Theor Appl Genet (2002) 104:482–489

R. Cronn · M. Cedroni · T. Haselkorn · C. Grover
J.F. Wendel

PCR-mediated recombination in amplification products derived from polyploid cotton

Received: 28 February 2001 / Accepted: 8 June 2001

Abstract *PCR recombination* describes a process of in vitro chimera formation from non-identical templates. The key requirement of this process is the inclusion of two partially homologous templates in one reaction, a condition met when amplifying any locus from polyploid organisms and members of multigene families from diploid organisms. Because polyploids possess two or more divergent genomes ("homoeologues") in a common nucleus, intergenic chimeras can form during the PCR amplification of any gene. Here we report a high frequency of PCR-induced recombination for four low-copy genes from allotetraploid cotton (*Gossypium hirsutum*). Amplification products from these genes (*Myb3, Myb5, G1262* and *CesA1*) range in length from 860 to 4,050 bp. Intergenomic recombinants were formed frequently, accounting for 23 of the 74 (31.1%) amplicons evaluated, with the frequency of recombination in individual reactions ranging from 0% to approximately 89%. Inspection of the putative recombination zones failed to reveal sequence-specific attributes that promote recombination. The high levels of observed in vitro recombination indicate that the tacit assumption of exclusive amplification of target templates may often be violated, particularly from polyploid genomes. This conclusion has profound implications for population and evolutionary genetic studies, where unrecognized artifactually recombinant molecules may bias results or alter interpretations.

Keywords Cotton · Gossypium · Polyploidy · WA Sequencing · Phylogeny

Communicated by J. Dvorak

R. Cronn · M. Cedroni · T. Haselkorn · C. Grover
J.F. Wendel (✉)
Department of Botany, Iowa State University, Ames, IA 50011-1024, USA
e-mail: jfw@iastate.edu
Tel.: +1-515-294-7172, Fax: +1-515-294-1337

Introduction

PCR recombination describes the process of in vitro chimera formation between amplification products from two or more templates. This process requires that a minimum of two highly similar, non-identical target sequences be present in a single PCR amplification reaction, a condition met by heterozygosity at a single locus or by genic redundancy. The presence of chimeric molecules primarily has been attributed to the periodic formation of incompletely extended PCR products (Saiki et al. 1988; Myerhans et al. 1990; Shammas et al. 2001). In the presence of two similar templates, prematurely terminated products can anneal to non-identical templates and be extended to completion in the next cycle. The resulting recombinant molecules are propagated during subsequent rounds of PCR, where they are subject to additional rounds of recombination. In vitro recombination can also occur via polymerase template switching in the absence of temperature cycling (Odelberg et al. 1995; Shammas et al. 2001). While PCR recombination has been most frequently evaluated using experimental templates (Judo et al. 1998; Myerhans et al. 1990; Odelberg et al. 1995; Yang et al. 1996; Shammas et al. 2001), the process also has been demonstrated for natural templates, such as heterozygous *Adh* loci from pocket gophers (Bradley and Hillis 1997) and non-homologous polyadenylated transcripts amplified via RT-PCR (Zaphiropoulos 1998). Chimera formation appears to be minimally influenced by polymerase choice, as the Klenow fragment of *Escherichia coli pol*I, *Taq*, and proofreading polymerases such as *Vent* (Bradley and Hillis 1997; Judo et al. 1998) reportedly yield chimeric products. While large PCR products appear especially prone to PCR recombination, targets as small as 242 bp (a partial tat sequence from HIV-1; Yang et al. 1996) demonstrate PCR recombination frequencies of 5% or more.

In light of the propensity of PCR to generate chimeras when presented to two related templates, it seems likely that PCR recombination could be a frequent outcome of amplification from complex eukaryotic genomes, given

their typically high levels of genic redundancy. The problem might be especially acute in polyploids, where two or more genomes ("homoeologues") are united in a single nucleus, thereby generating instantaneous doubling of all genes (Soltis and Soltis 2000; Wendel 2000). PCR amplification of low-copy genes from allopolyploid genomes typically results in the amplification of both homoeologous copies (Cronn and Wendel 1998; Small et al. 1998, 1999; Small and Wendel 2000; Liu et al. 2001). This commonly observed lack of amplification specificity is due to the limited nucleotide divergence characteristic of homoeologous loci in many allotetraploid species. For example, pairs of low-copy genes duplicated via polyploidy ("homoeologues") in AD-genome allotetraploid cotton (*Gossypiam hirsutum* L.) differ by an average of about 2.2% at the nucleotide level (Cronn et al. 1999).

In the course of isolating and sequencing genes implicated in cotton fiber development, we recently recovered amplification products that were determined by sequence analysis to be intergenic recombinants. Given the potential significance of in vitro chimera formation, we explored its extent using four different target sequences, each representing low-copy genes (*Myb3*, *Myb5*, *G1262* and *CesA1*) duplicated by relatively recent allopolyploidy (1–2 million years ago; Wendel 1989). Because models of the diploid progenitors are extant and were included in the study, the non-recombinant or "ancestral" conditions could be confidently determined in each case. Our results show that PCR-mediated intergenic recombination is common, with chimeras representing 23 of 74 (31.1%) of the amplicons screened, and comprising between 0% to approximately 89% of the amplicons in individual reactions.

Materials and methods

To estimate the frequency of chimera formation between homoeologous genes, we used standard PCR conditions (Table 1) to amplify four different targets from allotetraploid (AD-genome) *G. hirsutum* (race Palmeri) and one locus from the related allotetraploid *Gossypium barbadense* (cultivar K101 Riberalta). These included a 916-bp partial sequence from a putative P-glycoprotein gene (discovered by sequencing an anonymous PstI mapping probe, G1262; Cronn and Wendel 1998), partial sequences (*Myb3*: 860-bp; *Myb5*: 1,130-bp) from two genes encoding putative Myb transcription factors (Loguercio et al. 1999), and a 4,050-bp genomic sequence of the homoeologous genes encoding cellulose synthase A1(*CesA1*, formerly *CelA1*; Pear et al. 1996). Stringent Southern-hybridization profiles for each of these genes reveal a single hybridizing locus in the diploid model progenitor species *Gossypium herbaceum* (A-genome) and *Gossypium raimondii* (D-genome); these bands are additive in the allotetraploid derivatives *G. hirsutum* and *G. barbadense* (Cronn and Wendel 1998; Cronn, Cedroni, Liu and Wendel, unpublished).

Amplification primers designed for these four genes are specific to their respective homoeologous targets, and they amplify identically sized PCR products from A- and D-genome diploid progenitor cottons under optimized amplification conditions (Table 1). In all cases, the PCR pools from allotetraploid cottons were heterogeneous, resulting from amplification of both members of the homoeologous gene pair. This was evidenced by additivity of nucleotide polymorphisms at sites diagnostic for the orthologous sequences from the A- and D-genome diploid species. To decompose this additivity, PCR products from allotetraploid cottons were gel-isolated, ligated into pGemT-Easy (Promega Corporation,

Table 1 PCR amplification and compositional properties of low-copy nuclear genes from *Gossypium*

Locus		*G1262*	*Myb3*	*Myb5*	*CesA1*
F Primer (5′–3′)		GGCGGCAGGCTAAGCACTTCY	GGGCCACTAAAGAATGGAGCA	GCCTCTCCGACTGTAATTAACC	GATGGAATCTGGGGTTCCTGTTTGC
R primer (5′–3′)		CGGAGGTCATACTTCCAGCCTY	GCTACAGTTCACTATGTCGG	ACGATTACGAATTCATGTGG	AGCTCTGTGACACGGTGGTCGYTA
Amplicon length (bp)		916	860	1130	4050
Nucleotide composition	%G	21.2	24.1	24.1	21.0
	%A	30.1	31.7	29.4	27.5
	%T	26.8	28.3	26.0	32.0
	%C	21.7	15.9	20.6	19.5
Initial cycles (n)		10	10	10	10
Denaturation		1′ @ 94°C	1′ @ 94°C	1′ @ 94°C	1′ @ 94°C
Annealing		1′ @ 60°C, –0.6°C/cycle	1′ @ 48°C	1′ @ 48°C	1′ @ 60°C, –0.6°C/cycle
Extension		1′ @ 72°C	1′ @ 72°C	1′ @ 72°C	1′ @ 72°C
Final cycles (n)		25	20	20	25
Denaturation		1′ @ 94°C	1′ @ 94°C	1′ @ 94°C	1′ @ 94°C
Annealing		1′ @ 54°C	1′ @ 48°C	1′ @ 48°C	1′ @ 54°C
Extension		1′ @ 72°C	1′ @ 72°C	1′ @ 72°C	5′ @ 72°C
GenBank #		*	AF377305-AF377315	AF377316-AF377318	*

Wis.) and transformed into Top10F′ *E. coli* cells (Stratagene, LaJolla, Calif.) using standard heat-shock transformation protocols (Sambrook et al. 1989). Transformants containing cloned inserts were identified by colony PCR using gene-specific primers and the amplification conditions shown in Table 1.

Clones containing each gene of interest were characterized for homoeologue identity along the entire length of the insert using either direct sequencing (*Myb3, Myb5, G1262)* or diagnostic restriction-site analysis (*CesA1*). For all genes, we initially screened clones by sequencing the 5′ end of each insert with the "forward" gene-specific primer (Myb3F, Myb5F, G1262F, CesAF), and then compared sequences to those obtained from A- and D-genome diploids. Following this preliminary genome identification, we directly sequenced 18 (*Myb3, Myb5*) or 20 (*G1262*) clones per locus, selecting an approximately equal number of A- and D-genome clones for analysis. To identify chimeric PCR products, we compared the sequence obtained from these clones (all derived from allotetraploid *G. hirsutum*) to non-recombined sequences from the diploid species *G. herbaceum* (A-genome) and *G. raimondii* (D-genome). Exemplar sequences have been deposited in GenBank (Table 1), and alignments are available at http://www.botany.iastate.edu/~ifw/HomePage/jfwdata_sets.html.

To characterize the homoeologous 4.05-kb *CesA1* genes, we PCR-amplified and sequenced correspondin g regions from the diploid cottons *G. herbaceum* and *G. raimondii*. Based on these sequences, we identified four restriction sites that diagnosed polymorphisms between orthologous genes from the A-genome and D-genome diploids (see Figs. 1 and 2A). These include *EcoRI*, which cuts the D-genome homoeologue at position 649 (g/aattc versus gaattcc in the A-genome); *HindIII*, which cuts the A-genome homoeologue at position (A/AGCTT vs. GAGCTT) 1643; *BamHI*, which cuts the A-genome homoeologue at position 2,400 (g/gatcc versus gaatcc); and *HinPI*, which cuts the A-genome homoeologue at position 3,612 (g/cgc versus gccc). *CesA1* amplicons were isolated from *G. hirsutum* and *G. barbadense* in separate PCR and cloning reactions, and nine clones were selected from each of these species for restriction analysis (18 clones total).

To focus our attention on PCR chimeras at homoeologue-specific restriction sites, we amplified inserts from each *CesA1* clone using three different primer combinations. The *EcoRI* polymorphism at position 649 was evaluated by amplifying *CesA1*-cloned inserts with the CesAF primer, and a different reverse primer (CelAR, 5′-GGGAA CTGAT CCAAC ACCCA G-3′; Cronn et al. 1999) that amplifies the 5′-most 1.1 kb of the insert. Digestion with *EcoRI* yields a D-genome diagnostic band 0.63-kb in length, and two A-genome diagnostic bands 0.47-kb and 0.16-kb in length. To evaluate the internal restriction-site differences revealed by *Hind*III and *Bam*HI, inserts were amplified using two internal primers (Ces1F3, 5′-GGATG CTTCC CAGCC CCTCT CGACT A-3′; Ces1R4C, 5′-CCTCC ATTCT CCATT AGTGT AGAT-3′) that amplify the 1.9-kb region between nucleotides 911–2,818 of the 4.05-kb *CesA1* amplicon. Digestion with *Hind*III yields a D-genome diagnostic band 1.7-kb in size, and two A-genome diagnostic bands with lengths of 1.0-kb and 0.70-kb; digestion of the same product with *Bam*HI yields a D-genome diagnostic band 1.4-kb in length and two A-genome diagnostic bands 1.0-kb and 0.4-kb in length. Finally, the *Hin*PI polymorphism at position 3,619 was evaluated by amplifying *CesA1* cloned inserts with primer CesF5 (5′-CCCAT CAATC TGTCT GATCG GTTGC ACCA-3′) and the original CesAR amplification primer (Table 1). This primer combination amplifies the 3′-most 0.9-kb of the insert, and digestion with *Eco*RI yields one band 0.9-kb in size in the D-genome sequences but two bands of lengths 0.43-kb and 0.47-kb from A-genome clones. In all cases, amplification products were ethanol-precipitated to de-salt and concentrate the DNA. Digestion reactions (10 µl) included between 200 and 500 ng of DNA, 1 unit of the diagnostic restriction enzyme (Gibco-BRL; Gaithersburg, Md.), and the appropriate buffer. *Eco*RI and *Bam*HI digestion products were run on 4% NuSieve 3:1 agarose gels (BioWhittaker Inc., Portland, Me.), while the *Hin*dIII and *Hin*PI digestion products were resolved using 3% NuSieve 3:1.

Results

Genome-specific nucleotides (Fig. 1) were revealed by sequence analysis of the homoeologous copies from tetraploid cotton (A- and D-genomes of *G. hirsutum* and *G. barbadense*) and the orthologous loci from the diploid

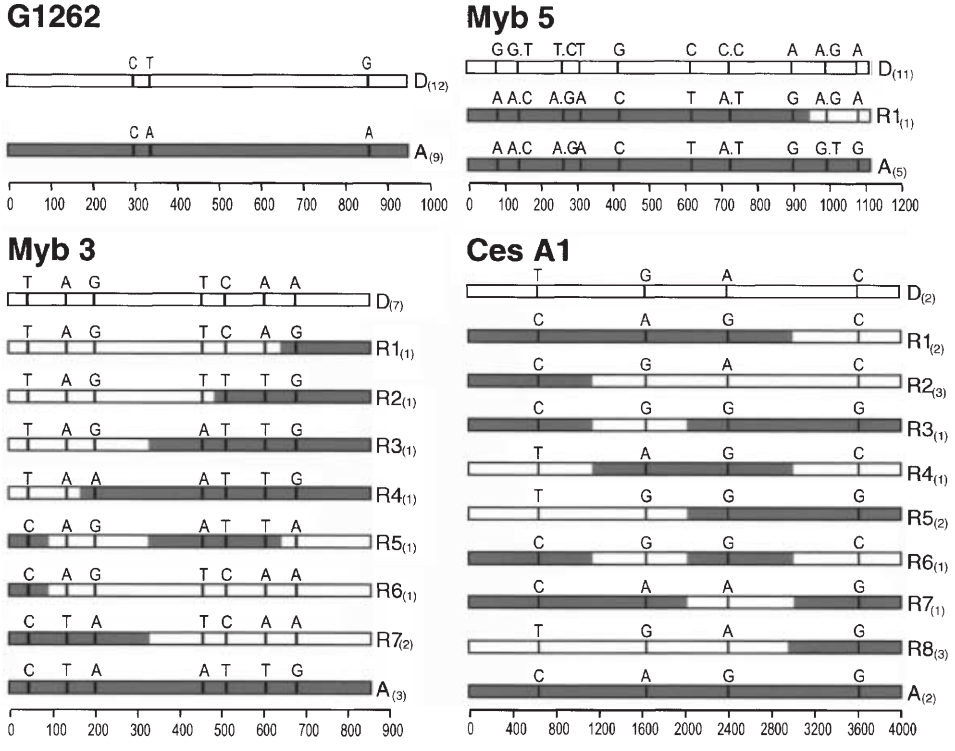

Fig. 1A–D PCR-induced recombination in homoeologous genes from *Gossypium* allotetraploids. Illustrated are non-recombinant ("A", "D") and chimeric cloned products ("R") recovered from amplification pools from allotetraploid cotton, for (**A**) *G1262*, (**B**) *Myb5*, (**C**) *Myb3*, and (**D**) *CesA1*. Recombination points are arbitrarily shown as located mid-way between flanking diagnostic nucleotide polymorphisms, which are identified above each schematic. The number of clones found for the non-recombinant and recombinant classes are indicated *in parentheses*

progenitors (A-genome=*G. herbaceum*; D-genome=*G. raimondii*). In the absence of PCR recombination, PCR-derived clones from tetraploid cotton should fall into two discrete classes exhibiting either exclusively A-genome or D-genome diagnostic nucleotides. However, if PCR recombination occurred between templates during amplification, new classes of amplification products would be formed, yielding chimeric patterns for the diagnostic nucleotides. This outcome was evidenced in our data: of a total of 74 amplification products screened in the allotetraploids, 23 were inferred to be recombinant, giving an overall recombination frequency of 31.1%. The propensity to recombine varied widely among the four gene systems. Of 20 *G1262* clones examined, none were recombinant; while for *Myb5* only a single chimeric product was revealed among the 18 sequences (5.6%). In contrast, higher levels of recombination were observed for *Myb3* (8 of 18=44.4%) and *CesA1* (14 of 18=77.8%).

Sequencing of *Myb5* clones revealed only a single recombinant product. This product (R1 in Fig. 1) contains approximately 700 bp from the 5′ end of the A-genome homoeologue and approximately 200 bp from the 3′ end of the D-genome homoeologue. Inspection of the recombination region between diagnostic nucleotides at positions 725 and 900 revealed no evidence of features that might promote recombination (Judo et al. 1998), such as regions of low complexity (simple-sequence repeats, runs of single-nucleotides) or unusually high G+C content.

As shown in Fig. 1, *Myb3* recombinants comprise seven classes of chimeras, most represented by a single clone. In nearly all cases, recombination was inferred to have involved a single template-switching event, such as 5′ A-genome→3′ D-genome (e.g., recombinants R6 and R7) or vice-versa (recombinants R1–R4). These recombinations were distributed along the length of the *Myb3* region sequenced; i.e., there were no recombination hotspots. In addition, there was no discernible evidence in the regions surrounding the recombination events of features that might promote recombination. One recombinant (R5) exhibited a more-complicated distribution of nucleotides indicative of multiple recombination events: the nucleotide at position 50 is diagnostic for the A-genome, as are nucleotides at positions 450, 520 and 610; these flank nucleotides at positions 130 and 200 which are diagnostic for the D-genome, as is the nucleotide at position 690. This distribution can be explained either by three intergenomic PCR recombination events or by a combination of *Taq* error at positions 50 and/or 690 and recombination. Given the improbability of the misincorporation of coincidentally diagnostic nucleotides at two different positions, our favored interpretation is that this clone reveals multiple in vitro recombination events. *Myb3* amplification products had the highest A+T% (60.0%) among the genes examined, but this was only marginally higher than that for *G1262* (A+T%=57.1%) or *Myb5* (A+T%=55.3%).

Restriction-site analysis of the *CesA1* amplicons showed a minimum of 14 recombinant clones among the 18 *CesA1* clones screened (77.7%). As indicated in Fig. 1, *CesA1* recombinant molecules comprise eight different classes of chimeras (R1 through R8), many of which were represented by a single clone. In five cases, recombination events involved a single switch between homoeologues, such as 5′ A-genome→3′ D-genome (e.g., recombinants R1 and R2; Fig. 2) or the converse

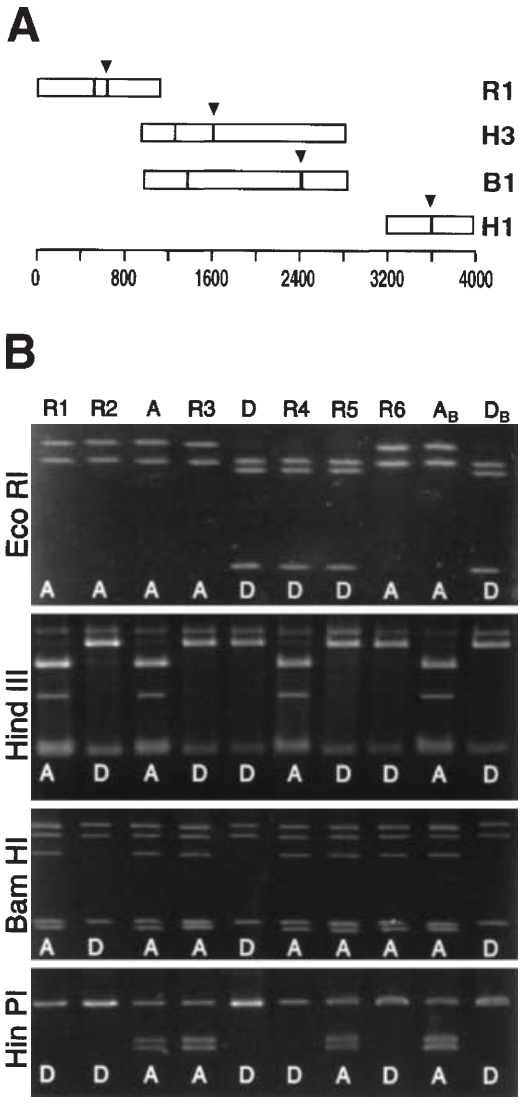

Fig. 2A, B Restriction digestion of cloned *CesA1* PCR amplification products, and inference of homoeologue recombination. **A** Inserts from individual clones were amplified using three different PCR primer combinations; CesAF+CelAR, Ces1F3+Ces1R4c, and CesF5+CesAR5. Amplification products were digested with restriction enzymes that reveal homoeologue identity; *EcoRI* (abbreviated "*R1*"; amplified with CesAF+CelAR), *Hin*dIII, *Bam*HI ("*H3*" and "*B1*"; Ces1F3+Ces1R4c) and *Hin*PI ("*H1*"; CesF5+CesAR5). The diagram shows the location of all restriction sites (vertical lines) with homoeologue-specific sites identified by *arrows*. **B** Restriction digestion patterns for eight PCR-derived clones, including six recombinant PCR products (*R1–R6*) and two putatively non-recombinant PCR products ("*A*" and "*D*"). The corresponding A- and D-genome *CesA1* inserts ("A_B", "D_B") were also amplified from *G. barbadense* BAC clones and digested with the same enzymes. Inferred genome identities resulting from each restriction digestion are indicated by *white letters*; conflict across enzymes reveals PCR recombinants

(R5, R8). The remaining *CesA1* recombinant products showed more-complicated nucleotide distributions, indicative of double (R3, R4, R7) and possibly triple (R6) recombination events. For example, restriction digestion patterns for six recombinant products (R1–R6) and two non-recombinant products (A, D) are illustrated in Fig. 2. Since recombination in *CesA1* was evaluated using restriction-digestion analysis, it is possible that nucleotide misincorporation (via *Taq* error) may have led to one or more erroneous inferences of recombination. While *Taq* error may lead to false inferences of recombination across these 4-kb amplicons, the four restriction enzymes used to diagnose genome identity sampled 22 bp, or less than 0.6% of the total sequence. For this reason, we infer that *Taq* error accounts for few, possibly none, of the estimated recombination events.

Of the 18 *CesA1* clones screened from tetraploid cotton, only four appeared non-recombinant. Because the restriction sites used to diagnose genomic origin were separated by 800–1200 bp, it is possible that putative non-recombinants possessed chimeric regions that were not sampled. To evaluate this possibility, we sequenced two of the four non-recombinant *CesA1* clones (one from each genome) in their entirety, and compared these sequences to orthologous sequences from the A- and D-genome diploid progenitors. Results from these sequence comparisons indicate that the two putative non-recombinants harbored regions of chimeric sequence that can only be explained by PCR recombination (data not shown). Hence, the combination of restriction digestion and direct sequencing show that the actual number of recombinant PCR products in the *CesA1* amplicon pool is minimally 16 of 18 (88.9%). It is possible that the two remaining "non-recombinant" clones also harbor chimeric segments; in any case, it is clear that PCR recombinants comprise the majority of products in the pool of *CesA1* amplicons.

Discussion

The polymerase chain reaction ranks among the most-common and important laboratory techniques, finding its way into diverse applications across the full spectrum of biological disciplines. In most investigations PCR-amplified products are sequenced, thereby providing data for genomic or genic characterization, descriptions of population genetic dynamics, or insights into phylogenetic history. An important assumption underlying these applications is that the isolated sequence faithfully replicates (within the limits of polymerase fidelity) the target of interest. PCR recombination violates this central assumption by generating chimeric sequences derived from two or more homologous templates. Here we have shown that in vitro chimera formation is not a rare or exceptional phenomenon, but instead is a high-frequency event leading to intergenic PCR-derived recombinants. Although our conclusions are based on a small sampling of genes from a single plant genus, they would appear to be generally applicable to any organism containing either heterozygous alleles at a single locus or genic redundancy. This would seem to encompass most eukaryotes (Sidow 1996; Spring 1997; Wolfe and Shields 1997; Postlethwait et al. 1998; Lynch and Conery 2000; Wendel 2000).

Minimizing the frequency of PCR recombination

Given the potential problems caused by PCR recombination, it becomes necessary to consider means to minimize its prevalence. The frequency of PCR recombination has been reported to be reduced by adopting long extension times (up to 6 min per kb; Judo et al. 1998), and by reducing the number of amplification cycles to the absolute minimum required for evaluation (Odelberg et al. 1995). Recently, Shammas et al. (2001) demonstrated that commonly used PCR additives such as betaine and dimethylsulfoxide (DMSO) were effective in reducing the frequency of recombination. Although each of these steps serve to minimize the overall frequency of in vitro PCR chimera formation, none of these modifications entirely eliminate PCR recombination.

At present, the best approach for isolating sequences completely free from chimeric segments may be to bypass PCR amplification from genomic DNA, and instead amplify target loci from genomic libraries. In the present study, we compared the *Myb3*, *Myb5* and *CesA1* PCR products derived from allopolyploid *G. hirsutum* to PCR-amplified sequences isolated from a *G. barbadense* BAC library (data not shown). This approach permits the selective amplification of individual homoeologues, and also facilitates direct sequencing of PCR products. Since the A- and D-genome homoeologues from *G. barbadense* are nearly identical to the corresponding sequences from *G. hirsutum* (≥99.9% at the nucleotide level), we could easily corroborate nucleotide positions diagnostic for the A- and D-genomes, and also re-confirm the chimeric nature of PCR amplicons derived from allotetraploid genomic DNA (Fig. 1). While direct comparison with BAC-derived sequences was possible for the small number of loci studied here, this approach is impractical for many species or in cases when it is necessary to sample multiple alleles or genes.

In the absence of library screening, other approaches may minimize PCR recombination. One possibility is to selectively amplify a specific target in the absence of alternative homologous templates. In allopolyploid plants, for example, three strategies for selectively isolating one (Doyle et al. 2000) or both (Cronn and Wendel 1998) homoeologues have been reported. All methods require prior sequence information, as well as evidence of orthology and homoeology (e.g., comparative linkage mapping and/or sequence analysis with diploid relatives). Two of the methods utilize restriction-enzyme-based methods that selectively remove one homoeologue (Doyle et al. 2000) from the amplification pool, or separate homoeologues into different amplification pools (Cronn and Wendel 1998). While these methods may reduce recom-

bination frequency, chimeras may be still evident if PCR products are cloned prior to sequencing. The first method (selective digestion of one homoeologue) permits amplification of single-stranded PCR products from the selectively digested template. Since prematurely terminated PCR products are believed to play a major role in PCR recombination (Judo et al. 1998), this approach may actually serve to *stimulate* PCR recombination. In the second method (amplification from size-fractionated pools), shearing or partial digestion of DNA allows the non-target template to migrate into unexpected size fractions, and may permit divergent, amplifiable templates to recombine during amplification. The final approach relies upon the design of genome-specific amplification primers that selectively amplify a single homoeologue (Small et al. 1999). Each of these approaches has merit, but their requirements for substantial preliminary information may limit their utility in some instances.

As an alternative to selective amplification, recombinant PCR sequences may in some cases be deconstructed into "parental" sequences if a suitably large number of clones is examined. This approach may be facilitated in allopolyploids by the existence of model diploid progenitors, but more-generally the success of sequence deconstruction will depend on recovery of a high proportion of clones that are unrecombined. In the present study, for example, unrecombined sequences comprised the majority of clones sequenced for three of the four genes included; only *CesA1* exhibited a sufficiently high recombination rate as to impede deconstruction. In all cases, the deconstruction process will be rendered more challenging by allelic heterozygosity and/or an increased number of homologous templates, as these features proportionally increase the number of nucleotide states at potentially diagnostic sites.

Irrespective of the manner used to isolate PCR amplified genes from polyploids, our data provide support for earlier predictions (Yang et al. 1996) of an increased-likelihood for long amplicons to undergo PCR recombination. Based on our limited sampling, however, we are reluctant to suggest that "small" amplicons (e.g., <1 kb) are inherently refractory to PCR recombination, or that "long" amplicons (≥4 kb) are unusually prone to recombination. The variance associated with the frequency of PCR recombination may be substantial, and the final frequency of recombinants may be influenced significantly by other factors (such as the timing of the introduction of a recombinant into a PCR reaction) than solely the size of the amplicon. Nevertheless, long templates do provide greater opportunities for premature termination than short templates, so an additional strategy to reduce the frequency of recombination may be to amplify large genes in small, overlapping segments.

Implications for prior studies

The high frequency of PCR recombination reported here, when combined with the ubiquity of the polymerase chain reaction, suggests that databases will be found to contain appreciable levels of illegitimate sequences. We explored this possibility using our own data from earlier studies of allotetraploid cotton, involving a highly repetitive 5S rDNA locus (Cronn et al. 1996), five pairs of homoeologous *Adh* genes (Small et al. 1998; Small and Wendel 2000), and 14 additional pairs of homoeologous loci isolated from *G. hirsutum* (Cronn et al. 1999). To identify putative PCR chimeras, we searched for stretches of diagnosable genome-specific nucleotides, gleaned from orthologous sequences from A- and D-genome diploid cottons, or from multiple clones derived from a single polyploid, as described above.

In the 5S rDNA data set, 41 sequences (each approximately 300 bp) were from allotetraploid *Gossypium* species (Cronn et al. 1996). Examination of these sequences provides evidence for at least one PCR chimera, from *Gossypium tomentosum* (clone "*G. tomentosum 2*"). This clone displays A-genome specific nucleotides from aligned positions 1–270, but the 3′ end of the clone (aligned positions 282–316) exhibits nucleotides otherwise restricted to D-genome cottons. While it is possible that this clone represents a rare intergenomic gene-conversion event (contradicting one of the authors' main conclusions), a more-likely explanation would seem to be PCR-mediated recombination. We note that the inferred recombination frequency (1/41=2.4%) is within the range expected from small amplification products (Myerhans et al. 1990; Odelberg et al. 1995; Yang et al. 1996; Judo et al. 1998).

Re-examination of the low-copy genes reported in Cronn et al. (1999) showed that five loci, *A1286*, *A1550*, *AdhC*, *CesA1* and *G1262*, exhibit substitution patterns that conceivably could reflect PCR recombination. At four of these loci (*A1286*, *AdhC*, *CesA1* and *G1262*) the sole evidence for the potential involvement of PCR recombination is a single nucleotide that is uniquely shared by both homoeologues. In each of these four cases, however, this nucleotide is flanked by sequences that contradict an interpretation of PCR recombination; hence, the most-likely explanation for the shared nucleotide is parallel (homoplasious) substitution. At locus *A1550*, two nucleotide positions from the A-genome homoeologue (aligned nucleotides 740 and 842) exhibit a similar pattern of unique shared mutation with the D-genome homoeologue. These homoplasious substitutions are separated by one additional informative nucleotide (aligned position 774) where the sequence from each homoeologue shares the expected nucleotide with their diploid relative (A-genomes exhibit a "G", while D-genomes exhibit a "T"). Since four recombination events are necessary to create this potentially chimeric molecule, it seems unparsimonious to exclusively invoke PCR recombination. In general, the low PCR recombination rate from this sample of homoeologous nuclear genes seems to contradict the findings of our present report. This discrepancy, however, most-likely reflects the selective fractionation methods used to isolate the 14 duplicate loci (Cronn and Wendel 1998). Briefly, this meth-

od combines restriction digestion and fractionation by gel electrophoresis to physically separate homoeologues into different amplification pools. Such a strategy minimizes the opportunity for PCR recombination, and facilitates direct sequencing of individual homoeologues without subsequent cloning steps. Our re-analysis here shows that this filter was effective in minimizing PCR recombination, a finding that illustrates the power of homoeologue-specific isolation methods.

In addition to the 14 loci of Cronn et al. (1999), there appears to have been minimal PCR recombination for the five alcohol dehydrogenase genes described from allotetraploid cottons (*AdhA–AdhE*; Small et al. 1998; Small and Wendel 2000), even though these genes were amplified and isolated *without* selective isolation methods. This finding contrasts with the present study, where comparably sized amplification products (e.g., *Mybs 3* and *5*) were isolated using identical methods. One factor that may have led to a reduced detection of PCR recombination for *Adh* genes is that *pools* of cloned PCR products were sequenced, typically using equimolar mixtures of ten independent clones (Small et al. 1998). This strategy is likely to yield the correct (non-recombinant) sequence because of the low-frequency of alternative (recombinant) nucleotides at any given position. The effect of this sequencing strategy can be illustrated with the recombinant clones isolated from *Myb3* (Fig. 1). If the polymorphic nucleotide at position 45 had been used to assign clones into homoeologue-specific pools, the D-genome pool would have contained 11 clones (seven non-recombinant, four recombinant) and the A-genome pool would have contained seven clones (three non-recombinant, four recombinant). Direct sequencing of the 5′ half (approximately 400 bp) of these two pools would have yielded the correct nucleotide sequence, as "non-recombined" nucleotides comprise a minimum of 10 of 11 diagnostic sites in the D-genome pool, and 5 of 7 diagnostic sites in the A-genome pool. Notice, however, that sequencing success from the 3′ half of the A-genome *Myb3* pool would have been unlikely since the majority of diagnostic nucleotides (4 of 7) were recombinant and represent the alternative homoeologue. Clearly, the success of these pooling strategies is dependent on the overall frequency of recombination, as well as the distribution of recombinant segments.

Implications of PCR recombination

Our results indicate that the central assumption of target specificity during PCR amplification may not always be robust. A variety of factors conceivably influence the degree of in vitro chimera formation during PCR amplification, including target sequence length, the number of partially homologous templates resident in the genome, cycling parameters such as extension time and number of cycles, structural features of the templates, and possibly nucleotide composition and the degree of divergence among potentially interacting loci. Because PCR-induced chimera formation may be a widespread phenomenon, sequences derived from cloned PCR products may be more confidently utilized when supporting evidence indicates an absence of recombination. This is particularly important in applications where recombinant molecules violate underlying assumptions of subsequent analyses, as in many population and evolutionary genetic studies and in explorations of phylo genetic history. Hence, for many applications it seems important that evidence of sequence integrity be garnered, using any combination of the methods outlined above.

Finally, we note that PCR recombination may be problematic not only for low-copy sequences but also for highly reiterated sequences such as rDNA, retrotransposons, or other classes of repetitive elements when these are sampled by PCR. Since these sequences typically share substantial sequence similarity and are present in high-copy number, they are likely to generate complex patterns of recombinants, not only in polyploid organisms, but in diploid organisms as well. Since the likelihood of chimera production increases with target length, PCR recombination will likely be most-significant during the amplification of larger templates, such as the commonly used 26S and 18S ribosomal genes. When levels of sequence homogenization by concerted evolutionary phenomena are high, the effects of PCR recombination may be minimal. If, however, genomes harbor divergent yet homologous templates and sequences are derived from clones, chimera formation is a possibility that should be recognized. One of the more vexing problems in the latter case arises from the difficulty of verifying that any given sequence is recombinant. Due of the complexity of large gene families, it may be impractical to recover a single member of a large gene family. This means that PCR-derived clones exhibiting evidence of intergenic recombination (such as the 5S clone *G. tomentosum 2*) are difficult to verify as real (a product of biological recombination) or artifactual (a product of PCR recombination).

Acknowledgements The authors thank Rick Noyes and Andrew Paterson for providing BAC clones for *Myb3*, *Myb5* and *CesA1*; Randy Small and Roberta Mason-Gamer for insightful comments on this manuscript and helpful discussion regarding the dynamics of PCR chimera formation; Curt Brubaker, Mike Hardig, Magnus Popp and Kara Shockey for sharing unpublished observations on the potential for PCR recombination in low-copy genes and rDNA; and Mike Hardig for finding evidence of PCR recombination in our previously-published 5S rDNA sequences. This work was supported by the National Science Foundation.

References

Bradley RD, Hillis DM (1997) Recombinant DNA sequences generated by PCR amplification. Mol Biol Evol 14:592–593

Cronn RC, Wendel JF (1998) Simple methods for isolating homoeologous loci from allopolyploid genomes. Genome 41: 756–762

Cronn RC, Zhao X, Paterson AH, Wendel JF (1996) Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. J Mol Evol 42:685–705

Cronn RC, Small RL, Wendel JF (1999) Duplicated genes evolve independently following polyploid formation in cotton. Proc Natl Acad Sci USA 96:14406–14411

Doyle JJ, Doyle JL, Brown AHD (2000) Origins, colonization, and lineage recombination in a widespread perennial soybean polyploid complex. Proc Natl Acad Sci USA 96:10741–10745

Judo MSB, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26:1819–1825

Liu Q, Brubaker CL, Green AG, Marshall DR, Sharp PJ, Singh SP (2001) Evolution of the *FAD2–1* fatty acid desaturase 5′ UTR intron and the molecular systematics of *Gossypium* (Malvaceae). Am J Bot 88:92–102

Loguercio LL, Zhang J-Q, Wilkins TA (1999) Differential regulation of six novel *MYB*-domain genes defines two distinct expression patterns in allotetraploid cotton (*Gossypium hirsutum* L.). Mol Gen Genet 261:660–671

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

Myerhans A, Vartanian J-P, Wain-Hobson S (1990) DNA recombination during PCR. Nucleic Acids Res 18:1687–1691

Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. Nucleic Acids Res 23:2049–2057

Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM (1996) Higher plants contain homologues of the bacterial *celA* genes encoding the catalytic subunit of cellulose synthase. Proc Natl Acad Sci USA 93:12637–12642

Postlethwait JH, et al. (1998) Vertebrate genome evolution and the zebrafish genetic map. Nature Genet 18:345–349

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491

Sambrook JE, Fritsch F, Maniatis T (1989) Molecular cloning. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Shammas FV, Heikkila R, Osland A (2001) Fluorescence-based method for measuring and determining the mechanisms of recombination in quantitative PCR. Clin Chim Acta 304:19–28

Sidow A (1996) Gen(ome) duplications in the evolution of early vertebrates. Curr Opin Genet Dev 6:715–722

Small RL, Wendel JF (2000) Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155:1913–1926

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85: 1301–1315

Small RL, Ryburn JA, Wendel JF (1999) Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Mol Biol Evol 16:491–501

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci USA 97:7051–7057

Spring J (1997) Vertebrate evolution by interspecific hybridization – are we polyploid? FEBS Lett 400:2–8

Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci USA 86:4132–4136

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Yang YL, Wang G, Dorman K, Kaplan AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high rate. AIDS Res Hum Retroviruses 12:303–306

Zaphiropoulos PG (1998) Non-homologous recombination mediated by *Thermus aquaticus* DNA polymerase I. Evidence supporting a copy choice mechanism. Nucleic Acids Res 26: 2843–2848