Differential Evolutionary Dynamics of Duplicated Paralogous Adh Loci in Allotetraploid Cotton (Gossypium)

Randall L. Small* and Jonathan F. Wendel†

*Department of Botany, The University of Tennessee, Knoxville; and †Department of Botany, Iowa State University

Levels and patterns of nucleotide diversity vary widely among lineages. Because allopolyploid species contain duplicated (homoeologous) genes, studies of nucleotide diversity at homoeologous loci may facilitate insight into the evolutionary dynamics of duplicated loci. In this study, we describe patterns of sequence diversity from an alcohol dehydrogenase homoeologous locus pair (AdhC) in allotetraploid cotton (Gossypium, Malvaceae). These data are compared with equivalent information from another homoeologous alcohol dehydrogenase gene pair (AdhA, Small, Ryburn, and Wendel 1999, Mol. Biol. Evol. 16:491-501) which has an overall slower evolutionary rate than AdhC. As expected from the predicted correlation between nucleotide diversity and evolutionary rate, nucleotide diversity was higher for AdhC than for AdhA. In addition, nucleotide diversity is higher in the D-subgenome of allotetraploid cotton for AdhC, confirming earlier observations for AdhA. These observations indicate that for these two pairs of Adh loci, the null hypothesis of equivalent evolutionary dynamics for duplicated genes in allotetraploid cotton is rejected.

Introduction

Levels of diversity and patterns of substitution in genes are the footprints of the evolutionary processes that have shaped extant gene pools. Analyses of these patterns can provide insights into how the evolutionary process differs between lineages and within loci within lineages (Clegg 1997; Clegg, Cummings, and Durbin 1997). In the absence of differential evolutionary pressures or genetic mechanisms, diversity is expected to be equivalent among loci, both in comparisons of orthologous genes between species and paralogous genes within species. Deviations from this expectation are the rule rather than the exception and may arise from myriad external and internal forces. Examples include variation in life history characteristics (e.g., self-pollination vs. outcrossing in plants [Liu, Zhang, and Charlesworth 1998; Savolainen et al. 2000]) and various forms of natural selection (e.g., primate ribonuclease genes [Zhang, Rosenberg, and Nei 1998]; gastropod toxin genes [Duda and Palumbi 1999]; plant self-incompatibility loci [Richman and Kohn 1999]; vertebrate MHC loci [Klein et al. 1998]; fungal mating type loci [May et al. 1999]). Additionally, it has been shown that nucleotide diversity is positively correlated with both evolutionary rates (Hudson, Kreitman, and Aguade 1995; May et al. 1999). Interestingly, it is necessary to obtain data from multiple loci within a given phylogenetic framework.

Gossypium L. (Malvaceae) has become a useful model system for studying molecular evolution (Wendel, Schnabel, and Seelanan 1995; Cronn et al. 1996; Cronn, Small, and Wendel 1999; Small, Ryburn, and Wendel 1999; Small and Wendel 2000a, 2000b) and especially for studying the molecular evolutionary consequences of polyploidy (Wendel, Schnabel, and Seelanan 1995; Wendel et al. 1999; Wendel 2000; Liu et al. 2001). The phylogenetic relationships of the ca. 50 diploid and 5 allotetraploid species of Gossypium are well characterized (Wendel and Albert 1992; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998; Wendel et al. 1999; Cronn et al. 2002). The five allotetraploid Gossypium species (designated AD-genome) diverged from a single recent allopolyploidization event (Wendel 1989; Small et al. 1998; Cronn, Small, and Wendel 1999), and the parental diploids are represented by the extant species Gossypium herbaceum L. (diploid A-genome) and Gossypium raimondii Ulbrich (diploid D-genome); thus the two component genomes of the allotetraploids are designated A- and D-subgenomes (or A' and D') to indicate their diploid origin. This well-understood organismal history facilitates the identification and comparison of orthologous and homoeologous loci (see e.g., Cronn and Wendel 1998; Small et al. 1998; Cronn, Small, and Wendel 1999; Small and Wendel 2000a).

A previous study (Small, Ryburn, and Wendel 1999) examined levels of nucleotide diversity for homoeologous AdhA loci in two allotetraploid species, Gossypium hirsutum L. and Gossypium barbadense L. Whereas that study revealed low diversity in both homoeologs, it also showed that the D-subgenome harbored greater nucleotide and allelic diversity than did the A-subgenome in both species. In concert with these data, a second study (Small et al. 1998) found that for a second alcohol dehydrogenase locus (AdhC), sequences from the D-subgenome homoeologs of all five allotetraploid species were evolving at a rate significantly greater than the rate in the A-subgenome homoeologs, again suggesting differential evolutionary pressures act-
ing on the two subgenomes. Finally, in evaluating the relative rates for the entire Adh gene family in Gossypium, we found that AdhC has higher evolutionary rates at both silent and nonsynonymous sites than AdhA (Small and Wendel 2000a). Thus, evolutionary rates for AdhA are low, relative to those for AdhC. Because evolutionary rates and levels of nucleotide diversity are positively correlated (Hudson, Kreitman, and Aguade 1987), these data predict that nucleotide diversity for it is also found for AdhC.

The purpose of this study then was to test these predictions for AdhC. Specifically we asked if: (1) nucleotide diversity is elevated for AdhC relative to AdhA, as predicted by the correlation between relative rates and nucleotide diversity; and (2) the pattern of higher diversity in the D-subgenome of the allotetraploids found for AdhA might similarly be elevated for AdhC. The purpose of this study then was to test these predictions for AdhC. Specifically, we asked if: (1) nucleotide diversity is elevated for AdhC relative to AdhA, as predicted by the correlation between relative rates and nucleotide diversity; and (2) the pattern of higher diversity in the D-subgenome of the allotetraploids found for AdhA might similarly be elevated for AdhC. The purpose of this study then was to test these predictions for AdhC. Specifically we asked if: (1) nucleotide diversity is elevated for AdhC relative to AdhA, as predicted by the correlation between relative rates and nucleotide diversity; and (2) the pattern of higher diversity in the D-subgenome of the allotetraploids found for AdhA might similarly be elevated for AdhC.

Materials and Methods

Plant Materials

Individual plants representing 22 accessions of G. hirsutum and six accessions of G. barbadense were included in this study. Each accession is representative of a wild-collected population or cultivar. These accessions are identical to those included in our previous study of AdhA (Small, Ryburn, and Wendel 1999), with the addition of a single G. barbadense accession (K101) which had been included in a previous study of AdhC (Small et al. 1998). Accessions were chosen to span the genetic and geographical variation encompassed by G. hirsutum and were originally selected based on the study of Brubaker and Wendel (1994), as described (Small, Ryburn, and Wendel 1999). Gossypium species as a general rule, and the cultivated allotetraploids in particular, are strongly selfing, intrapopulation variation is low, and heterozygosity is rare (Brubaker and Wendel 1994). Thus, our study was designed to maximize between-population, rather than within-population, sampling.

PCR Amplification and DNA Sequencing

To isolate AdhC sequences from specific duplicated genes in allotetraploid cotton, we designed two pairs of homoeolog-specific PCR amplification primers. Primer sequences were based on data from AdhC for all five allotetraploid species (Small et al. 1998) and were designed so that the final 3' nucleotide of each primer, as well as one other nucleotide within the primer, were specific for either the A- or D-subgenome homoeolog. The forward primers span the exon 2-intron 2 boundary, whereas the reverse primers span the intron 7-exon 8 boundary (fig. 1). To achieve homoeolog-specific amplification, a two-step procedure was used. The first step involved a 10-μl PCR amplification using 0.5 μl of template DNA, 1× Taq buffer (Promega), 200 μM each dNTP, 1.5 mM MgCl2, 0.2 μM each primer (either ADHCX2I2-D + ADHCX8I7-D to amplify the D-subgenome sequences or ADHCX2I2-A + ADHCX8I7-A to amplify the A-subgenome sequences). Cycling parameters used a touchdown approach (Don et al. 1991) that facilitates highly specific amplification. Initial annealing temperatures are set 5°C higher than the annealing temperature of the primers, so only amplification of the specific target is accomplished (in this case, the annealing temperature of the primers was 48–50°C, so the initial annealing temperature was set to 55°C). During the first 10 cycles, the annealing temperature is dropped by 0.5°C per cycle so that by the 11th cycle the programmed annealing temperature is down to the primer annealing temperature (50°C). An additional 15 cycles were then performed for a total of 25 cycles of 94°C for 1 min, 55–50°C for 1 min, and 72°C for 2 min, followed by a final 5 min 72°C extension step. The second step of the amplification process used 5 μl of the PCR product from the first step as a template for a second 25 μl PCR reaction with the same reaction components as above. Cycling conditions were 25 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 2 min, followed by a final 5 min 72°C extension step. These PCR products were subjected to agarose gel electrophoresis, excised from the gel, and eluted from the gel using GeneClean (BIO 101).

Purified PCR products were sequenced directly either using the ThermoSequenase 32P-radiolabeled terminator cycle sequencing kit (Amersham) followed by electrophoresis on a 5%–6% Long Ranger sequencing gel (FMC) or using the ABI Prism BigDye Terminator Cycle Sequencing Kit (Perkin-Elmer) followed by electrophoresis and detection on an ABI Prism 377 DNA sequencer at the Iowa State University DNA Sequencing and Synthesis Facility.

Because of the possibility that some mutations detected may be caused by nucleotide misincorporation by Taq polymerase during PCR, all singleton nucleotides were confirmed by reamplification and resequencing. In
all cases, the initial sequences inferred were corroborated by resequencing. In the case of heterozygous sequences, the amplification products were cloned into pGEM-T and sequenced as described previously (Small et al. 1998) to establish linkage relationships among polymorphic nucleotides.

Analyses

As in our previous study (Small, Ryburn, and Wendel 1999), we assumed that for each homoeolog both alleles were amplified. In a number of cases this assumption was validated by the presence of nucleotide polymorphism in the sequencing ladder and electropherograms, indicative of two products underlying the sequence (i.e., heterozygosity). The sequence uniformity detected for most accessions is assumed to be the result of homozygosity, the predominant condition in allotetraploid cottons (Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994; Small, Ryburn, and Wendel 1999). Removing identical sequences inferred from homozygous individuals from the analyses has little quantitative effect on the results and does not change the qualitative conclusions.

The sequences generated fell into four subsets that were analyzed separately and in combination when appropriate. These data sets include the A-subgenome of *G. barbadense* (6 accessions, 12 alleles), the D-subgenome of *G. barbadense* (6 accessions, 12 alleles), the A-subgenome of *G. hirsutum* (22 accessions, 44 alleles), and the D-subgenome of *G. hirsutum* (22 accessions, 44 alleles).

Relationships among the haplotypes of the *AdhC* sequences from *G. hirsutum* and *G. barbadense* were inferred using the software TCS (Clement, Posada, and Crandall 2000), which implements a statistical parsimony approach to estimating gene genealogies (Posada and Crandall 2001). Genealogies were inferred separately for the A-subgenome sequences and D-subgenome sequences, and the resulting genealogies are depicted in figures 2 and 3. The A-subgenome network (fig. 2) reveals considerable haplotype diversity in *G. hirsutum* inferred, and the resulting genealogies are depicted in figures 2 and 3. The A-subgenome network (fig. 2) reveals that *AdhC* sequences from *G. hirsutum* and *G. barbadense* are differentiated from each other by at least four mutation events. Rooting this network with the diploid A-genome species *G. arboreum* places the root between the *G. barbadense* and *G. hirsutum* sequences. As discussed subsequently, no allelic diversity was detected in the A-subgenome of *G. barbadense*—all sequences were identical. Seven different haplotypes were recovered from the A-subgenome of *G. hirsutum*. Four of these haplotypes were found only in single homozygous individuals; the remaining three haplotypes were represented multiple times in the sample.

In contrast to the A-subgenome network, the D-subgenome network (fig. 3) reveals considerable haplotype diversity in *G. hirsutum*, where a total of 24 distinct haplotypes were observed, and *G. barbadense*, where three haplotypes were observed. Rooting this network with the diploid D-genome species *G. raimondii* divides the network into two neighborhoods, one above and one below the root (fig. 3). The neighborhood above the root includes eight different haplotypes. One of these haplotypes was found in both *G. hirsutum* and *G. barbadense*. Three of the six *G. barbadense* accessions sampled were homozygous for this shared haplotype,
Fig. 2.—Gene genealogy of *G. hirsutum* and *G. barbadense* A-subgenome *AdhC* sequences. Haplotypes observed in our study are represented by open circles (*G. hirsutum*) or open squares (*G. barbadense*); the number of times each haplotype was observed is indicated by the number inside the circle or square. Lines connecting the haplotypes represent a single nucleotide substitution with filled circles representing inferred mutational steps not observed. The position of the root (indicated by the arrow) is inferred by outgroup comparison with the diploid A-genome species *G. arboreum*. The dashed line indicates ambiguity in the connection of the root to the haplotypes because of missing data in the outgroup.

Fig. 3.—Gene genealogy of *G. hirsutum* and *G. barbadense* D-subgenome *AdhC* sequences. Haplotypes observed in our study are represented by open circles (*G. hirsutum*), open squares (*G. barbadense*), or a filled square (haplotype found in both *G. hirsutum* and *G. barbadense*); the number of times each haplotype was observed is indicated by the number inside the circle or square. Lines connecting the haplotypes represent a single nucleotide substitution with filled circles representing inferred mutational steps not observed. The position of the root (indicated by the arrow) is inferred by outgroup comparison with the diploid D-genome species *G. raimondii*. The loop connecting the four haplotypes depicts recombination relationships among those haplotypes.

and it was also found in two of the *G. hirsutum* accessions, once in the homozygous condition and once as a part of a heterozygote. Five additional haplotypes found only in *G. hirsutum* accessions were also placed in this neighborhood (fig. 3). The remaining *G. barbadense* sequences were represented by two different haplotypes, both also found in this neighborhood. The neighborhood below the root includes 18 different haplotypes, including the majority of the *G. hirsutum* alleles; no *G. barbadense* alleles were found in this neighborhood. A number of low-frequency haplotypes (represented only once in the sample) are found along the branch leading to this neighborhood, but the majority of the sequences, both in terms of numbers of different alleles and raw numbers of sequences, are found at the tip of the network. This includes one haplotype represented by eight sequences, three haplotypes represented by four sequences, and two haplotypes represented by two sequences (fig. 3). In addition, the recombination detected in these sequences (see subsequently) is indicated by the loop of four haplotypes in this neighborhood.

Nucleotide Polymorphism

For each data set we calculated two overall measures of nucleotide diversity, $\pi$ and $\theta_w$ and a 95% confidence interval around $\theta_w$ (reported on a per base pair basis; table 1, fig. 4). $\pi$ was also calculated separately for intron, synonymous, silent, and nonsynonymous sites. In addition, we determined the number of distinct alleles recovered in each data set and the minimum number of recombination events. The same values were calculated for the *AdhA* data (Small, Ryburn, and Wendel 1999) and are presented in table 1 for comparison. Similar to the observations for *AdhA*, nucleotide diversity varies widely among genomes for *AdhC*. Values ranged from $\theta_w = 0/\pi = 0$ (*G. barbadense* A-sub-
Adh Diversity in Cotton  

<table>
<thead>
<tr>
<th>Adh</th>
<th>AdhA References</th>
<th>AdhC References</th>
</tr>
</thead>
</table>

### Table 1: Estimates of Nucleotide Diversity Per Base Pair and Tests of Neutral Evolution for AdhA and AdhC. A’ and D’. Refer to the A- and D-Subgenomes of the Allotetraploids, Respectively.

<table>
<thead>
<tr>
<th>Loci</th>
<th>N</th>
<th>θw</th>
<th>95% CI</th>
<th>θw</th>
<th>95% CI</th>
<th>θw</th>
<th>95% CI</th>
<th>θw</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdhA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. hirsutum A’</td>
<td>44</td>
<td>0.00105</td>
<td>0.00001 - 0.00207</td>
<td>0.00011</td>
<td>0.00000 - 0.00045</td>
<td>0.00023</td>
<td>0.00000 - 0.00123</td>
<td>0.00003</td>
<td>0.00000 - 0.00018</td>
</tr>
<tr>
<td>G. hirsutum A’</td>
<td>12</td>
<td>0.00458</td>
<td>0.00446 - 0.00469</td>
<td>0.00500</td>
<td>0.00487 - 0.00513</td>
<td>0.00532</td>
<td>0.00519 - 0.00545</td>
<td>0.00536</td>
<td>0.00523 - 0.00550</td>
</tr>
<tr>
<td>G. barbadense D’</td>
<td>44</td>
<td>0.00016</td>
<td>0.00014 - 0.00019</td>
<td>0.00016</td>
<td>0.00014 - 0.00018</td>
<td>0.00022</td>
<td>0.00019 - 0.00025</td>
<td>0.00025</td>
<td>0.00022 - 0.00028</td>
</tr>
<tr>
<td>G. barbadense D’</td>
<td>10</td>
<td>0.00016</td>
<td>0.00014 - 0.00019</td>
<td>0.00016</td>
<td>0.00014 - 0.00018</td>
<td>0.00022</td>
<td>0.00019 - 0.00025</td>
<td>0.00025</td>
<td>0.00022 - 0.00028</td>
</tr>
</tbody>
</table>

* Tajima (1989), no results are significant.
* Data from Small, Ryburn and Wendel (1999).

The minimum number of recombination events per data set was inferred using the method of Hudson and
Kaplan (1985). Recombination was detected only in the AdhC G. hirsutum D-subgenome data set. As noted above, these recombination events are restricted to a set of four closely related haplotypes (fig. 3). The tests of neutral evolution of Tajima (1989) and Fu and Li (1993) were performed for each data set, including subsets of each data set (introns and exons). None of these tests revealed significant departures from neutral expectations for any data set. We also performed the HKA test (Hudson, Kreitman, and Aguade 1987) and the MK test (McDonald and Kreitman 1991). For the HKA test, the data sets were partitioned as follows: the intraspecific comparison was between AdhC from the A-subgenome of G. hirsutum and AdhC from the D-subgenome of G. hirsutum; the interspecific comparison was provided by the AdhC sequences from the A- and D-subgenomes of G. barbadense. This test did not reveal any departure from neutrality ($\chi^2 = 0.84$, $P = 0.36$). The MK test was performed by tabulating numbers of fixed and polymorphic substitutions in exons for all four data sets and performing a $G$-test (with Williams correction) of independence. The MK test did reveal a significant departure from neutrality ($G = 6.924$, $P = 0.0085$, fig. 5) caused by an excess of polymorphic replacement substitutions.

Discussion

Levels and patterns of nucleotide diversity in Adh genes of allotetraploid Gossypium species differ markedly in several ways. First, the gene genealogies of AdhC in the A- and D-subgenomes differ in their patterns of haplotype distribution. Second, comparisons between homoeologous locus pairs (AdhA vs. AdhC) reveal that AdhC has consistently higher nucleotide and allelic diversity than AdhA. Finally, comparisons between subgenomes show that the D-subgenomes of G. hirsutum and G. barbadense harbor greater nucleotide and allelic diversity than the A-subgenomes. The sum of these observations indicates that the evolutionary forces that have shaped the nucleotide diversity at these loci have differed, both between loci and between subgenomes.

Gene Genealogies and Coalescence

AdhC gene genealogies were constructed separately for the A- and D-subgenome sequences of G. hirsutum and G. barbadense (figs. 2 and 3). These genealogies reveal different patterns of haplotype distribution in the two subgenomes. The A-subgenome sequences reflect a simple underlying pattern (fig. 2): G. hirsutum and G. barbadense sequences are separated on the genealogy by at least four substitutions. If this tree is rooted with an AdhC sequence from an A-genome diploid species (G. arboreum), the root falls on the branch separating the G. hirsutum and G. barbadense sequences; i.e., AdhC alleles coalesce within species. All G. barbadense sequences fell into a single haplotype. Gossypium hirsutum sequences were represented by seven different haplotypes; three of these were at intermediate frequency and four were found as homozygotes in single individuals. No recombination was detected among these sequences.

The pattern depicted in the genealogy of the D-subgenome sequences, on the other hand, is more complex (fig. 3). If this tree is rooted with an AdhC sequence from a D-genome diploid species (G. raimondii), the
FIG. 5.—Data and results for the McDonald-Kreitman test. The complete data set was trimmed to include only exon sequences, and only those sites that are polymorphic within that data set are shown. The nucleotide position numbers reflect the positions within the trimmed data set. All sequences are shown relative to the top sequence (pfx1A) with a "." indicating identity with the reference sequence. The status of each polymorphism (synonymous vs. replacement; fixed vs. polymorphic) is indicated at the bottom of each column. These values are compiled into a $2 \times 2$ contingency table and analyzed with a $G$-test (with Williams correction). The resulting $P = 0.0085$ indicates a significant departure from neutral expectations caused by an excess of polymorphic replacement substitutions. The amino acid changes are shown for each of the replacement substitutions.

<table>
<thead>
<tr>
<th>Nucleotide Position</th>
<th>Replacement</th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>9</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

**G. hirsutum**

A-subgenome

- pfx1A
- TX1_1A
- TX1_2D
- TX2_1D
- TX2_2D
- TX3_1D
- TX3_2D
- TX5_1D
- TX5_2D
- TX8_1D
- TX8_2D
- TX9_1D
- TX9_2D
- TXA_1A
- TXA_2A
- TXB_1A
- TXB_2A
- TXC_1A
- TXC_2A
- TXD_1A
- TXD_2A
- TXE_1A
- TXE_2A
- TXF_1A
- TXF_2A
- TXG_1A
- TXG_2A
- TXH_1A
- TXH_2A
- TXI_1A
- TXI_2A
- TXJ_1A
- TXJ_2A
- TXK_1A
- TXK_2A
- TXL_1A
- TXL_2A
- TXM_1A
- TXM_2A
- TXN_1A
- TXN_2A
- TXO_1A
- TXO_2A
- TXP_1A
- TXP_2A
- TXQ_1A
- TXQ_2A
- TXR_1A
- TXR_2A
- TXS_1A
- TXS_2A
- TXT_1A
- TXT_2A
- TXU_1A
- TXU_2A
- TXV_1A
- TXV_2A
- TXW_1A
- TXW_2A
- TXX_1A
- TXX_2A
- TXY_1A
- TXY_2A
- TXZ_1A
- TXZ_2A

---

**G. barbadense**

A-subgenome

- PIMA_1A
- B106_1A
- B250_1A
- B444_1A
- B559_1A

---

**G. hirsutum**

D-subgenome

- pfx1D
- TX1_1D
- TX1_2D
- TX2_1D
- TX2_2D
- TX4_1D
- TX4_2D
- TX5_1D
- TX5_2D
- TX9_1D
- TX9_2D
- TXA_1D
- TXA_2D
- TXB_1D
- TXB_2D
- TXC_1D
- TXC_2D
- TXD_1D
- TXD_2D
- TXE_1D
- TXE_2D
- TXF_1D
- TXF_2D
- TXG_1D
- TXG_2D
- TXH_1D
- TXH_2D
- TXI_1D
- TXI_2D
- TXJ_1D
- TXJ_2D
- TXK_1D
- TXK_2D
- TXL_1D
- TXL_2D
- TXM_1D
- TXM_2D
- TXN_1D
- TXN_2D
- TXO_1D
- TXO_2D
- TXP_1D
- TXP_2D
- TXQ_1D
- TXQ_2D
- TXR_1D
- TXR_2D
- TXS_1D
- TXS_2D
- TXT_1D
- TXT_2D
- TXU_1D
- TXU_2D
- TXV_1D
- TXV_2D
- TXW_1D
- TXW_2D
- TXX_1D
- TXX_2D
- TXY_1D
- TXY_2D
- TXZ_1D
- TXZ_2D

---

**G. barbadense**

D-subgenome

- K101_1D
- K101_2D
- B106_1D
- B250_1D
- B444_1D
- B559_1D

---

Fig. 5.—Data and results for the McDonald-Kreitman test. The complete data set was trimmed to include only exon sequences, and only those sites that are polymorphic within that data set are shown. The nucleotide position numbers reflect the positions within the trimmed data set. All sequences are shown relative to the top sequence (pfx1A) with a "." indicating identity with the reference sequence. The status of each polymorphism (synonymous vs. replacement; fixed vs. polymorphic) is indicated at the bottom of each column. These values are compiled into a $2 \times 2$ contingency table and analyzed with a $G$-test (with Williams correction). The resulting $P = 0.0085$ indicates a significant departure from neutral expectations caused by an excess of polymorphic replacement substitutions. The amino acid changes are shown for each of the replacement substitutions.
Comparative Evolutionary Dynamics of AdhA versus AdhC

Our previous study demonstrated that nucleotide substitution rates are higher for AdhC than for AdhA at both silent and nonsynonymous sites (Small and Wendel 2000a). Neutral theory predicts that evolutionary rates and nucleotide diversity will be positively correlated (Hudson, Kreitman, and Aguadé 1987), suggesting that nucleotide diversity should be higher for AdhC than for AdhA. That expectation is confirmed by our data, where on a per genome basis nucleotide diversity is higher for AdhC in every comparison. Furthermore, allelic diversity is consistently higher for AdhC than AdhA with 26 unique haplotypes recovered for AdhC (24 in the D-subgenome of G. hirsutum) as opposed to six haplotypes for AdhA (a maximum of four in any single genome, again in the D-subgenome of G. hirsutum). In addition, haplotype diversity is more widely dispersed on the gene genealogy in the D-subgenome than in the A-subgenome. Gossypium hirsutum D-subgenome sequences differ by up to 27 nucleotide substitutions (fig. 3), whereas the most divergent A-subgenome sequences differ by only four nucleotide substitutions (fig. 2).

In addition to the higher overall diversity of AdhC relative to AdhA, the patterns of silent and nonsynonymous diversity for the two loci are different. Specifically, no nonsynonymous diversity was detected for the AdhA genes, whereas nonsynonymous diversity accounts for a significant portion of the diversity at AdhC (table 1). Variation in silent versus nonsynonymous evolutionary rates has previously been described in plant genomes. For example, Gaut (1998) examined rate variation among nine nuclear genes for a rice-maize comparison and found that synonymous rates varied over a 2.4-fold range, and nonsynonymous rates varied over a 10-fold range. More relevant to the present study, five loci of the Gossypium Adh gene family have synonymous rates that vary over a 2.9-fold range and nonsynonymous rates that vary over a 3.3-fold range (Small and Wendel 2000a). The source of this variation in relative rates may be caused by either genomic processes that differentially affect the two loci, differential selective pressures on the two loci, or a combination of these factors. Recent evidence from extensive analyses of mammalian genomes suggests that evolutionary rates vary by genomic region, with genes from the same region showing similar synonymous rates (Matassi, Sharp, and Gautier 1999). Alternatively, different selection pressures on the two loci may be responsible for the observed differences in silent and nonsynonymous diversity. Support for this hypothesis is provided by the results of the MK test, which reveals that the patterns of synonymous and replacement substitutions at AdhC are not in accordance with neutral expectations. Specifically, there is an excess of polymorphic replacement substitutions, the majority of which (12/14) are polymorphic in the D-subgenomes of G. hirsutum or G. barbadense (or both). This observation contrasts with the lack of replacement substitutions in AdhA, suggesting differential selective pressures on AdhA and AdhC, ei-
ther purifying selection on AdhA, relaxed selection on AdhC, or a combination of the two. The lack of significant results for the Tajima or Fu and Li neutrality tests suggests that there is no disruption of the pattern of a neutral array of nucleotide substitutions, although the power of these tests is notoriously low (Simonsen, Churchill, and Aquadro 1995), and the effect of deviations from the tests assumptions (e.g., random mating—Gossypium species are strongly selfing) is unknown.

Thus, our observation of consistently greater nucleotide diversity and elevated nonsynonymous substitution rates in comparing AdhA and AdhC may be accounted for either by differential genomic context of the two genes, differential selective pressures on the two genes, or a combination of the two phenomena. Evidence is presented for differential selective pressures; the influence of genomic context, however, cannot be evaluated until similar data are available for genes in the same genomic context as AdhA and AdhC.

Comparative Evolutionary Dynamics of A-subgenome versus D-subgenome Sequences

Whereas our data clearly show that evolutionary dynamics differ between loci (AdhA vs. AdhC), the data also suggest differential patterns of evolution between sequences from the A- and D-subgenomes of allotetraploid Gossypium. In all pairwise comparisons of nucleotide diversity between subgenomes within a species (e.g., G. hirsutum A-subgenome vs. G. hirsutum D-subgenome for AdhC), nucleotide diversity is consistently higher in the D-subgenome (table 1, fig. 4). Likewise, the number of haplotypes recovered in each data set is consistently higher in the D-subgenome, with ratios ranging from 2:1 to 3.4:1 (table 1). Further, relative rate tests for AdhC have shown that the D-subgenome sequences are evolving at a significantly faster rate than A-subgenome sequences (Small et al. 1998; Small and Wendel 2000a). Finally, the excess of polymorphic replacement substitutions at AdhC, as evidenced both by the MK test and the high nonsynonymous diversity in the D-subgenome of G. hirsutum, suggests relaxed selection on the D-subgenomes of G. hirsutum and G. barbadense, at least for AdhC.

The genetic redundancy created by allopolyploidy or the large Adh gene family in Gossypium (at least seven loci in the diploid species [Small and Wendel 2000a]) may have allowed relaxed selection in the D-subgenome, whereas purifying selection maintained a narrower array of A-subgenome sequences. This hypothesis is consistent, with respect to both AdhA and AdhC, with the elevated evolutionary rate of the D-subgenome sequences over the A-subgenome sequences (Small et al. 1998) and the higher diversity in the D-subgenomes relative to the A-subgenomes. In addition, the presence of an intron-splice site mutation segregating in the G. hirsutum AdhC gene further suggests that, at least in G. hirsutum, this locus may be in the process of becoming a pseudogene.

Collectively, these observations might suggest an overall rate acceleration in the D-subgenome relative to the A-subgenome of allotetraploid Gossypium. Evolutionary rate analyses of 14 other nuclear loci, however, fail to reinforce this conclusion (Cronn, Small, and Wendel 1999). The apparent contradiction between the results of Cronn, Small, and Wendel (1999) and the present study indicates that either the Adh data are unusual, attributed perhaps to stochastic factors; the power of the relative rate tests alone are insufficient to detect subtle inequalities between subgenomes (as opposed to a combination of relative rate and nucleotide diversity analyses); or that something specific to these Adh loci promotes greater diversity in one of the two cotton subgenomes. We are currently unable to discriminate between these possibilities. Expression data and comparable studies of additional duplicated loci may provide critical clues in unraveling this conundrum.

Acknowledgments

We thank Julie Ryburn for technical assistance during the course of this investigation; Brandon Gaut and two anonymous reviewers provided helpful comments on an earlier draft of this paper; and funding was provided by the National Science Foundation to J.F.W.

LITERATURE CITED


BRANDON GAUT, reviewing editor

Accepted October 19, 2001