The Cytonuclear Dimension of Allopolyploid Evolution: An Example from Cotton Using Rubisco

Lei Gong,1 Armel Salmon,1,2 Mi-Jeong Yoo,1 Kara K. Grupp,1 Zining Wang,3 Andrew H. Paterson,3 and Jonathan F. Wendel*1

1Department of Ecology, Evolution and Organismal Biology, Iowa State University
2UMR CNRS 6553 Ecobio, Universite de Rennes 1, Rennes, France
3Plant Genome Mapping Laboratory (Department #398), University of Georgia

*Corresponding author: E-mail: jfw@iastate.edu.
Associate editor: Naoki Takebayashi

Abstract

During allopolyploid speciation, two divergent nuclear genomes merge, yet only one (usually the maternal) of the two sets of progenitor organellar genomes is maintained. Rubisco (1,5-bisphosphate carboxylase/oxygenase) is composed of nuclear-encoded small subunits (SSUs) and plastome-encoded large subunits (LSUs), providing an ideal system to explore the evolutionary process of cytonuclear accommodation. Here, we take initial steps in this direction, using Gossypium allopolyploids as our model. SSU copies from divergent (5–10 My) progenitor diploids ("A" and "D" genomes) were combined at the time of polyploid formation 1–2 Ma, with the LSU encoded by the maternal A-genome parent. LSU genes from A- and D-genome diploids and AD-genome allopolyploids were sequenced, revealing several nonsynonymous substitutions and suggesting the possibility of differential selection on the nuclear-encoded rbcS partner following allopolyploid formation. Sequence data for the rbcS gene family revealed nonreciprocal homoeologous recombination between A- and D-rbcS homoeologs in all polyploid species but not in a synthetic intergenomic F1 hybrid, demonstrating "gene conversion" during allopolyploid evolution. All progenitor rbcS genes are retained and expressed in the five extant allopolyploid species, but analysis of the leaf transcriptome showed that A-homoeologs are preferentially expressed in both the allopolyploid and hybrid, consistent with the maternal origin of rbcL. Although rbcS genes from both progenitor genomes are expressed, some appear to have experienced mutations that may represent cytonuclear coevolution.

Key words: polyploid, Gossypium, cytonuclear accommodation, rubisco, homoeologous, recombination, selective expression.

Introduction

Polyploid speciation involves the presence of two or more diploid parental genome sets within an organism without a change in the accompanying plastid and mitochondrial genomes. Polyploidy may occur via autopolyploidization, by multiplying a single genome, or via allopolyploidization, by combining divergent genomes (Soltis PS and Soltis DE 2000; Wendel 2000; Wendel and Doyle 2005). Recent analyses have demonstrated that all angiosperms have experienced at least one episode of polyploidy in their past, with many lineages having undergone repeated cycles of genome doubling (Wendel 2000; Cui et al. 2006; Jiao et al. 2011). In contrast to autopolyploids, which often exhibit subtle genomic and transcriptomic changes (Ozkan et al. 2006; Stupar et al. 2007; Parisod et al. 2010), allopolyploids often experience a diverse array of fascinating and complex genomic alterations, including homoeologous exchanges, non-Mendelian loss of genes or nongenic DNA, and epigenetic modifications (Liu and Wendel 2003; Adams and Wendel 2005b; Chen 2007; Doyle et al. 2008; Leitch AR and Leitch IJ 2008; Jackson and Chen 2010). At the gene expression level, allopolyploidy is accompanied by various forms of nonadditivity, genomic dominance, and biased expression (Adams and Wendel 2005a; Gaeta et al. 2007; Ha et al. 2009; Rapp et al. 2009; Flagel and Wendel 2010).

In addition to the genomic and transcriptomic alterations that occur in response to genome merger and doubling, cytonuclear accommodation would seem to be an important and understudied aspect of allopolyploid evolution, given the complex coordination between the nuclear, plastid, and mitochondrial genomes (Taylor 1989; Leon et al. 1998). During allopolyploid speciation, two different nuclear genomes merge, yet only one (usually the maternal) of the two sets of progenitor organellar genomes is maintained. Thus, the stoichiometry between organellar and nuclear genes is altered, and two divergent regulatory hierarchies become combined, potentially resulting in physiological disruptions. Relatively, little is known about this cytonuclear dimension of polyploid evolution, although based on gene balance theory (Birchler and Veitia 2007, 2010) and the common observation of cytoplasmic male sterility in interspecific hybrids (Schnable and Wise 1998), it seems likely that cytonuclear coevolution is a significant aspect of the stabilization and evolution of allopolyploid lineages. Rubisco (1,5-bisphosphate carboxylase/oxygenase), which derives from nuclear-encoded small subunits (SSUs)
and plastome-encoded large subunits (LSUs), provides a useful model for exploring the evolutionary process of cytonuclear accommodation. SSUs encoded by a nuclear rbcS multigene family are translated in the cytoplasm and are imported into the chloroplast, where rbcL is transcribed and translated as the LSU (Roderme et al. 1996). Assembled octamer holoenzymes of SSU and LSU are responsible for carbon fixation during photosynthesis. Here, we describe aspects of the rubisco system and cytonuclear coordination in allopolyploid cotton (Gossypium). Gossypium is an excellent model for studying this cytonuclear “shock” because of the well-documented formation of natural allopolyploid (AD genomes; 2n = 52) cottons ~ 1 to 2 Ma from hybridization between an A genome (2n = 26), African species much like modern Gossypium arboreum (A2), and a D genome (2n = 26), American species similar to modern G. rai-

modii (D5) (Senchina et al. 2003; Wendel and Cronn 2003; Wendel et al. 2009, 2010). Five allotetra-

polyploid cotton species are widely recognized, including the two cultivated species G. hirsutum (AD1), or “Upland cotton,” and G. barbadense (AD2), or “Pima” cotton. In addition to these two economically important species, three wild species are known: G. tomentosum (AD3) from Hawaii, G. mustelinum (AD4) from NE Brazil, and G. darwinii (AD5) from the Galapagos Islands. At the time of polyploid formation, rbcS copies from two divergent genomes (“A” and “D”) were combined, yet with only the maternal A LSU copy in the cytoplasm. Gossypium arboreum and G. rai-

modii, the extant diploids most similar to the ancient A- and D-parents (Wendel et al. 2009, 2010), respectively, are also available to facilitate identification of the duplic-
ated A- and D-specific gene copies and transcripts (homo-
eologs) in the natural allopolyploids.

Here, we characterize the rubisco system in diploid and allopolyploid cotton. By analyzing the genomic composition and expression of rbcS in hybrid and allopolyploid Gos-
sypium, we demonstrate post-polyploidy gene conversion of nuclear homoeologs and biased rbcS expression in the direction predicted by the organellar composition of allo-

polyploid cotton. We show that biased accumulation of rbcS transcripts from the maternal nuclear genome occurs in all wild and cultivated allopolyploids as well as in a syn-

thetic intergenomic diploid F1 hybrid.

Materials and Methods

Plant Materials

Fully expanded leaves were collected from six diploid cottons (G. herbaceum [A1], G. arboreum [A2], G. armouri-
num [D2-1], G. davidsonii [D3-d], G. raimondii [D5], and G. gossypioides [D6]), a laboratory synthesized A genome × D genome F1 hybrid (A2 × D5), and from five allopolyploid cottons (G. hirsutum cv. Acala Maxxa [AD1], G. barbadense cv. Pima S6 [AD2], G. tomentosum WT936 [AD3], G. mustelinum 15C [AD4], and G. darwinii PW45 [AD5]). For purposes of phylogenetic reconstruction, we included the outgroup species Gossypioides kirki (Seelanan et al. 1997). All lines were grown in the Pohl Conservatory at Iowa State University. After washing with diethylpyrocarbonate-treated water, each sample was divided into three parts, one for DNA extraction, one for RNA extraction, and the third for isolation of chloroplasts.

DNA and RNA Extraction and cDNA Synthesis

DNA was extracted using Qiagen DNeasy Plant Maxi Kit, following the manufacturer’s recommended protocol (Cat. No. 68163). RNA was extracted from leaves of some of these accessions (excluding A1, D2-1, D3-d, and D6), using the Concert Plant RNA Reagent (Cat. No. 12322-012), followed by an additional equal volume of Phenol:Chloro-

form (Ambion Cat. No. 9720) purification and RNA precip-
itation with isopropyl alcohol. DNase treatment was performed on 10 μg of each RNA sample as specified in the protocol provided for DNase I digestion (Cat. No. M0303S; New England Biolabs Inc.). RNA samples were reverse-transcribed into single-stranded cDNA by utilizing the SuperScript III First-Strand Synthesis System for reverse transcription-polymerase chain reaction (RT-PCR) pro-
vided by Invitrogen (Cat. No. 18080-051).

Primer Design, Cloning, Sequencing, and Genomic Mapping

rbcL Cloning in A2, AD1, D3, D5, and Gossypioides kirki

rbcL sequence data were obtained from GenBank for G. barbadense (NC_008641.1), G. hirsutum (NC_007944.1), and G. thurberi (NC_015204.1). From these data, conserved 5'-untranslated region (UTR) and 3'-UTR regions were re-

covered, and primers were designed to amplify full-length rbcL sequences (rbcL-sense: 5'-ATGAGTTGTAGGGAGG-

GA-3'; rbcL-anti: 5'-GATGGGCTAGAAAGATCTTG-3').

Each 20 μl PCR reaction contained 1× Taq DNA poly-

merase buffer (Invitrogen), 1.5 mM of MgCl2, 200 μM of deoxynucleoside triphosphates, 0.5 μM of each primer, 1 unit of Taq DNA polymerase (Invitrogen), and 100 ng of total DNA (containing the chloroplast genome).

Amplification used 94 °C for 5 min, 35 cycles of 94 °C for

30 s, 50 °C for 30 s and 72 °C for 1 min 30 s, and a final

extension at 72 °C for 10 min.

Amplified products were visualized via agarose (1%) gel electrophoresis. Bands of the expected size were excised from the gel and purified using the Wizard SV Gel and PCR Clean-Up System (Cat. No. A9285; Promega). Gel-purified samples were cloned using the TOPO TA cloning kit for Sequencing (Cat. No. K4575-01; Invitrogen). Since the expected size of rbcL (ca. 1,500 bp) is longer than one Sanger read, sequencing reactions were performed from both ends of the ligated rbcL fragments in the plasmid, using M13 reverse and T7 primers.

rbcS Cloning and Genomic Mapping

At present, rbcS has been sequenced and characterized from a number of species including several Gossypium spe-

cies (Meagher et al. 1989; Sagliocco et al. 1991). Sequence information from the coding sequences (CDS) region of the rbcS subunit in G. hirsutum (X5409) was downloaded from
NCBI. It was utilized as query sequences, with BlastN, to search for homologous contigs from our Cotton-32 expressed sequence tag (EST) contig database (Udall et al. 2006) using cutoff e value \(< 1 \times 10^{-10}\). Based on the ClustalW alignment of those contigs with corresponding ESTs (supplementary fig. S1, Supplementary Material online), degenerate primers located at start and stop codons were designed (sense 1–23: 5′-ATGGCCTCCTMATGATH-\(\text{WHATC-3}′\) and antisense 528–549: 5′-TTADDANCCTK-BAGGSTGKAG-3′).

We utilized the same PCR reaction system as specified above to amplify the \(rbcS\) copies from all diploid species (A1, A2, D2-1, D3-d, D5, D6, and \(Gossypioideae\ \text{kirkii}\)). In addition, a touchdown gradient annealing program was adopted to ensure specific amplification. An initial denaturation at 94 °C for 5 min was followed by 10 cycles of touchdown annealing: 94 °C for 30 s, 60 °C for 30 s for first cycle (minus 1 °C per cycle until the tenth cycle at 51 °C for 30 s) and 72 °C for 1 min 30 s. Another 25 cycles at 94 °C for 30 s, 51 °C for 30 s, and 72 °C for 1 min 30 s were followed by a final extension at 72 °C for 10 min. Sequence data were obtained by Sanger sequencing, performed at the Iowa State University Sequencing Facility.

As the cloned \(rbcS\) copies were not full-length, it was necessary to synthesize another set of primers targeting the 5′- and 3′-UTRs. These primers were obtained from ongoing sequence capture experiments in our laboratory (Salmon A, unpublished data). Array-based and solution-based sequence capture technologies rely on small oligo “baits,” to “capture” and isolate homologous sequence from libraries of sheared genomic DNA (http://www.nimblegen.com/products/seqcap; http://www.mycroarray.com/products/myselect.html) and provide an elegant solution for producing full-length targeted gene sequence data. Sequence capture bait probes were designed using the cloned partial \(rbcS\) genomic sequences, targeting \(rbcS\) in wild (accession TX2094) and domesticated (cv. Acala Maxxa) \(G. \text{hirsutum}\). Following capture and subsequent sequencing, primers were designed to amplify full-length \(rbcS\) from all other species used in this study except A1, D2-1, D3-d, and D6. PCR conditions were as described, with the exception that the number of cycles was decreased to 22 cycles, to help to limit PCR recombination (Judo et al. 1998; Cronn et al. 2002).

A high-quality draft assembly of the \(G. \text{raimondii}\) (DS) genome sequence (Paterson AH, unpublished data), to which all available genetically mapped DNA markers had been anchored based on a consensus genetic map described elsewhere (Rong et al. 2005), was used to determine the approximate chromosomal locations of full-length \(rbcS\) genes. The gene sequences were used in Blast searches against the D-genome sequence and were assigned locations in the D-genome chromosomes based on having a query length \(\geq 95\%\), an identity \(\geq 97.81\%\), and an e value \(\leq 1 \times 10^{-61}\).

**rbcS Transcription Cloning**

\(rbcS\) transcripts were cloned from their RT-PCR products using single-stranded cDNA as the template (10 ng per reaction), and the previous described primers designed for the full-length genomic \(rbcS\). Samples were gel-purified and cloned prior to obtaining sequence data via Sanger sequencing at the Iowa State University DNA sequencing facility.

**Sequence Alignment, Calculation of Substitution Rate, and Phylogenetic Reconstruction**

Alignments of all cloned \(rbcL\) CDS and of all full-length or partial \(rbcS\) copies were completed by using ClustalW program. These data were used as input matrices for calculation of substitution rate and phylogenetic analysis using maximum parsimony. Synonymous and nonsynonymous substitution rates \((K_s\ \text{and} \ K_a)\) were calculated using DnaSP 5.00.07 (Librado and Rozas 2009).

**Detection of Homoeologous Single Nucleotide Polymorphisms and Nonreciprocal Homoeologous Recombination Events**

Species-specific and genome-diagnostic single nucleotide polymorphisms (SNPs) were inferred from the alignment of full-length genomic \(rbcS\) orthologs in parental diploids and homoeologs in hybrid and allopolyploids. As shown by Salmon et al. (2010), genome-diagnostic SNPs (here termed homoeo-SNPs) are useful for detecting putative homoeologous recombination events in hybrid and allopolyploid cottons (Salmon et al. 2010). A custom perl script was written to identify regions of recombination. Only recombinants occurring in at least 25% of the total cloned sequences were accepted as true “gene conversion” events, with confirmation from independent amplification reactions to eliminate the possibility of PCR-recombination artifacts. The conversion regions or points that occurred in only one direction (“from A to D homoeolog” or “from D to A homoeolog”) would be considered nonreciprocal recombinations.

**SSU Transcript Levels Based on RNAseq**

Three biological replicates were prepared for A2, D5, the F1 hybrid, and AD1. Each replicate contained three 3-cm length, seventh leaves from newly germinated seedlings of each species except F1 hybrid. As F1 hybrid is sterile, we used the young leaf tissue (also 3-cm length) collected from vegetatively reproduced F1 hybrid. RNAs of each sample were extracted using the RNeasy Plant Mini Kit (Cat. No. 74904; Qiagen), and their quality and concentration were determined using the Agilent RNA 6000 Nano kit (Cat. No. 5067-1511; Agilent) on a 2100 Bioanalyzer (Agilent). High-quality RNAs were reverse transcribed into cDNA via the random primer amplification method. Indexed DNA sequencing libraries were constructed following the suggested protocol (Nagalakshmi et al. 2010). The libraries were sequenced on the Illumina Genome Analyzer II-x sequencers with 80 nt of single read at the Genomics Core Facility at the University of Oregon. Raw reads were classified into the correct species group according to their indexed flanking nucleotides. After trimming off the
adapters, fastaq-formatted reads from each library (corresponding to certain replicate of certain species) were readied for mapping.

We manually made three reference genome files containing full-length rbcS cDNA sequences obtained from sequenced clones of A2, D5, the F1 hybrid, and the allotetraploid AD1, respectively. The raw reads of each replicate from each species were mapped onto the corresponding reference file using the Burrows-Wheeler Alignment Tool (Li and Durbin 2009). We adopted the algorithm designed for mapping short query reads up to ~200 bp by using “bwa aln” command with zero mismatch to allow for BWA alignment. All other commands were input following the BWA tools manual with default parameters.

To quantify the expression of rbcS orthologs and homoeologs, we utilized the sequence alignment/map format tools (Li et al. 2009) to manipulate the SAM files generated by the BWA mapping tool. The SAM files were converted into bam files, which only included the reads with mapping quality values (MAPO) equal or larger than 30.

For the A2 and D5 libraries, the total number of mapped reads to the corresponding rbcS reference file for each library was used as the expression estimation of rbcS in the diploids. For the F1 hybrid and AD1 libraries, we used the pileup command implemented in SAMtools and summarized the read number at specific nucleotide sites, based on homoeo-SNPs detected in the F1 hybrid and AD1. This procedure allowed us to calculate the number of reads covering A-genome homoeo-SNPs and D-genome homoeo-SNPs. These values represented the A- and D-rbcS homoeolog expression levels, respectively, in hybrid and allotetraploid species.

**Statistical Analysis**

In both the F1 hybrid and allotetraploid cotton, the expression level of A- (or D-) rbcS homoeologs can be modeled as the number of A- (or D-) genome reads in a sequence of “n” independent yes/no A- (or D-) rbcS experiments, each of which yields probability p for A-rbcS and q for D-rbcS. Here, the number “n” represents the total read number of all rbcS homoeologs in each sample. The probabilities p and q represent the expression proportions of A-rbcS and D-rbcS at the time when allopolyploidization occurred. These probabilities of expressed A-rbcS or D-rbcS are expected to follow the binomial distribution. By utilizing the A- or D-homoeolog read numbers as the observed numbers of successful events, we can calculate significance levels associated with deviations from expectations. In other words, we can analyze whether there is any biased or selective expression for A-rbcS or D-rbcS in the intergenomic F1 hybrid and in allopolyploid cotton. We assumed that the expression ratio of rbcS between A- and D-diploid progenitors was additive (or maintained) when allopolyploids initially formed. Because we do not know the expression status of A- and D-rbcS in the actual progenitors, we use extant A- and D-diploids as proxies. Thus, the null hypothesis for the expression ratio of A- and D-rbcS homoeologs is determined based on the expression levels observed in extant A- and D-genome diploids. We calculated the percentage of expressed rbcS orthologs in the transcriptomes of A2 and D5, a t-test based on those percentages was performed to check whether p (A2 expression) is significantly different from q (D5 expression). If p is not significantly different than q, then we can use p = q = 0.5 as the null hypothesis to test for biased expression of A- or D-homoeologs in the hybrid and allopolyploid, which means we can assume equal expression levels.

**Results**

**Divergence between A- and D-Genome LSU Gene Sequences**

Overall, an average of 19 clones for each species was bidirectionally sequenced (supplementary table S1, Supplementary Material online), leading to the alignment shown in supplementary figure S2 (Supplementary Material online). Sequences were deposited in NCBI GenBank with accession numbers JQ034247–JQ034251. Synonymous substitution rates (not shown) varied from 0.42% (A2 vs. AD1) to 0.83% (A2 vs. D3 or D5 and AD1 vs. D3 or D5). As expected, nonsynonymous substitution rates were lower, ranging from 0.0% (A2 vs. AD1) to 2.08% (A2 vs. D5); the average nonsynonymous substitution rate of A- versus D-rbcL sequences was 1.88% (A2 vs. D3 = 1.67%; A2 vs. D5 = 2.08%) (table 1). Nonsynonymous substitutions were clustered at the 3’-end of the genes, as shown (table 1 and supplementary fig. S2, Supplementary Material online).

A maximum parsimony phylogenetic tree of rbcL was constructed illustrating the amino acid substitutions that have accompanied organismal divergence (fig. 1). Notably, the diverged amino acids were clustered at the C- and N-terminals of the LSU (positions 478, 479, and 480 at the C-terminal end and sites 28 and 86 at the N-terminal end), which are near the α/β barrel active sites. At the region 20 Å away from the active site in which SSUs can interact with the LSU (around positions 221–290), divergent amino acids were also clustered, which thereby could potentially influence LSU catalysis (Spreitzer et al. 2005; Genkov and Spreitzer 2009). Thus, in both the diploid F1 hybrid and the allopolyploid, if the A-genome and D-genome SSUs have diverged in a manner that influences this key protein interaction, we might expect evolutionary compensation in either or both the SSU and the LSU, such that efficient rubisco function is maintained.

<table>
<thead>
<tr>
<th>Species</th>
<th>A2</th>
<th>AD1</th>
<th>D3-d</th>
<th>D5</th>
<th>G. kirkii</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>D</td>
<td>E</td>
<td>Q</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>E</td>
<td>V</td>
<td>H</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>I</td>
<td>F</td>
<td>V</td>
<td>K</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>F</td>
<td>M</td>
<td>Y</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>M</td>
<td>H</td>
<td>P</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>I</td>
<td>Y</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>A</td>
<td>S</td>
<td>E</td>
<td>K</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>P</td>
<td>A</td>
<td>M</td>
<td>H</td>
</tr>
</tbody>
</table>

*Species listed are Gossypium arboreum (A2), Gossypium hirsutum cv. Acala Maxxa (AD1), Gossypium davidsonii accession (D3-d), Gossypium raimondii (D5), and Gossypioideae kirkii (G. kirkii). Positions 226, 255, and 282 reside at the interface between SSU and LSU, which are 20 Å away from the active site. Positions 478, 479, and 480 are in the α/β barrel active site.
Partial and Full-Length Genomic rbcS Genes

Partial rbcS sequences from diploid cottons (supplementary table S2, Supplementary Material online) were aligned, as shown in supplementary figure S3 (Supplementary Material online) (GenBank accession numbers JQ034293–JQ034304). In accordance with the conserved exon/intron structure of plant rbcS genes (Clegg et al. 1997), all Gossypium rbcS genes have two introns and three exons. As expected, most indels and substitutions occur in the introns (supplementary fig. S3, Supplementary Material online). In each species, there was either one or two rbcS genes that have a 9-nt gap in the first exon. These genes are hereafter referred to as short rbcS copies, with sequences not having the gap comprising the long copies.

Contigs (supplementary table S3, Supplementary Material online) assembled from the sequence capture reads in G. hirsutum (cv. Acala Maxxa and accession TX2094) were aligned and are shown in supplementary figure S3 (Supplementary Material online), where the 5′-UTR and 3′-UTR regions are also annotated. Full-length genomic rbcS genes from A2 and D5 were aligned and are shown in figure 2. All previously cloned partial rbcS copies were detected in the full length–cloned rbcS sequences (GenBank accession numbers JQ034252–JQ034256); that is, no novel rbcS genes were detected by the primers amplifying full-length rbcS copies. As we previously observed in the alignment of the partial rbcS orthologs, most genomic variation occurred within the two introns (nucleotide positions 186–286 and 421–512). All comparisons between the long and short rbcS copies exhibited more substitutions than within-group (long or short) rbcS genes (table 2), consistent with the phylogeny modeled in figure 3. Only a few substitutions were observed within groups, even among different species of cotton. For example, there were 76 polymorphic intron positions between A2 long and short paralogs. However, between A2 short paralogs, only 17 polymorphic positions were observed. Even between A2 and D5 short orthologs, we found only 28 polymorphisms.

All predicted parental rbcS orthologs were maintained in the synthetic diploid F1 hybrid without any genomic changes. Conversely, in all five A-D allotetraploids, a number of substitutions were observed in the A- and D-homoeologs, discussed below under “nonreciprocal homoeologous recombination (NRHR).”

rbcS Gene Family Structure in A- and D-Diploid Species

We mapped each rbcS gene in G. raimondii to its genome sequence using BlastN. A single long rbcS was detected (chromosome 11, 372649–373390). However, evidence below from cloned sequences suggests that each diploid actually contains two long copies. It is possible that a second identical long rbcS copy exists and was not detected in the current genome assembly, due to insufficient read coverage or ambiguous assembly around identical or nearly identical tandemly repeated long rbcS genes. Alternatively, one long rbcS gene was lost in G. raimondii subsequent to its divergence from the actual D-genome donor (which contained two long copies) to allopolyploid cotton. A final possibility is that a single D-genome copy exists in both the G. raimondii and the actual D-genome donor, with the independent duplication of this copy following polyploid formation. Of these three scenarios, we view the first as most parsimonious and hence used in the interpretation in figure 3 (duplicate long rbcS genes in red, indicated as being homogenized by ongoing gene conversion).

Mapping data for the short rbcS genes were consistent with the sequence data and the gene tree modeled in figure 3. Specifically, two copies are tandemly arranged on chromosome 1 (5890424–5889715 and 5909906–5910615). This indicates that two short-lineage rbcS copies exist in D5 genome and that these too have been subjected to sequence homogenization via a concerted evolutionary process (Elder and Turner 1995), as previously reported for rbcS in other species (Meagher et al. 1989).

From the foregoing, we infer that there are four basal rbcS genes (two short and two long) in the rbcS gene family in Gossypium (fig. 3). Given the absence of detection of the short copies in the outgroup, Gossypioides kirkii (supplementary table S2, Supplementary Material online), a likely scenario (but not the only one) is that the two short rbcS copies arose by an ancient gene duplication with the subsequent accumulation of the deletions that diagnose the “short” subfamily. Because of frequent gene homogenization, the duplicates apparently have evolved “in concert,” rendering difficulty in their independent detection from sequencing genic clones. Occasional escapes from concerted evolution, whereby sequence divergence surmounts sequence homogenization, permit the detection of the two short (or long) paralogs. This, in fact, is our interpretation of the situation for the two detected short rbcS genes in G. arboresum (A2), G. herbaceum (A1), and in G. davidsonii (D3-d).

Transcripts of rbcS Genes in Diploid G. arboresum (A2) and G. raimondii (DS)

Sequences from rbcS cDNAs are aligned and shown in figure 2, confirming the inferred intron positions. All detected rbcS orthologs found in A2 and D5 (fig. 2 and supplementary table S4, Supplementary Material online), all expected
Fig. 2. Alignment of genomic and cDNA rbcS genes in A2 and D5. Red and yellow lines represent the exons and introns, respectively. Within each exon, polymorphic sites are denoted by green or blue characters. Homoeo-SNPs at positions 540 and 623 are annotated in purple, where all A2 orthologs had a G but D5 orthologs had an A.
Table 2. Pairwise Numbers of Nucleotide Substitutions among rbcS Paralogs and Orthologs in Introns.

<table>
<thead>
<tr>
<th>Introns (supplementary table S4, Supplementary Material online) were compared within two intronic regions (186–286 and 421–512) of full-length rbcS genomic sequences from Gossypium arboreum (A2) and Gossypium raimondii (D5).</th>
<th>A2-Long</th>
<th>A2-Short1</th>
<th>A2-Short2</th>
<th>D5-Long</th>
<th>D5-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2-Long</td>
<td>ID</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A2-Short1</td>
<td>76</td>
<td>ID</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A2-Short2</td>
<td>76</td>
<td>17</td>
<td>ID</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>D5-Long</td>
<td>11</td>
<td>75</td>
<td>75</td>
<td>ID</td>
<td>—</td>
</tr>
<tr>
<td>D5-Short</td>
<td>81</td>
<td>28</td>
<td>28</td>
<td>80</td>
<td>ID</td>
</tr>
</tbody>
</table>

Nonreciprocal Homoeologous Recombination

As expected based on the interpretation of figure 3, we found four A-genome rbcS homoeologs (two long and two short) and four D-genome rbcS homoeologs (two long and two short) in the allotetraploids (GenBank accession numbers JQ034257–JQ034292). Using homoeo-SNPs that distinguish orthologous genes from the diploid parents, it was possible to determine the parental origin of each rbcS homoeolog in the allotetraploids (fig. 6), and as an extension of this, the amount and type of genetic changes that have taken place since polyploid formation.

Compared with G. arboreum (A2) and G. raimondii (D5), two additional long homoeologs were detected in the allotetraploids. As discussed above, our favored interpretation of this observation is that they escaped detection in the diploids due to a high frequency of gene conversion; the alternative of multiple duplication following polyploidy is considered highly improbable, given the largely additive nature of polyploid cotton with respect to its diploid progenitors (Senchina et al. 2003; Grover et al. 2004).

Interestingly, nucleotide substitutions in the long rbcS A homoeologs in the allotetraploids were diagnostic of the paternal D-genome, based on their homoeo-SNPs (fig. 6a). Conversely, all of the D-genome rbcS homoeologs exhibited a fair amount of genetic variation compared with the parental diploid rbcS copies around regions IV, V, VI, and VII listed in table 4. All of these regions contained A-genome homoeo-SNPs (fig. 6b and c). This pattern of substitutions is characteristic of a gene conversion process, or more formally, NRHR, as has recently been reported for cotton (Salmon et al. 2010). Since the gene conversion regions are diagnosed based on homoeo-SNPs identified from the diploid parents, it is not possible to pinpoint conversion endpoints or potential conversions that occurred in conserved regions. Thus, the break points of our detected converted region or points represent the minimal estimate with significant NRHR signal. As shown in table 4, except for the conversions around positions 457 and 731, all conversions were to A2 and D5, respectively, with significant NRHR signal. As shown in supplementary table S4, Supplementary Material online, the amount and type of genetic changes that have taken place since polyploid formation.

It has been demonstrated that the loop between the βA and βB strands in the SSU is the region where the SSU and LSU interact. This region, therefore, has an important impact on rubisco CO2/O2 specificity and other catalytic properties (Wasmann et al. 1989; Spreitzer et al. 2005; Genkov and Spreitzer 2009). Using the spinach βA and βB regions as our reference SSU, we located the loop between β-strands A and B (fig. 5) in Gossypium rbcS. One of the species-specific amino acid sites, amino acid position 116, was found to be located within the loop between β-strands A and B. Except for this specific amino acid difference, other parts of the loop were identical in all orthologs of A2 and D5.

Nonreciprocal Homoeologous Recombination

As expected based on the interpretation of figure 3, we found four A-genome rbcS homoeologs (two long and two short) and four D-genome rbcS homoeologs (two long and two short) in the allotetraploids (GenBank accession numbers JQ034257–JQ034292). Using homoeo-SNPs that distinguish orthologous genes from the diploid parents, it was possible to determine the parental origin of each rbcS homoeolog in the allotetraploids (fig. 6), and as an extension of this, the amount and type of genetic changes that have taken place since polyploid formation.

Compared with G. arboreum (A2) and G. raimondii (D5), two additional long homoeologs were detected in the allotetraploids. As discussed above, our favored interpretation of this observation is that they escaped detection in the diploids due to a high frequency of gene conversion; the alternative of multiple duplication following polyploidy is considered highly improbable, given the largely additive nature of polyploid cotton with respect to its diploid progenitors (Senchina et al. 2003; Grover et al. 2004).

Interestingly, nucleotide substitutions in the long rbcS A homoeologs in the allotetraploids were diagnostic of the paternal D-genome, based on their homoeo-SNPs (fig. 6a). Conversely, all of the D-genome rbcS homoeologs exhibited a fair amount of genetic variation compared with the parental diploid rbcS copies around regions IV, V, VI, and VII listed in table 4. All of these regions contained A-genome homoeo-SNPs (fig. 6b and c). This pattern of substitutions is characteristic of a gene conversion process, or more formally, NRHR, as has recently been reported for cotton (Salmon et al. 2010). Since the gene conversion regions are diagnosed based on homoeo-SNPs identified from the diploid parents, it is not possible to pinpoint conversion endpoints or potential conversions that occurred in conserved regions. Thus, the break points of our detected converted region or points represent the minimal estimate with significant NRHR signal. As shown in table 4, except for the conversions around positions 457 and 731, all conversions were to A2 and D5, respectively, with significant NRHR signal. As shown in supplementary table S4, Supplementary Material online, the amount and type of genetic changes that have taken place since polyploid formation.
all other recombination events were nonreciprocal, as reported previously for other genes (Salmon et al. 2010).

A notable feature of the gene conversion process is that most inferred events occurred between homeologs of the same group, either long or short (fig. 6a–c). Again using G. hirsutum as an example, for the two long copies in the A-genome of polyploid cotton (AD1-A-Long 1 and AD1-A-Long 2), six NRHR events out of a total of seven NRHR events (85.71%) occurred between A-long and D-long copies, without the involvement of a short homeolog (fig. 6a). The same phenomenon was observed in other gene conversion events in other homeologs of AD1 (supplementary table S5, Supplementary Material online).

Gene conversion was also discovered to have generated nonsynonymous amino acid substitutions (fig. 7). For instance, at the loop between βA/βB strands in all allopolyploids, the D homeolog-specific amino acid asparagine was converted to serine, which was the amino acid at this location in the A-genome. This type of consistent conversion from one parental type to the other was rare, as no additional instances of this type of conversion were detected. The amino acid conversion is the result of genome-specific SNP conversion in all D-homeologs from “A” (resulting in an asparagine residue) to “G” (resulting in a serine residue). Notably, this same conversion of D homeologs to A homeologs type at the βA/βB loop region was observed in all other allotetraploids (supplementary table S6, Supplementary Material online), suggesting that it occurred early in allopolyploid evolution.

To assess whether NRHR events were restricted to the early stages of allopolyploid genome stabilization or, instead, occurred more evenly during and after allopolyploid speciation, we evaluated the NRHR events represented in table 4 and supplementary table S6 (Supplementary Material online) in a phylogenetic context (fig. 8), using the well-established phylogeny of allopolyploid cottons (Wendel and Cronn 2003; Wendel et al. 2009, 2010). This showed that most of the NRHR events were shared by all or most species. Conversion events I, VI, and VII were detected in all five polyploid species, event II was shared by AD1, AD3, and AD5, and event III was shared by AD1, AD2, AD3, and AD5. A few NRHR events were not detected in the AD2 and AD5 lineages. Specifically, II was not observed in AD2, V was not observed in AD5, and IV was not observed in both AD2 and AD5.

**Expression of Duplicated SSU Genes in Hybrid and Allotetraploid Cottons**

To understand how duplicated SSU genes are utilized following genome merger and doubling, reads of three RNA-seq replicates from the two progenitor genomes A2 and D5 were first mapped to our Cotton-46 EST contig database. By mapping the raw reads to the full-length A2 and D5 rbcS reference files, we were able to estimate the expression level of rbcS orthologs in each species, from which percentages of the expressed rbcS orthologs in each species were calculated (table 5).

No significant difference in rbcS expression was observed between the two diploids A2 and D5 (t-test P value = 0.38). This suggests that rbcS genes were transcribed at comparable levels in the A- and D-genome progenitors that gave rise to modern allotetraploid cottons. Accordingly, we used equal expression probabilities (see Materials and Methods) as the null hypothesis in statistical tests for biased expression of rbcS orthologs in hybrid and allotetraploid cottons.

As previously mentioned, hybrid cotton was additive with respect to genomic composition of rbcS compared with the orthologs in its parental A2 and D5 species. Two species-specific SNPs were detected (fig. 2): one at

---

**Table 3. Pairwise Sequence Differences among rbcS cDNAs.**

<table>
<thead>
<tr>
<th>cDNAs</th>
<th>A2-Long</th>
<th>A2-Short1</th>
<th>A2-Short2</th>
<th>D5-Long</th>
<th>D5-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2-Long</td>
<td>ID</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A2-Short1</td>
<td>35 (25 + 10); 0.1059</td>
<td>ID</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A2-Short2</td>
<td>36 (26 + 10); 0.1011</td>
<td>6 (6 + 0); 0</td>
<td>ID</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>D5-Long</td>
<td>17 (12 + 5)</td>
<td>37 (27 + 10)</td>
<td>38 (28 + 10)</td>
<td>ID</td>
<td>—</td>
</tr>
<tr>
<td>D5-Short</td>
<td>39 (28 + 11)</td>
<td>10 (7 + 3)</td>
<td>12 (9 + 3)</td>
<td>37 (30 + 7); 0.0594</td>
<td>ID</td>
</tr>
</tbody>
</table>

注: 每个单元格中的总对数差异（同源性和非同源性差异）。表中Ks/Ka比值表明两种相同的物种之间。

---

**Fig. 4. Alignment of translated CDS regions of rbcS genes cloned in A2 and D5. Conserved amino acids are shown in faded gray. Nonconserved amino acids are shown in black. The group-specific amino acid loci (shared amino acid within long or short rbcS group) are marked with diamonds. Circles denote species-specific amino acids (shared within species, even among long and short groups). The stars "*" at the end of the aligned sequences denote stop codons.**
position 347 in the cDNA sequence (corresponding to position 540 in the genomic alignment), where A2 orthologs have a G but D5 orthologs have an A; and a second at position 430 locus in the cDNA sequence (corresponding to position 623 in the genomic alignment), where A2 orthologs have a G but D5 orthologs have an A. Importantly, these SNPs distinguish all paralogs, both long and short, from the A and D genomes, and thus, these two species-diagnostic SNPs can be used to estimate homoeolog expression in the F1 hybrid. In all three replicates, the observed total expression of A-homoeologs (abbreviated as A subtotal in table 6) was significantly higher than that of D-homoeologs (abbreviated as D subtotal in table 6) in the F1 hybrid. Thus, there was a significant genomic bias in rbcS expression in diploid hybrid cotton and in the direction of its maternal genome donor.

To evaluate the possibility of biased expression in allopolyploid cotton, we took advantage of the observation that position 430 in the cDNA sequence (623 in the genomic sequence) may be used for homoeolog expression diagnosis. Consistent with the findings for the F1 hybrid, the same phenomena of biased expression of the A-homoeologs was also found in all three replicates of the G. hirsutum RNAseq libraries (table 6).

Discussion

This study offers a description of the evolutionary history of rbcS genes in the context of allopolyploid speciation. The data demonstrate an ongoing pattern of gene duplication and interaction among rbcS genes in diploid and allopolyploid cotton, nonreciprocal recombination between homoeologous rbcS partners, and selective or biased expression of homoeologs from the maternal allopolyploid progenitor. Each of these aspects is discussed briefly below.

rbcS Gene Family Evolution

In angiosperms studied to date, the rbcS gene family contains two to eight genes. Concerted evolution and expansions and contractions in the number of gene copies have been reported (Meagher et al. 1989; Clegg et al. 1997), perhaps facilitated by tandem duplication and slipped-strand mispairing (Dean et al. 1989). These processes are evident in Gossypium, where ancient gene duplications are suggested to have given rise to the short rbcS genes from long rbcS ancestors, and where sequence homogenization among similar genes (short and long) appears to be common. Additional insights into the dynamics of these processes should emerge as additional whole-genome sequence information becomes available.

Because homogenization via gene conversion appears to be common for rbcS evolution (Meagher et al. 1989; Clegg et al. 1997), it may be difficult to assess copy number. For example, in G. raimondii, two identical D5 short rbcS genes are tandemly located in chromosome 1. It also is apparent from our data, however, that occasionally sequence substitutions escape sequence homogenization, perhaps due to their recency or for unknown mechanistic reasons. Regardless of the underlying process and evolutionary dynamics, this escape from homogenization (Meagher et al. 1989) permits the discernment of “novel” genes. In our data, this is exemplified by the two short rbcS copies detected in G. herbaeceum (A1), G. arboreum (A2), and G. davidsonii (D3-d).

One interesting dimension of our data is the apparent near-independence between the long and the short rbcS lineages in both A- and D-genome diploid cottons. Such relative independence is likely related to their different genomic locations, which we show here to be on chromosome 11 (long) and 1 (short), respectively. Within each cluster in each species, tandemly duplicated long or short copies are inferred to maintain sequence identity or near-identity via gene conversion, which must occur with sufficient frequency to homogenize variants that arise through normal mutational processes. Occasionally, as mentioned above, tandem duplicates may diverge from one another (fig. 3), perhaps when homogenization rates fail to match mutation rates, or perhaps if new paralogs become dispersed elsewhere in the genome. This latter process may underlie the origin of the long and short rbcS arrays.
(a)

<table>
<thead>
<tr>
<th></th>
<th>A2-Short 1</th>
<th>A2-Short 2</th>
<th>A2-Long</th>
<th>AD1-A-Long 1</th>
<th>AD1-A-Long 2</th>
<th>D5-Long</th>
<th>D5-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>40</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>60</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>80</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>100</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>120</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>140</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th></th>
<th>A2-Short 1</th>
<th>A2-Short 2</th>
<th>A2-Long</th>
<th>AD1-D-Long 1</th>
<th>AD1-D-Long 2</th>
<th>D5-Long</th>
<th>D5-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>40</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>60</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>80</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>100</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>120</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>140</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>160</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

(c)

<table>
<thead>
<tr>
<th></th>
<th>A2-Short 1</th>
<th>A2-Short 2</th>
<th>A2-Long</th>
<th>AD1-D-Short 1</th>
<th>AD1-D-Short 2</th>
<th>D5-Long</th>
<th>D5-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>40</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>60</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>80</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>100</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>120</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>140</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>160</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Downloaded From http://mbe.oxfordjournals.org/ at Iowa State University on October 17, 2012
reported here, an event tracing to prior to the origin of *Gossypium*. The newly dispersed copies thereby become more protected from gene conversion processes, an interpretation supported by the hierarchical similarity of rbcS genes in other species (Meagher et al. 1989; Clegg et al. 1997). Yet, despite the spatial separation and long divergence of the long and short rbcS arrays, occasional gene conversion between those arrays is implicated by the presence of two species-specific SNPs in the exons (at positions of 347 and 430 in the cDNAs; fig. 2). Although we cannot rule out the formal alternative of coincidental and reciprocal nucleotide substitutions among long- and short-gene families in each species, this seems like a less parsimonious interpretation, particularly in light of the proven propensity of rbcS to evolve under concerted evolutionary pressure.

Thus, the picture that emerges for rbcS is that of a gene family that is characterized by a repeated history of duplication, divergence, and homogenization, with a subtle balance between the latter two opposing forces. One issue that remains unclear is the nature of the long and short rbcS arrays, occasional gene conversion between those arrays is implicated by the presence of two species-specific SNPs in the exons (at positions of 347 and 430 in the cDNAs; fig. 2). Although we cannot rule out the formal alternative of coincidental and reciprocal nucleotide substitutions among long- and short-gene families in each species, this seems like a less parsimonious interpretation, particularly in light of the proven propensity of rbcS to evolve under concerted evolutionary pressure.

Cytonuclear Evolution in Allopolyploids

According to cytonuclear epistasis and maternal–offspring coadaptation theory (Wolf and Hager 2006; Wolf 2009), selection should favor optimal interactions between genes encoded in the cytoplasmic and nuclear compartments. With respect to rubisco in *Gossypium* F1 hybrid and allopolyploids, the suggestion is that differences that evolved during diploid divergence would be selectively optimized or altered following the merger of the two diverged diploid genomes at the time of polyploid formation. At this time, two suites of rbcS genes from divergent parents became combined in the same nucleus, yet only in the A-genome cytoplasm. One might surmise that the initial mismatch between large and small rubisco SSUs would generate selection pressure for optimizing function.

One evolutionary response to this mismatch may have been intergenic sequence homogenization. Here, we observed preferential NRHR between homoeologs of the same group, long or short. We propose that when the ancient neopolyploids first formed, rbcS homoeologs of A- and D-genome belonging to the same group (long or short group) could pair more easily than those between groups because of similar homoeologous chromosome locations. This evidence for homoeologous exchange is consistent with classic cytagenetic experiments, which demonstrate that homoeologous pairing is more frequent in synthetic allopolyploidy has been accompanied by new opportunities for intergenic interaction, as shown in figure 6, table 4, and supplementary table S6 (Supplementary Material online). This further emphasizes the dynamic nature of rbcS evolution in plant genomes.

---

**Table 4.** NRHR among rbcS Genes in *Gossypium hirsutum*.

<table>
<thead>
<tr>
<th>G. hirsutum (AD1)</th>
<th>Conversion from A to D&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Conversion from D to A&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Long Type</strong></td>
<td><strong>Number of Sequences</strong></td>
<td><strong>Percentage</strong></td>
</tr>
<tr>
<td><strong>region or point</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I&lt;sup&gt;a&lt;/sup&gt;. 232</td>
<td>41</td>
<td>100</td>
</tr>
<tr>
<td>II. 332–344,&lt;br&gt; 422–427,&lt;br&gt; 453–457</td>
<td>14</td>
<td>34.15&lt;br&gt; No conversion signal</td>
</tr>
<tr>
<td>III. 731</td>
<td>14</td>
<td>34.15</td>
</tr>
<tr>
<td><strong>Short Type</strong></td>
<td><strong>Number of Sequences</strong></td>
<td><strong>Percentage</strong></td>
</tr>
<tr>
<td><strong>region or point</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V. 595–604, 635–637, 731</td>
<td>53</td>
<td>100&lt;br&gt; VI. 540, 583</td>
</tr>
</tbody>
</table>

<sup>a</sup> Recombined regions are categorized into “conversion from A to D” and “conversion from D to A.”
<sup>b</sup> Roman numerals represent events that include one to several conversion regions that appear in all five allopolyploids (AD1–AD5), as shown in figure 8.
<sup>c</sup> Break points that occurred in exons are shown in underlined numbers. Break points that occurred in introns are in normal numbers.

---

**Fig. 6.** Gene conversion in *Gossypium hirsutum* (AD1) rbcS genes. Conserved nucleotides are shown in light gray. Homoeo-SNPs are denoted by blue diamonds, whereas uninformative SNPs are marked with white diamonds. An autoapomorphic nucleotide is highlighted in purple and marked with a purple diamond. Panels (a, b, and c) show the recombination pattern in A-long homoeologs, D-long homoeologs, and D-short homoeologs, respectively. Red and green bars represent the A- and D-homoeologous regions, respectively. Mosaic bars denote regions that include homoeologous recombination. Two conversions are not shown in the “a” panel due to space constraints; these are the A-long SNP converted to a D-long SNP at position 232 (from T to C), and at position 731 (from A to G).
Gossypium neopolyploids than in modern evolutionarily stabilized allopolyploids (Endrizzi et al. 1985). In addition, most of the conversion events detected in \(rbcS\) genes phylogenetically map to a basal position in the cotton allopolyploid radiation, in contrast to previous studies which rarely found basal gene conversion events in nuclear-encoded genes (Salmon et al. 2010). We suggest the possibility that this prevalence of phylogenetically basal \(rbcS\) NRHR events reflects a selective response favoring enhanced cytonuclear coordination for achieving efficient photosynthesis following allopolyploidy.

A second evolutionary response to allopolyploidy may have been gene conversion that became selectively stabilized because of its relationship to rubisco holoenzyme activity. Specifically, in the loop between \(\beta_a/\beta_b\) in the SSU that interacts with the LSU, there is only one species-specific amino acid difference between A2 and D5. In the ancient neopolyploid, this difference may have affected rubisco catalysis when both A- and D-SSU homoeologs were assembled into the holoenzyme complex (Spreitzer et al. 2005). This observation raises the possibility that the gene conversion observed in the D-genome SSU homoeologs in the \(\beta_a/\beta_b\) loop region, toward their A-genome counterpart forms (generating A-genome–like D-SSU around the loop region), had functional significance and hence became stabilized as a possible consequence of cytonuclear selection. A caveat to this interpretation is the possibility that the actual D-genome donor to the allopolyploid lacked the SSU difference (from the A-genome SSUs) exhibited by modern \(G.\) raimondii. Although this is a formal possibility, \(G.\) raimondii is closely related to the actual D-genome donor, as shown by all sequence data (Senchina et al. 2003; Wendel et al. 2010). In addition, the extensive and ongoing pattern of NRHR observed in the study implies that the mechanistic machinery was in place for the suggested cytonuclear accommodation.

A third evolutionary response to the merger of diverged genomes may have been selection for biased gene expression. Here, we have shown that at the transcription level, A-genome \(rbcS\) homoeologs are preferentially expressed, even in the diploid hybrid as well as in all natural allopolyploids. This is an intriguing observation, particularly in light of the evidence that the species-specific \(\beta_a/\beta_b\) regions in D-genome SSUs were homogenized into forms mimicking those from the A-genome, as discussed above. We note that other parts of the A- and D-SSUs remain distinct, suggesting that these might differentially affect the catalytic activity of the hybrid holoenzyme (A-genome LSUs assembled with D-genome SSUs) (Spreitzer et al. 2005). In addition, it may be that selection has operated at the promoter level, at sequence regions, and for functional partners not studied here.

Although our results are consistent with the hypothesis of cytonuclear accommodation, both NRHR (Salmon et al. 2010) and expression biases (Hovav et al. 2008; Flagel and Wendel 2010) have been observed for genes not...
obviously related to the cytonuclear dimension of allopolyploidy. Thus, our observations might best be considered as intriguing suggestions rather than proof of cytonuclear coevolution during allopolyploid evolution.

**Conclusion**

We have shown that the rubisco protein and its regulation provide an intriguing intergenomic model for the exploration of the process of cytonuclear coevolution following a prominent mode of speciation in plants. The use of a well-established phylogenetic framework and the existence of living models of the ancestral diploid donors to the new allopolyploid lineage make *Gossypium* a useful system for this analysis. We report evidence for several forms of potential evolutionary response to this biological recombination of diverged diploid genomes, including biased gene conversion and biased transcription of SSU genes from the nuclear genome of the maternal parent of the allopolyploids. Future research in other angiosperm polyploid systems and using additional protein models will provide an interesting comparison with the present work and shed light on this potentially significant aspect of allopolyploid speciation and evolution in plants. Additional insights into the process of cytonuclear coevolution will likely emerge from experimental studies of function, from analyses of mechanisms of SSU and LSU regulation at several levels (transcription, translation, import, and turnover), and from analysis of other cytoplasmic and nuclear partner genes.

### Table 5. Expression Levels of *rbcS* in Diploid Cottons (A2 and D5), as determined by mRNA sequencing.

<table>
<thead>
<tr>
<th>Homoeolog</th>
<th>A2 <em>rbcS</em> expression</th>
<th>D5 <em>rbcS</em> expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mapped <em>rbcS</em> reads A2-1</td>
<td>Mapped <em>rbcS</em> reads D5-1</td>
</tr>
<tr>
<td></td>
<td>Number of reads mapped to A2 <em>rbcS</em> genes</td>
<td>Number of reads mapped to D5 <em>rbcS</em> genes</td>
</tr>
<tr>
<td></td>
<td>Total mapped reads mapped to Cotton-46 EST Contig Database</td>
<td>Total mapped reads mapped to Cotton-46 EST Contig Database</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Homoeolog</th>
<th>A2 <em>rbcS</em> expression</th>
<th>D5 <em>rbcS</em> expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mapped <em>rbcS</em> reads A2-2</td>
<td>Mapped <em>rbcS</em> reads D5-2</td>
</tr>
<tr>
<td></td>
<td>Mapped <em>rbcS</em> reads A2-3</td>
<td>Mapped <em>rbcS</em> reads D5-3</td>
</tr>
<tr>
<td></td>
<td>Number of reads mapped to A2 <em>rbcS</em> genes</td>
<td>Number of reads mapped to D5 <em>rbcS</em> genes</td>
</tr>
<tr>
<td></td>
<td>Total mapped reads mapped to Cotton-46 EST Contig Database</td>
<td>Total mapped reads mapped to Cotton-46 EST Contig Database</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
</tbody>
</table>

**Supplementary Material**

Supplementary tables S1–S6 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

This work was supported by grants from the National Science Foundation Plant Genome Program (to J.F.W. and A.H.P.) and from Cotton Incorporated.

**References**


