Science Foundation (IIP-0917856, IIP-1127755 to AHP); USDA (ARS-58-6402-7-241, 58-6402-1-644 and 58-6402-1-645 to DGP, ARS 6402-21310-003-00 to BES); Bayer Crop Science, The Consortium for Plant Biotechnology Research, and Cotton, Inc (AHP).
Table 1. Number and proportion (%) of conversion types in HC and LC.

<table>
<thead>
<tr>
<th></th>
<th>HC (At-&gt;Dt)</th>
<th>HC (Dt-&gt;At)</th>
<th>LC (At-&gt;Dt)</th>
<th>LC (Dt-&gt;At)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A -&gt; G</td>
<td>6932(19.63)</td>
<td>1047(16.62)</td>
<td>2147(15.90)</td>
<td>2908(18.88)</td>
</tr>
<tr>
<td>T -&gt; C</td>
<td>6828(19.33)</td>
<td>993(15.76)</td>
<td>2249(16.65)</td>
<td>3020(19.61)</td>
</tr>
<tr>
<td>G -&gt; A</td>
<td>5067(14.35)</td>
<td>989(15.70)</td>
<td>2034(15.06)</td>
<td>1843(11.97)</td>
</tr>
<tr>
<td>C -&gt; T</td>
<td>5029(14.24)</td>
<td>935(14.84)</td>
<td>2091(15.48)</td>
<td>1821(11.82)</td>
</tr>
<tr>
<td>T -&gt; G</td>
<td>2011(5.69)</td>
<td>244(3.87)</td>
<td>772(5.72)</td>
<td>675(4.38)</td>
</tr>
<tr>
<td>A -&gt; C</td>
<td>1902(5.39)</td>
<td>233(3.70)</td>
<td>770(5.70)</td>
<td>690(4.48)</td>
</tr>
<tr>
<td>G -&gt; C</td>
<td>1771(5.01)</td>
<td>407(6.46)</td>
<td>822(6.09)</td>
<td>927(6.02)</td>
</tr>
<tr>
<td>C -&gt; G</td>
<td>1755(4.97)</td>
<td>373(5.92)</td>
<td>780(5.78)</td>
<td>935(6.07)</td>
</tr>
<tr>
<td>C -&gt; A</td>
<td>1091(3.09)</td>
<td>352(5.59)</td>
<td>483(3.58)</td>
<td>656(4.26)</td>
</tr>
<tr>
<td>G -&gt; T</td>
<td>1064(3.01)</td>
<td>323(5.13)</td>
<td>541(4.01)</td>
<td>688(4.47)</td>
</tr>
<tr>
<td>A -&gt; T</td>
<td>936(2.65)</td>
<td>200(3.18)</td>
<td>415(3.07)</td>
<td>603(3.92)</td>
</tr>
<tr>
<td>T -&gt; A</td>
<td>929(2.63)</td>
<td>203(3.22)</td>
<td>402(2.98)</td>
<td>636(4.13)</td>
</tr>
</tbody>
</table>
Fig. 1. Allelic changes in polyploid cotton. (A) Allelic changes causing striking inferred protein changes between A₁ and D₁ subgenomes of tetraploid (*G. hirsutum*) Acala Maxxa. (B) Evolutionary history of inter-genomic conversion in polyploid cotton. The proportion of A₁ and D₁ neutral (intergenic) mutations and conversions in both HC and LC are shown in each window in time. The star marks AD allopolyploidization occurring ~1-2 million years ago. See Fig. S1 for raw data.
Fig. 2. Distributions of biased conversion, heterozygosity and GC-content across chromosomes 4 and 9. Chromosomes are divided into 1Mb bins using sliding windows with 100Kb step size. (A) Heatmap of gene and transposon density across chromosomes. Log-ratio of Z-scores of gene versus transposon density in each bin is shown, with blue marking transposon-rich and red marking gene-rich regions. Red asterisks mark inferred heterochromatin boundaries based on densities of genes and transposons. (B) Genomic distribution of log-likelihood ratio of At to Dt converted alleles to Dt to At converted alleles in tetraploid cotton, calculated for each bin. Red line denotes log-likelihood ratio
of zero. (C) Comparison of GC% between cotton A and D genomes. Blue bins indicate GC% of D > A genome, other bins are red. (D) Genomic distribution of heteroallelic density (i.e. with at least one allele differing from the other three) in tetraploid cotton (Acala Maxxa). Heteroallelic density is calculated for each bin, corrected by read coverage, with below-average density indicated by red lines and average or higher density by blue lines. See Fig. S3 for other chromosomes.
Fig. 3. Genome-wide correlations between conversion, GC% ratio of D to A genome, and transposon frequency (per bp), for 1 Mbp bins using 100 kbp sliding windows.

Conversion ratio is measured by log-likelihood ratio following Fig. 2, B. GC% ratio and transposon density are calculated for each bin. (A) Correlation between conversion and GC% ratios ($R^2=0.506$). Data are best fitted with the quadratic function $y = -0.02392x^2 + 0.07992x + 0.9831$. (B) Correlation between conversion ratios and transposon frequency ($R^2=0.684$). Data are best fitted with the quadratic function $y = -0.1322x^2 + 0.5126x + 0.2981$. 

Fig. 4. Inter-genomic conversion of genes and transposons. (A) Proportion (% by total bp length) of each transposon type in HC and LC. Transposon types are: RLX, LTR-retrotransposon; RLG, Gypsy; TXX, transposon (either transposon or retrotransposon); RLC, Copia; RXX, retrotransposons; DTM, mutator; DTC, CACTA; DXX, DNA transposons; RIX, LINE; Non, others; DTA, hAT; DTT, Tc1-Mariner; DHH, Helitron. See Fig. S4 for counts. (B) Proportion of A,T to G,C conversion in transposons in HC versus LC. (C) Phylogenetic distribution of the log-ratio of nonsynonymous to synonymous conversions in HC and LC. Conversions are more gene-enriched in HC than LC. See Fig. S5 for raw data. (D) Density of converted alleles in each functional region.
References


Chester M., J. P. Gallagher, V. V. Symonds, A. Veruska Cruz da Silva, E. V. Mavrodiev et al., 2012 Extensive chromosomal variation generated in a recently formed


Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078-2079.


Lu P, X. Han, J. Qi, J. Yang, J. Wijeratne Asela et al., 2012 Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Research 22:508-518.


McClintock B., 1983 The significance of responses of the genome to challenge. Nobel lectures physiology or medicine.


Rong J., E. A. Feltus, V. N. Waghmare, G. J. Pierce, P. W. Chee 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.


Wijnker E, G. V. James, J. Ding, F. Becker, J. R. Klasen et al., 2013 The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. eLife 2:e01426.


Zhao X. P., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price et al., 1998 Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res. 8(6):682-682.