Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing

Justin T. Page  
*Brigham Young University - Provo*

Mark D. Huynh  
*Brigham Young University - Provo*

Zach S. Liechty  
*Brigham Young University - Provo*

Kara Grupp  
*Iowa State University*

David M. Stelly  
*Texas AgriLife Research*

Follow this and additional works at: [http://lib.dr.iastate.edu/eeob_ag_pubs](http://lib.dr.iastate.edu/eeob_ag_pubs)

Part of the Botany Commons, Ecology and Evolutionary Biology Commons, Genomics Commons, and the Plant Breeding and Genetics Commons

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/eeob_ag_pubs/65](http://lib.dr.iastate.edu/eeob_ag_pubs/65). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing

Justin T. Page,* Mark D. Huynh,† Zach S. Liechty,† Kara Grupp,§ David Stelly,§ Amanda M. Hulse,* Hamid Ashrafi,** Allen Van Deynze,** Jonathan F. Wendel,§ and Joshua A. Udall*,†,†

*Biology Department, Brigham Young University, Provo, Utah 84602, †Plant and Wildlife Science Department, Brigham Young University, Provo, Utah 84602, ‡Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, §Department of Soil and Crop Sciences, Texas AgriLife Research, College Station, Texas 77843, and **Seed Biotechnology Center, University of California-Davis, Davis, California 95616

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-genome 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 (Gossypium herbaceum), A2 (Gossypium arboreum), or D3 (Gossypium raimondii) genomes. We conducted a comparative analysis using deep re-sequencing of multiple accessions of each diploid species and identified approximately 24 million SNPs between the A-diploid and D-diploid genomes. These analyses facilitated the construction of a robust index of conserved SNPs between the A-genomes 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 approximately 25,400 deleted regions included more than 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.

KEYWORDS cotton fiber comparative genomics molecular evolution allopolyploid

We aligned re-sequencing reads to the existing D-genome sequence and discovered novel changes between the A-genomes 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 approximately 900 Kbp of sequence in the polyploid genome have 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.

The genus Gossypium (cotton) includes approximately 45 diploid species that are divided into eight monophyletic groups, each designated by a single letter (“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 million years ago (Wendel 1989). Two of the descendant allopolyploid species—Gossypium hirsutum (AD1) and Gossypium barbadense (AD2)—as well as two African-Asian A diploids—Gossypium herbaceum (A1) and Gossypium arboreum (A2)—were each independently domesticated for their long, spinnable, epidermal seed trichomes. These four species collectively provide the world’s cotton fiber production, with more than 90% of this total being attributable 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 three-fold 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 because of 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-genomes 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-genomes and D-genomes diverged approximately 5–10 million years ago. The F-genome of *Gossypium 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, AT and DT ("T" denotes tetraploid), because allopolyploidization is thought to have occurred during the mid-Pleistocene era, or 1–2 million years ago (Wendel 1989). Consequently, genome differences between genomes AT and DT serve as a fair approximation of the differences between AT and DT 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 (*Gossypium 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-diploids and AD-polyploids to the *G. raimondii* (D5) 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, because each alignment of reads has identical relative positions (Page et al. 2013). In considering the relationships among sequences from A-genome, D-genome, and AD-genome cotton species, it is useful to distinguish between two classes of SNPs. Briefly, homoec-SNPs are fixed differences that distinguish (and hence diagnose) the A-genomes and D-genomes. Allele-SNPs, however, 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 accesses. Allele-SNPs are those historically used by breeders to improve cotton cultivars in marker-assisted or genomic selection methods. Homoec-SNPs add another layer to practical utility of allele-SNPs in that they provide a genomic feature to distinguish between duplicate gene copies. Homoec-SNPs are also useful in an evolutionary context because their analysis offers insights into 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-genome and D-genome diploids. Our first objective was to determine all of the homoec-SNPs between the A-genomes 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 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 spinable fiber in A-genome diploid species. Our third objective was to 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, A2-73), and Texas A&M University (A2-255). DNA was extracted from four accessions of *G. raimondii* (D5-2, D5-4, D5-31, D5-53), two accessions of *G. herbaceum* (A1-73, A1-155), and four 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-genomes and D-genomes. Approximately 40-times the 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).

**Homoec-SNP index**

An index of homoec-SNPs between the A-genomes and D-genomes was produced by comparing sequence data from nine *Gossypium* diploids (A1-97, A1-155, A2-34, A2-255, A2-1011 vs. D5-2, D5-4, D5-31, D5-53). 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-genomes and D-genomes. Approximately 40-times the 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).

**Table 1 Transitions and transversions in the homoec-SNP index**

<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>2,495,527</td>
<td>626,075</td>
<td>1,003,583</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>2,547,739</td>
<td>353,034</td>
<td>647,148</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>644,840</td>
<td>352,239</td>
<td>2,544,261</td>
<td></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>

Rows = A allele. Columns = D allele. There was an overall transition/transversion ratio of ~1.92 and GC fractions of 45.4% (A genome) and 45.1% (D genome).
D_{5}-31, D_{5}-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 D_{5} 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, a 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/Research/Software/BamBam.aspx) to call SNPs with at least 10-times coverage and a minimum minor allele frequency of 40%. Because the alignments from diploids used the same reference genome, homologous loci in the A-genomes 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 homozygous, all A-genome diploids had the same base, and 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 nine diploids, this time using the SNP-tolerant mapping (“-v”) option of GSNAP. We also mapped reads for three additional diploids (F1-1, A1-73, A2-4) and one allopolyploid (G. hirsutum cv. Maxxa). GSNAP and SAMtools were otherwise used as noted. Reads from the tetraploid Maxxa were assigned to the A_{5}-genomes and D_{1}-genomes using PolyCat (Page et al. 2013). All SNPs (homoeo-SNPs and allele-SNPs) were called by InterSnp between the 13 resulting BAM files, one for each A or D diploid, one for the A_{5}-genome of Maxxa, and one for the D_{5}-genome of Maxxa. The number of heterozygous loci in each individual was summarized after filtering loci within conserved duplications. We constructed a neighbor-joining tree for the various diploid accessions, as well as the A_{5}-genomes and D_{5}-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 D_{5} reference. We did the same with the Maxxa homoeo-SNPs to make "pseudo-A_{5}r" and "pseudo-D_{5}r" genomes. Although these pseudo-genomes did not have indels or structural variations that are present in the actual A, A_{5}, and D_{5} genomes, the majority of gene sequences were conserved (Flagel et al. 2012; Paterson et al. 2012). Thus, these pseudo-genomes served to characterize the location of allele-SNPs within genes and other conserved noncoding sequences, and having each genome on the same "scale" greatly simplifies genome comparisons.

**Duplications and deletions**

We detected putative duplications (relative to the D_{5} reference genome) in the other diploids 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-genome and F-genome diploids after alignment to the D_{5} reference genome. Assuming the libraries from both A-genomes and D-genomes would be equally biased, reads from D_{5}-53 served as a control sample, estimating the expected coverage pattern. Peaks in A-genome coverage relative to D_{5}-53 represent putative duplicated sites that were sampled at a higher frequency during sequencing. To filter out false-positives, we also called coverage peaks in D_{5}-2, D_{5}-4, and D_{5}-31, relative to D_{5}-53. We used bedtools (Quinlan and Hall 2010) to compare peaks among and between the datasets and to identify their position relative to gene annotations in the D_{5} version 2.1 (Paterson et al. 2012).

Similarly, putative deletions in the test diploids were called if the reference sequence D_{5}-53 had 20-times 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 (<3-times) at every point in that block. This detection was performed by Gapfall, part of the BamBam package (http://udall-lab.byu.edu). Blast2Go was used for an enrichment analysis (using the Fisher exact test) on genes duplicated or deleted in the A diploids (Conesa et al. 2005). Default B2G parameters were used.

**Polyploid conversion events**

We used two methods to identify possible nonreciprocal homoeologous or "gene conversion" events between the A_{5}-genomes and D_{5}-genomes of G. hirsutum cv. Maxxa. We first identified individual converted loci based on homoeo-SNPs, where reads from the A_{5}-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 15-times the minimum coverage for the duplicated genome and less than four-times the coverage for the "deleted" genome. If a region spanning at least 1 Kbp was "duplicated" in the A_{5}-genome relative to the A diploids and "deleted" in the D_{5}-genome relative to the D diploids, then an A_{5}-biased conversion event was inferred. Similarly, an A_{5}-genome deletion and D_{5}-genome duplication suggested a D_{5}-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-genome 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-genomes 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 because of the high frequency of C→T mutations caused by deamination of cytosines.

The genome-wide SNP index (SNP index 2.0) was based on comparisons of deep sequence coverage between multiple diploid A-genomes 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 the SNP-tolerant mapping of GSNAP, more than 77% of A-genome reads mapped, reflecting a mapping improvement of approximately 15% compared to mapping without the SNP-index (Table 2, Supporting Information, Figure S1). D-genome mapping was unaffected (~95%). The error rate of categorization of WGS reads was less than 2%, as estimated by categorizing the diploid reads and looking for incorrectly categorized reads. Although 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-genomes and D-genomes (Figure S2). 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 accesses with no SNP differences. A Blast2Go enrichment analysis of these genes identified the following three enriched GO terms: NADH dehydrogenase (ubiquinone) activity (GO:0008137); NADH dehydrogenase (quinone) activity (GO:0005136); 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 SD for the 378 genes vs. 89 to 51,174 bp with mean 3249 ± 2806 SD for all 37,223 genes; Figure S2) (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 more than 70% of mapped polyploid reads to the A1-genomes or D1-genomes (Figure S3). For the tetraploid Maxxa, a greater percentage of reads were assigned to the A1-genome than to the D1-genome, despite the fact that categorization only occurred in regions shared by the two genomes. This later criteria preempted the larger A-genome from an A1 categorization bias. The unexpectedly higher categorization rate of A1 reads may be partially explained by the fact that A1 and A2 diploids are a two-fold better approximation of the A1-genome than D5 is of the D1-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 A1-genome, resulting in the observed bias. To a much lesser degree, the A2 categorization bias may also be partially attributed to duplicated loci in the A1-genome mapping to a single locus in the D5 reference, although these artifacts were largely avoided by the detection of duplications.

Because of their recent common ancestry, many of the identified differences between the A-genome and D-genome diploids were retained between the A1-genomes and D1-genomes as homoeo-SNPs. A total of 20,828,020 homoeo-SNPs were identified between the A1-genomes and D1-genomes of the allopolyploid cultivar Maxxa. The difference between the number of approximately 20 million SNPs in the polyploid and the “retained ancestral” homoeo-SNPs (~16 million; 75.8%) were autapomorphic SNPs that were derived after the divergence of the A1-genome and D1-genome from the A-genome and D-genome common ancestor, respectively. This portion of the homoeo-SNPs (5,046,151; 24.2%) was only identified between the genomes in the allopolyploid and 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 autapomorphic SNPs will be identified as more polyploid genomes are re-sequenced.

For all of the annotated genes in the D-genome reference, an alignment of A, A1, D1, and D-genomes was created, from which the amount of molecular evolution between the A-genomes 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 SEs were much smaller because of the much larger dataset. We found slightly less divergence (dN and dS) between the polyploid genomes (n = 28,317) than between the diploid genomes (n = 30,874), although the difference is not significant. The different totals between the diploids and polyploids suggested that more than 2500 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...
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 genes, and almost all were associated with peptidylprolyl cis-trans isomerase activity. However, only 106 and 42 genes were detected to overlap "conversion regions" in the diploids and tetraploid genome.

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., nonsynonymous) compared to the reference sequence. Of these, there were 1651 genes with SNPs that resulted in premature stop codons in the pseudo-translation of A-genome transcripts, 1802 genes with premature stop codons in A1-genomes, 703 in AT, and 709 genes with premature stop codons in DT (Figure 2). These genes were not excluded in estimates of molecular evolution. None of these gene sets had any enriched GO terms. The low level of DT genes were not excluded in estimates of molecular evolution. None of these alternative stop codons were within 10% of the transcript length, and none were found near the annotated start codon.

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 reference sequence. Of these, there were 1651 genes with SNPs that resulted in premature stop codons in the pseudo-translation of A-genome transcripts, 1802 genes with premature stop codons in A1-genomes, 703 in AT, and 709 genes with premature stop codons in DT (Figure 2). These genes were not excluded in estimates of molecular evolution. None of these alternative stop codons were within 10% of the transcript length, and none were found near the annotated start codon.

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 A1, 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 A1, and 437 in DT, resulting in longer peptide sequences. Several GO terms (~20 in each genome) were enriched within these genes, and almost all were associated with photosynthesis (Table S1). 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-genomes and D-genomes, we detected unique nucleotide variation within and between individual accessions. Within a genome type (i.e., A or D), these types 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 two accessions, whereas a similar pairwise comparison between the A1-genomes 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; Table S2). There were approximately twice as many SNP between the A1-genomes 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* (J. F. 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 approximately 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. Interestingly, heterozygous loci in the D-genome diploids were equally common in genic and nongenic

### Table 2 Summary of nine diploid WGS re-sequencing libraries that were re-sequenced in this study and additional libraries (A1_97, F1_1, and Maxxa) 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_5_4_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>530966</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 Amount of molecular evolution between the A and D genomes of cotton

<table>
<thead>
<tr>
<th>Comparison</th>
<th>dN</th>
<th>dS</th>
<th>dN/dS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. D</td>
<td>0.0094</td>
<td>0.0276</td>
<td>0.3726</td>
</tr>
<tr>
<td>n = 24,462</td>
<td>Median</td>
<td>0.0068</td>
<td>0.0256</td>
</tr>
<tr>
<td>SD</td>
<td>0.0106</td>
<td>0.0225</td>
<td>0.4236</td>
</tr>
<tr>
<td>A1 vs. A2</td>
<td>0.0092</td>
<td>0.0266</td>
<td>0.3772</td>
</tr>
<tr>
<td>n = 26,156</td>
<td>Median</td>
<td>0.0066</td>
<td>0.0237</td>
</tr>
<tr>
<td>SD</td>
<td>0.0104</td>
<td>0.0228</td>
<td>0.4156</td>
</tr>
</tbody>
</table>
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, A1 and D1 were phylogenetically sister to the common ancestor of the [A1 + A2] clade and the D2 clade, respectively. Our results agreed with previous reports that the A1 or A2 diploids were approximately twice as good of an approximation of the A1-genome as the D2 diploids were of the D1-genome (Wendel and Cronn 2003; Senchina et al. 2003). The distance from D1 to D2 was 0.14, and the distance from A1 to the common ancestor of A1 and A2 was 0.08. However, the distances from A1 to A1 and A2 were 0.10 and 0.11, respectively. Moreover, the distance from A2 to any individual A-genome diploid was 0.14, similar to the distance between D1 and any D-genome diploid. Although the group of A-genome diploids provided an approximation of A1 that was two-times better than a group of D2 diploids did of D1, any individual A or D diploid appeared to be equally similar to its A1 or D1 counterpart. The exceptionally low diversity among D2 diploids explained the fact that a group of D2-genome diploids was not significantly better for approximation of the D1-genome than was a single D2. However, multiple accessions of A diploids (A1 and/or A2) did provide a substantial (nearly two-times better) improvement in construction of the A1-genome pseudo-sequence.

Duplications and deletions

Duplications were detected in A-genome diploids as coverage peaked 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 million years ago (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, because the methodology used makes no distinction regarding genomic location of duplicated segments. There were 30,709 duplications in all A-genome diploids but no duplication in D-genome diploids. These duplicated blocks overlapped 1007 genes, with a minimum overlap of 50% of the gene length. Only one 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 40-times the 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 (Figure S4). 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 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) (Table S3). 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-genome diploids, meaning that the genic region was spanned by a single deletion block.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Whole Genome n</th>
<th>%</th>
<th>Genic Loci Only n</th>
<th>%</th>
<th>Nongenic Loci Only n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1_1</td>
<td>9,968,998</td>
<td>17.2</td>
<td>332,247</td>
<td>6.1</td>
<td>9,636,751</td>
<td>18.4</td>
</tr>
<tr>
<td>A1_73</td>
<td>2,963,374</td>
<td>7.1</td>
<td>126,260</td>
<td>2.6</td>
<td>2,837,114</td>
<td>7.7</td>
</tr>
<tr>
<td>A1_97</td>
<td>6,504,768</td>
<td>12.4</td>
<td>265,607</td>
<td>5.0</td>
<td>6,239,161</td>
<td>13.2</td>
</tr>
<tr>
<td>A1_155</td>
<td>7,549,531</td>
<td>13.9</td>
<td>322,095</td>
<td>6.0</td>
<td>7,227,436</td>
<td>14.8</td>
</tr>
<tr>
<td>A2_4</td>
<td>7,061,224</td>
<td>13.2</td>
<td>283,151</td>
<td>5.3</td>
<td>6,778,073</td>
<td>14.1</td>
</tr>
<tr>
<td>A2_34</td>
<td>6,826,660</td>
<td>12.9</td>
<td>270,384</td>
<td>5.1</td>
<td>6,556,276</td>
<td>13.7</td>
</tr>
<tr>
<td>A2_255</td>
<td>5,898,387</td>
<td>11.6</td>
<td>236,113</td>
<td>4.5</td>
<td>5,662,274</td>
<td>12.4</td>
</tr>
<tr>
<td>A2_1011</td>
<td>6,878,801</td>
<td>13.1</td>
<td>252,230</td>
<td>4.9</td>
<td>6,626,571</td>
<td>14.1</td>
</tr>
<tr>
<td>D5_2</td>
<td>193,418</td>
<td>0.3</td>
<td>20,536</td>
<td>0.4</td>
<td>172,882</td>
<td>0.3</td>
</tr>
<tr>
<td>D5_4</td>
<td>257,399</td>
<td>0.4</td>
<td>25,370</td>
<td>0.5</td>
<td>232,029</td>
<td>0.4</td>
</tr>
<tr>
<td>D5_31</td>
<td>178,290</td>
<td>0.3</td>
<td>20,723</td>
<td>0.4</td>
<td>157,567</td>
<td>0.3</td>
</tr>
<tr>
<td>D5_53</td>
<td>181,224</td>
<td>0.3</td>
<td>20,665</td>
<td>0.4</td>
<td>160,559</td>
<td>0.3</td>
</tr>
<tr>
<td>Maxxa.A</td>
<td>4,465,088</td>
<td>9.2</td>
<td>198,477</td>
<td>3.9</td>
<td>4,266,611</td>
<td>9.8</td>
</tr>
<tr>
<td>Maxxa.D</td>
<td>686,674</td>
<td>1.3</td>
<td>47,861</td>
<td>1.0</td>
<td>638,813</td>
<td>1.4</td>
</tr>
</tbody>
</table>
**Polyplod conversion events**

We used two 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 polyplod 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. 2010; Flagel et al. 2012). The second method identified converted regions based on coverage patterns in the A1- genomes and D1-genomes relative to their respective diploid relatives. Here, we consider the NRHR events as "conversion" events regardless of whether they occur in coding or in noncoding sequences (sensu amplio of gene conversion).

Based on homoeo-SNPs (first method), 1,748,889 conserved SNPs in 29,576 genes in the diploid genomes suggested an A1- biased allele conversion (a D1 nucleotide converted to the A1 nucleotide). In contrast, a total of 361,795 SNPs in 12,346 genes suggested a D1- biased conversion. These data suggest a nearly five-fold bias based on homoeo-SNPs and a two-and-one-half-fold bias based on genes in favor of A1-biased conversion, in stark contrast to the two-fold bias in favor of D1-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 5470 bp, with a median of 337 bp (Figure S5). Two hundred 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) and as gff files at the Udall laboratory web site (http://udall-lab.byu.edu).

**DISCUSSION**

**Genome resources for Gossypium**

Using the diploid re-sequencing data, we created several useful resources for the Gossypium genome. First, a genome-wide map of the SNPs between the diploid A-genomes 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 replicates of the A-genome 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) and as gff files at the Udall laboratory web site (http://udall-lab.byu.edu).

A second resource is the set of alignments of gene and protein sequences of the A-genomes, A1- genomes, and D1-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
Table 5 SNPs attributable to specific areas of the phylogeny

<table>
<thead>
<tr>
<th>Genome</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>3809</td>
</tr>
<tr>
<td>A2</td>
<td>1,152,825</td>
<td>2941</td>
</tr>
<tr>
<td>AT</td>
<td>1,472,900</td>
<td>5247</td>
</tr>
<tr>
<td>All D</td>
<td>14,601,331</td>
<td>0</td>
</tr>
<tr>
<td>DT</td>
<td>3,563,979</td>
<td>4518</td>
</tr>
</tbody>
</table>

As shown in Figure 3. Because of possible conversion events, it is not possible to determine how many SNPs were shared by all A or D diploids.

differences between genomes. Because we have made this a public resource, any researcher investigating cotton now has homoeo-SNPs 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-genomes 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 of 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 1000 bp, we found 47,399 unique loci where sequence reads of the A1-genome and D1-genome were indistinguishable when compared to each other and to the reference genome. Assuming sequence read lengths less than 500 bp, 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-genome and D-genome divergence. Because of the lack of an outgroup sequence, none of the “duplications” or “deletions” described here is polarized, so their duplicate or deleted status is only relative to the single D-genome reference. Moreover, the methods used do not yield insights into the mechanistic underpinnings of the indels, which may conceivably entail a full spectrum of deleterious 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 (2850/17,102) 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-genomes and D-genomes. Although 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 before polyploidization and include parts of approximately 1300 genes per accession. If these A-genome accessions were used as a "parental genome reference" for investigations of polyploidy, then the deletions common to the A1-genomes, A2-genomes, and ancestral A-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 (Table S3). 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 attributable to horizontal transfer of chloroplast genes to the nucleus because only four have high similarity to chloroplast genes. Although 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 because 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 attributable to differences between the A-genomes, D-genomes, A1-genomes, and D1-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 genomic 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 nonreciprocal “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 intergenic aspect of allopolyploid speciation.

This first method contains an inherent bias in favor of A2-biased conversion because of the greater genetic distance between D2 and D3 compared to the distance between A2 and A1 or A2. This bias, however, should only be approximately 50% based on our understanding of the genetic distances between A and A2 vs. D and D2 (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 indicative of a conversion is indistinguishable from that caused by an occurred after the divergence of AT but before the A1–A2 split, suggest-}

branch in the 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 more than 12-times (0.05385/0.00447) as many SNPs that look like A2-biased conversion—because they are shared by A diploids but not by AT—as SNPs that look like D2-biased conversion. The difference between the 12-fold expected value for A2-biased conversions as visualized by the branch lengths to the five-fold observed value could be explained by a bias toward D2–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, but we consider 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 A1- and D1- 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.

ACKNOWLEDGMENTS

We thank Brian Dalley of the Huntsman Cancer Institute for his assistance with DNA shearing and his generous insights regarding the Illumina HiSeq sequencing. We thank the Fulton SuperComputing Laboratory (FSL) at Brigham Young University for their generous financial support of this work. We thank the National Science Foundation and Cotton Inc. for their financial support of this work.

LITERATURE CITED


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


Communicating editor: J. A. Scott