Copy Number Lability and Evolutionary Dynamics of the *Adh* Gene Family in Diploid and Tetraploid Cotton (Gossypium)

Randall L. Small* and Jonathan F. Wendel†

**Department of Botany, University of Tennessee, Knoxville, Tennessee 37996 and †Department of Botany, Iowa State University, Ames, Iowa 50011*

Manuscript received January 18, 2000
Accepted for publication April 21, 2000

ABSTRACT

Nuclear-encoded genes exist in families of various sizes. To further our understanding of the evolutionary dynamics of nuclear gene families we present a characterization of the structure and evolution of the alcohol dehydrogenase (*Adh*) gene family in diploid and tetraploid members of the cotton genus (Gossypium, Malvaceae). A PCR-based approach was employed to isolate and sequence multiple *Adh* gene family members, and Southern hybridization analyses were used to document variation in gene copy number. *Adh* gene copy number varies among Gossypium species, with diploids containing at least seven *Adh* loci in two primary gene lineages. Allotetraploid Gossypium species are inferred to contain at least 14 loci. Intron lengths vary markedly between loci, and one locus has lost two introns usually found in other plant *Adh* genes. Multiple examples of apparent gene duplication events were observed and at least one case of pseudogenization and one case of gene elimination were also found. Thus, *Adh* gene family structure is dynamic within this single plant genus. Evolutionary rate estimates differ between loci and in some cases between organismal lineages at the same locus. We suggest that dynamic fluctuation in copy number will prove common for nuclear genes, and we discuss the implications of this perspective for inferences of orthology and functional evolution.

NUCLEAR genes are generally part of gene families—multiple genes of common origin that encode products of the same or similar function. These gene families vary from small families with few loci (*e.g.*, many metabolic enzymes such as *Adh*, *Pgi*, *rbcS*; Clegg *et al.* 1997) to large families with hundreds of loci (*e.g.*, heat-shock proteins; Waters 1995). The evolutionary processes that control the structure and dynamics of such gene families are relatively poorly understood (reviewed by Clegg *et al.* 1997). The majority of molecular evolutionary studies have focused either on a single locus within a single species (*e.g.*, *Adh1* in maize; Gaut and Clegg 1993) or on an entire gene family across a broad phylogenetic spectrum (*e.g.*, *Adh* in eukaryotes; Yokoyama and Harry 1993). While both scales of study provide essential and complementary perspectives, the fine-scale dynamics of gene family evolution may best be revealed through analyses of model gene families within a well-characterized phylogenetic framework.

In this article we provide an example using the cotton genus, Gossypium (Malvaceae), a phylogenetically well-understood group, and alcohol dehydrogenase (*Adh*) as a model gene family with a relatively low copy number. Gossypium has a number of attributes that make it favorable for molecular evolutionary studies. Most importantly, the genus has been extensively studied from many perspectives, and phylogenetic analyses have been conducted using multiple molecular data sets (Figure 1; Wendel and Albert 1992; Seelanan *et al.* 1997; Small *et al.* 1998; R. C. Cronn, R. L. Small, T. Haselkorn and J. F. Wendel, unpublished data). Additionally, a number of molecular evolutionary studies have been published using the insights provided by this well-understood comparative framework (*e.g.*, VanderWiel *et al.* 1993; Wendel *et al.* 1995a,b; Small *et al.* 1998, 1999; Cronn *et al.* 1996, 1999; Liu *et al.* 2000; Small and Wendel 2000).

Adh is among the best-studied plant nuclear-encoded gene families, in terms of both molecular biological and molecular evolutionary investigations (reviewed by Clegg *et al.* 1997). *Adh* genes generally are of a convenient size for study (Figure 2; 2–3 kb in length with ~1100 nucleotides of coding sequence), usually have 10 exons and 9 introns, and generally exist as members of small gene families (often only two or three loci). The ADH enzyme is important primarily in response to hypoxic conditions, under which its expression is highly induced (Dolferus *et al.* 1997a). Additionally, ADH may be important during seedling development, fruit ripening, and pollen development (Freeling and Bennett 1985; Dolferus *et al.* 1997a). Molecular evolutionary studies of *Adh* genes have been performed in a number of plants, *e.g.*, maize (Eyre-Walker *et al.* 1998), barley (Cummings and Clegg 1998), Arabidopsis (Innan *et al.* 1996), Leavenworthia (Charlesworth *et al.*

Corresponding author: Randall Small, Department of Botany, 437 Hesler Biology, University of Tennessee, Knoxville, TN 37996-1100. E-mail: rsmall@utk.edu

1998), cotton (Small *et al.* 1999), palms (Gaut *et al.* 1996), and grasses (Gaut *et al.* 1999). While *Adh* is generally found in small gene families, phylogenetic analyses of available plant sequences suggest that this is due to repeated inflation and shrinkage of the gene family in different organismal lineages throughout plant evolution (Gaut *et al.* 1996; Morton *et al.* 1996; Clegg *et al.* 1997). However, distinguishing a history of repeated gene duplication and loss from incomplete sampling and other possible explanations requires detailed analysis in model plant groups.

The purpose of this article is to describe the *Adh* gene family of diploid and allotetraploid species of Gossypium. Our goals were (1) to unravel the apparent copy number complexity and history of gene duplication and divergence among *Adh* gene family members and (2) to provide a comparative analysis of the evolutionary dynamics of the gene family members. The data demonstrate that the *Adh* gene family in Gossypium is both complex and evolutionarily labile, having been subjected to gene duplication, pseudogenization, and intron loss events.

MATERIALS AND METHODS

Plant materials: Diploid species of Gossypium are divided into genome groups (A–K; see Figure 1; Table 1) on the basis of cytogenetic and crossing data, and phylogenetic analyses indicate that each genome group is monophyletic (Wendel and Albert 1992; Seelanan *et al.* 1997). These groups of species exist in three primary centers of diversity: the A-, B-, E-, and F-genomes in Africa and Asia; the C-, G-, and K-genomes in Australia; and the D-genome in North, Central, and South America (Wendel 1995). In addition to the diploid species, there are five allotetraploid (AD-genome) Gossypium species, all apparently derived from a single allopolyploidization between A- and D-genome diploids that occurred <2 mya (Wendel 1989; Seelanan *et al.* 1997; Small *et al.* 1998). The parents of the allopolyploids are best represented by the extant species *Gossypium herbaceum* L. (A-genome, African species) and *G. raimondii* Ulbrich (D-genome, South American species).

We focused on three diploid species, one representing each of the primary centers of diversity, as well as the parents of the polyploids, and one of the allotetraploid species. Specifically, we included *G. robinsonii* F. Mueller (Australian C-genome), *G. herbaceum* (African-Asian A-genome), *G. raimondii* (New World D-genome), and *G. hirsutum* L. ("upland cotton"; AD-genome allotetraploid). As outgroups we included either *Gossypioides kirkii* (Mast.) J. B. Hutch. or *Kokia drynarioides* (Seemann) Lewton. These two genera collectively compose the sister lineage of Gossypium (Seelanan *et al.* 1997). All species sampled and locations of voucher materials are listed in Table 1.

Isolation of *Adh* sequences: Some information on the *Adh* gene family in Gossypium has been published previously. Isozyme surveys (*e.g.*, Wendel and Percival 1990; Wendel *et al.* 1992; Millar *et al.* 1994) suggested that the *Adh* gene family included at least two loci and, in some species, a third (Millar *et al.* 1994; J. F. Wendel, unpublished data). Molecular genetic analyses of *Adh* have been conducted in *G. hirsutum* (Millar *et al.* 1994; Millar and Dennis 1996a,b). These analyses focused on a group of loci induced by hypoxic conditions and revealed at least five classes of sequences, termed *Adh1* and *Adh2a-Adh2d* by Millar and Dennis (1996a).

To isolate additional *Adh* sequences we employed a PCR-based approach. We used *Adh* primers P1 and P2 (sequences of all PCR primers used in this study are given in the legend of Figure 2) homologous to regions of exons 2 and 9 (Figure 2) to amplify *Adh* sequences from all species studied. PCR reaction conditions were as follows: a 50-μl reaction with 1 unit *Taq* polymerase (Promega, Madison, WI), 1× buffer (Promega), 200 μm each dNTP, 2.0 mm $MgCl_2$, 10 pmol each primer, and 1 μl template DNA ($\backsim$ 10–100 ng). Amplification was accomplished using a program of 30 cycles of denaturation at 94° for 1 min, annealing at 50° for 1 min, and extension at 72° for 2 min, followed by a final 5-min extension at 72°; all amplifications were performed in MJ Research (Watertown, MA) thermocyclers. These reactions resulted in amplification of multiple *Adh* sequences, as evidenced by agarose gel resolution of multiple bands ranging in size from 1.2 to 1.8 kb. To isolate individual PCR products we cloned the heterogeneous PCR product pool into pGEM-T (Promega) and screened colonies for *Adh* inserts as described (Small *et al.* 1998).

On the basis of data generated from the above procedure we designed sets of locus-specific PCR amplification primers (Figure 2). These primer pairs permitted selective amplification of one locus at a time, which in turn allowed us to sequence PCR products directly.

To make valid evolutionary comparisons, it is necessary to show that the sequences being compared are orthologous (related by speciation), rather than paralogous (related by gene duplication). Evidence that *Adh* sequences from different species are orthologous derived from a number of sources. Initially, orthology was inferred from retention of gene size, structure, and sequence similarity across species. Subsequently, orthology was verified by phylogenetic analyses and comparative genetic mapping. Given the well-supported phylogeny for the species of Gossypium (Wendel and Albert 1992; Seelanan *et al.* 1997), phylogenetic analysis can help establish orthology if the organismal phylogeny is recovered from the putatively orthologous sequences. Comparative genetic mapping data may provide the strongest evidence for orthology by showing retention of a shared genomic location of presumptively orthologous sequences. Shared map location is expected for orthologous loci, while paralogous loci may reside in different regions of the genome.

DNA sequencing: Sequencing was performed either by automated DNA sequencing (ABI Prism) at the Iowa State University DNA Sequencing and Synthesis Facility or by using a ^{33}P-labeled dideoxy terminator cycle sequencing kit (Amersham, Arlington Heights, IL) with electrophoresis on 5–6% Long Ranger gels (FMC, Rockland, ME). Because Gossypium species are selfing and, therefore, usually homozygous (*e.g.*, Wendel *et al.* 1992; Brubaker and Wendel 1994; Small *et al.* 1999), direct sequencing of PCR products generally resulted in a monomorphic sequence.

Southern hybridization analyses: Southern blot analysis was used for restriction fragment length polymorphism (RFLP) mapping experiments, whereby the *Adh* loci resolved in this study were included in previously published genetic maps for the A- and D-genome diploid species groups (Brubaker *et al.* 1999) and the AD-genome allotetraploid species group (Reinisch *et al.* 1994). We also used Southern blots to estimate copy number of each of the sequence types isolated. Generally, Southern hybridization provides an estimate of gene copy number, with the number of hybridizing bands roughly equivalent to the number of loci. However, digestions with enzymes that cut within the probe region can result in two hybridizing bands for a single locus, an effect that can be amplified when using longer probes. Thus we reasoned that with small ($\sim$500

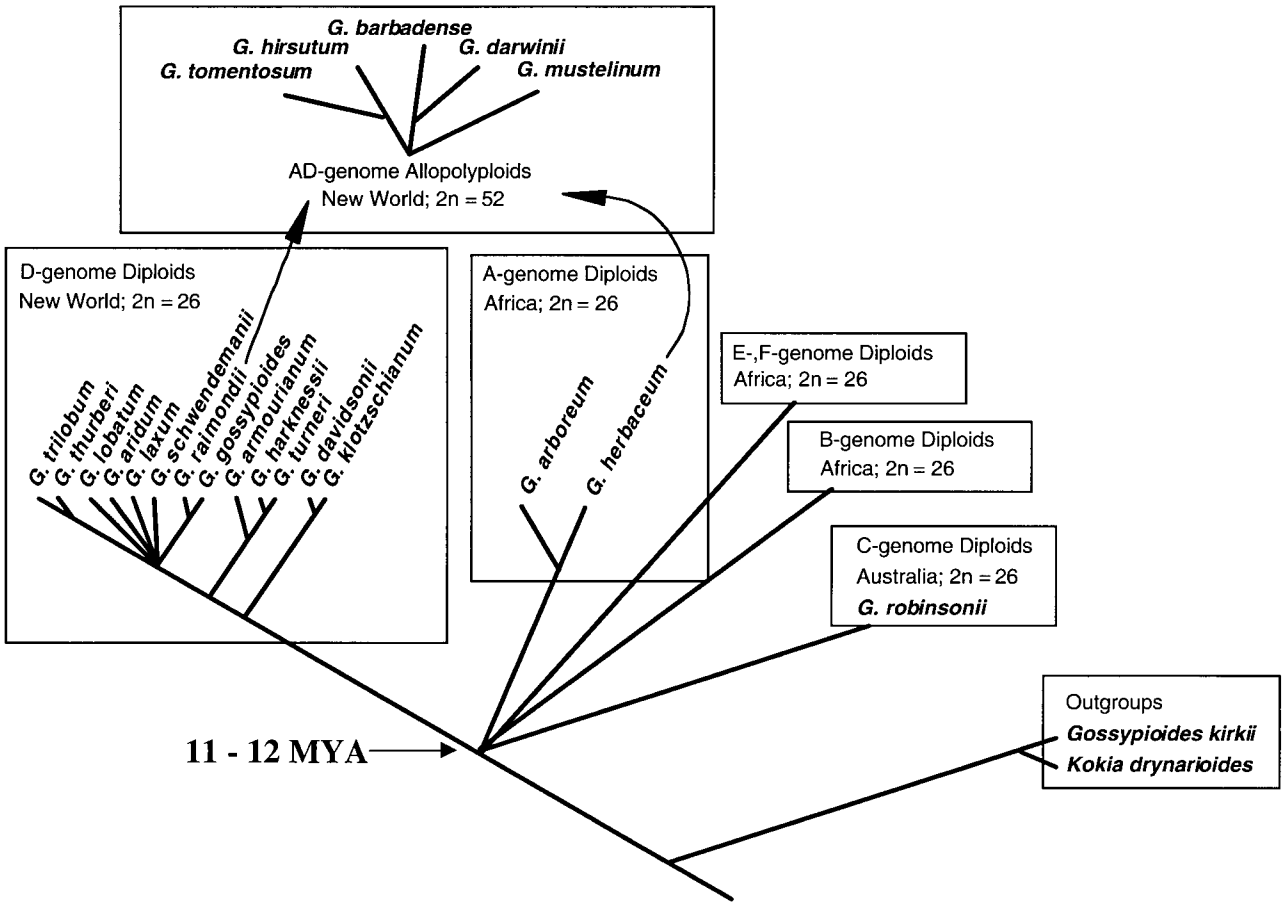

Figure 1.—Phylogenetic hypothesis for the genus Gossypium and outgroups, showing relationships among the diploid ($2n$ = 26) species, the origin of the allotetraploid ($2n$ = 52) species, and estimates of the timing of the initial divergence within the genus (Wendel and Albert 1992; Seelanan *et al.* 1997; Small *et al.* 1998; R. C. Cronn, R. L. Small, T. Haselkorn and J. F. Wendel, unpublished data).

bp) probes, each hybridizing band should be equivalent to a single locus if there are no restriction sites within the probe region and if the plant is homozygous. Heterozygosity, though rarely observed in Gossypium (Wendel *et al.* 1992; Brubaker and Wendel 1994; Small *et al.* 1999), can be distinguished from gene duplication by using multiple enzyme digestions, because heterozygosity is expected to be detected with one or a few enzymes while gene duplication would be expected

TABLE 1

Plant materials

Taxon	Accession	Voucher
Outgroups		
Gossypioides kirkii (Mast.) J. B. Hutch.		TS 3
Kokia drynarioides (Seemann) Lewton		TS 6
C-genome diploid		
Gossypium robinsonii F. Mueller	AZ-50	TS 12
D-genome diploid		
G. raimondii Ulbrich	#436	JFW and TDC 127
A-genome diploids		
G. herbaceum L.	A_1-73	JFW 539
G. arboreum L.	A_2-74	JFW and TDC 312
AD-genome tetraploid		
G. hirsutum L.	"Palmeri"	JFW and TDC 632

All voucher specimens are deposited at the Iowa State University Ada Hayden Herbarium (ISC). TS, Tosak Seelanan; JFW and TDC, J. F. Wendel and T. D. Couch.

to be revealed with most or all enzymes. To distinguish between these alternatives, DNAs (~5 μg) of the diploids *G. robinsonii*, *G. herbaceum*, and *G. raimondii* and the allotetraploid *G. hirsutum* were digested individually with the restriction enzymes *Eco*RI, *Eco*RV, *Hin*dIII, and *Xba*I, electrophoresed in 0.8% agarose gels, and transferred to nylon membranes.

Hybridization probes generally consisted of gene fragments representing the intron 3/exon 4 region from the *G. robinsonii* gene for each locus (Figure 2); these probes were generated by PCR amplification using cloned *G. robinsonii* fragments of the appropriate locus and primers Fex3 (ATG A[A/G]G C[C/T]G GAG GGT) and Bex4-3′ (CA[A/G] AC[C/T] TT[A/G] TC[A/G] AG) (provided by B. Gaut, U.C. Irvine). Preliminary Southern hybridization analyses showed that under stringent hybridization conditions (65°, 6× SSC followed by washing at 65° in 0.1× SSC, 0.5% SDS) probes did not cross-hybridize. In some cases alternative probes were used, including individual intron fragments, or the 3′ untranslated region (UTR) of cDNAs (generously provided by A. Millar, M. Ellis, and E. Dennis, CSIRO, Australia and described in Millar and Dennis 1996a); these probes were produced by restriction digestion of cloned DNA fragments. Probes were radiolabeled via random primer labeling (GIBCO-BRL, Gaithersburg, MD). Hybridization and washing conditions were as described above.

Genetic mapping: All mapping analyses used segregating F_2 populations described by Reinisch *et al.* (1994) and Brubaker *et al.* (1999). Previously described restriction-digested membrane-bound DNAs were probed with locus-specific *Adh* probes generated as described above.

In cases where RFLP analysis did not reveal polymorphism we employed alternate techniques to generate segregation data. In some cases, PCR-RFLP was used, whereby PCR products were digested with restriction enzymes that reveal a polymorphism between parental lines and, thus, segregation in the F_2 population. Single-stranded conformational polymorphism (SSCP) analysis was performed as described (Pokorny *et al.* 1997). Similar to SSCP, known length differences between PCR products from the two parents could be used in mapping through incorporation of [^{32}P]dCTP into PCR amplifications of F_2 individuals, followed by resolution on sequencing gels.

Genetic mapping procedures followed Reinisch *et al.* (1994) and Brubaker *et al.* (1999) using MapMaker version 2.0 (Lander *et al.* 1987). Mapping data are reported in terms of homoeologous assemblages of Brubaker *et al.* (1999), who compared genetic maps of the AD-genome allotetraploids (*G. hirsutum* × *G. barbadense*) with representatives of its diploid progenitors, the A-genome (*G. herbaceum* × *G. arboreum*) and the D-genome (*G. trilobum* × *G. raimondii*). Thus each homoeologous assemblage consists of four linkage groups—one from each diploid group (A, D) and two (A′, D′) from the allotetraploid.

Molecular evolutionary and phylogenetic analyses: *Adh* genes isolated from Gossypium were subjected to phylogenetic analysis along with plant *Adh* genes available from GenBank. *Adh* coding regions were aligned and subjected to neighbor-joining analysis (Saitou and Nei 1987) using Kimura two-parameter distances as implemented in PAUP* (Swofford 1999).

For each locus we performed phylogenetic and evolutionary rate analyses. Phylogenetic analysis (maximum parsimony) was performed for each locus using sequences from *G. kirkii* or *K. drynarioides* as the outgroup. In addition we performed relative rate tests (Tajima 1993) for all pairs of sequences (C *vs.* A, C *vs.* A′, C *vs.* D, C *vs.* D′, A *vs.* D, A *vs.* A′, D *vs.* D′, A′ *vs.* D′) using outgroup sequences. We also calculated Jukes-Cantor corrected synonymous (*Ksyn*) and nonsynonymous (*Ka*) substitution rates according to Nei and Gojobori (1986), as well as a Jukes-Cantor corrected silent (*Ksil*; calculated from synonymous and intron sites) and intron (*Ki*) rates. All relative rate values (*Ksyn*, *Ksil*, *Ka*, *Ki*) were calculated as the mean of all pairwise comparisons between sequences of the three diploid species (C-genome: *G. robinsonii*; D-genome: *G. raimondii*; A-genome: *G. herbaceum* or *G. arboreum*) because recent analyses have shown that these three lineages diverged from each other nearly simultaneously (Seelanan *et al.* 1997; Liu *et al.* 2000; R. C. Cronn, R. L. Small, T. Haselkorn and J. F. Wendel, unpublished data). Finally, we calculated absolute synonymous substitution rates for each locus. These estimates were calculated as the *Ksyn* (as above) divided by twice the estimated time of divergence of 11–12 million years. These divergence times are based on chloroplast *ndh*F sequence data (Seelanan *et al.* 1997) that resulted in estimated divergences of 11 mya for the A-D genome split and 12 mya for the D-C and A-C genome splits. The above calculations were expedited by the software programs Tajima93 (T. Seelanan, unpublished software), DnaSP (Rozas and Rozas 1999), and PAUP* (Swofford 1999).

RESULTS

Characterization of the *Adh* gene family: To elucidate *Adh* gene family complexity in Gossypium we undertook a PCR survey of representative diploid and allopolyploid Gossypium species as well as outgroup species. This resulted in amplification of four distinct size classes of PCR products, ranging from 1.2 to 1.8 kb. These PCR product pools were cloned and examples from each size class were identified and sequenced. We subsequently determined (see below) that each of these sequence classes represented different genetic loci (or sets of loci) and have termed them *AdhA*, *AdhB*, *AdhC*, and *AdhD*. An additional locus was later isolated (see below) and has been denoted *AdhE*. Each of these loci was sequenced from four representative Gossypium species (*G. raimondii*, *G. herbaceum*, *G. robinsonii*, *G. hirsutum*) and at least one of the outgroup species (*G. kirkii*, *K. drynarioides*) and was subjected to copy number estimation, genetic mapping experiments, and phylogenetic analysis. Absolute and relative evolutionary rates were also calculated for each locus and are presented in Table 2. Orthology of the sequences from different species was inferred from (1) shared gene structure (Figure 2) and sequence similarity, (2) genetic mapping data that show retention of genomic location across species (Figure 3), and (3) recovery of organismal relationships from phylogenetic analysis of each locus (Figure 4).

***AdhA*:** The Gossypium *AdhA* locus is unusual in that it lacks two of the introns (4 and 7) typically found in plant *Adh* genes (Figure 2; Small *et al.* 1999; Small and Wendel 2000). The introns that remain are also short relative to other Gossypium *Adh* genes (Figure 2) making *AdhA* the shortest Gossypium *Adh* gene. *AdhA* sequences are deposited in GenBank under accession nos. AF085064, AF090146, AF136457–AF136459, and AF201888.

Southern hybridization analysis indicates that *AdhA* exists in one copy per diploid genome, as a single band

TABLE 2

Patterns of nucleotide substitution within and among loci and lineages

Locus[a]	*Ki*[b]	*Ksyn*[c]	*Ksil*[d]	*Ka*[e]	*Ksyn*:*Ka* ratio	Estimated absolute synonymous substitution rate (synonymous substitutions/synonymous site/year)[f]
AdhA	0.023	0.039	0.030	0.004	9.8:1	$1.63–1.77 \times 10^{-9}$
AdhB	0.025	0.014	0.023	0.006	2.3:1	$0.58–0.64 \times 10^{-9}$
AdhC[g]	0.057	0.031	0.052	0.013	2.4:1	$1.29–1.41 \times 10^{-9}$
AdhD	0.031	0.040	0.032	0.008	5.0:1	$1.67–1.82 \times 10^{-9}$
AdhE[h]	0.032	0.032	0.032	0.010	3.2:1	$1.33–1.45 \times 10^{-9}$

[a] Shown is the mean of all pairwise comparisons between sequences of the three diploid species (C-genome, *G. robinsonii*; A-genome, *G. herbaceum* or *G. arboreum*; D-genome, *G. raimondii).*

[b] Number of substitutions per site for intron sites only.

[c] Number of synonymous substitutions per synonymous site in coding sequences (Nei and Gojobori 1986).

[d] Number of substitutions per site including intron and synonymous sites.

[e] Number of nonsynonymous substitutions per nonsynonymous site in coding sequences (Nei and Gojobori 1986).

[f] Calculated as the mean synonymous distance (*Ks*) divided by twice the time since divergence.

[g] This comparison includes the *G. arboreum AdhC* pseudogene.

[h] Because only a short fragment of the A-genome diploid sequence for *AdhE* was recovered, this comparison uses the A-subgenome sequence of *G. hirsutum.*

is observed in all digests of diploids and two bands are seen in the allotetraploid (Small *et al.* 1999). The sole exception to this is with the *Eco*RV digest of *G. herbaceum*, which displays two bands (not shown). Using the *AdhA* intron 3/exon 4 probe in Southern hybridization analysis of F_2 populations, we were able to genetically map *AdhA* to homoeologous assemblage 8C of Brubaker *et al.* (1999) in both of the diploid populations and in the D-subgenome of the allotetraploid (Figure 3).

Phylogenetic analysis of *AdhA* sequences (Figure 4) revealed the topology expected from our understanding of relationships among the species studied, with the sequence from the A-genome diploid being sister to its counterpart from the A-subgenome of the allotetraploid

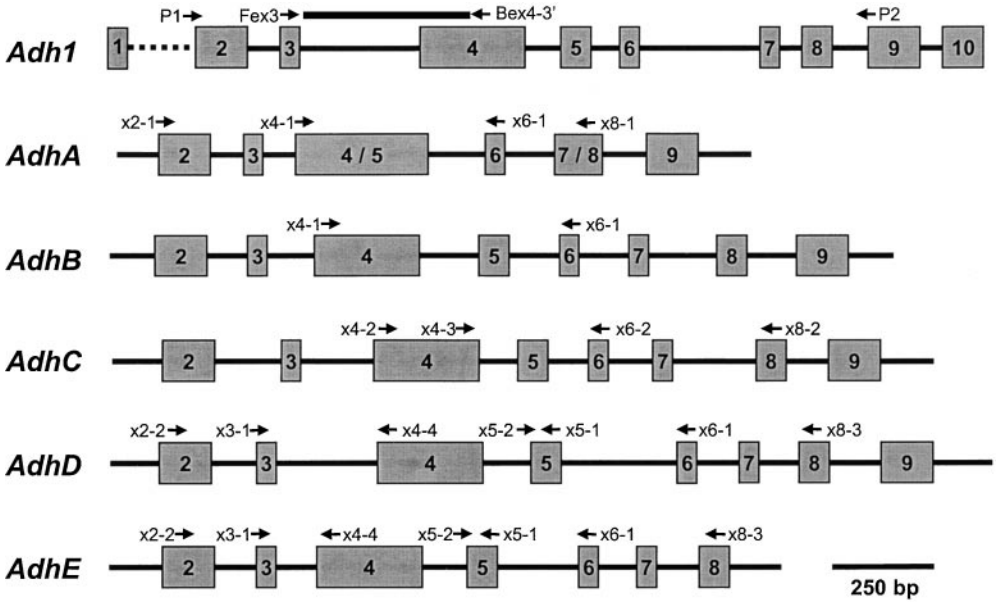

Figure 2.—Schematic representations of the maize *Adh1* and Gossypium *AdhA–E* genes. Numbered boxes represent exons and intervening lines represent introns. Amplification and sequencing primers are shown in their approximate locations; the thick line above the intron 3/exon 4 region of the maize *Adh1* gene indicates the region used as a probe in Southern hybridization experiments; intron 1 of the maize *Adh1* gene is represented by a dashed line to indicate that its full length (~530 bp) is not shown; a 250-bp scale is shown at bottom right for reference. Introns 4 and 7 are missing from *AdhA*. Primer sequences (written 5′ to 3′) are as follows with forward primers denoted by (f) and reverse primers by (r): P1 (f) (CTGCKGTKGCATGGGARGCAGGGAAGCC); P2 (r) (GCACAGCCACACCCCAACCCTG); x2-1 (f) (CTTCACTGCTTTATGTCACACT); x2-2 (f) (GCAATGGAGGTTCGTCTG); x3-1 (f) (ACTCCATTATTTCCTCGTAT); x4-1 (f) (TCATGTTCTCCCTATCTTCAC); x4-2 (f) (GTGGAGAGTGTAGGTGAAGG); x4-3 (f) (GGGCAGACTAGGTTTTCCAAAG); x4-4 (r) (ACCTCACCCACACTCTCAAC); x5-1 (r) (GCCACAGTTGAACCTTTG); x5-2 (f) (AATAATTTTCGAGGTCTTGG); x6-1 (r) (ATCAACACCAATAATCCTAGAA); x6-2 (r) (TCAATACCAATGATCCTAGAA); x8-1 (r) (GGACGCTCCCTGTACTCC); x8-2 (r) (GAAACCATGGCCTGGGTG); x8-3 (r) (GATCATGGCATTAATGTTTC).

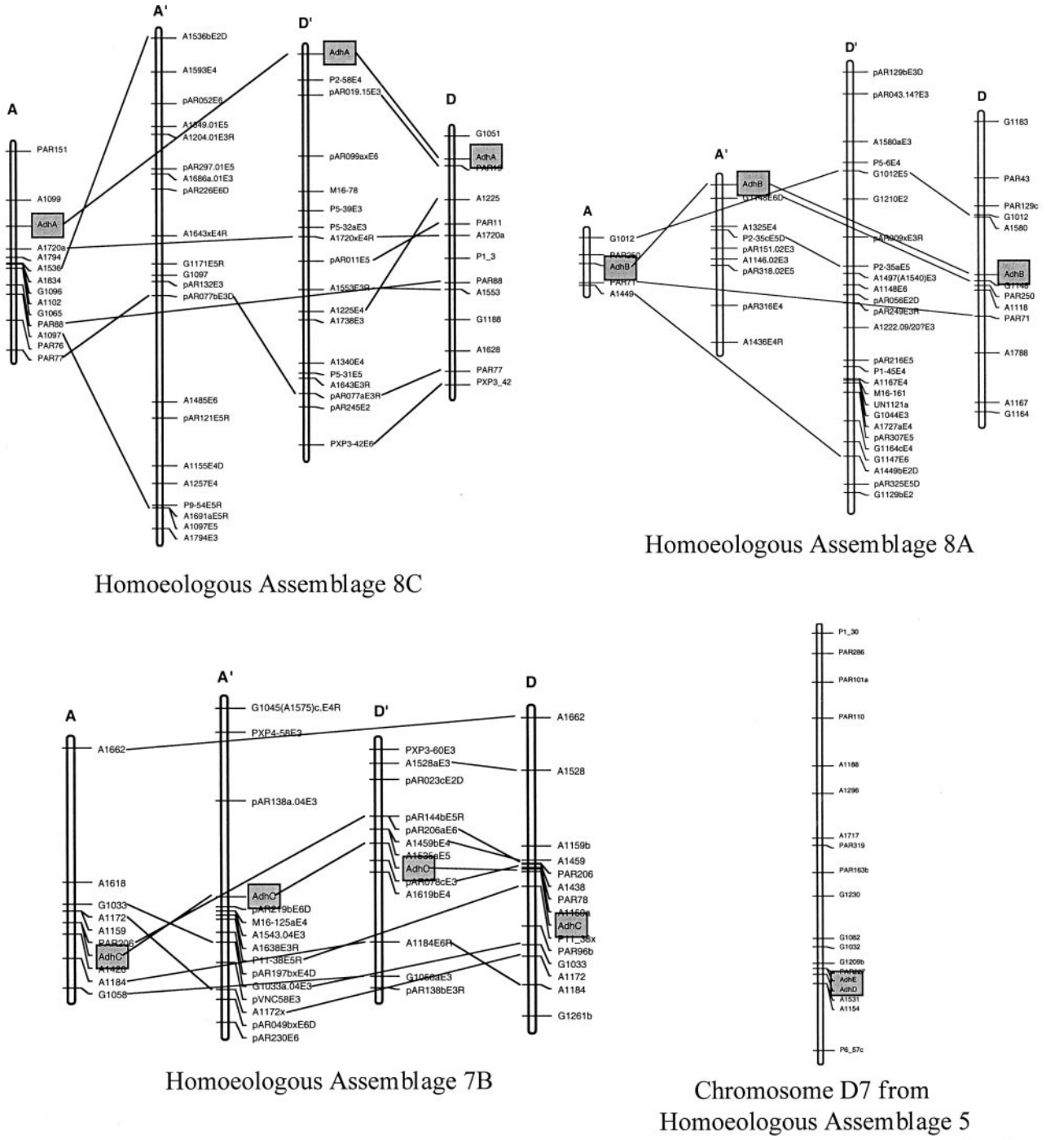

Figure 3.—Comparative genetic mapping of Gossypium *Adh* loci as in Brubaker *et al.* (1999). *AdhA* maps to homoeologous assemblage 8C in both A- and D-genome diploids and in the D-subgenome of the allotetraploid. *AdhB* maps to homoeologous assemblage 8A in both A- and D-genome diploids and in the A-subgenome of the allotetraploid. *AdhC* maps to homoeologous assemblage 7B in both A- and D-genome diploids and in both subgenomes of the allotetraploid. *AdhD* and *AdhE* are closely linked on chromosome D7 (D-genome diploid) in homoeologous assemblage 5.

and the sequence from the D-genome diploid being sister to its counterpart from the D-subgenome of the allotetraploid. The C-genome sequence was resolved as sister to the A-genome *AdhA* gene, which was not unanticipated given recent analyses that often support this resolution (Seelanan *et al.* 1997; Liu *et al.* 2000; R. C. Cronn, R. L. Small, T. Haselkorn and J. F. Wendel, unpublished data).

Using the estimated divergence times of 11–12 mya (Seelanan *et al.* 1997), we calculated an absolute synonymous substitution rate (using only exon sequences) for *AdhA* of 1.63–1.77 $\times 10^{-9}$ synonymous substitutions/

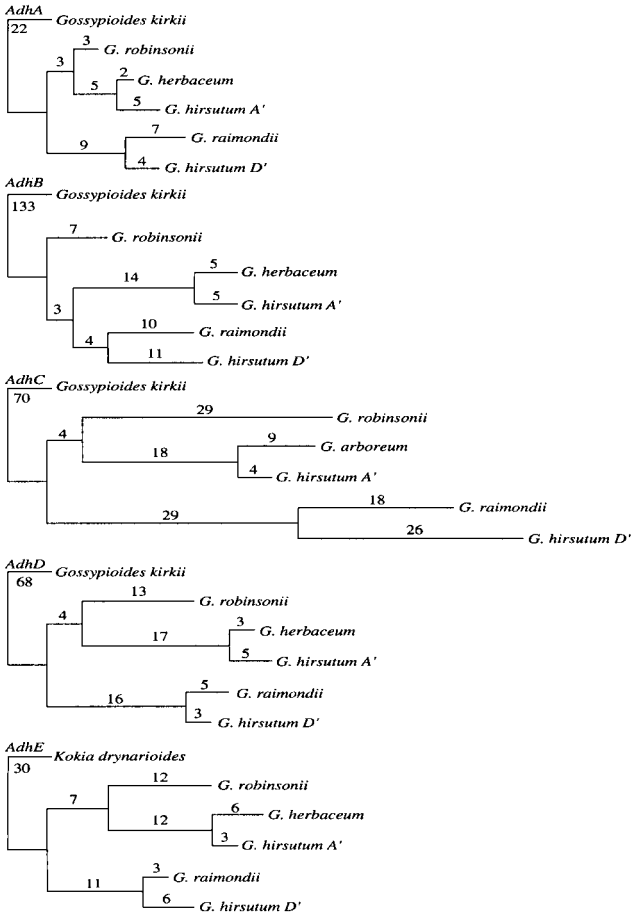

Figure 4.—Phylogenetic trees resulting from parsimony analysis of sequences of *AdhA*, *AdhB*, *AdhC*, *AdhD*, and *AdhE*, respectively, rooted with a *G. kirkii* or *K. drynarioides* sequence. Branch lengths are given above each branch. The A- and D-subgenomic sequences of *G. hirsutum* are designated *G. hirsutum* A′ and D′, respectively. For each tree the following information is provided: tree length including autapomorphies (L), consistency index (CI), and retention index (RI). *AdhA*: single most parsimonious tree, L, 60; CI, 0.97; RI, 0.90; *AdhB*: one of two equally parsimonious trees, L, 192; CI, 0.97; RI, 0.77; *AdhC*: one of two equally parsimonious trees, L, 207; CI, 0.97; RI, 0.89; *AdhD*: one of two equally parsimonious trees, L, 134; CI, 0.97; RI, 0.90; *AdhE*: single most parsimonious tree, L, 90; CI, 0.98; RI, 0.91.

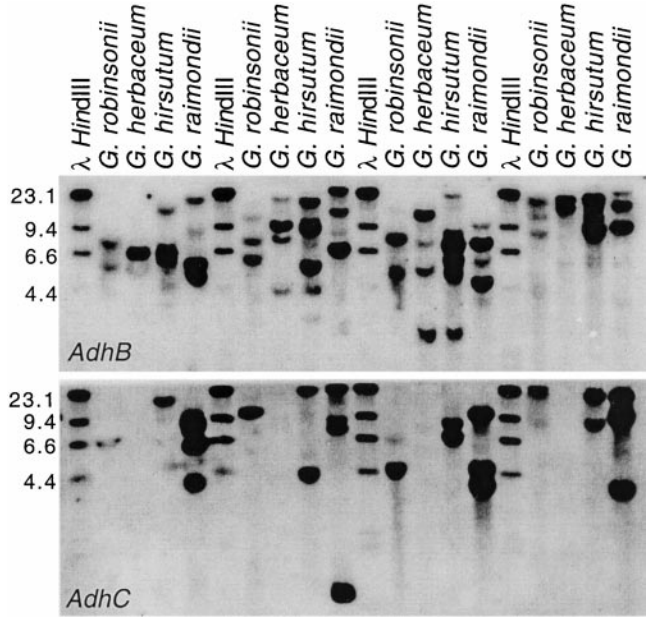

Figure 5.—Southern hybridization analysis of a portion of the *Adh* gene family in Gossypium. DNA of each species (*G. robinsonii*, *G. herbaceum*, *G. hirsutum*, and *G. raimondii*) was digested with four enzymes (*Eco*RI, *Eco*RV, *Hin*dIII, and *Xba*I). Each panel of digestions is separated by a λ *Hin*dIII-digested marker lane. (Top) Probed with an *AdhB* intron 3/exon 4 probe. (Bottom) Probed with an *AdhC* intron 3/exon 4 probe.

synonymous site/year. This estimate differs slightly from the previously published estimate (Small *et al.* 1999) of 1.5–2.1 $\times$ 10^{-9} because it was calculated as the mean of all pairwise comparisons divided by two separate divergence times as opposed to two different point estimates in the previous article. This different approach was taken because of the apparent nearly simultaneous branching of the lineages represented by modern C-, A-, and D-genome cottons.

***AdhB*:** The Gossypium *AdhB* locus maintains a 10 exon/9 intron structure typical of most angiosperm *Adh* genes (Figure 2), as do all other Gossypium *Adh* genes. On the basis of phylogenetic analysis (see below) we found this locus to be closely related to the *Adh2* genes reported by Millar and Dennis (1996a). *AdhB* sequences are deposited in GenBank under accession nos. AF226630–AF226636.

Southern blots revealed a complex pattern when probed with the *AdhB* intron 3/exon 4 probe (Figure 5), yet the *AdhB* probe does not cross-hybridize to fragments detected using *AdhA*, *AdhC*, *AdhD*, or *AdhE* probes. Diploid species displayed from two to four bands per digest while the tetraploid displayed up to six hybridizing bands (Figure 5). Sequence alignment of *AdhB* with the *Adh2* genes of Millar and Dennis (1996a) shows that there is retention of significant sequence homology between these genes, even in the introns, such that they would cross-hybridize under our experimental conditions. We were able to genetically map *AdhB*-like loci in three of the four linkage groups of homoeologous assemblage 8A (Figure 3). In addition to segregating bands observed with the *AdhB* probe, we mapped *Adh2a* of Millar and Dennis (1996a) using the 3′ UTR of a cDNA. This locus is tightly linked to *AdhB*, suggesting that the *AdhB/Adh2* gene "subfamily" evolved via a process of tandem gene duplication.

Phylogenetic analysis of the *AdhB* sequences again resulted in the expected topology (Figure 4) and relative rate tests detect no departures from rate homogeneity. As noted above, the *Adh2* sequences of Millar and Dennis (1996a) appear closely related to our *AdhB* sequences, on the basis of (1) overall nucleotide similarity in the coding regions and (2) the ability to confidently

align intron sequences (intron sequences are unalignable in most other interlocus comparisons, although see discussion of *AdhD/E* below). Inclusion of all *AdhB* and *Adh2* sequences in a phylogenetic analysis (data not shown) reveals that (1) the *Adh2b* sequence (Millar and Dennis 1996a) is probably orthologous to the *AdhB* sequences we report here as it is sister to the *AdhB* sequence from the D-subgenome of *G. hirsutum* and (2) the *Adh2a* and *Adh2d* sequences appear to represent loci that are distinct both from our *AdhB* and from each other, as also noted by Millar and Dennis (1996a). Our present estimate is that there are a minimum of three *Adh* sequences in the diploids that retain sufficient sequence homology to cross-hybridize with our *AdhB* clone and that this class represents an *AdhB/Adh2* subfamily of genes (*AdhB* = *Adh2b*, *Adh2a*, and *Adh2d*). An additional sequence isolated by Millar and Dennis (1996a), *Adh2c*, is a cDNA that consists of the 3′ end of a gene and the 3′ UTR. There is no overlap between this sequence and other *Adh* sequences isolated so we are unable to determine whether *Adh2c* corresponds to any previously isolated sequences; thus, *Adh2c* may represent either an additional or a previously isolated gene. Sequences isolated from the outgroups *G. kirkii* and *K. drynarioides* appear to represent paralogs, rather than orthologs of *AdhB* as they are relatively divergent from the *AdhB* sequences (data not shown).

***AdhC*:** Sequence data for *AdhC* were reported previously in the context of a phylogenetic analysis of the allotetraploid species of Gossypium (Small *et al.* 1998). *AdhC* sequences have been deposited in GenBank under accession nos. AF036567–AF036569, AF036574, AF036575, and AF169254.

Southern blots show that *G. hirsutum* displays two bands per digest, as expected for a single locus duplicated as a consequence of polyploidization (Figure 5). Unexpectedly, however, the D-genome diploid *G. raimondii* displays three bands per digest, indicative of one or more gene duplications, whereas the A-genome diploid *G. herbaceum* does not hybridize at all to the *AdhC* probe (Figure 5). As reported previously (Small *et al.* 1998), we were able to isolate an *AdhC* fragment from *G. arboreum*, the only other extant A-genome taxon, and this fragment clearly represents a pseudogene as it contains an internal stop codon and large deletions (one of which removes all of exon 6 plus regions of the flanking introns). We were able to genetically map *AdhC* to homoeologous assemblage 7B on both diploid maps and in both subgenomes of the allotetraploid map (Figure 3). Because *AdhC* is missing from *G. herbaceum*, it was mapped as a dominant marker in the *G. herbaceum* × *G. arboreum* mapping population.

Phylogenetic analysis of *AdhC* sequences results in the expected topology (Figure 4) and also reveals the rate heterogeneity previously described (Small *et al.* 1998). The deviation from rate homogeneity is due to an apparent rate acceleration in the lineage leading to *G. raimondii* and the D-subgenome of the allotetraploids, relative to the A- and C-genome lineages.

***AdhD*:** *AdhD* is the largest of the Gossypium *Adh* genes reported here, owing primarily to the length of introns 3 and 5 (Figure 2). Phylogenetic analysis (see below) indicates that this locus is probably orthologous to the *Adh1* sequence reported by Millar and Dennis (1996a). *AdhD* sequences have been deposited in GenBank under accession nos. AF059418 and AF250201–AF250205.

Southern hybridizations revealed strong hybridization to a single band in the diploid species and two bands in the allotetraploid species, in addition to weaker hybridization to one or more bands in some digests. This suggested that an additional locus closely related to *AdhD* was present in the Gossypium genome, a suspicion subsequently confirmed. For the phylogenetic study of Seelanan *et al.* (1999), PCR primers were developed that were intended to be locus-specific for *AdhD*; these primers were homologous to regions in exons 2 and 8 (Figure 2). Amplification using these primers, however, resulted in two distinct products—*AdhD* and a second, heretofore undiscovered locus; this second locus was termed *AdhE* and is discussed below. *AdhE* is similar to *AdhD*, both in exon (Table 3) as well as in most intron sequences, which explains the cross-hybridization noted above. *AdhD* and *AdhE* are distinguishable at the PCR amplicon level, however, because they differ in size due to length differences (primarily) in introns 3 and 5. Due to a lack of polymorphism at the RFLP level for *AdhD* we were able to map this locus only by using SSCP, where *AdhD* and *AdhE* were readily distinguished by size. This allowed us to map *AdhD* in the D-genome mapping population where it mapped to chromosome D7 (Figure 3). Notably, *AdhD* and *AdhE* mapped to positions close to each other on this linkage group, suggesting a history of local, perhaps tandem duplication.

Phylogenetic analysis leads to the expected topology (Figure 4) and rate homogeneity. Inclusion of the *Adh1* cDNA sequence of Millar and Dennis (1996a) indicates that *Adh1* is probably orthologous to *AdhD*, as the *Adh1* cDNA sequence comes out as sister to the *AdhD* sequence from the A-subgenome of *G. hirsutum* (data not shown). This is bolstered by Southern hybridization analysis using the 3′ UTR of the *Adh1* cDNA as a probe (data not shown). The Southern hybridization pattern of *Adh1* was a subset of the patterns shown using the *AdhD* intron 3/exon 4 probe. Presumably the 3′ UTR of the *Adh1* cDNA is sufficiently diverged from that of *AdhE* that they do not cross-hybridize. Thus, we identified the *AdhE* bands by subtraction.

***AdhE*:** This locus was isolated using PCR primers homologous to regions in exons 2 and 8 (see above); thus, the genomic sequence data generated is shorter than that for the other Gossypium *Adh* loci. PCR amplifications yielded *AdhE* amplicons from the D-genome diploid and from both subgenomes of the allotetraploid, but no products were amplified from either of the two

TABLE 3

Percentage identity among *G. robinsonii*, *Zea mays*, and *A. thaliana Adh* coding regions

	1	2	3	4	5	6	7	8
1. Arabidopsis *Adh*[a]	—	81.3	79.4	80.8	85.3	80.5	84.1	82.0
2. Zea *Adh1*[b]	74.6	—	87.1	80.1	83.8	80.5	89.3	88.4
3. Zea *Adh2*[c]	72.1	82.0	—	79.3	80.8	76.7	84.5	84.1
4. Gossypium *AdhA*	73.5	73.2	69.3	—	85.3	80.8	82.0	82.0
5. Gossypium *AdhB*	76.8	75.6	71.6	80.6	—	86.8	83.3	82.8
6. Gossypium *AdhC*	73.8	72.8	68.5	80.1	85.5	—	81.5	80.7
7. Gossypium *AdhD*	76.6	76.0	74.1	75.1	75.0	75.0	—	92.7
8. Gossypium *AdhE*	76.0	77.3	74.6	75.3	75.6	75.3	93.4	—

Nucleotide identity is below the diagonal; amino acid identity is above the diagonal.
[a] GenBank accession no. X77943.
[b] GenBank accession no. X00580.
[c] GenBank accession no. X01965.

extant A-genome diploids. Additional PCR experiments using internal primers (x5-2 and x6-1, Figure 2) allowed us to amplify a ~300-bp fragment that included a portion of exon 5, all of intron 5, and a portion of exon 6. *AdhE* sequences have been deposited in GenBank under accession nos. AF250206–AF250211.

As noted above, sequences of *AdhD* and *AdhE* have high identity, resulting in cross-hybridization on Southern blots. We deciphered the relationships among these genes with a combination of Southern hybridizations (data not shown). Intron + exon probes from either *AdhD* or *AdhE* hybridized to both loci and thus revealed identical patterns. Use of an *Adh1* (*AdhD*) cDNA 3′ UTR probe, however, revealed a hybridization pattern that was a subset of the fragments revealed with the exon + intron probes. Presumably then, those bands that hybridized to exon + intron probes as well as the *Adh1* 3′ UTR represent *AdhD*, while those bands that hybridize only to the exon + intron probes represent *AdhE*.

AdhE displayed an RFLP polymorphism in the parents of the D-genome diploid mapping population. Analysis of the segregation data showed that *AdhE* maps on chromosome 7, tightly linked to *AdhD*. Phylogenetic analysis of *AdhE* sequence revealed the expected topology (Figure 4).

DISCUSSION

Interlocus comparisons of evolutionary dynamics: An advantage of studying a small gene family in a phylogenetically understood, closely related group of species is that a number of intra- and interlocus comparisons may be drawn regarding processes and patterns of evolution. For *Adh* in Gossypium, these may be illuminated by interlocus comparisons of sequence divergence for exons and introns, variation in intron presence, variation in evolutionary rates between loci and lineages for each locus, and variation in gene copy number. Each of these is discussed in turn.

Exon and intron divergence: Table 3 presents a comparison of divergence in coding sequences (for both nucleotide and amino acid sequences) among the Gossypium *Adh* loci. For perspective we also include comparisons between Gossypium loci and other model *Adh* loci: maize *Adh1* and *Adh2* and *Arabidopsis thaliana Adh*. Divergence amounts among the Gossypium *Adh* genes reflect their phylogenetic relationships (see below), in that Gossypium *AdhA*, *AdhB*, and *AdhC* are all more similar to each other than any one of them is to *AdhD* or *AdhE*, and vice versa. Nucleotide identities among the Gossypium sequences reflect the two gene lineages, exceeding 80% for all comparisons within the *AdhA-AdhB-AdhC* group and being 93.4% for the *AdhD-AdhE* comparison. These numbers have close parallels in the amino acid identity matrix (Table 3). Notably, *Adh* genes from Arabidopsis and Zea are not dramatically more divergent from the Gossypium sequences, nor from each other, than are sequences from intergenic comparisons within Gossypium. For example, nucleotide and amino acid identities among the three genes from Arabidopsis and Zea fall within relatively narrow ranges (72–82% and 79–87%, respectively) that are similar to those observed among Gossypium sequences. In addition, intergeneric divergences are not impressively higher than those calculated among genes within Gossypium, with identities between Arabidopsis and Gossypium genes being only incrementally higher than those between Zea and Gossypium. Collectively, these data fail to suggest a close relationship between any of the Gossypium sequences and the model *Adh* genes from the other taxa. Thus, it is not possible to support any inference of orthology among any of these loci in intergeneric comparisons. Instead, the data suggest relatively ancient paralogy among these *Adh* genes and, by extension, a complex history of gene duplication and loss.

Intron number variation: Most plant *Adh* sequences have a 10 exon/9 intron structure (Figure 2), with introns found at identical sites. The Pinus genomic sequences

isolated also have this structure (Perry and Furnier 1996), suggesting that it is the ancestral condition in seed plant *Adh* genes. Intron loss from nuclear genes is not uncommon (Drouin and Moniz de Sá 1997; Frugoli *et al.* 1998; Loguercio and Wilkins 1998), however. Several cases of missing introns have been reported in *Adh* genes, including several members of the Brassicaceae: Arabidopsis (Chang and Meyerowitz 1986), Arabis (Miyashita *et al.* 1996), and Leavenworthia (Charlesworth *et al.* 1998), as well as in barley (Trick *et al.* 1988). While the mechanism(s) of intron loss have not been demonstrated, they presumably involve interaction between an intact gene and a processed pseudogene or reverse-transcribed cDNA (Drouin and Moniz de Sá 1997; Frugoli *et al.* 1998; Loguercio and Wilkins 1998).

All Gossypium *Adh* genes have the normally found introns in the same positions as in other plant *Adh* genes, with the exception of *AdhA*, which has lost two introns (Figure 2) as previously reported (Small *et al.* 1999; Small and Wendel 2000). The absent introns are those between exons 4 and 5 and exons 7 and 8. It is intriguing that these are two of the three introns missing from the Brassicaceae *Adh* genes and that phylogenetic analysis shows that this shared loss is not due to inheritance of an intronless gene from a common ancestor (see below). This situation may be analogous to repeated independent loss of introns from chloroplast genes (*e.g.*, Downie *et al.* 1991; Lai *et al.* 1997).

Intron sequence divergence between loci presumably is a measure of evolutionary distance between loci, but the possibility exists for interlocus interactions and gene conversion events. In most comparisons between Gossypium *Adh* loci, intron sequences are unalignable and intron lengths differ. These data constitute compelling evidence for an absence of interlocus interactions. The sole exception may be for the *AdhB*/*Adh2* gene cluster. *AdhB*/*Adh2* sequences are alignable throughout their length, although a number of insertions and deletions (indels) must be introduced in the introns. Also, these loci map very close to each other in their respective linkage groups, suggesting a history of recent tandem gene duplication. Millar and Dennis (1996a) noted the potential recombinant origin of one of the *Adh2* sequences they isolated; such a scenario makes sense in light of the tandem arrangement of the genes and the potential for genic interactions to occur.

Rate variation: Mean absolute evolutionary rate values for plant nuclear genes have been estimated (Wolfe *et al.* 1987; Gaut 1998) to range from a low of 1.5×10^{-9} synonymous substitutions/synonymous site/year (Small *et al.* 1999; Vieira *et al.* 1999) to a high of 30×10^{-9} synonymous substitutions/synonymous site/year (Wolfe *et al.* 1987), although this upper value probably reflects an inflated, paralogous comparison. Although rates vary widely among loci and plant lineages, a mean rate, based on a comparison of nine nuclear genes in rice and maize, has been calculated at 6.0×10^{-9} synonymous substitutions/synonymous site/year (Gaut 1998). Our results from *Adh* in Gossypium exemplify this rate variation between loci and between lineages.

Rate variation among loci is evident from comparisons of both absolute and relative rates. First, using an independently estimated calibration point (Figure 1), we estimated absolute synonymous substitution rates for all five loci. These estimates range from 0.58×10^{-9} (*AdhB*) to 1.82×10^{-9} (*AdhD*) synonymous substitutions/synonymous site/year, a greater than 3-fold difference among loci. Such variation was also noted by Gaut (1998) in a comparison of nine nuclear genes between rice and maize. We note that although similar levels of synonymous rate variation were observed (3.1-fold difference in Gossypium, 2.4-fold in grasses; Gaut 1998), rates in Gossypium are much lower. It should be noted, however, that absolute rate estimation is particularly dependent on divergence time estimates. Thus, variation seen among lineages may reflect different rates in different lineages or, alternatively, may reflect relative accuracy of dating divergence times.

Rate variation among loci is also apparent when comparing synonymous (*Ksyn*) and nonsynonymous (*Ka*) relative rates (Table 2). Because these rates are calculated on a per site basis, they can be directly compared (within a given phylogenetic context) despite the fact that they are derived from sequences of different lengths. Synonymous rates range from *Ksyn* = 0.014 (*AdhB*) to *Ksyn* = 0.040 (*AdhD*), a 2.9-fold difference. Average nonsynonymous rates range from *Ka* = 0.004 (*AdhA*) to *Ka* = 0.013 (*AdhC*), a 3.3-fold difference. These observations are again consistent with those of Gaut (1998), who noted greater variation in nonsynonymous than in synonymous rates.

Rate equivalence among lineages was evaluated using the Tajima (1993) relative rate tests. Significant deviation from rate homogeneity was detected only for *AdhC*, as previously reported (Small *et al.* 1998), where sequences from the D-(sub)genomes are accumulating nucleotide substitutions at a higher rate than are the other sampled genes. It is provocative that this rate acceleration is accompanied by increased nucleotide polymorphism in the D-subgenome of the allotetraploids *G. hirsutum* and *G. barbadense* (R. L. Small and J. F. Wendel, unpublished data). Observations of the same bias, but to a lesser extent, were reported for *AdhA* (Small *et al.* 1999). Together, these observations suggest that the D-subgenome lineage may be subject to different intragenomic evolutionary pressures than the A-subgenome (but see Cronn *et al.* 1999).

History of *Adh* duplication and divergence: A central finding of the present work is that the *Adh* gene family is not only complex, but is evolutionarily labile with respect to gene copy number, even within a single angiosperm genus. For example, although Southern hybridization analysis indicates that *AdhA* exists in a single

copy per diploid genome in most species, a broader sampling of taxa revealed a gene duplication in a group of four Mexican Gossypium species (Small and Wendel 2000). Similarly, Southern analysis of an *AdhB* fragment revealed two to four hybridizing fragments in all diploid genomes, suggestive of a recent history of gene duplication (Figure 5). The *AdhB* loci we resolved also matched the sequences of *Adh2* genes described from *G. hirsutum* (Millar and Dennis 1996a). Phylogenetic analysis of these sequences suggests a minimum of three *AdhB*/*Adh2*-like loci, with a fourth (*Adh2c*) suggested by the work of Millar and Dennis (1996a). Mapping data indicate that these loci are tightly linked and are probably the result of local gene duplications. An additional example of recent gene duplication involves *AdhD* and *AdhE*, which cross-hybridize at the Southern level. Each appears to be represented by a single locus per diploid genome, tightly linked to each other.

Gene duplication is only one of the phenomena creating *Adh* gene family complexity in Gossypium. *AdhC* reveals in a microcosm several phenomena impacting *Adh* evolution, including not only gene duplication, but also pseudogenization and deletion, each in different species. Southern blots (Figure 5) reveal three hybridizing bands in the D-genome species, *G. raimondii*, suggesting gene duplication(s). This same figure shows that *AdhC* does not hybridize to any sequence in the genome of *G. herbaceum*, an A-genome diploid species; attempts to PCR amplify *AdhC* from *G. herbaceum* were also unsuccessful. Hybridization of *AdhC* to the other extant A-genome species, *G. arboreum*, did result in a single hybridizing band (data not shown) and we were able to isolate an *AdhC* gene fragment from *G. arboreum* via PCR (Small *et al.* 1998). This gene fragment, however, clearly represents a pseudogene, as it contains both an internal stop codon and a large deletion that removes the entirety of exon 6 as well as portions of the surrounding introns. Despite the lack of an intact *AdhC* in either of the extant A-genome diploid species, the A-subgenome of all five allotetraploid species contains what appears to be fully intact *AdhC* sequences (Small *et al.* 1998). This indicates that pseudogenization and loss of *AdhC* from *G. arboreum* and *G. herbaceum*, respectively, occurred after the split of these species from the taxon that was involved in the origin of the allotetraploids. Furthermore, mutations in intron splice site sequences and deletions in some *AdhC* sequences from the D-subgenome of the allotetraploid species suggest that these loci may also be pseudogenes.

The sum of these observations indicates that while the *Adh* gene family in angiosperms may seem stable in terms of copy number (Clegg *et al.* 1997), analysis of the gene family in a group of closely related species reveals dynamic fluctuations in gene copy number (Morton *et al.* 1996; Clegg *et al.* 1997; Gaut *et al.* 1999). These fluctuations are due to both the origin of new genes via gene duplication events (often due to local duplications) and to the loss of genes through pseudogenization and gene deletion.

As noted above, most angiosperms are reported to have two or three *Adh* loci (*e.g.*, Gottlieb 1982; Dennis *et al.* 1984, 1985), although it is rare that the goal of a study is to document the total number of genes within a gene family in a species. Thus prior estimates may reflect either an actual small gene family size or an absence of thorough searching for additional genes. For example, isozyme analysis indicated that diploid Gossypium contained two (*e.g.*, Suiter 1988) or, rarely, three *Adh* loci (Millar *et al.* 1994; J. F. Wendel, unpublished data). The molecular genetic analysis of Millar and Dennis (1996a) documented five potential loci. The present study, however, indicates that there are at least seven *Adh* loci in diploid Gossypium and, thus, a minimum of 14 in the allotetraploids.

Variation in gene number from other species has been documented previously. For example, three loci have been reported from a number of species, *e.g.*, Hordeum (Trick *et al.* 1988), Sorghum (Ellstrand *et al.* 1983), some accessions of maize (Osterman and Dennis 1989), some palms (Morton *et al.* 1996), some Paeonia species (Sang *et al.* 1997), and Leavenworthia (Charlesworth *et al.* 1998). Other species, notably some members of the Brassicaceae (Arabidopsis, Arabis; Chang and Meyerowitz 1986; Miyashita *et al.* 1996), have but a single *Adh* locus. The largest plant *Adh* gene family previously reported is from a gymnosperm, *Pinus banksiana*, which contains at least seven expressed *Adh* loci (Perry and Furnier 1996). Gossypium contains the largest *Adh* gene family yet described in angiosperms with at least 7 genes in the diploids and 14 in the allotetraploids, thus equaling the largest *Adh* gene family described from any plant. The functional significance of this observation is, at present, unknown, but it is interesting to note that cultivated cotton is relatively intolerant to flooding despite the large *Adh* gene family and the fact that ADH expression is induced severalfold in anaerobically induced cotton plants (Millar *et al.* 1994; Millar and Dennis 1996a,b).

The foregoing discussion documents the complexity and lability of the *Adh* gene family in plants. A logical extension is that the use of terms such as "*Adh1*," and "*Adh2*," erroneously perpetuates the myth that all plant *Adh1* genes are more closely related to each other than any are to *Adh2* genes. This unjustified assumption of orthology appears to be responsible, at least in part, for the use of the term *Adh1* to refer to genes expressed early during development and constitutively at low levels throughout the plant, while genes called *Adh2* are often expressed primarily when induced by hypoxia or other environmental stresses.

We conducted phylogenetic analysis of all reported plant *Adh* sequences and generated the topology shown in Figure 6. Similar analyses have been performed previously, although with fewer plant *Adh* sequences (Sun

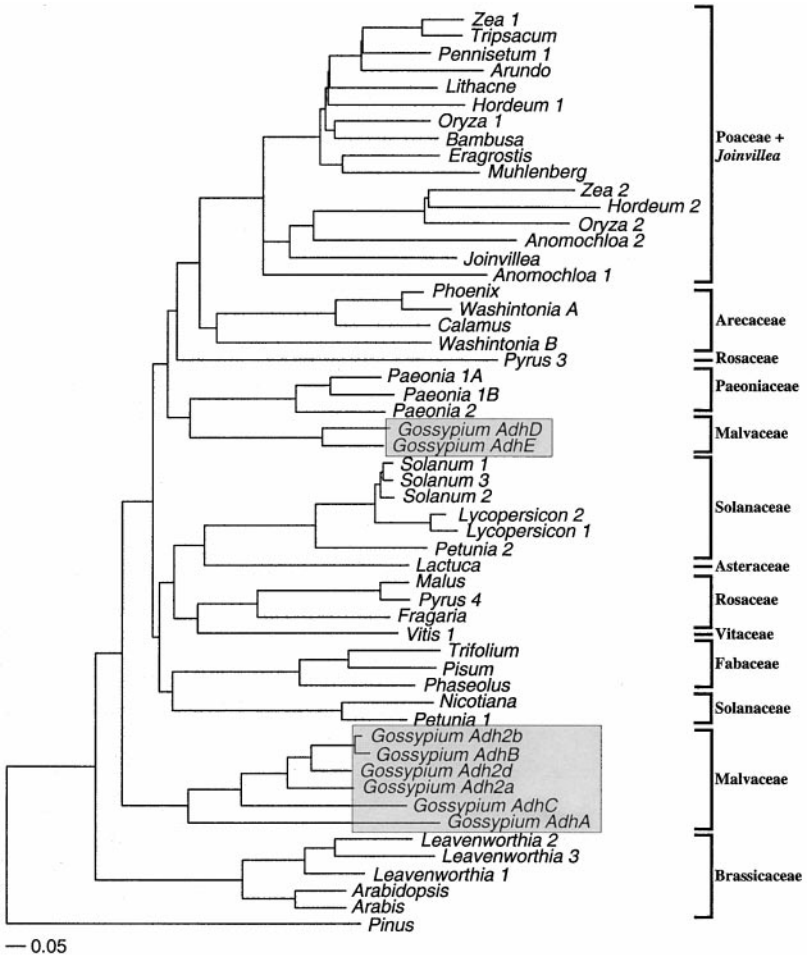

Figure 6.—Phylogenetic analysis (neighbor-joining based on Kimura two-parameter distances) of plant *Adh* genes; rooted with a *P. banksiana Adh* sequence.

and Plapp 1992; Yokoyama and Harry 1993; Dolferus *et al.* 1997b). Several conclusions may be drawn from this analysis. First, *Adh* sequences do not fall into two primary clades as predicted by the ancient gene duplication hypothesis. In fact, the topology of the tree shows that gene duplications have occurred at multiple levels within the tree, *i.e.*, at various times during evolution. Examples of relatively old duplications include sequences from the plant family Solanaceae (Lycopersicon, Nicotiana, Petunia, and Solanum), which occur on two clades separated by a number of other groups (Figure 6). A similar history is evident for sequences from the Rosaceae (Fragaria, Malus, and Pyrus *Adh4 vs.* Pyrus *Adh3*). More recent gene duplications are also evident. For example, the *Adh1* and *Adh2* sequences of the grass family are more closely related to each other than they are to other monocot sequences, indicating that a recent gene duplication is responsible for this arrangement (Gaut *et al.* 1999). Similar results have been obtained for Paeonia, where one recent gene duplication gave rise to *Adh1* and *Adh2* and a second gave rise to *Adh1a* and *Adh1b* in a subset of species (Sang *et al.* 1997).

The phylogenetic analysis only hints at the complexity of the history of gene duplication and divergence that must have occurred on a global level. This history is reflected within the microcosm of the single genus Gossypium, which shows evidence of both ancient and recent gene duplication events. Figure 6 shows that the two primary gene lineages, *AdhA*/*B*/*C* and *AdhD*/*E*, diverged from one another near the base of the tree, suggesting an ancient duplication. Other, more recent duplication events are also apparent in Gossypium, *e.g.*, the duplications giving rise to *AdhA*, *AdhB*, and *AdhC* in one lineage and *AdhD* and *AdhE* in the other lineage. Finally, even more recent duplications became apparent with wider sampling of Gossypium species, as noted above.

Previous studies have documented variation in *Adh* copy number and noted that the *Adh* gene tree is not consistent with a simple ancient gene duplication hypothesis (Morton *et al.* 1996; Clegg *et al.* 1997). It has not been clear, however, how often *Adh* copy number has changed in angiosperm evolution, nor the taxonomic scale at which copy number fluctuation has occurred. Although comparable studies are lacking, we have no reason to suspect that Gossypium is unusual with respect to *Adh* gene family evolution, and we suggest that dynamic copy number fluctuation will turn out to be common not only for *Adh* but for many, if not

most, gene families (Morton *et al.* 1996; Clegg *et al.* 1997). To the extent that this is true, it affects our ability to accurately infer orthology relationships among genes from disparate taxa, which has important implications for phylogenetic analyses as well as in studies of functional conservation and diversification. As noted above, plant *Adh* genes are often grouped into *Adh1*-like genes that are expressed under certain developmental conditions, or *Adh2*-like genes that are inducible under hypoxic conditions. If *Adh1* genes are not orthologous (derived from a common *Adh1* gene), this suggests convergent evolution toward both developmentally regulated and inducible members and that this condition has evolved multiple times. Refinements in our understanding of regulation and expression patterns of *Adh* genes in different species should shed light on this issue.

We thank A. Millar, M. Ellis, and E. Dennis of the Commonwealth Scientific and Industrial Research Organization, Australia for providing *G. hirsutum Adh* clones and sequences; J. Ryburn and T. Haselkorn for technical assistance; C. Brubaker for assistance with the genetic mapping; B. Gaut for numerous discussions, providing primers, and suggestions that improved the manuscript; K. Schierenbeck for providing primers; an anonymous reviewer for suggestions that improved the manuscript; and the National Science Foundation for financial support (to J.F.W.).

LITERATURE CITED

Brubaker, C. L., and J. F. Wendel, 1994 Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Am. J. Bot. **81:** 1309–1326.

Brubaker, C. L., A. H. Paterson and J. F. Wendel, 1999 Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome **42:** 184–203.

Chang, C., and E. M. Meyerowitz, 1986 Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA **83:** 1408–1412.

Charlesworth, D., F. Liu and L. Zhang, 1998 The evolution of the alcohol dehydrogenase gene family in plants of the genus *Leavenworthia* (Brassicaceae): loss of introns, and an intronless gene. Mol. Biol. Evol. **15:** 552–559.

Clegg, M. T., M. P. Cummings and M. L. Durbin, 1997 The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. USA **94:** 7791–7798.

Cronn, R. C., X. Zhao, A. H. Paterson and J. F. Wendel, 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, R. C., R. L. Small and J. F. Wendel, 1999 Duplicated genes evolve independently after polyploid formation in cotton. Proc. Natl. Acad. Sci. USA **96:** 14406–14411.

Cummings, M. P., and M. T. Clegg, 1998 Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA **95:** 5637–5642.

Dennis, E. S., W. L. Gerlach, A. J. Pryor, J. L. Bennetzen, A. Inglis *et al.*, 1984 Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. Nucleic Acids Res. **12:** 3983–4000.

Dennis, E. S., M. M. Sachs, W. L. Gerlach, E. J. Finnegan and W. J. Peacock, 1985 Molecular analysis of the alcohol dehydrogenase 2 (*Adh2*) gene of maize. Nucleic Acids Res. **13:** 727–743.

Dolferus, R., M. Ellis, G. D. Bruxelles, B. Trevaskis, F. Hoeren *et al.*, 1997a Strategies of gene action in *Arabidopsis* during hypoxia. Ann. Bot. **79:** 21–31.

Dolferus, R., J. C. Osterman, W. J. Peacock and E. S. Dennis, 1997b Cloning of the Arabidopsis and rice formaldehyde dehydrogenase genes: implications for the origin of plant ADH enzymes. Genetics **146:** 1131–1141.

Downie, S. R., R. G. Olmstead, G. Zurawski, D. E. Soltis, P. S. Soltis *et al.*, 1991 Six independent losses of the choroplast DNA *rpl2* intron in dicotyledons: molecular and phylogenetic implications. Evolution **45:** 1245–1259.

Drouin, G., and M. Moniz de Sá, 1997 Loss of introns in the pollen-specific actin gene subfamily members of potato and tomato. J. Mol. Evol. **45:** 509–513.

Ellstrand, N. C., J. M. Lee and K. W. Foster, 1983 Alcohol dehydrogenase isozymes in grain sorghum (*Sorghum bicolor*): evidence for a gene duplication. Biochem. Genet. **21:** 147–154.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to domestication of maize. Proc. Natl. Acad. Sci. USA **95:** 4441–4446.

Freeling, M., and D. C. Bennett, 1985 Maize *Adh1*. Annu. Rev. Genet. **19:** 297–323.

Frugoli, J. A., M. A. McPeek, T. L. Thomas and C. R. McClung, 1998 Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics **149:** 355–365.

Gaut, B. S., 1998 Molecular clocks and nucleotide substitution rates in higher plants. Evol. Biol. **30:** 93–120.

Gaut, B. S., and M. T. Clegg, 1993 Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA **90:** 5095–5099.

Gaut, B. S., B. R. Morton, B. C. McCaig and M. T. Clegg, 1996 Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. USA **93:** 10274–10279.

Gaut, B. S., A. S. Peek, B. R. Morton and M. T. Clegg, 1999 Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. **16:** 1086–1097.

Gottlieb, L. D., 1982 Conservation and duplication of isozymes in plants. Science **216:** 373–380.

Innan, H., F. Tajima, R. Terauchi and N. T. Miyashita, 1996 Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics **143:** 1761–1770.

Lai, M., J. Sceppa, J. A. Ballenger, J. J. Doyle and R. P. Wunderlin, 1997 Polymorphism for the presence of the *rpl2* intron in chloroplast genomes of *Bauhinia* (Leguminosae). Syst. Bot. **22:** 519–528.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181.

Liu, Q., C. L. Brubaker, A. G. Green, D. R. Marshall, P. Sharp *et al.*, 2000 Evolution of the *FAD2-1* fatty acid desaturase 5′ UTR intron and the molecular systematics of *Gossypium* (Malvaceae). Am. J. Bot. (in press).

Loguercio, L. L., and T. A. Wilkins, 1998 Structural analysis of a hmg-coA-reductase pseudogene: insights into evolutionary processes affecting the *hmgr* gene family in allotetraploid cotton (*Gossypium hirsutum* L.). Curr. Genet. **34:** 241–249.

Millar, A. A., and E. S. Dennis, 1996a The alcohol dehydrogenase genes of cotton. Plant Mol. Biol. **31:** 897–904.

Millar, A. A., and E. S. Dennis, 1996b Protein synthesis during oxygen deprivation in cotton. Aust. J. Plant Physiol. **23:** 341–348.

Millar, A. A., M. R. Olive and E. S. Dennis, 1994 The expression and anaerobic induction of alcohol dehydrogenase in cotton. Biochem. Genet. **32:** 279–300.

Miyashita, N. T., H. Innan and R. Terauchi, 1996 Intra- and interspecific variation of the alcohol dehydrogenase locus region in wild plants *Arabis gemmifera* and *Arabidopsis thaliana*. Mol. Biol. Evol. **13:** 433–436.

Morton, B. R., B. S. Gaut and M. T. Clegg, 1996 Evolution of alcohol dehydrogenase genes in the Palm and Grass families. Proc. Natl. Acad. Sci. USA **93:** 11735–11739.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Osterman, J. C., and E. S. Dennis, 1989 Molecular analysis of the *ADH1-C*^{m} allele of maize. Plant Mol. Biol. **13:** 203–212.

Perry, D. J., and G. R. Furnier, 1996 *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups. Proc. Natl. Acad. Sci. USA **93:** 13020–13023.

Pokorny, R. M., A. B. Dietz, S. Galandiuk and H. L. Neibergs, 1997 Improved resolution of asymmetric-PCR SSCP products. Biotechniques **22:** 606–608.

Reinisch, A. J., J. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel *et al.*, 1994 A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics **138:** 829–847.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Sang, T., M. J. Donoghue and D. Zhang, 1997 Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. **14:** 994–1007.

Seelanan, T., A. Schnabel and J. F. Wendel, 1997 Congruence and consensus in the cotton tribe (Malvaceae). Syst. Bot. **22:** 259–290.

Seelanan, T., C. L. Brubaker, J. M. Stewart, L. A. Craven and J. F. Wendel, 1999 Molecular systematics of Australian *Gossypium* section *Grandicalyx* (Malvaceae). Syst. Bot. **24:** 183–208.

Small, R. L., and J. F. Wendel, 2000 Phylogeny, duplication, and intraspecific variation of *Adh* sequences in New World diploid cottons (*Gossypium*, Malvaceae). Mol. Phylogenet. Evol. (in press).

Small, R. L., J. A. Ryburn, R. C. Cronn, T. Seelanan and J. F. Wendel, 1998 The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogenetic reconstruction in a recently diverged plant group. Am. J. Bot. **85:** 1301–1315.

Small, R. L., J. A. Ryburn and J. F. Wendel, 1999 Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.). Mol. Biol. Evol. **16:** 491–501.

Suiter, K. A., 1988 Genetics of allozyme variation in *Gossypium arboreum* L. and *Gossypium herbaceum* L. (Malvaceae). Theor. Appl. Genet. **75:** 259–271.

Sun, H.-W., and B. V. Plapp, 1992 Progressive sequence alignment and molecular evolution of the Zn-containing alcohol dehydrogenase family. J. Mol. Evol. **34:** 522–535.

Swofford, D. L., 1999 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Tajima, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135:** 599–607.

Trick, M., E. S. Dennis, K. J. R. Edwards and W. J. Peacock, 1988 Molecular analysis of the alcohol dehydrogenase gene family of barley. Plant Mol. Biol. **11:** 147–160.

VanderWiel, P. L., D. F. Voytas and J. F. Wendel, 1993 Copia-like retrotransposable element evolution in diploid and polyploid cotton (*Gossypium* L.). J. Mol. Evol. **36:** 429–447.

Vieira, C. P., J. Vieira and D. Charlesworth, 1999 Evolution of the cycloidea gene family in *Antirrhinum* and *Misopates*. Mol. Biol. Evol. **16:** 1474–1483.

Waters, E. R., 1995 The molecular evolution of the small heat-shock proteins in plants. Genetics **141:** 785–795.

Wendel, J. F., 1989 New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. USA **86:** 4132–4136.

Wendel, J. F., 1995 Cotton, pp. 358–366 in *Evolution of Crop Plants*, edited by N. Simmonds and J. Smartt. Longman, London.

Wendel, J. F., and V. A. Albert, 1992 Phylogenetics of the cotton genus (*Gossypium* L.): character-state weighted parsimony analysis of chloroplast DNA restriction site data and its systematic and biogeographic implications. Syst. Bot. **17:** 115–143.

Wendel, J. F., and A. E. Percival, 1990 Molecular divergence in the Galapagos Islands—Baja California species pair, *Gossypium klotzschianum* and *G. davidsonii* (Malvaceae). Plant Syst. Evol. **171:** 99–115.

Wendel, J. F., C. L. Brubaker and A. E. Percival, 1992 Genetic diversity in Gossypium hirsutum and the origin of upland cotton. Am. J. Bot. **79:** 1291–1310.

Wendel, J. F., A. Schnabel and T. Seelanan, 1995a Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA **92:** 280–284.

Wendel, J. F., A. Schnabel and T. Seelanan, 1995b An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol. Phylogenet. Evol. **4:** 298–313.

Wolfe, K. H., W.-H. Li and P. M. Sharp, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84:** 9054–9058.

Yokoyama, S., and D. E. Harry, 1993 Molecular phylogeny and evolutionary rates of alcohol dehydrogenases in vertebrates and plants. Mol. Biol. Evol. **10:** 1215–1226.

Communicating editor: M. K. Uyenoyama