Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing

Justin T. Page1, Mark D. Huynh2, Zach S. Liechty2, Kara Grupp3, David Stelly4, Amanda M. Hulse4, Hamid Ashrafi5, Allen Van Deynze5, Jonathan F. Wendel3, and Joshua A. Udall1,2

1Biology Department, Brigham Young University, Provo, UT, 84602
2Plant and Wildlife Science Department, Brigham Young University, Provo, UT, 84602
3Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, 50011
4Department of Soil & Crop Sciences, Texas AgriLife Research, College Station, TX, 77843
5Seed Biotechnology Center, University of California-Davis, Davis, CA, 95616

© The Author(s) 2013. Published by the Genetics Society of America.
Abstract

Understanding the composition, evolution, and function of the *Gossypium hirsutum* (cotton) genome is complicated by the joint presence of two genomes in its nucleus (A₁ and D₁ genomes). These two genomes were derived from progenitor A- and D-genome diploids involved in ancestral allopolyploidization. To better understand the allopolyploid genome, we re-sequenced the genomes of extant diploid relatives that contain the A₁ (*G. herbaceum*), A₂ (*G. arboreum*), or D₅ (*G. raimondii*) genomes. We conducted a comparative analysis using deep re-sequencing of multiple accessions of each diploid species and identified 24M SNPs between the A- and D-diploid genomes. These analyses facilitated the construction of a robust index of conserved SNPs between the A- and D-genomes at all detected polymorphic loci. This index is widely applicable for read mapping efforts of other diploid and allopolyploid *Gossypium* accessions. Further analysis also revealed locations of putative duplications and deletions in the A-genome relative to the D-genome reference sequence. The ~25,400 deleted regions included >50% deletion of 978 genes, including many involved with starch synthesis. In the polyploid genome, we also detected 1,472 conversion events between homoeologous chromosomes including events that overlapped 113 genes. Continued characterization of the *Gossypium* genomes will further enhance our ability to manipulate fiber and agronomic production of cotton.

**Corresponding Author:** Joshua A. Udall, 295 WIDB, Plant and Wildlife Science Department, Brigham Young University, Provo, UT, 84602, USA

**Keywords:** cotton fiber, comparative genomics, molecular evolution, allopolyploid
Article Summary

We aligned resequencing reads to the existing D-genome sequence and discovered novel changes between the A- and D-genomes in both diploid and polyploid plants. We identified single base differences throughout the genome between the diploid genomes and discovered that 978 genes of the D-genome reference sequence are consistently deleted in the A-genome. We discovered that ~900Kbp of sequence in the polyploid genome has been converted from one genome to another in separate conversion events scattered across the genome. These discoveries help us better understand the dynamic nature of polyploid genomes and provide many avenues for further genomic research in cotton.
Introduction

The genus *Gossypium* (cotton) includes approximately 45 diploid species that are divided into eight monophyletic groups, each designated by a single letter denoting (viz., “A” through “G”, and “K”, hereafter referred to as genome groups (Wendel et al. 2012)). Ancient hybridization between A and D diploids resulted in a new allopolyploid (AD) lineage in the New World approximately 1-2 mya (Wendel 1989). Two of the descendant allopolyploid species — *G. hirsutum* (AD₁) and *G. barbadense* (AD₂) — as well as two African-Asian, A diploids — *G. herbaceum* (A₁) and *G. arboreum* (A₂) — were each independently domesticated for their long, spinnable, seed epidermal trichomes. These four species collectively provide the world’s cotton fiber production, with over 90% of this total being due to the cultivation of “upland cotton”, *G. hirsutum* (Wendel and Cronn 2003). Understanding the cotton genome is important for facilitating advances in crop variety development and utilization. In addition, insights into polyploid evolution in cotton may further our understanding of other polyploid crops.

Molecular studies and comparisons between diploid cotton species have revealed a genus with extraordinary genome dynamics. For example, there is a nearly threefold variation in genome sizes among diploids (Wendel and Cronn 2003; Wendel et al. 2012), with the A-genome (1.7 Gbp) being nearly twice the size of the D-genome (0.9 Gbp), largely due to the proliferation of GORGE3 gypsy-like retrotransposons (Hawkins et al. 2006). Despite this size difference, comparative mapping studies have indicated that gene order and colinearity have been largely conserved between the diploid A- and D-genomes (Brubaker et al. 1999), with the corollary that most genome size diversity reflects variation in the rates of proliferation and deletion of repetitive elements (Hawkins et al. 2009; Grover and Wendel 2010). Molecular phylogenetic and
dating studies indicate that the A- and D-genomes diverged ~5-10 million years ago (mya). The F-genome of *G. longicalyx* diverged from the A-genome after the A-D divergence, making it a suitable outgroup for a comparison of the A-genome diploids.

The respective A and D diploid genomes are closely related to the two homoeologous genomes in allopolyploid cotton, A_T and D_T (‘T’ denotes tetraploid), because allopolyploidization is thought to have occurred in the mid-Pleistocene, or 1-2 mya (Wendel 1989). Consequently, genome differences between diploids A_2 and D_5 serve as a fair approximation of the differences between A_T and D_T tetraploid genomes (Udall 2006; Flagel *et al.* 2012). Thus, the existence of models of the diploid progenitors of allopolyploid cotton provide powerful reference points for inference of homoeology (*e.g.*, of genes, transcripts, RNA-seq reads) in allopolyploid cotton. The recent publication of the genome sequence of the D-genome diploid (*G. raimondii*; Paterson *et al.* 2012) allows for the development of new analytical and comparative approaches for the genomics of both diploid and polyploid cotton. For example, a tool was recently created to assign the sequence reads of allopolyploid cotton to their respective genome after mapping reads from A-diploid and AD-polyploid to the *G. raimondii* (D_5) reference sequence (Page *et al.* 2013).

Combined with the rapid increase in available sequence data, these new genomic approaches may facilitate molecular and traditional improvement efforts of cotton. For example, analysis of reads from diploids mapped to a single genome reference provides a straightforward method to identify single nucleotide polymorphisms (SNPs) between and within genomes, since each alignment of reads has identical, relative positions (Page *et al.* 2013). In considering the
relationships among sequences from A-, D-, and AD-genome cotton species, it is useful to
distinguish between two classes of SNPs. Briefly, homoeo-SNPs are fixed differences that
distinguish (and hence diagnose) the A- and D-genomes. Allele-SNPs, on the other hand, are
traditional segregating polymorphisms within a single genome - between the two alleles of an
individual accession (i.e. heterozygosity) or between the corresponding homozygous alleles of
different accessions. Allele-SNPs are those historically used by breeders to improve cotton
cultivars in marker-assisted or genomic selection methods. Homoeo-SNPs add another layer to
practical utility of allele-SNPs, in that they provide a genomic feature to distinguish between
duplicate gene copies. Homoeo-SNPs are also useful in an evolutionary context because their
analysis offers insights the molecular evolutionary properties of allopolyploid cotton, and more
generally, allopolyploid genomes.

To better understand both diploid and allopolyploid cotton genomes, we performed deep, whole
genome re-sequencing of several diploid accessions of both A- and D-genome diploids. Our first
objective was to determine all of the homoeo-SNPs between the A- and D-genomes. Using reads
from these diploids and publicly available reads from diploid and allopolyploid cottons, we
compiled a database of SNPs between the various genomes studied. Our second objective was to
describe genome evolution between the genomes that could be characterized by read coverage.
We also examined loci that are either duplicated or deleted in the A-genome species, based on
coverage of A-genome reads mapped to the D5-genome reference. Where those duplications or
deletions overlap with genes, they may provide insight into the evolutionary basis for the
phenotypic differences among diploids, including the production of spinnable fiber in A-genome
diploid species. Out third objective was document the extent of genome interaction based on
sequence data in the polyploid (i.e. conversion events). A robust description of conversion events throughout the cotton genomes will serve as a bioinformatic aide to future genomic analyses of allopolyploid cotton.
MATERIALS AND METHODS

Plant material. Plant material was grown and harvested from greenhouses at Brigham Young University (D5-2, D5-31, A1-155, A2-34, A2-1011), Iowa State University (D5-4, D5-53, A2-4, A1-73), and Texas A&M (A2-255). DNA was extracted from 4 accessions of *G. raimondii* (D5-2, D5-4, D5-31, D5-53), 2 accessions of *G. herbaceum* (A1-73, A1-155), and 4 accessions of *G. arboreum* (A2-4, A2-34, A2-255, A2-1011) using a Qiagen DNeasy plant kit.

Acquisition of DNA sequence. After shearing DNA with a Covaris instrument at the Huntsman Cancer Institute (Salt Lake City, UT), DNA libraries were prepared with the Illumina TruSeq V3 kit and sequenced by Beijing Genome Institute (BGI, Sacramento, CA), producing 100 bp paired-end reads. We assumed that the Illumina library construction process would perform equally well on high quality DNA of the A- and D-genomes. Approximately 40x genomic coverage was obtained for each library (Table 1). Reads from the diploids have been deposited in the NCBI Sequence Read Archive (SRA) under the following entries: PRJNA202235, PRJNA202236, and PRJNA202239 for *G. arboreum*, *G. herbaceum*, and *G. raimondii*, respectively. Additional genomic sequence reads for *G. longicalyx* (F1-1; SRR617255), *G. herbaceum* (A1-97; SRR617256, SRR617284, SRR617704), and *G. hirsutum* cv. Maxxa (SRR617482) were obtained from the Sequence Read Archive. All reads were trimmed for quality with Sickle, using a minimum phred quality threshold of 20 (https://github.com/najoshi/sickle).
Homoeo-SNP index. An index of homoeo-SNPs between the A- and D-genomes was produced by comparing sequence data from 9 *Gossypium* diploids (A1-97, A1-155, A2-34, A2-255, A2-1011 versus D5-2, D5-4, D5-31, D5-53). First, all reads were mapped with GSNAP (Wu and Nacu 2010), using the options “-n1 –Q” (requiring unique best mapping for each read), to the 13 chromosomes of the D5 reference sequence (Paterson *et al.* 2012). Second, alignment files were processed with SAMtools to produce sorted BAM files (Li *et al.* 2009). Third, we used InterSnp, custom code built on the BAMtools API (https://github.com/pezmaster31/bamtools) and available as part of the BamBam package (http://udall-lab.byu.edu) to call SNPs with at least 10x coverage and a minimum minor allele frequency of 40%. Since the alignments from diploids used the same reference genome, homologous loci in the A- and D-genomes were readily compared to identify SNPs between genomes (homoeo-SNPs). Homoeo-SNPs were called at a locus position of the D-genome reference when all diploid genomes with coverage at that particular locus were 1) homozygous, 2) all A-genome diploids had the same base, and 3) all D-genome diploids had the same base, different from the A-genome diploids. Finally, loci with identified homoeo-SNPs were tabulated into a text file that was converted into a homoeo-SNP index for use by GSNAP and PolyCat.

SNP identification and diversity analysis. Using the homoeo-SNP index, we again mapped the sequence reads from the 9 diploids, this time using GSNAP’s SNP-tolerant mapping (“-v” option). We also mapped reads for 3 additional diploids (F1-1, A1-73, A2-4) and 1 allopolyploid (*G. hirsutum cv. Maxxa*). GSNAP and SAMtools were otherwise used as above. Subsequently, reads from the tetraploid Maxxa were assigned to the A_T- and D_T-genomes using PolyCat (Page *et al.* 2013). All SNPs (homoeo- and allele-SNPs) were called by InterSnp between the 13
resulting BAM files, one for each A or D diploid, one for Maxxa’s Ar-genome, and one for
Maxxa’s Dr-genome. The number of heterozygous loci in each individual was summarized after
filtering loci within conserved duplications (see below). We constructed a neighbor-joining tree
for the various diploid accessions, as well as the Ar- and Dr-genomes, using the PHYLIP
(Felsenstein 1989) program “neighbor” and default settings. The distance matrix consisted of the
percentage of aligned sites that differed in pairwise comparisons.

We used homoeo-SNPs between the diploids to generate a “pseudo-A” genome, with the A
alleles substituted into the Ds reference. We did the same with the Maxxa homoeo-SNPs to make
“pseudo-Ar” and “pseudo-Dr” genomes. While these pseudo-genomes did not have indels or
structural variations that are present in the actual A, Ar, and Dr genomes, the majority of gene
sequences have been conserved (Flagel et al. 2012; Paterson et al. 2012). Thus, these pseudo-
genomes served to characterize the location of SNPs within genes and other conserved non-
coding sequences and having each genome on the same ‘scale’ greatly simplifies genome
comparisons.

**Duplications and deletions.** We detected putative duplications (relative to the Ds reference
genome) in the other diploids studied using MACS (with default settings), a commonly used tool
for ChIP-seq analysis (Zhang et al. 2008). It empirically models peaks in coverage of ChIP-seq
reads using a dynamic Poisson distribution, thereby estimating the location of a DNA binding
molecule. Here we used MACS to call coverage peaks within WGS reads from the A- and F-
genome diploids after alignment to the D-genome reference. Assuming the libraries from both A-
and D- genomes would be equally biased, reads from Ds-53 served as a control sample,
estimating the expected coverage pattern. Peaks in A-genome coverage relative to Ds-53 represent putative duplicated sites that were, consequently, sampled at a higher frequency during sequencing. To filter out false positives, we also called coverage peaks in Ds-2, Ds-4, and Ds-31, relative to Ds-53. We used bedtools (Quinlan and Hall 2010) to compare peaks among and between the datasets and to identify their position relative to genes annotations in the Ds version 2.1 (Paterson et al. 2012).

Similarly, putative deletions in the test diploids were called if the reference sequence Ds-53 had 20x or higher coverage at both ends, as well as at an additional point at least 200 bp from either end of a region of at least 1000 bp, and if the test diploid had near zero coverage (<3x) at every point in that block. This detection was carried out by Gapfall, part of the BamBam package (http://udall-lab.byu.edu). Blast2Go was used for an Enrichment Analysis (using the Fisher’s exact test) on genes duplicated or deleted in the A diploids (Conesa et al. 2005). Default parameters were used.

**Polyploid conversion events.** We used 2 methods to identify possible non-reciprocal homoeologous or “gene conversion” events between the At- and Dt-genomes of *G. hirsutum cv. Maxxa*. We first identified individual converted loci based on homoeo-SNPs, where reads from the At-genome carried the D-genome nucleotide, or vice versa. The second method used duplications and deletions to identify regions of conversion, but for deletion detection using a 15x minimum coverage for the duplicated genome and <4x coverage for the ‘deleted’ genome. If a region spanning at least 1 Kbp was “duplicated” in the At-genome relative to the A diploids and “deleted” in the Dt-genome relative to the D diploids, an At-biased conversion event was
inferred. Similarly, an At-genome deletion and Dt-genome duplication suggested a Dt-biased conversion. These analyses of the polyploid genome were limited to regions that were present in both diploid genomes (A and D) because homoeo-SNPs could only be predicted in such regions.
RESULTS

Intergenomic SNPs. Similar to previous work on the detection and frequency of SNPs in genic regions (Page et al. 2013), we produced a robust index of 23,859,893 homoeo-SNPs between the genomes of diploid A- and D-genome cotton. These SNPs covered the genome of the D-genome reference sequence at a density of one SNP per 32.3 bases (Figure 1). This total number of SNPs is a dramatic increase from the number previously reported with the D-genome sequence (Paterson et al. 2012) and in genic sequences (Page et al. 2013). The index had a transition/transversion ratio of 1.92 (Table 1), similar to the Maize HapMap2 (Chia et al. 2012). This genome-wide SNP analysis confirmed our speculation that the previous ratio was downwardly biased in our gene-focused index. Across polymorphic nucleotide positions, there was not a significant difference between the GC biases of the A- and D-genomes (45.4% and 45.1%, respectively). However, these values were higher than the genome-wide GC content, suggesting an increased likelihood for SNPs at G or C nucleotides, possibly due to the high frequency of C->T mutations caused by de-amination of cytosines.

The genome-wide SNP index (SNP-index 2.0) was based on comparisons of deep sequence coverage between multiple diploid A- and D-genomes, so we anticipated it would be more robust and widely applicable for read mapping efforts of other diploids and allopolyploids than the previous index. The improved index increased mapping efficiency of A-genome reads to the D-genome reference sequence. With GSNAP’s SNP-tolerant mapping, over 77% of A-genome reads mapped, reflecting an approximately 15% mapping improvement compared to mapping without the SNP-index (Table 2; Supplementary Figure 1). D-genome mapping was unaffected (~95%). The error rate of categorization of WGS reads was <2%, as estimated by categorizing
the diploid reads and looking for incorrectly categorized reads. While this error rate is slightly higher than that estimated for the original PolyCat index (Page et al. 2013), it is still an acceptable rate considering the increased fraction of reads that overlap a SNP between genomes (~70%, up from <50%), and considering WGS reads mapping to less conserved intergenic regions.

Within the diploid index, genes had a median intergenomic SNP per base rate of 2.2% (range 0 – 16.1%). Notably, there were 593 genes that had no unambiguous homoeo-SNPs between diploid A- and D-genomes (Supplementary Figure 2). Of these, 215 genes had one or more allele-SNPs (i.e. one diploid genome had two nucleotides, one matching the second diploid genome and the other nucleotide being novel). The remaining 378 genes were completely conserved across all accessions with no SNP differences. A Blast2Go enrichment analysis of these genes identified 3 enriched GO terms: NADH dehydrogenase (ubiquinone) activity (GO:0008137), NADH dehydrogenase (quinone) activity (GO:0050136) and NADH dehydrogenase activity (GO:0003954). Most of these genes were shorter than the average gene within the D-genome annotations (95 to 8113 bp with mean 810±786 s.d. for the 378 genes vs. 89 to 51,174 bp with mean 3249±2806 s.d. for all 37,223 genes; Supplementary Figure 2) (Paterson et al. 2012).

In the polyploid, improved categorization of reads into its two separate genomes was enabled by the genomic SNP-index. Using the SNPs from the diploids and the D-genome reference, PolyCat assigned over 70% of mapped polyploid reads to the A_T- or D_T-genomes (Supplementary Figure 3). For the tetraploid Maxxa, a greater percentage of reads were assigned to the A_T-genome than
to the Dr-genome, despite the fact that categorization only occurred in regions shared by the 2 genomes. This later criteria pre-empted the larger sized A-genome from an AT categorization bias. The unexpectedly higher categorization rate of AT reads may be partially explained by the fact that A1 and A2 diploids are a 2x better approximation of the AT-genome than D5 is of the DT-genome. For example, nucleotide diversity appears to play a role in mapping efficiency among the diploid A-genome species. The most divergent line (A1-73) had the lowest mapping percentage of any of the of the A-genome diploid accessions. Because sequence divergence is less between the A-genome diploid and polyploid than the D-genome and polyploid, read categorization based on SNPs between the diploids would be more effective for the AT-genome, resulting in the observed bias. To a much lesser degree, the AT categorization bias may also be partially attributed to duplicated loci in the AT-genome mapping to a single locus in the D5 reference, though these artifacts were largely avoided by the detection of duplications (see below).

Because of their recent common ancestry, many of the identified differences between the A- and D-genome diploids were retained between the AT- and DT-genomes as homoeo-SNPs. A total of 20,828,020 homoeo-SNPs were identified between the AT- and DT-genomes of the allopolyploid cultivar Maxxa. The difference between the number of ~20M SNPs in the polyploid and the “retained ancestral” homoeo-SNPs (~16M; 75.8%) were autapomorphic SNPs that were derived after the divergence of the AT- and Dr-genomes from the A- and D-genome common ancestor. This portion of the homoeo-SNPs (5,046,151; 24.2%) were only identified between the genomes in the allopolyploid, but not in the comparison of the diploid genome sequences. These unique, homoeo-SNPs were found throughout the genome, in 34,810 of the 37,223 annotated genes. We
anticipate that additional polyploid autopomorphic SNPs will be identified as more polyploid
genomes are resequenced.

For all of the annotated genes in the D-genome reference, an alignment of A, A_{\text{T}}, D_{\text{T}}, and D-
genomes was created, from which the amount of molecular evolution between the A and D
genomes of cotton was calculated (Table 3). The results of this effort concurred with our
previously published work based on aligned EST contigs (Flagel et al. 2012), although standard
errors were much smaller because of the much larger data set. We found slightly less divergence
(dN and dS) between the polyploid genomes (n = 28,317) than between the diploid genomes (n =
30,874), though the difference is not significant. The different totals between the diploids and
polyploids suggested that >2,500 genes in the polyploid did not have sufficient polymorphism
for an appropriate estimation of molecular evolution. We further investigated the alignments of
these genes to ascertain whether their close sequence similarity was the result of gene conversion
between homoeologous genomes. Of the genes without dN/dS estimates, 759 were found to only
have sufficient polymorphisms between the tetraploid genomes (and not between the diploid
genomes), 3,316 were found to have sufficient polymorphisms between the diploids (but not
between the tetraploid genomes). This cumulative large difference between ploidy levels further
suggested that gene conversions may play a role in reducing genetic diversity between genomes.
However, only 106 and 42 genes were detected to overlap ‘conversion regions’ in the diploids
and tetraploid genome (see below).

Of the 2,817,991 SNPs between diploids that fell within genes, 486,514 were inferred to be in
exonic positions, including 248,599 that caused amino acid changes (i.e., non-synonymous)
compared to the reference sequence. Of these, there were 1,651 genes with SNPs that resulted in premature stop codons in the pseudo-translation of A-genome transcripts, 1,802 genes with premature stop codons in AT, and 709 genes with pre-mature stop codons in DT (Figure 2). These genes were not excluded in estimates of molecular evolution (above). None of these gene sets had any enriched GO terms. The low level of DT premature stops may simply reflect an ascertainment bias of the annotated reference genome that was based on a diploid D genome. Many of the putative stop codons were found near the annotated end of the gene suggesting that they might have only a minimal impact on protein function. Alternatively, their inference may reflect bioinformatic artifacts, such as imperfect gene annotation of the D-genome, or alternative stops that independently evolved in the A-genomes. Most of these alternative stops codons were within 10% of the 3’ end of the gene. If one ignores the premature stop codons within the last 10% of the annotated genes, 803 premature stops were shared between the diploid A-genomes and the AT-genome (Figure 2). This result was marginally less than previously reported (Paterson et al. 2012), because we had the added power of multiple A-genome re-sequencing efforts.

Other SNPs disrupted a start or stop codon. We identified 806 genes with disrupted start codons (i.e., resulting in an amino acid distinct from that in the D5 reference) in the A genome, 703 in AT, and 684 in DT. These genes could have a longer or shorter coding sequence than as originally annotated. No GO terms were enriched in these gene sets. We also identified 831 genes with altered stop codons in the A genome, 693 in AT, and 437 in DT, resulting in longer peptide sequences. Several GO terms (~20 in each genome) were enriched within these genes, almost all associated with photosynthesis (Supplementary Table 1). There were also 406 genes without a
stop codon within the D5 gene annotation, with the same photosynthesis GO terms being enriched.

**Diversity and heterozygosity.** In addition to creating an index of nucleotide differences between the diploid A- and D-genomes, we detected unique nucleotide variation within and between individual accessions. Within a genome type (*i.e.* A or D), these type of SNPs are called allele-SNPs. The allelic genotype of each diploid and both genomes of the allopolyploid Maxxa were determined at all polymorphic loci. A pairwise comparison between accessions found that the D5 diploids had extremely low nucleotide diversity (< 1 million SNPs) between any 2 accessions, while a similar pairwise comparison between the A1 and A2 genomes found that accessions were more diverse (4-5 million SNPs within A1 or A2; 6-8 million SNPs between A1 and A2; Supplementary Table 2). There were approximately twice as many SNPs between the A1 and A2 genomes as within either of the two species. These results are not unexpected given the exceptionally low diversity found in a survey of allozyme diversity in *G. raimondii* (Wendel, unpublished data), and the appreciable levels of diversity in the chosen accessions of *G. arboreum* and *G. herbaceum* (Wendel *et al.* 1989).

In addition to having more fixed allele-SNPs between accessions, the A-genome diploids were more heterozygous than the D-genome diploids (~13% and <1%, respectively; Table 4). In the A-genome diploids, heterozygous loci were about twice as frequent outside than inside of genes. This was not surprising, given the expectation of more intense purifying selection on coding sequences. Of course, these estimates of heterozygosity excluded loci that were duplicated in the A-genome (see below). Interestingly, heterozygous loci in the D-genome diploids were equally
common in genic and non-genic regions. This genomic difference likely reflects both the exceptionally low genetic diversity within the D-genome and a high level of generalized inbreeding. In this respect we note that *G. raimondii* has a narrow natural range, and presently exists as only scattered populations with very low effective population sizes.

A neighbor-joining tree was constructed and rooted based on the known relationship between the A/F-genome clade and the D-genome clade (Figure 3). Many fixed allele-SNPs (i.e. not heterozygous within a line) could be attributed to mutations occurring along specific branches of the phylogeny (Table 5). The tree correctly reconstructed the accepted relationships of the diploids and their relationships to the two genomes of allopolyploid cotton (*Senchina et al.* 2003; Grover *et al.* 2012; Wendel *et al.* 2012). Specifically, and unsurprisingly, A_T and D_T were phylogenetically sister to the common ancestor of the [A_1 + A_2] clade and the D_5 clade, respectively. Our results agreed with previous reports that the A_1 or A_2 diploids were about 2x as good an approximation of the A_T-genome as the D_5 diploids were of the D_T-genome (Wendel and Cronn 2003; *Senchina et al.* 2003). The distance from D_T to D_5 was 0.14, and the distance from A_T to the common ancestor of A_1 and A_2 was 0.08 (0.14/0.08 ~ 2x). However, the distances from A_T to A_1 and A_2 was 0.10 and 0.11, respectively. Moreover, the distance from A_T to any individual A-genome diploid was 0.14, similar to the distance between D_T and any D-genome diploid. While group of A-genome diploid provided a 2x better approximation of A_T than a group of D_5 diploids does of D_T, any individual A or D diploid appeared to be equally similar to its A_T or D_T counterpart. The exceptionally low diversity among D_5 diploids explains the fact that a group of D_5-genome diploids is not significantly better for approximation of the D_T-genome than is a single D_5. On the other hand, multiple accessions of A diploids (A_1 and/or A_2)
do provide a substantial (nearly 2x) improvement in construction of the A\textit{r}-genome pseudo-sequence.


**Duplications and deletions.** Duplications were detected in A-genome diploids as coverage peaks across the D-genome reference sequence (Figure 1). Because these duplications were detected relative to the D-genome diploids, they represent events that occurred after the split of these two clades 5-10 mya (Wendel and Cronn, 2003; Wendel et al., 2012). Thus, peaks shared by all A-genome accessions represent pieces of the A-genome that are duplicated relative to the D-genome. These coverage peaks represent a mix of tandem and dispersed duplications, as the methodology employed makes no distinction regarding genomic location of duplicated segments.

There were 30,709 regions duplicated in all A-genome diploids but not duplication in D-genome diploids. These duplicated blocks overlapped 1,007 genes, with a minimum overlap of 50% of the gene length. Only 1 GO term was enriched among these genes: structural constituent of ribosome (GO: 0003735).

In contrast to duplications, putatively deleted regions of the A-genome were detected with a higher degree of certainty because their diagnosis is based on lack of coverage rather than a quantitative difference in coverage (Figure 1). Some regions of the D-genome reference genome did not have any A-genome reads mapped to them, despite a 40x WGS coverage based on the number of produced A-genome reads and correctly mapping D-genome reads to the same region. Each accession had a unique set of deleted regions, including genes (Supplementary Figure 4).

There were 25,408 regions deleted in all A-diploid genomes. The genomic regions included 978 annotated D-genome reference genes where the deleted region minimally overlapped of 50% of
the gene length. Among the genes within deleted regions, 118 GO terms were enriched compared
to the population of GO terms within the annotated gene set (Paterson et al. 2012, Supplementary
table 3). Most of these deletion terms were associated with starch synthesis, tRNAs, or DNA
repair mechanisms. Three hundred seventy-eight genes were completely deleted in the A-
genome diploids relative to the D diploids, meaning that the genic region was spanned by a
single deletion block.

**Polyploid conversion events.** We used 2 different methods to detect conversion events in the
allopolyploid genome of Acala Maxxa (*G. hirsutum*). The first method identified historical
nonreciprocal homoeologous recombination events (NRHR) at individual loci based on homoeo-
SNPs within a polyploid genome. Earlier analyses in cotton used this method to detect
conversion events (or NRHR events) based on comparative analysis of assembled EST
sequences from diploid and allopolyploid cotton (Salmon et al., 2009; Flagel et al., 2012). The
second method identified converted regions based on coverage patterns in the Aτ- and Dτ-
genomes relative to their respective diploid relatives. Here we consider the NRHR events as
“conversion” events regardless if they occur in coding or non-coding sequences (*sensu amplo* of
gene conversion).

Based on homoeo-SNPs (first method), 1,748,889 conserved SNPs in 29,576 genes in the diploid
genomes suggested an Aτ-biased allele conversion (a Dτ nucleotide converted to the Aτ
nucleotide). In contrast, a total of 361,795 SNPs in 12,346 genes suggested a Dτ-biased
conversion. These data suggest a nearly 5x bias based on homoeo-SNPs and a 2.5x bias based on
genes in favor of $A_T$-biased conversion, in stark contrast to the 2x bias in favor of $D_T$-biased conversion reported previously (Paterson et al. 2012).

Based on coverage/deletion information (second method), conversion events were found in both directions across 882 Kbp of the D-genome reference sequence (Figure 1). These events ranged from 1 to 5470bp with a median of 337bp (Supplementary Figure 5). Two hundred and fifty-nine regions suggested an $A_T$-biased conversion. These regions spanned 275 Kbp and overlapped the coding sequence of 19 genes. They also included 12,696 putative homoeo-SNPs (based on the diploids), none of which were detected within the tetraploid. On the other hand, 1,213 regions showed a $D_T$-biased conversion. These regions spanned 607 Kbp and overlapped the coding sequence of 94 genes. They included 21,142 putative homoeo-SNPs, of which only 3 were also detected within the tetraploid. The genes overlapped by these regions had no enriched GO terms and indicated a conversion bias in both directions, but with the $D_T$ direction most prominent, similar to that previously reported (Paterson et al. 2012). The events detected by the second method only included 1,375 of the possibly conversion-related SNPs identified by the first method.
**DISCUSSION**

**Genome resources for *Gossypium*** Using the diploid resequencing data, we created several useful resources for the *Gossypium* genome. First, a genome-wide map of the SNPs between the diploid A- and D-genomes of cotton was created. Re-sequencing multiple accessions of each diploid enabled us to distinguish bases that were specific to a single accession from bases that are more representative of one diploid genome or another. It also allowed us to identify conserved genomic features shared by all A-genome or D-genome species and accessions. In that sense, the multiple accessions of each species acted as re-sequencing replications of the A- or D-genome ‘treatments’. We have demonstrated that most SNPs identified between diploid genomes can be directly extrapolated to differences between the descendant allopolyploid genomes (*i.e.* homoeo-SNPs) because of their recent common ancestor. In addition, we also identified several million homoeo-SNPs that were unique to the Maxxa allopolyploid genome. These documented SNPs can be used for genome identification of individual sequence reads (Udall 2006; Flagel *et al.* 2012) or the development of genotyping assays (Van Deynze *et al.* 2009; Byers *et al.* 2012). With an index of homoeo-SNPs, read-mapping efficiency was significantly improved and future false positive allele-SNPs can be filtered out of marker sets resulting in more reliable allele-SNP assessment. In addition to the homoeo-SNPs, we also identified allele-SNPs in the diploid cotton accessions. These SNP sets are available as Gbrowse tracks at CottonGen ([http://www.cottongen.org](http://www.cottongen.org)), and as gff files at the Udall lab website ([http://udall-lab.byu.edu](http://udall-lab.byu.edu)).

A second resource is the set of alignments of gene and protein sequences of the A-, At-, Dt- genomes to accompany the previously published *G. raimondii* annotations (Paterson *et al.* 2012). These alignments were used to identify SNPs and to further refine our understanding of the
molecular evolutionary differences between genomes. Because we have made this a public resource, any researcher investigating cotton now has homoeo- and allele-SNPs information for any target gene already identified. This simple, yet tedious task has been a common obstacle of genetic research in polyploid cotton.

A third resource, and one that we suggest will be a fruitful topic for further investigation, is the description of putative duplications and deletions that distinguish the A- and D-genomes, and hence originated subsequent to their divergence from a common ancestor. These localized structural variations offer a rich source of sequences to mine for possible functional consequences, and to further our understanding the mechanisms of copy number variation during genome evolution in plants.

Through our read mapping efforts, we noticed that the limited number and the stochastic distribution of homoeo-SNPs could have implications for de novo genome assembly of polyploid cotton. Although 70% of the reads from allopolyploid cotton could be assigned to one of its two co-resident genomes, 30% of the reads that mapped to the D-genome reference did not overlap a homoeo-SNP, and hence could not be categorized. Using an arbitrary length of 1000bp, we found 47,399 unique loci where sequence reads of the A_T-genome and D_T-genome were indistinguishable when compared to each other and to the reference genome. Assuming sequence read lengths <500bp, these regions would likely co-assemble during a de novo whole genome shotgun assembly with current read lengths. Consequently, co-assembled segments will create unique challenges of graph structure bifurcation (or higher branching) during the contig construction steps of de novo assembly. Part of this challenge could be addressed by generating
reads with a greater likelihood of overlapping homoeo-SNPs, i.e. longer reads. Present data, however, suggest that *de novo* assembly of the allopolyploid cotton genome would not be successful if based on contemporary read lengths.

**Insights into the genome biology of Gossypium.** One of the intriguing results of this study is the insight it provides into the origin and frequency of indels during A- and D-genome divergence. Because of the lack of an outgroup sequence, none of the “duplications” or “deletions” described here are polarized, so their duplicate or deleted status is only relative to the single D-genome reference. Moreover, the methods employed do not yield insights into the mechanistic underpinnings of the indels, which may conceivably entail a full spectrum of deletional mechanisms and processes of tandem and dispersed duplication.

Notwithstanding, the present study does reveal the scope and scale of the indel generating process during 5-10 million years of diploid evolution. Additionally intriguing are the genomic distributions of the duplicated and deleted regions. For example, Chromosome 13 is notable for its high frequency of duplications, containing one sixth (2850/17102) of the total number of conserved duplications in the A-genome, yet only 2.9% (174/6072) of the deletions. Given that Chromosome 13 comprises a mere 7.8% of the *G. raimondii* reference sequence (Paterson et al., 2012), the suggestion arises that there has been exceptional expansion and/or contraction of this chromosome during the evolution of the two-fold size difference that distinguishes the A- and D-genomes. While some of this difference certainly reflects the expansion of transposable elements in the A-genome or possible contraction in the D-genome (Hawkins *et al.* 2009), it is unclear
what genomic features have allowed more numerous rearrangements in chromosome 13 than
within other chromosomes in the genome.

In addition to this broad-scale view of the contributions of duplications and deletions to
Gossypium genome evolution, the data presented here offer a rich database that can be mined for
potentially significant gene duplication and deletion. For example, gene loss has been associated
with polyploidization (Shaked et al. 2001; Ozkan 2003; Han et al. 2005; Tate et al. 2009), but
the deletions we have described in the A-genome occurred prior to polyploidization and include
parts of ~1,300 genes per accession. If these A-genome accessions were used as a ‘parental
genome reference’ for investigations of polyploidy, the deletions common to the A1-, A2-, and
ancestral A1-genomes would be confounded with any putative deletions that occurred as a result
of polyploidization. Thus, this initial database of duplications and deletions will be a useful
research tool for investigations of the evolution of the Gossypium genome.

We observed that several genes involved in starch synthesis were deleted in the A-genome
diploids, including seven genes with 1,4-alpha-glucan branching enzyme activity
(Supplementary Table 3). It is tempting to speculate that these deletions increased the amount of
glucose available in the A-genome diploid for cellulose synthesis and thereby played a role in the
increased length of mature, A-genome cotton fibers. Previous studies have documented altered
carbon partitioning (Yong-Ling Ruan 1998) and altered starch accumulation (Chaohua et al.
2005) in fiberless G. hirsutum mutants. The deletion of starch genes in the A-genome may have
been associated with the opposite effect resulting in more carbon being allocated to cellulose
production and less to starch production. We caution that many of the deleted genes are members
of gene families and the remaining paralogs may partially or fully compensate for their deletion in the A-genome diploids. Nevertheless, the deleted genes discovered in this study offer interesting avenues of future research of gene duplication and functional compensation.

Among genes with altered stop codons, we detected an enriched number of genes having photosynthesis-related functions. It appears unlikely that the altered stop codons are due to horizontal transfer of chloroplast genes to the nucleus since only 4 have high similarity to chloroplast genes. While a biological explanation for this enrichment remained a mystery, it was likely that a portion of the enrichment of photosynthesis GO terms was an artifact of the gene annotation process. For example, the actual stop codon location may have been ambiguous since the original annotation of these genes actually had no stop codon. Perhaps, the initial gene annotation effort was simply unable to identify the full coding sequence and subsequent updates will include corrections to the original annotations. Regardless of the initial annotation, the enrichment of photosynthesis GO terms in genes with altered stop codons was interesting, but it was not due to differences between the A-, D-, AT-, and Dr-genomes.

We also identified homogenized genome regions from conversion events between the two homoeologous genomes of *G. hirsutum*. Gene conversion has been defined as the nonreciprocal transfer of genetic information between homologous sequences, leading to homogenization during meiotic or mitotic recombination (Szostak *et al.* 1983; Chen *et al.* 2007; Hsu *et al.* 2010; Jacquemin *et al.* 2011). Unlike analyses of pairs of genes (Drouin 2002; Mondragon-Palomino and Gaut 2005; Xu *et al.* 2008), whole genome analysis of gene conversion has been pioneered in rice (Xu *et al.* 2008; Wang *et al.* 2009; Jacquemin *et al.* 2011), a sequenced diploid genome
with many closely related species also with sequenced genomes. By comparison of the diploid re-sequencing data to the publicly available WGS data of Maxxa Acala, we were able to identify conversion events between homoeologous sequences within the polyploid *Gossypium* genome using two different methods. The two methods of detecting homoeologous conversion events resulted in different directional biases.

Using the first method, we had previously used SNPs to estimate that up to 5% of the polyploid transcriptome had experienced ‘homoeologous gene conversion’ (Salmon *et al.* 2010; Flagel *et al.* 2012). In both previous studies, identification of autapomorphic SNPs was not possible because of limited diploid sequencing data. Based on our current data, the presence of autapomorphic SNPs (and a liberal method of identification) appeared to have caused an overestimate in the amount of homoeologous conversion in genic regions. Thus, the genome sequence of a definitive outgroup is needed to unambiguously identify regions of conversion using SNP information alone. One dimension of the conversion events and the multi-alignment resource for all genes is the identification of loci where one of the two allopolyploid genomes has “overwritten” the other, via a mechanism of reciprocal or non-reciprocal “gene conversion”. At present, the functional consequences of these observations remain unexplored, but it is intriguing to ask whether these conversion events are functionally insignificant (which might, for example, be the case when only synonymous sites are involved), or if instead, specific genes or regulatory sequences have been selectively “doubled” or “eliminated” by this unusual, intergenomic aspect of allopolyploid speciation.
This first method contains an inherent bias in favor of $\text{At}$-biased conversion because of the greater genetic distance between $\text{Dt}$ and $\text{Ds}$, compared to the distance between $\text{At}$ and $\text{A}_1$ or $\text{A}_2$. This bias, however, should only be about 50% based on our understanding of the genetic distances between $\text{A}$ and $\text{At}$ vs. $\text{D}$ and $\text{Dt}$ (the latter is 50% greater; Flagel et al., 2012). In addition, the genotype pattern indicative of a conversion is indistinguishable from that caused by an autapomorphic SNP in the diploid species. For example, suppose an autapomorphic SNP in the $\text{A}$-genome ancestor of $\text{A}_1$ and $\text{A}_2$ (not shared with $\text{At}$) changes a $\text{C}$ to a $\text{T}$ at a given locus. Consequently, the $\text{D}$-genome diploids, $\text{Dt}$-genome, and $\text{At}$-genome would all have a $\text{C}$, while the $\text{A}$-genome diploids would have a $\text{T}$. So the $\text{At}$ nucleotide would appear the same as $\text{D}$ and $\text{Dt}$, suggesting a $\text{Dt}$-biased allele conversion event, even though they simply shared the ancestral allele. These confounding autapomorphic SNPs would have occurred after the divergence of $\text{At}$ but before the $\text{A}_1$-$\text{A}_2$ split suggesting a $\text{Dt}$-biased conversion. On the other hand, an $\text{At}$-biased conversion would be suggested by an autapomorphic SNP occurring after the divergence of $\text{Dt}$, but before the most recent common ancestor of the $\text{Ds}$ diploids. Because the $\text{A}$-genome diploid is approximately a 2x better approximation of the actual progenitor diploid than is the $\text{D}$ genome diploid (reviewed in Wendel et al., 2012), these branch lengths are very different (Figure 3). In fact, the phylogeny showed a distance of only 0.00447 for the branch corresponding to an autapomorphic SNP shared by all $\text{A}$ diploids but not by $\text{At}$. On the other hand, the equivalent branch in the $\text{D}$-genome clade has an autapomorphic SNP distance of 0.05385. These numbers suggest that, in the absence of any actual conversion events, there should be over 12x (0.05385/0.00447) as many SNPs that look like $\text{At}$-biased conversion—because they are shared by $\text{A}$ diploids but not by $\text{At}$—as SNPs that look like $\text{Dt}$-biased conversion. The difference between the 12x expected value for $\text{At}$-biased conversions as visualized by the branch lengths to
the 5x observed value could be explained by a bias towards Dr-biased conversion events, as reported elsewhere and as detected by the second, coverage-based method. Thus, we consider the conversion events detected by the SNP-based method (method 1) to be inaccurate based on autapomorphic SNPs, while the conversion events of the coverage method (method 2) to be a conservative, yet relatively accurate assessment of conversion between the polyploid *G. hirsutum* genomes.

The second method of conversion detection used deletion and coverage information to detect many separate events and the direction of bias agreed with previous reports (Paterson *et al.* 2012). This method was very conservative and may represent a minimum amount of conversion events in the polyploid genome because of the uncertainty of the actual endpoints of conversion and the additional amounts of conversion suggested between homoeologous gene copies in the dN/dS analysis. The conversion events resulted in a loss of genomic diversity between the AT- and Dr-genomes. Parts of at least 113 genes were included in conversion events between homoeologous chromosomes. Other investigations of genome evolution in rice have uncovered convergent evolution of ancient paralogs on Chr11 and Chr12 (~2.1 Mb) mediated by gene conversion including up to 180 genes (Jacquemin *et al.* 2011). The conversion events we have described were more recent (after polyploidization 1-2 Ma) and our inference space was limited to a single species of *G. hirsutum*. It will be interesting to see if other polyploid Gossypium genomes also have the same conversion events in their genomes and to estimate the rate of gene conversion between homoeologous genomes.
Whole genome re-sequencing of diploid *Gossypium* has identified insights into the genome evolution of cotton. These insights proved to be useful for characterization of the *G. hirsutum* genome via publicly available re-sequencing data. Additional *de novo* and re-sequencing efforts of polyploid Gossypium will continue to add to our understanding of the cotton genome thereby enhancing our ability to manipulate the fiber and agronomic characteristics of cotton.
Acknowledgements

We thank the National Science Foundation and Cotton Inc. for their financial support of this work. We thank the Fulton SuperComputing Lab (FSL) at BYU for their generous support using the supercomputer. The FSL provided server environment for all of the included bioinformatics analysis described in this work. We also thank Brian Dalley of the Huntsman Cancer Institute for his assistance with DNA shearing and his generous insights of the Illumina HiSeq sequencing.
References


Drouin, G., 2002 Characterization of the gene conversions between the multigene family members of the yeast genome. Journal of Molecular Evolution 55: 14–23


Han, F., G. Fedak, W. Guo, and B. Liu, 2005 Rapid and repeatable elimination of a parental genome-specific DNA repeat (pGc1R-1a) in newly synthesized wheat allopolyploids Genetics 170: 1239-1245

Han, M. V., and C. M. Zmasek, 2009 phyloXML: XML for evolutionary biology and comparative genomics BMC bioinformatics 10: 356

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


Shaked, H., K. Kashkush, H. Ozkan, M. Feldman, and A. A. Levy, 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


Yong-Ling Ruan, P. S. C., 1998 A fiberless seed mutation in cotton is associated with lack of fiber cell initiation in ovule epidermis and alterations in sucrose synthase expression and carbon partitioning in eveloping seeds Plant Physiology 118: 399

Table 1. Transitions and transversions in the homoeo-SNP index (rows = A allele, columns = D allele). There was an overall transition/transversion ratio of about 1.92 and GC fractions of 45.4% (A genome) and 45.1% (D genome).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>2,495,527</td>
<td>626,075</td>
<td>1,003,583</td>
</tr>
<tr>
<td>G</td>
<td>2,547,739</td>
<td>-</td>
<td>353,034</td>
<td>647,148</td>
</tr>
<tr>
<td>C</td>
<td>644,840</td>
<td>352,239</td>
<td>-</td>
<td>2,544,261</td>
</tr>
<tr>
<td>T</td>
<td>1,003,739</td>
<td>628,050</td>
<td>2,492,619</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2. Summary of nine diploid WGS resequencing libraries that were re-sequenced in this study and additional libraries (A1_97, F1_1, and Maxxa) were obtained from the SRA.

<table>
<thead>
<tr>
<th>Accession</th>
<th>PI</th>
<th>Raw Pairs</th>
<th>Trimmed Reads</th>
<th>Raw Mapping</th>
<th>Mapped % using SNP-index 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1_155</td>
<td>630024</td>
<td>385,657,228</td>
<td>761,269,884</td>
<td>65.3%</td>
<td>78.0%</td>
</tr>
<tr>
<td>A1_73</td>
<td>485587</td>
<td>202,723,343</td>
<td>238,035,929</td>
<td>53.1%</td>
<td>85.6%</td>
</tr>
<tr>
<td>A1_97</td>
<td>529670</td>
<td>328,713,056</td>
<td>652,350,335</td>
<td>65.0%</td>
<td>77.8%</td>
</tr>
<tr>
<td>A2_1011</td>
<td>629339</td>
<td>412,420,252</td>
<td>816,274,495</td>
<td>58.3%</td>
<td>73.8%</td>
</tr>
<tr>
<td>A2_255</td>
<td>615756</td>
<td>300,406,057</td>
<td>595,289,591</td>
<td>61.1%</td>
<td>75.5%</td>
</tr>
<tr>
<td>A2_34</td>
<td>183160</td>
<td>367,844,399</td>
<td>729,370,248</td>
<td>62.3%</td>
<td>76.5%</td>
</tr>
<tr>
<td>A2_44</td>
<td>185788</td>
<td>78,180,657</td>
<td>153,728,823</td>
<td>63.3%</td>
<td>76.9%</td>
</tr>
<tr>
<td>A2_4</td>
<td>529707</td>
<td>343,470,023</td>
<td>686,940,046</td>
<td>48.9%</td>
<td>82.4%</td>
</tr>
<tr>
<td>D5_2</td>
<td>530899</td>
<td>152,913,856</td>
<td>304,706,886</td>
<td>95.6%</td>
<td>95.3%</td>
</tr>
<tr>
<td>D5_31</td>
<td>530928</td>
<td>217,334,954</td>
<td>428,323,703</td>
<td>95.8%</td>
<td>95.5%</td>
</tr>
<tr>
<td>D5_4</td>
<td>530901</td>
<td>310,387,080</td>
<td>616,432,521</td>
<td>95.1%</td>
<td>94.8%</td>
</tr>
<tr>
<td>D5_53</td>
<td>530950</td>
<td>188,469,224</td>
<td>375,193,268</td>
<td>96.2%</td>
<td>96.0%</td>
</tr>
<tr>
<td>F1_1</td>
<td>530986</td>
<td>534,258,839</td>
<td>1,055,751,863</td>
<td>71.1%</td>
<td>79.1%</td>
</tr>
<tr>
<td>Maxxa Acala</td>
<td>540885</td>
<td>463,761,132</td>
<td>919,898,042</td>
<td>72.5%</td>
<td>79.8%</td>
</tr>
</tbody>
</table>
Table 3. Amounts of molecular evolution between the A and D genomes of cotton.

<table>
<thead>
<tr>
<th></th>
<th>dN</th>
<th>dS</th>
<th>dN/dS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A vs. D</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.0094</td>
<td>0.0276</td>
<td>0.3726</td>
</tr>
<tr>
<td>Median</td>
<td>0.0068</td>
<td>0.0256</td>
<td>0.2768</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0106</td>
<td>0.0225</td>
<td>0.4236</td>
</tr>
<tr>
<td>n = 28,462</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A_T vs. D_T</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.0092</td>
<td>0.0266</td>
<td>0.3772</td>
</tr>
<tr>
<td>Median</td>
<td>0.0066</td>
<td>0.0237</td>
<td>0.2843</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0104</td>
<td>0.0228</td>
<td>0.4156</td>
</tr>
<tr>
<td>n = 26,156</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Number of heterozygous loci in each accession, along with the percentage of total observable loci that were heterozygous.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Whole genome (%)</th>
<th>Genic loci only (%)</th>
<th>Non-genic loci only (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1_1</td>
<td>9,968,998, 17.2%</td>
<td>332,247, 6.1%</td>
<td>9,636,751, 18.4%</td>
</tr>
<tr>
<td>A1_73</td>
<td>2,963,374, 7.1%</td>
<td>126,260, 2.6%</td>
<td>2,837,114, 7.7%</td>
</tr>
<tr>
<td>A1_97</td>
<td>6,504,768, 12.4%</td>
<td>265,607, 5.0%</td>
<td>6,239,161, 13.2%</td>
</tr>
<tr>
<td>A1_155</td>
<td>7,549,531, 13.9%</td>
<td>322,095, 6.0%</td>
<td>7,227,436, 14.8%</td>
</tr>
<tr>
<td>A2_4</td>
<td>7,061,224, 13.2%</td>
<td>283,151, 5.3%</td>
<td>6,778,073, 14.1%</td>
</tr>
<tr>
<td>A2_34</td>
<td>6,826,660, 12.9%</td>
<td>270,384, 5.1%</td>
<td>6,556,276, 13.7%</td>
</tr>
<tr>
<td>A2_255</td>
<td>5,898,387, 11.6%</td>
<td>236,113, 4.5%</td>
<td>5,662,274, 12.4%</td>
</tr>
<tr>
<td>A2_1011</td>
<td>6,878,801, 13.1%</td>
<td>252,230, 4.9%</td>
<td>6,626,571, 14.1%</td>
</tr>
<tr>
<td>D5_2</td>
<td>193,418, 0.3%</td>
<td>20,536, 0.4%</td>
<td>172,882, 0.3%</td>
</tr>
<tr>
<td>D5_4</td>
<td>257,399, 0.4%</td>
<td>25,370, 0.5%</td>
<td>232,029, 0.4%</td>
</tr>
<tr>
<td>D5_31</td>
<td>178,290, 0.3%</td>
<td>20,723, 0.4%</td>
<td>157,567, 0.3%</td>
</tr>
<tr>
<td>D5_53</td>
<td>181,224, 0.3%</td>
<td>20,665, 0.4%</td>
<td>160,559, 0.3%</td>
</tr>
<tr>
<td>Maxxa.A</td>
<td>4,465,088, 9.2%</td>
<td>198,477, 3.9%</td>
<td>4,266,611, 9.8%</td>
</tr>
<tr>
<td>Maxxa.D</td>
<td>686,674, 1.3%</td>
<td>47,861, 1.0%</td>
<td>638,813, 1.4%</td>
</tr>
</tbody>
</table>
Table 5. SNPs attributable to specific areas of the phylogeny (Figure 3). Note that, because of possible conversion events, it is not possible to determine how many SNPs were shared by all A or D diploids.

<table>
<thead>
<tr>
<th>Genome(s)</th>
<th>SNPs</th>
<th>Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>All A</td>
<td>5,544,440</td>
<td>25,408</td>
</tr>
<tr>
<td>A₁</td>
<td>1,024,299</td>
<td>3,809</td>
</tr>
<tr>
<td>A₂</td>
<td>1,152,825</td>
<td>2,941</td>
</tr>
<tr>
<td>Aᵢ</td>
<td>1,472,900</td>
<td>5,247</td>
</tr>
<tr>
<td>All D</td>
<td>14,601,331</td>
<td>0</td>
</tr>
<tr>
<td>Dᵢ</td>
<td>3,563,979</td>
<td>4,518</td>
</tr>
</tbody>
</table>
**Figure Legends**

**Figure 1.** Plot of genes, homoeo-SNPs, duplications, deletions, and conversion events in the A-genomes, relative to the D5 reference sequence, produced by Circos (Krzywinski et al. 2009). Considering the concentric circles from the outside inward, the outermost (and first) green circle indicates the location of annotated genes. The next circle (red) is a histogram of the number of homoeo-SNPs in a 1 Mbp window throughout the genome. The next 2 red (high frequency) to yellow (low frequency) circles are heat maps showing the location of duplications in the A1 and A2 genomes as compared to the D5 genome (A2 interior). The next 2 blue (high frequency) to yellow (low frequency) circles are heat maps showing the location of deletions in the A1 and A2 genomes as compared to the D5 genome (A2 interior). The final 2 circles show conversion events in the tetraploid *G. hirsutum cv. Maxxa*. The first circle shows conversion of loci to the A nucleotide on a red to yellow scale, while the innermost (and last) circle shows conversion of loci to the D nucleotide on a blue to yellow scale.

**Figure 2.** Premature stop codons were found in each *Gossypium* genome. A) Premature stop codons (compared to the annotations of the D-reference genome) were found in the A, A_T, and D_T genomes. B) Common genes with premature stop codons in the first 90% of the gene.

**Figure 3.** Neighbor-joining tree built by PHYLIP [30], based on SNPs between genomes. Units (as measured by the indicated scale) are percentage of represented polymorphic sites that differed between 2 individuals. Image rendered by Archaeopteryx (Han and Zmasek 2009).
a) Number of Genes

<table>
<thead>
<tr>
<th>Normalized Gene Position</th>
<th>A</th>
<th>At</th>
<th>Dt</th>
</tr>
</thead>
<tbody>
<tr>
<td>359</td>
<td>803</td>
<td>548</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>24</td>
<td>482</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Venn Diagram

- A
- At
- Dt

Legend:
- Blue: A
- Red: At
- Green: Dt