INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Evolution of the alcohol dehydrogenase gene family in diploid and tetraploid *Gossypium* L.

by

Randall Lee Small

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Major Professor: Jonathan F. Wendel

Iowa State University

Ames, Iowa

1999

Graduate College

Iowa State University

This is to certify that the Doctoral dissertation of

Randall Lee Small

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

Research Objectives

Molecular data have had a profound effect on the field of plant evolutionary biology. These data have been applied profitably to analyses at multiple levels, from population-level studies of genetic diversity to large-scale analyses of plant relationships. These analyses thus far have relied, for the most part, on molecular data from either the chloroplast genome (cpDNA) or from nuclear-encoded ribosomal DNA (rDNA). The reasons behind this bias are primarily practical, in that both cpDNA and rDNA are present in high copy number and represent relatively simple genetic systems often making inferences of homology straightforward. Such inferences are imperative in that one of the critical assumptions of evolutionary analyses are that strictly homologous characters are being analyzed. As an alternative to cpDNA or rDNA, low-copy number nuclear-encoded genes offer a potentially vast array of sequences that could be co-opted for evolutionary studies. The difficulties associated with using low-copy sequences, however, are not insignificant. The primary obstacle is that nuclear genes often exist in multigene families, thereby requiring identification and isolation of absolute orthologues (genes related by speciation) as opposed to paralogues (genes related by duplication).

The goal of the research described herein is to use two well-developed model systems to explore the evolutionary dynamics of a nuclear-encoded gene family and to apply those foundational data in empirical studies of plant phylogeny and genetic diversity. The models I have exploited are the alcohol dehydrogenase (*Adh*) gene family as a model genetic system in selected species of the genus *Gossypium* L. (Malvaceae) as a model organismal system.

Among nuclear-encoded genes in plants, the *Adh* gene family is probably the best characterized from a phylogenetic and molecular evolutionary perspective (Clegg et al. 1997). *Adh* genes have been isolated from a large number of angiosperms and are generally encoded by a small number of loci (typically 2-3; e.g., Dennis et al. 1984, 1985). Gene structure (intron/exon number and position) is relatively conserved (generally 10 exons & 9 introns), although exceptions do exist (e.g., *Arabidopsis thaliana* has 7 exons and 6 introns). ADH enzymes are important metabolic components, especially in a plant's response to hypoxia (Freeling and Bennet 1985). The enzyme converts acetaldehyde to ethanol and in the process regenerates NAD^+ from NADH, thus allowing glycolysis to continue even in the absence of oxygen.

The cotton genus, *Gossypium* L., consists of approximately 50 pantropically distributed species of which the majority are diploid ($2n=2x=26$), but five are allotetraploids ($2n=4x=52$; Wendel 1995). *Gossypium* has been the subject of both evolutionary and genetic study including: cytogenetics (reviewed in Endrizzi et al. 1985); generic-level phylogenetic analyses based on

multiple plastid and nuclear markers (Cronn et al. 1996; Seelanan et al. 1997; Small et al. 1998; Wendel and Albert 1992,); analyses of the origin and relationships among the allotetraploid species (Brubaker et al. 1993; Brubaker and Wendel 1994; Small et al. 1998; Wendel et al. 1992); and RFLP genetic linkage maps of the tetraploid species (Reinisch et al. 1994) and the parental diploid species (Brubaker et al. 1999). This wealth of background information makes *Gossypium* a model system for examining molecular evolution in a well understood phylogenetic context.

This study was undertaken with three primary goals. First, to characterize the *Adh* gene family in representative species of *Gossypium*. These data provide the foundation for the following two goals: exploring the phylogenetic utility of nuclear-encoded genes, and examining rates and patterns of molecular evolution within and among loci and lineages. The three papers that describe original research in this dissertation include a paper that addresses each one of these goals.

Dissertation Organization

The dissertation is organized into five chapters. The introductory chapter provides an overview of the objectives of the research, outlines the model systems used in the research, and briefly reviews the relevant literature. The subsequent three chapters constitute original research papers that are either published, accepted for publication, or prepared for submission for publication. The first of these chapters reports on the organization of the *Adh* gene family in *Gossypium*. The second and third chapters report results from applications of the foundational data to problems in plant phylogenetics and genetic diversity, respectively. The final chapter provides general conclusions drawn from the research as a whole.

Chapter 2, entitled "Organization and evolution of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium* L.)," has been prepared for submission to the journal *Molecular Biology and Evolution*. This foundational paper provides data on the organization of the alcohol dehydrogenase (*Adh*) gene family in *Gossypium* including details on gene family size, gene structure, genetic mapping, and molecular evolutionary patterns.

Chapter 3, entitled "The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group," was published in the *American Journal of Botany* (Small et al. 1998). This paper compares the potential phylogenetic utility of a number of noncoding cpDNA sequences (introns and intergenic spacers) to that of a pair of homoeologous *Adh* loci in the allotetraploid species of *Gossypium*. The results of these analyses clearly show that nuclear-encoded genes such as *Adh* can provide significantly better resolution of phylogenetic problems than cpDNA sequences.

Chapter 4, entitled "Low levels of nucleotide diversity at homoeologous *Adh* loci in allotetraploid cotton (*Gossypium* L.)," was published in the journal *Molecular Biology and Evolution* (Small, Ryburn, and Wendel 1999). This paper reports a study of genetic diversity at the nucleotide level (nucleotide diversity) for a pair of homoeologous *Adh* loci in two allotetraploid cotton species, *Gossypium hirsutum* L. and *Gossypium barbadense* L. We found that nucleotide diversity at these loci is extraordinarily low, lower than any previous reports for plant nuclear-encoded genes. This appears to be due to a combination of factors including a selfing breeding system, repeated genetic bottlenecks, and a low rate of molecular evolution. In addition, we obtained qualitative evidence that the subgenomes of the allotetraploid are under different evolutionary pressures, as evidenced by an increase in nucleotide diversity and heterozygosity in one relative to the other.

Chapter 5 provides a summary of the conclusions reached in this research as well as suggesting additional areas of research that could be pursued.

Literature Review

Adh Evolution in Plants — *Adh* in angiosperms exists in nuclear-encoded, small gene families (Clegg et al. 1997; Gottlieb 1982; Sun and Plapp 1992; Yokoyama and Harry 1993). Estimates of *Adh* copy number have been derived primarily from isozyme studies, which indicate that most angiosperms have two or three expressed loci (Gottlieb 1982), and molecular genetic analyses have often corroborated these estimates (e.g., Dennis et al. 1984, 1985).

While there are numerous publications characterizing *Adh* genes and their expression patterns in a variety of plant species, few describe phylogenetic or molecular evolutionary analyses of sequence variation within or among species. Published studies are primarily from the grass family, especially maize (Gaut and Clegg 1991, 1993a,b; Gaut et al. 1996) or *Arabidopsis thaliana* and related taxa (Hanfstingl et al. 1994; Innan et al. 1996; Miyashita et al. 1996). These studies have highlighted several aspects of *Adh* evolution including variation in evolutionary rates, the importance of recombination in generating allelic diversity, and the prevalence of gene duplications and deletions.

Rates and patterns of evolution of *Adh* loci have been shown to vary, both between orthologous loci (genes related by speciation) and among paralogous loci (genes related by gene duplication). For example, in grasses, *Adh1* in maize is highly polymorphic relative to pearl millet (4.5-fold higher nucleotide diversity) and exhibits a nucleotide substitution rate estimated to be 1.7 times faster (Gaut and Clegg 1993a,b). Patterns have also been shown to differ between paralogous loci within species. Gaut et al. (1996) analyzed sequence data from *Adh1* and *Adh2* of three grasses and found that amino acid replacement rates are accelerated in *Adh2* without a

concomitant acceleration of synonymous rates. Clearly, patterns of molecular evolution for *Adh* genes vary both between closely related taxa at orthologous loci, and between paralogous loci within species.

A second observation derived from previous studies is that recombinational events are a potent force in generating allelic diversity at *Adh* loci. This has been shown in maize (Gaut and Clegg 1993a; Goloubinoff et al. 1993; Hanson et al. 1996) and *Arabidopsis* (Innan et al. 1996).

A final observation is that patterns of genomic change (gene duplication and deletion) are complex and dynamic. Global phylogenetic analyses of *Adh* sequences show that plant *Adh* sequences appear as a unique group distinct from other *Adh* sequences. The observed pattern, however, is not that predicted if all loci were the result of a single ancient duplication (e.g., Clegg et al. 1997; Gaut et al. 1996; Morton et al 1996; Shafqat et al. 1996). This latter assertion is supported by the growing number of plants that have been shown to contain more than two *Adh* loci. For example, most grasses have two loci, but barley, sorghum, and some accessions of maize have three (Trick et al. 1988; Ellestrand et al. 1983; Osterman and Dennis 1989, respectively). The palm, *Washingtonia*, has three loci (Morton et al. 1996), as do some peonies (Paeoniaceae; Sang et al. 1997), whereas diploid *Gossypium* have at least seven loci (Millar and Dennis 1996; Small and Wendel unpub. data).

Adh in *Gossypium* — *Adh* has been well-studied at the isozyme level in *Gossypium*. (Hancock 1982; Millar et al. 1994; Wendel and Percival 1990). These studies have shown that diploid *Gossypium* species have *at least* two expressed *Adh* loci as well as variation in the number of expressed loci among diploids. Preliminary characterization of the *Adh* gene family at the molecular level in *G. hirsutum* has also been published (Millar and Dennis 1996; Millar et al. 1994). Sequence and Southern blot analyses reveal a surprisingly complex pattern given the isozyme-derived hypothesis of a two-locus system in the diploids (and therefore presumably a four-locus system in the allotetraploids). A total of four "classes" of cDNAs were isolated and an additional "class" was isolated from a genomic clone. Southern blot analyses reveal patterns suggestive of at least six to seven loci.

Organismal Context — *Gossypium* L. consists of approximately 50 pantropically distributed species of which the majority are diploid (2N=2X=26), but five are allotetraploids (2N=4X=52; Wendel 1995). *Gossypium* has been the subject of both evolutionary and genetic study including: cytogenetics (reviewed in Endrizzi et al. 1985); generic-level phylogenetic analyses based on multiple plastid and nuclear markers (Seelanan et al. 1997; Wendel and Albert 1992); analyses of the origin and relationships among the allotetraploid species (Brubaker et al. 1993; Brubaker and Wendel 1994; DeJoode and Wendel 1992; Small et al. 1998; Wendel 1989; Wendel and Albert 1992); and RFLP genetic linkage maps of the tetraploid species (Reinisch et

al. 1994) and the parental diploid species (Brubaker et al. 1999). This wealth of background information makes *Gossypium* an model system for examining molecular evolution both among diploid lineages and within the common genome of the allotetraploids.

My analyses will focus on four species of *Gossypium*: the diploids *G. robinsonii* (C-genome Australian cotton as an outgroup for comparative purposes); *G. herbaceum* (African-Asian A-genome diploid, representative of parent of the tetraploid); *G. raimondii* (South American D-genome diploid, representative of parent of the tetraploid); and the allotetraploid *G. hirsutum* (AD-genome). These taxa were chosen specifically because, as described above, their phylogenetic relationships are well-resolved, the genomic relationships among diploid and allotetraploid are understood, and critically, genetic linkage maps have been constructed for the A, D and AD genome taxa, allowing us to clearly assess orthology of loci.

Literature Cited

BRUBAKER, C. L., J. A. KOONTZ, AND J. F. WENDEL. 1993. Bidirectional cytoplasmic and nuclear introgression in the New World cottons, *Gossypium barbadense* and *G. hirsutum* (Malvaceae). Amer. J. Bot. **80**:1203-1208.

BRUBAKER, C. L., AND J. F. WENDEL. 1994. Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Amer. 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 (in press).

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. Journal of Molecular Evolution **42**:685-705.

DEJOODE, D. R., AND J. F. WENDEL. 1992. Genetic diversity and origin of the Hawaiian islands cotton, *Gossypium tomentosum*. American Journal of Botany **79**:1311-1319.

DENNIS, E. S., W. L. GERLACH, A. J. PRYOR, J. L. BENNETSEN, A. INGLIS, D. LLEWELLYN, M. M. SACHS, R. J. FERL AND W. J. PEACOCK. 1984. Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. Nuc. 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. Nuc. Acids Res. **13**:727-743

ELLESTRAND, 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.

ENDRIZZI, J. D., E. L. TURCOTTE, AND R. J. KOHEL. 1985. Genetics, cytology, and evolution of *Gossypium*. Adv. Genet. **23**: 271-375.

FREELING, M., AND D. C. BENNETT. 1985. Maize *Adh1*. Annual Review of Genetics **19**:297-323.

GAUT, B. S., AND M. T. CLEGG. 1991. Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. Proc. Natl. Acad. Sci. USA **88**:2060-2064.

GAUT, B. S., AND M. T. CLEGG. 1993a. Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA **90**:5095-5099.

GAUT, B. S., AND M. T. CLEGG. 1993b. Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics **135**:1091-1097.

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.

GOLOUBINOFF, P., S. PÄÄBO, AND A. C. WILSON. 1993. Evolution of maize inferred from sequence diversity of an *Adh2* gene segment from archaeological specimens. Proc. Natl. Acad. Sci. USA **90**:1997-2001.

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

HANCOCK, J. F. 1982. Alcohol dehydrogenase isozymes in *Gossypium hirsutum* and its putative diploid progenitors: the biochemical consequences of enzyme multiplicity. Pl. Syst. Evol. **140**:141-149.

HANFSTINGL, U., A. BERRY, E. A. KELLOGG, J. T. COSTA III, W. RÜDIGER, AND F. M. AUSUBEL. 1994. Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138**:811-828.

HANSON, M. A., B. S. GAUT, A. O. STEC, S. I. FUERSTENBERG, M. M. GOODMAN, E. H. COE, AND J. F. DOEBLEY. 1996. Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics **143**:1395-1407.

INNAN, H., F. TAJIMA, R. TERAUCHI, AND N. MIYASHITA. 1996. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics **143**:1761-1770.

MILLAR, A. A., AND E. S. DENNIS. 1996. The alcohol dehydrogenase genes of cotton. Pl. Mol. Biol. **31**:897-904.

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 the alcohol dehydrogenase genes in the Palm and Grass families. Proc. Natl. Acad. Sci. USA. **93**:11735-11739.

OSTERMAN, J. C. AND E. S. DENNIS. 1989. Molecular analysis of the *Adh1-C*^{m} allele of maize. Plant Mol. Biol. **13**:203-212.

REINISCH, A. J., J. DONG, C. L. BRUBAKER, D. M. STELLY, J. F. WENDEL AND A. H. PATERSON. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics **138**:829-847.

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. Syst. Bot. **22**:259-290.

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 phylogeny reconstruction in a recently diverged plant group. Amer. 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

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

TRICK, M., E. S. DENNIS, K. J. R. EDWARDS, AND W. J. PEACOCK. 1988. Molecular analysis of the alcohol dehydrogenase gene family of barley. Pl. Mol. Biol. **11**:147-160.

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* N. Simmonds and J. Smartt, eds. Evolution of crop plants. 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*). Pl. 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. Amer. J. Bot. **79**:1291-1310.

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.

CHAPTER 2. ORGANIZATION AND EVOLUTION OF THE ALCOHOL DEHYDROGENASE GENE FAMILY IN DIPLOID AND TETRAPLOID COTTON (*GOSSYPIUM* L.)

A paper to be submitted to the journal *Molecular Biology and Evolution*

Randall L. Small[1] and Jonathan F. Wendel[1]

Abstract

Most plant nuclear-encoded genes exist in gene families of various sizes. 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*). A PCR-based approach was employed to isolate and sequence multiple *Adh* gene family members. We used Southern hybridization analyses to document variation in gene copy number in three diploids which represent the primary centers of *Gossypium* diversity (Australia — *G. robinsonii*, Africa — *G. herbaceum*, New World — *G. raimondii*), as well as one of the five allotetraploid species (*G. hirsutum*). The diploid species of *Gossypium* contain at least seven *Adh* loci in two primary gene lineages. One of these lineages contains two loci that are the result of a local duplication; the other lineage contains at least five loci. Sequence analysis reveals extensive variation in intron lengths between loci, and one locus has lost two introns usually found in other plant *Adh* genes. Evolutionary rate estimates differ between loci and in some cases between organismal lineages at the same locus. Finally, the *Adh* gene family appears relatively active in that multiple examples of apparent gene duplication events were found and at least one case of pseudogenization and one case of gene elimination were also found.

Key Words: *Adh*, cotton, genetic mapping, molecular evolution, gene family, polyploidy

Introduction

Plant nuclear genes are generally part of gene families — multiple genes encoding 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, Cummings, and Durbin 1997) to large families with hundreds of loci (e.g., heat shock proteins, Waters 1995). The

[1]Department of Botany, Iowa State University, Ames, IA 50011.

evolutionary processes that control the structure and dynamics of such gene families are relatively poorly understood (reviewed by Clegg, Cummings, and Durbin 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 1993a), or on a whole gene family in a wide range of species (e.g., *Adh* in eukaryotes, Yokoyama and Harry 1993). Focusing on a single species allows fine-scale dissection of evolutionary forces acting on a given locus at the population level, but lacks the perspective provided by a comparative approach. Analyses of a wide range of species and loci allow for more general conclusions, yet are rarely inclusive enough (in terms of both species and loci) to provide more than a coarse picture of evolutionary dynamics affecting a given gene family. A balanced approach that evaluates a gene family within a well-characterized phylogenetic framework allows for a focused analysis of the evolution of a whole gene family in a defined set of species.

To accomplish this we have employed two model systems: the cotton genus, *Gossypium*, as a model organismal framework, and the alcohol dehydrogenase (*Adh*) gene family as a model low-copy nuclear-encoded gene. *Gossypium* has a number of attributes that make it amenable to molecular evolutionary studies. First, there is a wealth of cytogenetic data for *Gossypium* that have resulted in the division of the species into "genome groups" (A-K, reviewed in Endrizzi, Turcotte, and Kohel 1985; Stewart 1995). Second, *Gossypium* contains both divergent diploid and reticulate allotetraploid species allowing evolutionary analyses at multiple levels. Third, a well-resolved and robustly supported phylogeny based on multiple molecular data sets exists for the whole genus (fig. 1; Wendel and Albert 1992; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998). The phylogenetic analyses correspond well with previous taxonomic (Fryxell 1992) and cytogenetic (Endrizzi, Turcotte, and Kohel 1985) studies. Fourth, genetic maps exist for both the allotetraploid species group (Reinisch et al. 1994) and its parental diploid species groups (Brubaker, Paterson, and Wendel 1999). Finally, a number of previous molecular evolutionary studies have been published within this framework (Cronn et al. 1996; Small et al. 1998, Small, Ryburn, and Wendel 1999; VanderWiel, Voytas, and Wendel 1993; Wendel, Schnabel, and Seelanan 1995a, 1995b).

Adh is among the best studied plant nuclear-encoded gene families, both in terms of molecular biological and molecular evolutionary investigations (reviewed by Clegg, Cummings, and Durbin 1997). *Adh* genes are generally of a convenient size for study (fig. 2; ca. 1100 nucleotides of coding sequence, and whole genes between 2-3 kb in length) with (usually) 10 exons and 9 introns, and generally exist as a relatively small gene family (often only two or three loci). The ADH enzyme is important primarily in response to hypoxic conditions where its expression is highly induced (Dolferus et al. 1997). Additionally, ADH may be important during

seedling development, fruit ripening, and pollen development (Freeling and Bennett 1985; Dolferus et al. 1997). Molecular evolutionary studies of *Adh* loci have been performed in a number of plant species (e.g., maize, Eyre-Walker et al. 1998; barley, Cummings and Clegg 1998; *Arabidopsis*, Innan et al. 1996; *Leavenworthia*, Charlesworth, Liu, and Zhang 1998; cotton, Small, Ryburn, and Wendel 1999; palms, Gaut et al. 1996) and have led to a number of conclusions. As noted above, *Adh* is generally found in small gene families, often with only two to three loci. Phylogenetic analyses of all available plant *Adh* sequences, however, indicate that this is not due to retention of the products of an ancient gene duplication, but apparently to repeated inflation and shrinkage of the gene family in different organismal lineages throughout plant evolution (Clegg, Cummings, and Durbin 1997; Gaut et al. 1996). A second insight gleaned from these analyses is that evolutionary rates may differ dramatically at a number of levels: nonsynonymous vs. synonymous rates (Gaut et al. 1996); absolute rate variation between different lineages (Gaut et al. 1996, Small et al. 1998, Small, Ryburn, and Wendel 1999) and between paralogous loci (Gaut et al. 1996). These differences in evolutionary rates, along with differences in other variables (such as life history traits) have led to vast differences in genetic diversity values among *Adh* loci in different species. In fact, in estimates of nucleotide diversity for plant nuclear-encoded genes with large sample sizes, *Adh* loci hold both the highest and the lowest published values (*Adh1* in *Zea mays* ssp. *parviglumis* θ_W=0.0245, Eyre-Walker et al. 1998; *AdhA* in the A-subgenome of *Gossypium hirsutum* θ_W=0.007, Small, Ryburn, and Wendel 1999, respectively). Clearly then, the gene family a given locus belongs to may not necessarily be predictive of its rate or pattern of molecular evolution.

The purpose of the present paper is to describe the *Adh* gene family of both diploid and allotetraploid species of *Gossypium* with the goal of better understanding the patterns and processes of the evolution of this gene family. Specifically, we wish to address the following questions. (1) How many *Adh* loci are there in *Gossypium* species and how many gene lineages do these loci fall into? (2) What is the structure of these genes? (3) What molecular evolutionary patterns can we detail?

Materials and Methods

Plant Materials: The diploid species of *Gossypium* have been divided into a number of genome groups (A-K; see fig. 1, table 1) based on cytogenetic data, and phylogenetic analyses indicate that these genome groups are monophyletic (Wendel and Albert 1992; Seelanan, Schnabel, and Wendel 1997). These groups 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 species of *Gossypium*, all derived from a single allopolyploidization event that occurred ca. 0.5-2 million years ago (Wendel 1989; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998). The parents of this allopolyploidization event are best represented by the extant species *G. herbaceum* L. (A-genome, African species) and *G. raimondii* (Ulbrich) (D-genome, South American species); thus the allotetraploids are termed the AD-genome and the individual subgenomes of the allotetraploid are referred to as A′ and D′. Relationships among these species groups are depicted in fig. 1.

We have chosen to focus 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, as a diploid outgroup from the Australian C-genome we included *Gossypium robinsonii* (F. Mueller); from the African-Asian A-genome clade we chose *G. herbaceum*; from the New World D-genome clade we included *G. raimondii*; and from the AD-genome allotetraploid species we included *Gossypium hirsutum* L. ("upland cotton"). All species sampled and locations of voucher materials are listed in table 1.

Isolation of *Adh* Sequences: Some baseline information on the *Adh* gene family in *Gossypium* has been published previously. Isozyme surveys have been conducted on a wide range of species (e.g., Wendel and Percival 1990; Wendel, Brubaker, and Percival 1992; Millar, Olive, and Dennis 1994) and suggested that the *Adh* gene family included at least two loci, and in some instances a third (Wendel, unpublished data; Millar, Olive, and Dennis 1994). Hancock (1982) estimated *Adh* isozyme number and performed biochemical analyses on a number of *Gossypium Adh* isozymes. Some molecular genetic analyses of *Adh* have been conducted in *G. hirsutum* (Millar et al. 1994; Millar and Dennis 1996a, 1996b). These analyses focused on a group of loci that were induced by hypoxic conditions and revealed at least five classes of such sequences, termed *Adh1* and *Adh2a-Adh2d* by Millar and Dennis (1996b).

To isolate additional *Adh* sequences we employed a PCR-based approach. We used *Adh* primers P1 and P2 (primers and PCR reaction conditions are described in Small et al. 1998, all PCR primers described are given in table 2) that are homologous to regions of exon 2 and exon 9 (fig. 2) to amplify *Adh* sequences from the *Gossypium* species of interest. These reactions resulted in amplification of multiple *Adh* sequences, as evidenced by multiple bands of sizes ranging from ca. 1.2-1.8 kb detected via agarose gel electrophoresis of the PCR products. To isolate individual PCR products for analysis we cloned this heterogeneous PCR product pool into pGEM-T (Promega) and screened colonies for *Adh* inserts as described (Small et al. 1998).

Based on data generated from the above procedure we were able to design sets of locus-specific PCR amplification primers (all PCR primers are described in table 2). These primer pairs

allowed us to selectively amplify sequences of only one locus at a time which in turn allowed us to sequence PCR products directly.

DNA Sequencing: DNA sequencing was performed in several ways. PCR products were either cloned into pGEM-T (Promega) or sequenced directly. Sequencing was performed either by automated DNA sequencing (ABI prism) at the Iowa State University DNA Sequencing and Synthesis Facility, or using the ^{33}P-labeled dideoxy terminator cycle sequencing kit (Amersham) and electrophoresed on 5-6% Long Ranger gels (FMC).

Southern Hybridization Analyses: Southern hybridizations were performed for two reasons. First, genetic maps for the A- and D-genome diploid species groups (Brubaker, Paterson, and Wendel 1999) and the AD-genome allotetraploid species group (Reinisch et al. 1994) were based on RFLP analyses of segregating F_2 populations. We performed additional RFLP analyses to add the *Adh* loci to these genetic maps. Secondly, we wished to estimate copy number of each of the sequence types isolated. We reasoned that with small (ca. 500 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 organism is homozygous. Heterozygosity, though rarely observed (e.g., Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994; Small, Ryburn, and Wendel 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 to be revealed with most or all enzymes.

Genetic Mapping: All mapping analyses used segregating F_2 populations described by Reinisch et al. (1994), and Brubaker, Paterson, and Wendel (1999). Previously described restriction digested membrane-bound DNAs (Reinisch et al. 1994, Brubaker, Paterson, and Wendel 1999) were probed with locus-specific *Adh* probes. Probes generally consisted of gene fragments representing the intron 3/exon 4 region from the *G. robinsonii* gene for each locus (fig. 2). Nucleotide divergence between most pairs of *Adh* loci is ca. 15-25% in exons and introns are unalignable in most interlocus comparisons (table 3). Preliminary Southern hybridization analyses showed that, under stringent hybridization conditions (65°C, 6X SSC followed by washing at 65°C in 0.1X SSC, 0.5%SDS) these probes do not cross hybridize. These probes were produced by PCR-amplifying the intron 3/exon 4 region from cloned PCR products of each individual locus using primers Fex3 and Bex4-3′ (provided by B. Gaut, University of California, Irvine; see fig. 2). In some cases alternative probes were used including individual intron fragments, or the 3′ UTR of cDNAs (generously provided by A. Millar, M. Ellis, and E. Dennis,

CSIRO, Australia and described in Millar and Dennis 1996b); these probes were produced by restriction digestion of cloned genomic DNA fragments. Probes were radiolabeled via random primer labeling (Gibco-BRL). Hybridization and washing was performed according to Sambrook et al. (1989), except that the 37°C wash was omitted.

In cases where RFLP analysis did not reveal polymorphism we used other techniques to generate segregation data. Specifically, we used PCR-RFLP, SSCP, and length polymorphism of PCR amplified fragments. PCR-RFLP is digestion of PCR products with restriction enzymes that reveal a polymorphism between parental lines, and thus segregates in the F_2 population. SSCP, or single stranded conformational polymorphism was performed as described (Pokorny et al. 1997). Similar to SSCP, we exploited known length differences between PCR products from the two parents by incorporating ^{32}P-dCTP into a PCR reaction and running the products on a 5% Long Ranger acrylamide gel.

Genetic mapping analyses of the F_2 segregation patterns follow Reinisch et al. (1994) and Brubaker, Paterson, and Wendel (1999) using MapMaker version 2.0 (Lander et al. 1987). Mapping data are reported in terms of homoeologous assemblages of Brubaker, Paterson, and Wendel (1999), who compared genetic maps of the AD-genome allotetraploids (*G. hirsutum* x *G. barbadense*) with representatives of its diploid progenitors, the A-genome (*G. herbaceum x G. arboreum*), and the D-genome (*G. trilobum* x *G. raimondii*). Thus each homoeologous assemblage consists of four linkage groups — one from each diploid group and two 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 2-parameter distances as implemented in PAUP* (Sinauer Assoc., Sunderland, MA).

For each individual locus we performed phylogenetic and evolutionary rate analyses on sequences of the four representative species. Phylogenetic analysis (parsimony) was performed for each locus using *G. robinsonii* as the outgroup. In addition, using *G. robinsonii* as the outgroup we performed relative rate tests (Tajima 1993) for all pairs of sequences (A vs. D, A vs. A′, D vs. D′, A′ vs. D′). We also calculated Jukes-Cantor corrected synonymous (*Ksyn*), non-synonymous (*Ka*) substitution rate according to Nei and Gojobori (1986), as well as a silent (*Ksil*; including both synonymous and intron sites) and an intron rate (*Ki*). All such values were calculated as the mean of all pairwise comparisons between ingroup and *G. robinsonii* outgroup sequences. Using previously published (Seelanan, Schnabel, and Wendel 1997) estimates of divergence times for two of the nodes within *Gossypium* (see fig. 1) we also estimated absolute

synonymous rates using only exon data. Many of the above calculations were expedited by the software programs Tajima93 (T. Seelanan, unpublished software), DnaSP (Rozas and Rozas 1997, 1999), and PAUP*.

Results

Initial Characterization of the *Adh* Gene Family: To begin to understand the *Adh* gene family in *Gossypium* we undertook a PCR survey of representative species, *G. robinsonii*, *G. herbaceum*, *G. raimondii*, and *G. hirsutum*. This resulted in amplification of four distinct size classes of PCR products from each species, ranging in size from ca. 1.2 to 1.8 kb as determined by agarose gel electrophoresis. These PCR product pools were cloned and examples from each size class from each species were identified and sequenced. We 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 has been isolated using a separate pair of PCR primers (see below) and has been denoted *AdhE*. Each of these loci was isolated and sequenced from the four representative species and was subjected to copy number estimation and phylogenetic analysis. We also attempted to genetically map each locus. Each locus is individually detailed below.

AdhA: The *Gossypium AdhA* locus is unique among the genes described in this paper in that it lacks two of the introns typically found in plant *Adh* genes, specifically introns four and seven (fig. 3). Those introns that remain are also short relative to other *Gossypium Adh* genes (table 4) making *AdhA* the shortest *Adh* gene in our sample.

Southern hybridization analysis indicates that *AdhA* exists in one copy per diploid genome (fig. 4a), as a single band is observed in all digests of diploids and two bands in the tetraploid. The sole exception to this is the *Eco*RV digest of *G. herbaceum*, which displays two bands (fig. 4a). Using the *AdhA* intron 3/exon 4 probe in Southern hybridization analysis of F_2 populations we were able to genetically map this locus to homoeologous assemblage 8C of Brubaker, Paterson, and Wendel (1999) in both of the diploid populations and in the D-subgenome of the allotetraploid (fig. 5).

Phylogenetic analysis of *AdhA* sequences revealed the expected topology where the sequence from the A-genome diploid is sister to its counterpart from the A-subgenome of the allotetraploid and the sequence from the D-genome diploid is sister to its counterpart from the D-subgenome of the allotetraploid (fig. 6a). Differential evolutionary rates, however, are suggested by the branch lengths. The branch leading to the *G. raimondii* and *G. hirsutum* D-subgenome sequences is 2.5 times longer than the branch leading to *G. herbaceum* and *G.*

hirsutum A-subgenome sequences. Relative rate tests, however, indicate that only the rate difference between the *G. raimondii* and *G. herbaceum* sequences is significant ($P < 0.05$). Nucleotide substitution rates were calculated as the mean of all pairwise comparisons of both diploid (A- and D-genome) and allotetraploid (A- and D-subgenome) sequences vs. the C-genome outgroup sequence. The mean *Ksyn* = 0.0369 while the mean *Ka* = 0.0015 resulting in a *Ksyn*:*Ka* ratio of 25:1; *Ksil* = 0.260 and *Ki* = 0.0198. In addition to relative rates we calculated an absolute synonymous rate for *AdhA* using only exon sequences (Small, Ryburn, and Wendel 1999), using two separate estimated times of divergence (Seelanan, Schnabel, and Wendel 1997). These resulted in an absolute rate estimate of 1.5 - 2.1 x 10^{-9} synonymous substitutions/synonymous site/year.

A previous study (Small, Ryburn, and Wendel 1999) explored levels of genetic diversity at this locus in both subgenomes of the allotetraploid cottons *G. hirsutum* and *G. barbadense*. Accompanying the slow substitution rate we found extraordinarily low levels of nucleotide diversity at these loci. Despite the low levels of diversity observed, preliminary evidence suggested that the D-subgenome harbored greater genetic diversity (both nucleotide and allelic) than the A-subgenome. This is suggestive of differential evolutionary dynamics affecting the two subgenomes. To follow up this observation we are exploring additional loci to determine whether or not this pattern is consistent across loci.

AdhB: The *Gossypium AdhB* locus maintains a ten exon/nine intron structure typical of most angiosperm *Adh* genes (fig. 3), as do all other *Gossypium Adh* genes. Based on phylogenetic analysis (see below) this locus is closely related to the *Adh2* genes reported by Millar and Dennis (1996b).

Southern blots revealed a complex pattern when probed with the *AdhB* intron 3/exon 4 probe (fig. 4b), yet the *AdhB* probe does not cross hybridize to *AdhA*, *AdhC*, *AdhD*, or *AdhE*. Diploid species displayed from two to four bands per digest while the tetraploid displayed up to six hybridizing bands. Sequence alignment of *AdhB* with the *Adh2* genes of Millar and Dennis (1996b) show 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 (fig. 7). In addition to segregating bands observed with the *AdhB* probe, we have mapped *Adh2a* of Millar and Dennis (1996) using the 3′ UTRs of cDNAs. This locus is tightly linked to *AdhB* suggesting the *AdhB/Adh2* gene "subfamily" has evolved via tandem gene duplication.

Phylogenetic analysis of the *AdhB* sequences again resulted in the expected topology (fig. 6b) and the relative rate tests detect no departures from rate homogeneity. As noted above, the *Adh2* sequences of Millar and Dennis (1996b) appear closely related to our *AdhB* sequences based on (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 these sequences in a phylogenetic analysis (fig. 8) reveals that (1) the *Adh2b* sequence of Millar and Dennis (1996b) is probably orthologous to the *AdhB* sequences we report here as it comes out sister to the *AdhB* sequence from the D-subgenome of *G. hirsutum* (fig. 8); and (2) the *Adh2a* and *Adh2d* sequences appear to represent loci that are distinct both from our *AdhB* and from each other based on levels of nucleotide similarity, as also noted by Millar and Dennis (1996b). 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.

Using only the *AdhB* sequences (i.e., not the *Adh2* sequences) we have estimated *Ksyn* = 0.0177 and *Ka* = 0.0045 resulting in a *Ksyn*:*Ka* ratio of 3.9:1; *Ksil* = 0.0217 and *Ki* = 0.0228. The absolute rate for *AdhB* was estimated at 0.6 - 0.7 x 10^{-9} synonymous substitutions/synonymous site/year.

AdhC: Sequence data for *AdhC* have been previously reported in the context of a phylogenetic analysis of the tetraploid species of *Gossypium* (Small et al. 1998). Analysis of *AdhC* highlights the dynamic nature of the *Adh* gene family in *Gossypium*. This locus displays evidence of gene duplication, pseudogenization, and gene loss in various species. Southern hybridization shows that the allotetraploid *G. hirsutum* displays two bands per digest as expected. On the other hand, the D-genome diploid, *G. raimondii*, displays three bands per digest indicative of gene duplication(s), whereas the A-genome diploid, *G. herbaceum*, does not hybridize at all to the *AdhC* probe. 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 six 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 (fig. 9). Because *AdhC* is missing from *G. herbaceum*, it was mapped as a dominant marker in the *G. herbaceum* x *G. arboreum* mapping population.

Phylogenetic analysis of these sequences results in the expected topology (fig. 6c) and also reveals the rate heterogeneity previously described (Small et al. 1998). This 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. Due to the rate heterogeneity we report absolute rates calculated for the A- and D-genome lineages separately. In each case we calculate the absolute rate as the mean of the comparisons of the C-genome to the diploid and the related subgenome of the allotetraploid. For the A-genome lineage *Ksyn* = 0.0230 and *Ka* = 0.0108 for a *Ksyn*:*Ka* ratio of 2.1:1; *Ksil* = 0.0356 and *Ki* = 0.0387. For the D-genome lineage *Ksyn* = 0.0511 and *Ka* = 0.0137 for a *Ksyn*:*Ka* ratio of 3.7:1; *Ksil* = 0.0611 and *Ki* = 0.0586. We estimated absolute synonymous substitution rates of 0.9×10^{-9} synonymous substitutions/synonymous site/year for the A-genome lineage and 2.1×10^{-9} synonymous substitutions/synonymous site/year for the D-genome lineage; thus the D-genome lineage appears to be evolving over twice as fast as the A-genome lineage.

AdhD: The *AdhD* gene is the largest of the *Gossypium Adh* genes reported here, owing primarily to the length of introns three and five (fig. 3). Sequence data from this locus have been used in a phylogenetic analysis of a group of Australian cottons (Seelanan et al. 1999). Based on phylogenetic analysis (see below) this locus is probably orthologous to the *Adh1* sequence reported by Millar and Dennis (1996b).

Southern hybridization analysis using an intron 3/exon 4 probe revealed strong hybridization to a single band in the diploid species, and two bands in the allotetraploid species, but also showed weaker hybridization to additional band(s) in some digests. This suggested that an additional locus closely related to *AdhD* was present in the *Gossypium* genome. This suspicion was subsequently confirmed. For the phylogenetic study of Seelanan et al. (1999) PCR primers were produced that were intended to be locus-specific for *AdhD*; these primers were homologous to regions in exons two and eight. 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. It is important, however, to note that *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 dramatically in size due to length differences primarily in introns three and five. Due to a lack of polymorphism at the RFLP level for *AdhD* we were able to map this locus only by using single-stranded conformational polymorphism (SSCP) techniques where we can readily distinguish *AdhD* from *AdhE* by size. This allowed us to map *AdhD* in the D-diploid mapping population where it mapped to Chromosome D7 (fig. 10). It is interesting to note, however, that *AdhD* and *AdhE* (see below) map to positions very close to each other on

this linkage group (2.2 cM; one recombinant between them out of 62 F_2 progeny) suggesting that these loci are the result of a local duplication.

Phylogenetic analysis of these sequences result in the expected topology (fig. 6d) and displays near-equality of branch lengths in the two clades. Inclusion of the *Adh1* cDNA sequence of Millar and Dennis (1996b) 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* (fig. 6d). This is bolstered by Southern hybridization analysis using the 3′ UTR of the *Adh1* cDNA as a probe (fig. 4d). 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 *AdhE* that they do not cross-hybridize. Thus, we can identify the *AdhE* bands by subtraction (see below).

Using the *AdhD* sequences we have estimated *Ksyn* = 0.0397 and *Ka* = 0.0095 resulting in a *Ksyn:Ka* ratio of 4.2:1; *Ksil* = 0.285 and *Ki* = 0.0266. The absolute rate for this locus was estimated at 1.7 - 1.8 x 10^{-9} synonymous substitutions/synonymous site/year.

AdhE: This locus was isolated using PCR primers homologous to regions in exons two and eight (see above); thus the genomic sequence available for this locus is shorter than for the other *Gossypium Adh* loci we isolated. For the sequence that is available, this locus appears to maintain the characteristic 10 exon/9 intron structure (fig. 3). These amplifications produced *AdhE* PCR products from the D-genome diploid, and from both subgenomes of the allotetraploid, but no amplification of *AdhE* from either of the A-genome diploid species.

As noted above, the 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 using separately: an *AdhD* intron 3/exon 4 probe; an *AdhE* exon 5/intron 5 probe, an *AdhD* intron 3 probe, and an *Adh1* (= *AdhD*) cDNA 3′ UTR probe. The *AdhD* intron 3/exon 4 probe and the *AdhE* exon 5/intron 5 probe revealed identical hybridization patterns with one to two hybridizing bands in the diploids and two to four hybridizing bands in the allotetraploid. Use of the *AdhD* intron 3 and *Adh1* 3′ UTR also revealed hybridization patterns identical to each other, and which were a subset of the fragments revealed with the exon + intron probes: a single fragment in each diploid and two to three in the allotetraploid. Presumably those bands that hybridize to both the *AdhD* and *AdhE* exon + intron probes as well as the *AdhD* intron 3 probe and the *Adh1* 3′ UTR represent *AdhD*, while those bands that hybridize only to the *AdhD* and *AdhE* exon + intron probes represent *AdhE*. These data also indicate that, despite our inability to PCR-amplify *AdhE* from an extant A-genome taxon, it does exist, at least in *G. herbaceum*.

One of these sets of *AdhE* bands was polymorphic in the parents of the D-genome diploid mapping population and we were able to map *AdhE* to Chromosome 7, tightly linked to *AdhD*.

Other than the inability to recover an *AdhE* sequence from an A-genome diploid species, phylogenetic analysis reveals the expected topology (fig. 6e). Estimates for relative substitution rates were *Ksyn* = 0.0295, and *Ka* = 0.0095 for a ratio of 3.1:1; *Ksil* = 0.0323 and *Ki* = 0.0330. The absolute rate was calculated at 1.1×10^{-9} synonymous substitutions/synonymous site/year.

Discussion

***Adh* Gene Family Evolution:** Early work on angiosperm nuclear gene families such as *Adh* suggested that many genes appeared to be encoded by a relatively small number of loci, often only two to three (e.g., Gottlieb 1982). One explanation for this observation is that extant gene copies are the result of a gene duplication that occurred prior to the origin of angiosperms (Gottlieb 1982; Morton, Gaut, and Clegg 1996). Such a scenario makes certain predictions about the relationships among extant copies of the genes: first, it predicts that all plant *Adh* genes should fall into one of two groups, corresponding to the two products of the ancient gene duplication; secondly it predicts that the relationships among the organisms that bear the copies of the genes should be similar in each of the two clades and that these relationships should reflect the organismal relationships. This scenario is perpetuated by the use of terms such as "*Adh1*," "*Adh2*," etc., that suggest, intentionally or unintentionally, that all *Adh1* genes are more closely related to each other than any are to *Adh2* genes. This unjustified assumption appears to be responsible, at least in part, for the use of the term *Adh1* to refer to genes expressed early during development and at low levels throughout the plant, while genes called *Adh2* are often only expressed when induced by hypoxia (or more likely they are so called by their discoverers because they display such tendencies). Recent work (e.g., Morton, Gaut, and Clegg 1996; Clegg, Cummings, and Durbin 1997; see below), however, has shown that the *Adh* gene family appears to be dynamic in the number of genes that exist in a given organism, and that its history has been characterized by many gene duplication and deletion events.

Phylogenetic analysis of plant *Adh* sequences available from GenBank combined with data presented in this paper results in the topology shown in fig. 11. Several noteworthy 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 that are separated by a number of other groups (fig. 11). A similar history is evident for the sequences

from the Rosaceae (*Fragaria*, *Malus*, and *Pyrus Adh4* vs. *Pyrus Adh3*). More recent gene duplications are also evident in the tree (fig. 11). 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, suggesting that a recent gene duplication is responsible for this arrangement. Similar results have been obtained for peonies (fig. 11) 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, Donoghue, and Zhang 1997).

A global phylogenetic analysis of plant *Adh* genes indicates a history of gene duplication and divergence on a global level. Such a history is also evident within the microcosm of the single genus *Gossypium* which shows evidence of both ancient and recent gene duplication events. For example, figure 11 shows that *Gossypium Adh* sequences are found in two primary gene lineages: *AdhA/B/C* and *AdhD/E*. The split between these lineages goes almost to the base of the tree, suggesting that this split was quite ancient. 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 are also apparent. Southern hybridization and genetic mapping evidence as well as phylogenetic analysis indicate that the *AdhB/Adh2* group of sequences are closely related and presumably the result of recent tandem gene duplication events. Southern hybridization data also suggest that the *AdhC* gene has become duplicated (or perhaps triplicated) in *G. raimondii* (fig. 4).

The significance of these observations is that *Adh* gene family evolution in plants is an ongoing and dynamic process with both gene duplication and gene deletion occurring at multiple levels within the phylogeny of angiosperms. This is important not only for our understanding of gene evolution, but also for our understanding of gene function. As noted above, plant *Adh* genes are often grouped into *Adh1*-like genes which are expressed under certain developmental conditions, or *Adh2*-like genes that are inducible under hypoxic conditions. If these generalizations are true, yet all *Adh1* genes are not orthologous (derived from a common *Adh1* gene) this suggests that there has been convergent evolution toward an *Adh* gene family that has 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 hypothesis.

Gene Family Size: 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 these estimates may reflect small gene family size, or alternatively, a lack 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 (Wendel, unpublished data; Millar, Olive, and Dennis 1994). The molecular genetic analysis of Millar and Dennis (1996) documented five distinct loci. The present study indicates that there are at least seven *Adh* loci in *Gossypium*.

Variation in gene number from other species has been documented; for example, three loci have been reported from a number of species (e.g., *Hordeum* — Trick et al. 1988; *Sorghum* — Ellestrand, Lee, and Foster 1983; maize, Osterman and Dennis 1989; some palms, Morton, Gaut, and Clegg 1996; some *Paeonia* species — Sang, Donoghue, and Zhang 1997; *Leavenworthia* — Charlesworth, Liu, and Zhang 1998). Other species, notably some members of the Brassicaceae (*Arabidopsis*, *Arabis*, Chang and Meyerowitz 1985; Miyashita, Innan, and Terauchi 1996), have but a single *Adh* locus. The largest plant *Adh* gene family yet reported is from a gymnosperm, *Pinus banksiana*, which contains at least seven expressed *Adh* loci (Perry and Furnier 1996). Taken in the context of the phylogenetic analysis discussed above, these data reinforce the dynamic nature of *Adh* gene family evolution. *Gossypium* contains the largest *Adh* gene family yet described in an angiosperms with at least seven genes, and equals the largest 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 several-fold in anaerobically induced cotton plants (Millar, Olive, and Dennis 1994; Millar and Dennis 1996a,b).

Interlocus Comparisons of Evolutionary Dynamics: One of the advantages of studying a small gene family in a phylogenetically well-understood and closely related group of species is that a number of intra- and interlocus comparisons may be drawn regarding processes and patterns of evolution. Specifically, for *Adh* in *Gossypium*, we note that there is variation degrees of sequence variation among loci for exons; variation in intron presence and degree of intron sequence divergence between loci; variation in evolutionary rates, both between loci and between lineages for some loci; and finally variation in gene copy number. Each of these is discussed below.

Exon Variation — Table 3 presents a comparison of genetic 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 system *Adh* loci: maize *Adh1*, maize *Adh2*, and *Arabidopsis thaliana Adh*. These data reflect the phylogenetic relationships among the sequences in that *Gossypium AdhA*, *AdhB*, and *AdhC* are all more similar to each other than any of them are *AdhD* or *AdhE* and vice versa. All nucleotide sequence percent

similarities fall within a relatively small range from 68.5% (*Gossypium AdhC* vs. maize *Adh2*) to 93.4% (*Gossypium AdhD* vs. *AdhE*). Amino acid percent identities cover a similarly small range from 76.7% (*Gossypium AdhC* vs. maize *Adh2*) to 92.7% (*Gossypium AdhD* vs. *AdhE*).

Intron Variation — The majority of plant *Adh* sequences have a ten exon/nine intron structure (fig. 2), with introns of various size and sequence found at identical sites within the gene. 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), and several cases of missing introns have been reported in *Adh* genes including from species of the Brassicaceae (*Arabidopsis* — Chang and Meyerowitz 1986; *Arabis* — Miyashita et al. 1996; *Leavenworthia* — Charlesworth, Liu, and Zhang 1998), and 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.

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 (fig. 3). The introns absent from this gene are those between exons 4 and 5 and exons 7 and 8. It is curious to note that these are two of the three introns that are missing from the Brassicaceae *Adh* genes and that phylogenetic analysis clearly shows that this shared loss is not due to inheritance of an intronless gene from a common ancestor (fig. 11). This situation may be analogous to repeated loss of introns from chloroplast genes (e.g., Downie et al. 1991; Lai et al. 1997).

Intron sequence divergence between loci is presumably a measure of evolutionary divergence between loci, but may also reflect the proximity of the loci to each other (dispersed vs. tandem) and therefore the possibility for interlocus interactions. In most comparisons between *Gossypium Adh* loci the intron sequences are unalignable and intron lengths differ (table 4). There are two exceptions to this: *AdhB/Adh2*, and *AdhD/AdhE* sequences.

The *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 all map either to identical sites or very close to each other. This suggests that they are the result of recent tandem gene duplication events. Millar and Dennis (1996b) 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 unequal crossing over to occur.

The *AdhD/AdhE* genes are also tandemly arranged, at least in the single genome in which they have both been mapped. Comparison of these two loci reveal that the intron sequences are alignable throughout most of the gene, although large indels are present in introns 3 and 5.

Rate Variation — Average absolute evolutionary rate values for plant nuclear genes have been examined by several authors (e.g., Wolfe et al. 1987, 1989; Gaut 1998) and range from a low of ca. 1.5 x 10^{-9} synonymous substitutions/synonymous site/year (Small, Ryburn, and Wendel 1999) to 30 x 10^{-9} synonymous substitutions/synonymous site/year (Wolfe et al. 1987), although this upper value probably reflects a paralogous comparison, and so is inflated. A mean rate based on a comparison of nine nuclear genes in rice and maize has been calculate at 6.0 x 10^{-9} synonymous substitutions/synonymous site/year (Gaut 1998). Clearly then, evolutionary rates vary among nuclear-encoded genes. This variation is apparent not only when comparing different organisms (e.g., palms evolve more slowly than grasses across loci, Gaut et al. 1996), but also when comparing different genes within a common organismal framework (e.g., *Adh2* has a faster nonsynonymous rate than *Adh1* in grasses, Gaut et al. 1996). Our results from *Adh* in *Gossypium* show rate variation both between loci and between lineages.

Rate variation between loci is evident from comparisons of both absolute and relative rates. First, using a common pair of calibration points (see fig. 1) we estimated absolute synonymous substitution rates for all five loci. These estimates range from ca. 1.0 x 10^{-9} (*AdhA*) to 2.7 x 10^{-9} (*AdhC*) synonymous substitutions/synonymous site/year, an almost 3-fold difference among loci. Such variation was also noted by Gaut (1998) in a comparison of nine nuclear genes between rice and maize from which he calculated an average rate of 6.0 x 10^{-9}; it is interesting to note that while similar levels of synonymous rate variation were observed (2.7-fold difference in *Gossypium*, 2.4-fold in grasses — Gaut 1998), the rates in *Gossypium* are much lower. Rate variation among loci is also apparent when comparing synonymous (*Ksyn*) and nonsynonymous (*Ka*) relative rates. 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. All *Ksyn* and *Ka* values are reported in table 5. Average synonymous rates per locus ranged from *Ksyn* = 0.0177 (*AdhB*) to *Ksyn* = 0.0397 (*AdhD*), a 2.2-fold difference. Average nonsynonymous rates ranged from *Ka* = 0.0020 (*AdhA*) to *Ka* = 0.0122 (*AdhC*), a 6.1-fold difference. These observations are again consistent with those of Gaut (1998) who noted that in the comparison of nine nuclear genes in rice and maize the synonymous rate varied only 2.4-fold, while the nonsynonymous rate varied over 10-fold.

Nucleotide substitution rate variation is also apparent between lineages for two *Adh* loci. Inspection of the phylogenetic trees constructed for these loci (fig. 6a-e) reveal apparent rate heterogeneity among sequences of *AdhA* (fig. 6a) and *AdhC* (fig. 6c). Application of the Tajima (1993) relative rate tests statistically support these observations in both cases. For *AdhA*, statistically significant rate heterogeneity is detected only between *G. raimondii* and *G.*

herbaceum with *G. raimondii* evolving at a faster rate (Small, Ryburn, and Wendel 1999). These data suggest that the clades to which these species belong are also evolving at different rates, despite the lack of statistical support. For *AdhC*, in all comparisons the D-(sub)genomes are accumulating nucleotide substitutions at a statistically significantly higher rate than the A-(sub)genomes (Small et al. 1998). Rate variation among lineages may be fueled by a number of phenomena including, for example, the generation-time effect, fidelity of DNA polymerases or repair enzymes, and selection (reviewed in Gaut 1998). It is unclear what processes have resulted in rate variation among *Gossypium* lineages, although it is provocative to note that in both cases of statistically supportable rate variation the D-(sub)genome lineage had an accelerated rate relative to the A-(sub)genome lineage. This is accompanied by increased nucleotide polymorphism in the D-subgenome of the allotetraploids *G. hirsutum* and *G. barbadense* for both *AdhA* and *AdhC* (Small, Ryburn, and Wendel 1999; unpublished data). The sum of these observations suggest that the D-(sub)genome lineage may be subject to different and accelerated evolutionary pressures relative to the A-(sub)genome. Further research is necessary to evaluate the generality of this observation and to address its underlying mechanism(s).

Copy Number Variation — The *Adh* loci described herein vary in their relative divergence from other loci, and in the number of genes observed in different species. For example, initial Southern hybridization analysis of *AdhA* indicated that it was single copy per diploid genome in all species sampled (fig. 4a). In the course of a phylogenetic study of a group of New World diploid species, however, we obtained evidence suggestive of a gene duplication confined to four of these species (unpublished data). Southern hybridization of an *AdhB* fragment revealed a number (2-4) of hybridizing fragments in all diploid genomes suggestive of a number of closely related loci. The *AdhB* loci also closely matched the sequences of *Adh2* genes described from *G. hirsutum* (Millar and Dennis 1996b) which had also been suggested to be in relatively high copy number. Unlike most interlocus comparisons, we were able to align even the intron sequences between *AdhB* and *Adh2*. Phylogenetic analysis of these sequences suggest a minimum of three *AdhB/Adh2*-like loci, with a fourth (their *Adh2c*) suggested by the work of Millar and Dennis (1996b). Mapping data indicate that these loci are all tightly linked and are probably the result of local gene duplications.

AdhC reveals in a microcosm many of the phenomena impacting *Adh* evolution on a global scale in plants. For *AdhC* we have evidence of gene duplication, pseudogenization, and deletion, all in different species. Southern hybridization analysis of *AdhC* (fig. 4c) revealed three hybridizing bands in the D-genome species, *G. raimondii*, suggesting gene duplication(s). This same figure shows that *AdhC* does not hybridize to anything 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 (see 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 six 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 contain what appear to be fully intact *AdhC* sequences (Small et al. 1998) indicating that the pseudogenization and loss of *AdhC* from *G. arboreum* and *G. herbaceum* occurred after the split of these species from the species that were involved in the origin of the allotetraploid species. 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.

While *AdhD* and *AdhE* cross-hybridize at the Southern level, they each appear to be represented by a single locus per diploid genome, although tightly linked to each other. Using an *AdhD* intron 3 probe Seelanan et al. (1999) have shown that *AdhD* is single copy in a number of wild diploid Australian *Gossypium* species. This same probe also reveals a single hybridizing band per diploid genome in the four species we sampled (data not shown). We have been unable to design an equivalent *AdhE*-specific probe, but have inferred based on subtraction of *AdhD* fragments from an *AdhD/AdhE* hybridization profile that *AdhE* is also single copy.

While the *Adh* gene family in angiosperms often seems to be stable in terms of copy number (Clegg, Cummings, and Durbin 1997), a detailed analysis of the whole gene family in a group of closely related species reveals that dynamic fluctuations in gene copy number are occurring. 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.

Conclusions: The study presented here was designed to elucidate the structure, organization, and evolution of the *Adh* gene family in the genus *Gossypium*. These data were produced for two primary reasons: (1) to facilitate an understanding of the patterns and processes of gene family evolution within a well-understood phylogenetic framework so that the inferences drawn from this work can be generalized to other species and gene families; and (2) to provide the foundation for more detailed studies of phylogeny, nucleotide diversity, and molecular evolution (Small et al. 1998; Small, Ryburn, and Wendel 1999). The data summarized here provide insights into gene family evolution. Specifically, as more species are sampled, it is becoming apparent that gene

family size is more variable than initially predicted, with *Gossypium* representing the largest *Adh* gene family yet described in an angiosperm. Secondly it is also apparent that the process of gene birth and death are also more dynamic than previously thought, with three of the *Adh* loci showing evidence of recent gene duplications as well as evidence for pseudogenization and gene loss. Finally, we provide evidence of evolutionary rate variation among *Adh* loci as well as among lineages. The absolute synonymous substitution rates we calculated are slower than published average nuclear gene rates (Wolfe, Li, and Sharp 1987; Wolfe, Sharp, and Li 1989; Gaut 1998). These data and analyses provide new insights as well as additional avenues in need of research.

Acknowledgments

We thank A. Millar, M. Ellis, and E. Dennis for providing us with *Gossypium hirsutum Adh* clones and sequences; Julie Ryburn for technical assistance; and the National Science Foundation for financial support (to JFW).

Literature Cited

BRUBAKER, C. L., A. H. PATERSON, AND J. F. WENDEL. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome (in press).

CHANG, C., AND E. 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.-L. LIU, AND L. ZHANG. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). 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. Journal of Molecular Evolution **42**:685-705.

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. BENNETSEN, A. INGLIS, D. LLEWELLYN, M. M. SACHS, R. J. FERL AND W. J. PEACOCK. 1984. Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. Nuc. 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. Nuc. Acids Res. **13**:727-743

DOLFERUS, R., M. ELLIS, G. DE BRUXELLES, B. TREVASKIS, F. HOEREN, E. S. DENNIS. AND W. J. PEACOCK. 1997. Strategies of gene action in *Arabidopsis* during hypoxia. Ann. Bot. **79**:21-31.

DOWNIE, S. R., R. G. OLMSTEAD, G. ZURAWSKI, D. E. SOLTIS, P. S. SOLTIS, J. C. WATSON, AND J. D. PALMER. 1991. Six independent losses of the chloroplast 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.

ELLESTRAND, 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.

ENDRIZZI, J. D., E. L. TURCOTTE, AND R. J. KOHEL. 1985. Genetics, cytology, and evolution of *Gossypium*. Adv. Genet. **23**: 271-375.

EYRE-WALKER, A., R. L. GAUT, H. HILTON, D. L. FELDMAN, AND B. S. GAUT. 1998. Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95**:4441-4446.

FREELING, M., AND D. C. BENNETT. 1985. Maize *Adh1*. Annual Review of Genetics **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.

FRYXELL, P. A. 1992. A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). Rheedea **2**:108-165.

FU, Y.-X., AND W.-H. LI. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693-709.

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.

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

HANCOCK, J. F. 1982. Alcohol dehydrogenase isozymes in *Gossypium hirsutum* and its putative diploid progenitors: the biochemical consequences of enzyme multiplicity. Pl. Syst. Evol. **140**:141-149.

INNAN, H., F. TAJIMA, R. TERAUCHI, AND N. 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.

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. Protein synthesis during oxygen deprivation in cotton. Aust. J. Plant Physiol. **23**:341-348.

MILLAR, A. A., AND E. S. DENNIS. 1996b. The alcohol dehydrogenase genes of cotton. Pl. Mol. Biol. **31**:897-904.

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.

MONIZ DE SÁ, M., AND G. DROUIN. 1996. Phylogeny and substitution rates of angiosperm actin genes. Molecular Biology and Evolution **13**:1198-1212.

MORTON, B. R., B. S. GAUT AND M. T. CLEGG. 1996. Evolution of the 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. Molecular Biology and Evolution 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.

PATERSON, A. H., C. L. BRUBAKER, AND J. F. WENDEL. 1993. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Molecular Biology Reporter **11**:122-127.

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 AND A. H. PATERSON. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics **138**:829-847.

ROZAS, J., AND R. ROZAS. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. **13**:307-311.

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. Syst. Bot. **22**:259-290.

SEELANAN, T., J. F. WENDEL, C. L. BRUBAKER, J. MCD. STEWART, AND L. A. CRAVEN. 1999. Molecular systematics of Australian *Gossypium* L. section *Grandicalyx* (Malvaceae). Syst. Bot. (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 phylogeny reconstruction in a recently diverged plant group. Amer. 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

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

STEWART, J. MCD. 1995. Potential for crop improvement with exotic germplasm and genetic engineering. Pp. 313-327 *in* G. A. Constable and N. W. Forrester, eds. Challenging the future: proceedings of the world cotton research conference-1. CSIRO, Melbourne.

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. Pl. 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.

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* N. Simmonds and J. Smartt, eds. Evolution of crop plants. 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*). Pl. 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. Amer. J. Bot. **79**:1291-1310.

WENDEL, J. F., A. SCHNABEL, AND T. SEELANAN. 1995a. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol. Phyl. Evol. **4**:298-313.

WENDEL, J. F., A. SCHNABEL, AND T. SEELANAN. 1995b. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA **92**:280-284.

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.

WOLFE, K. H., P. M. SHARP, AND W.-H. LI. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. **29**:208-211.

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.

Table 1. Plant materials. All voucher specimens are deposited at the Iowa State University Ada Hayden Herbarium (ISC). Voucher abbreviations are as follows: TS = Tosak Seelanan, JFW & TDC = J. F. Wendel and T. D. Couch

Taxon	Accession	Voucher
C-genome diploid		
Gossypium robinsonii F. Mueller	AZ-50	TS 12
D-genome diploid		
Gossypium raimondii Ulbrich	#436	JFW & TDC 127
A-genome diploids		
Gossypium herbaceum L.	A_1-73	JFW 539
Gossypium arboreum L.	A_2-74	JFW & TDC 312
AD-genome tetraploid		
Gossypium hirsutum L.	"Palmeri"	JFW & TDC 632

Table 2. PCR amplification and sequencing primers used in this study.

Locus	Primer[a]
Amplification primers	
Adh (all loci)	P1: CTG CKG TKG CAT GGG ARG CAG GGA AGC C (f)
	P2: GCA CAG CCA CAC CCC AAC CCT G (r)
AdhA	Adhx2-1: CTT CAC TGC TTT ATG TCA CAC T (f)
	Adhx8-1: GGA CGC TCC CTG TAC TCC (r)
AdhD/E	Adhx2-2: GCA ATG GAG GTT CGT CTG (f)
	Adhx8-3: GAT CAT GGC ATT AAT GTT TC (r)
Sequencing primers	
AdhA/B/C	Adhx4-1: TCA TGT TCT CCC TAT CTT CAC (f)
AdhC	Adhx4-2: GTG GAG AGT GTA GGT GAA GG (f)
	Adhx4-3: GGG CAG ACT AGG TTT TCC AAA G (f)
	Adhx6-2: TCA ATA CCA ATG ATC CTA GAA (r)
	Adhx8-2: GAA ACC ATG GCC TGG GTG (r)
AdhD/E	Adhx2-2: GCA ATG GAG GTT CGT CTG (f)
	Adhx3-1: ACT CCA TTA TTT CCT CGT AT (f)
	Adhx4-4: ACC TCA CCC ACA CTC TCA AC (r)
	Adhx5-1: GCC ACA GTT GAA CCT TTG (r)
	Adhx5-2: AAT AAT TTT CGA GGT CTT GG (f)
	Adhx6-1: ATC AAC ACC AAT AAT CCT AGA A (r)

[a]Primers all written 5′ to 3′; forward primers denoted (f), reverse primers denoted (r).

Table 3. Comparison of percent identity among *Gossypium*, maize, and *Arabidopsis Adh* sequences. Percent nucleotide identity is below the diagonal; percent amino acid identity is above the diagonal.

	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. maize *Adh1*[b]	74.6	—	87.1	80.1	83.8	80.5	89.3	88.4
3. maize *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	—

[a]GenBank accession X77943
[b]GenBank accession X00580
[c]GenBank accession X01965

Table 4. Intron sizes comparison between loci; given in base pairs (bp) for aligned sequences.

Intron Number	2	3	4	5	6	7	8	Aligned Length
AdhA	80	75	absent	85	99	absent	81	1,218 bp
AdhB	106	92	121	103	110	144	84	1,554 bp
AdhC	157	180	81	86	99	184	71	1,653 bp
AdhD	104	259	89	279	92	88	116	1,823 bp
AdhE	98	81	92	208	96	92	NA	1,362 bp

Table 5. Comparison of patterns of nucleotide substitution within and among loci and lineages.

Locus[a]	*Ki*[b]	*KsiI*[c]	*Ksyn*[d]	*Ka*[e]	*Ksyn:Ka* ratio
AdhA A′	0.0126	0.0190	0.0320	0.0010	32.0:1
AdhA D′	0.0270	0.0330	0.0418	0.0020	20.9:1
AdhA	0.0198	0.0260	0.0369	0.0015	24.6:1
AdhB A′	0.0265	0.0226	0.0081	0.0050	1.6:1
AdhB D′	0.0192	0.0209	0.0274	0.0041	6.7:1
AdhB	0.0228	0.0217	0.0177	0.0045	3.9:1
AdhC A′	0.0387	0.0356	0.0230	0.0108[f]	2.1:1
AdhC D′	0.0586	0.0611	0.0511	0.0137	3.7:1
AdhC	0.0512	0.0483	0.0371	0.0122	3.0:1
AdhD A′	0.0261	0.0285	0.0431	0.0095	4.5:1
AdhD D′	0.0272	0.0285	0.0364	0.0095	3.8:1
AdhD	0.0266	0.0285	0.0397	0.0095	4.2:1
AdhE A′[g]	0.0287	0.0284	0.0271	0.0066	4.1:1
AdhE D′	0.0352	0.0343	0.0307	0.0110	2.8:1
AdhE	0.0330	0.0323	0.0295	0.0095	3.1:1

[a]For each *Adh* locus the data are presented for three separate comparisons: (1) as the mean of all pairwise comparisons between the C-genome outgroup (*G. robinsonii*) and the A-genome diploid (*G. herbaceum* or *G. arboreum*) and the A-subgenome of the allotetraploid (*G. hirsutum*) (denoted A′); (2) as the mean of all pairwise comparisons between the C-genome outgroup (*G. robinsonii*) and the D-genome diploid (*G. raimondii*) and the D-subgenome of the allotetraploid (*G. hirsutum*) (denoted D′); and (3) as the mean of all pairwise comparisons between the C-genome outgroup and sequences of both A- and D-genome diploids and both subgenomes of the allotetraploid.

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

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

[d]Number of synonymous substitutions per synonymous site in coding sequences; calculated via the method of Nei and Gojobori (1986).

[e]Number of nonsynonymous substitutions per nonsynonymous site in coding sequences; calculated via the method of Nei and Gojobori (1986).

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

[g]Because the A-genome diploid sequence for *AdhE* is not available, these values represent a comparison of *G. robinsonii* and the A-subgenome sequence of *G. hirsutum*.

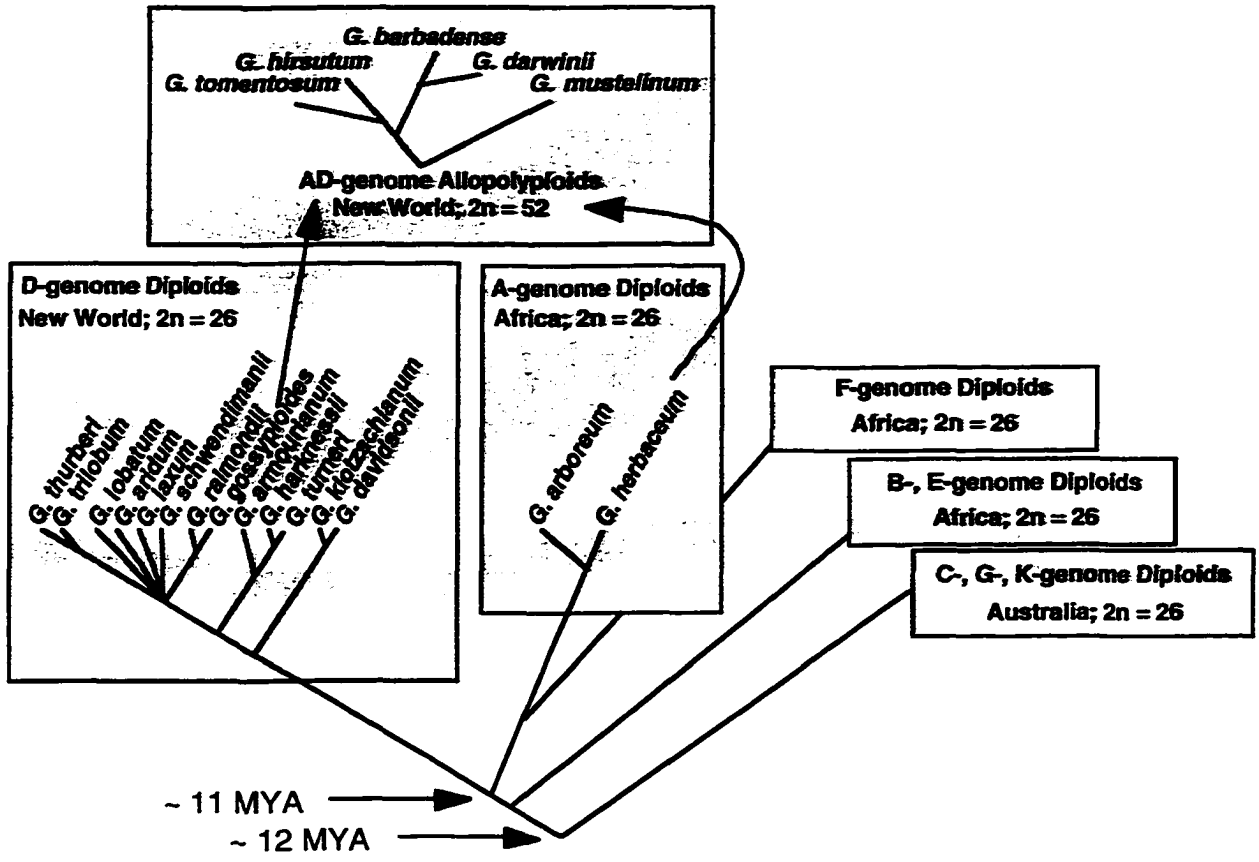


Figure 1. Phylogenetic hypothesis for the genus Gossypium showing relationships among the diploid (2n = 26) species, the origin of the allotetraploid (2n = 52) species, and estimates of the timing of the initial divergences within the genus (Wendel and Albert 1992; Seelanan et al. 1997; Small et al. 1998).

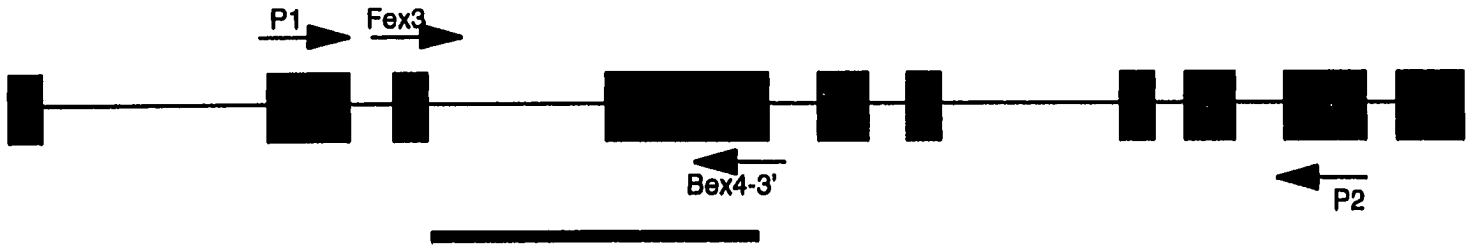


Figure 2. Structure of maize *Adh1* gene as an example of a plant *Adh* gene; numbered boxes represent exons; intervening lines represent introns. Approximate locations of PCR primers are shown as arrows; forward primers are shown above, reverse primers below. The bold line below represents the intron 3 / exon 4 region used as a probe in Southern hybridization analyses.

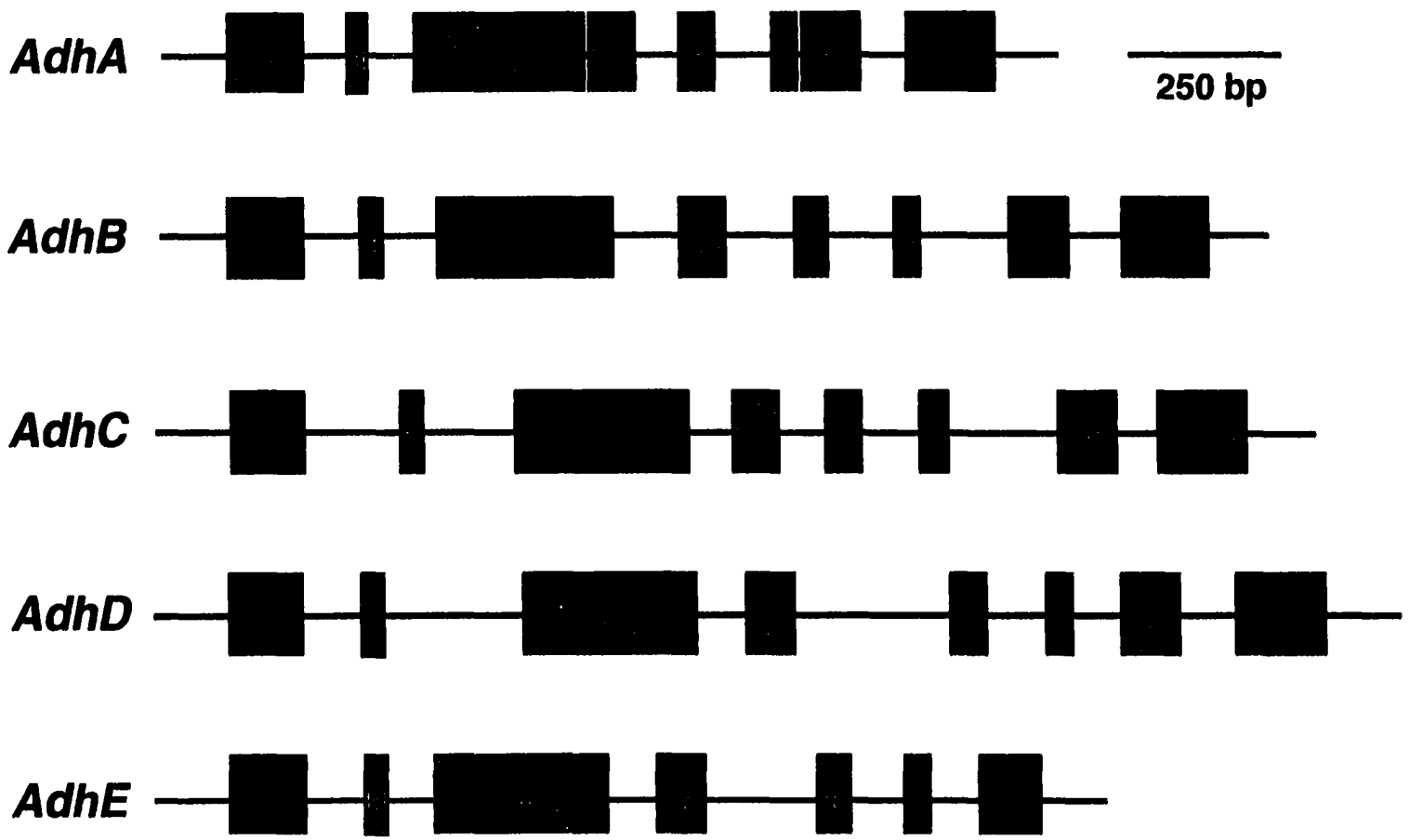


Figure 3. Schematic representation of *Gossypium Adh* genes. Numbered boxes represent exons; intervening lines represent introns. A 250 bp scale is shown for reference. Note that introns 4 and 7 are missing from *AdhA*.

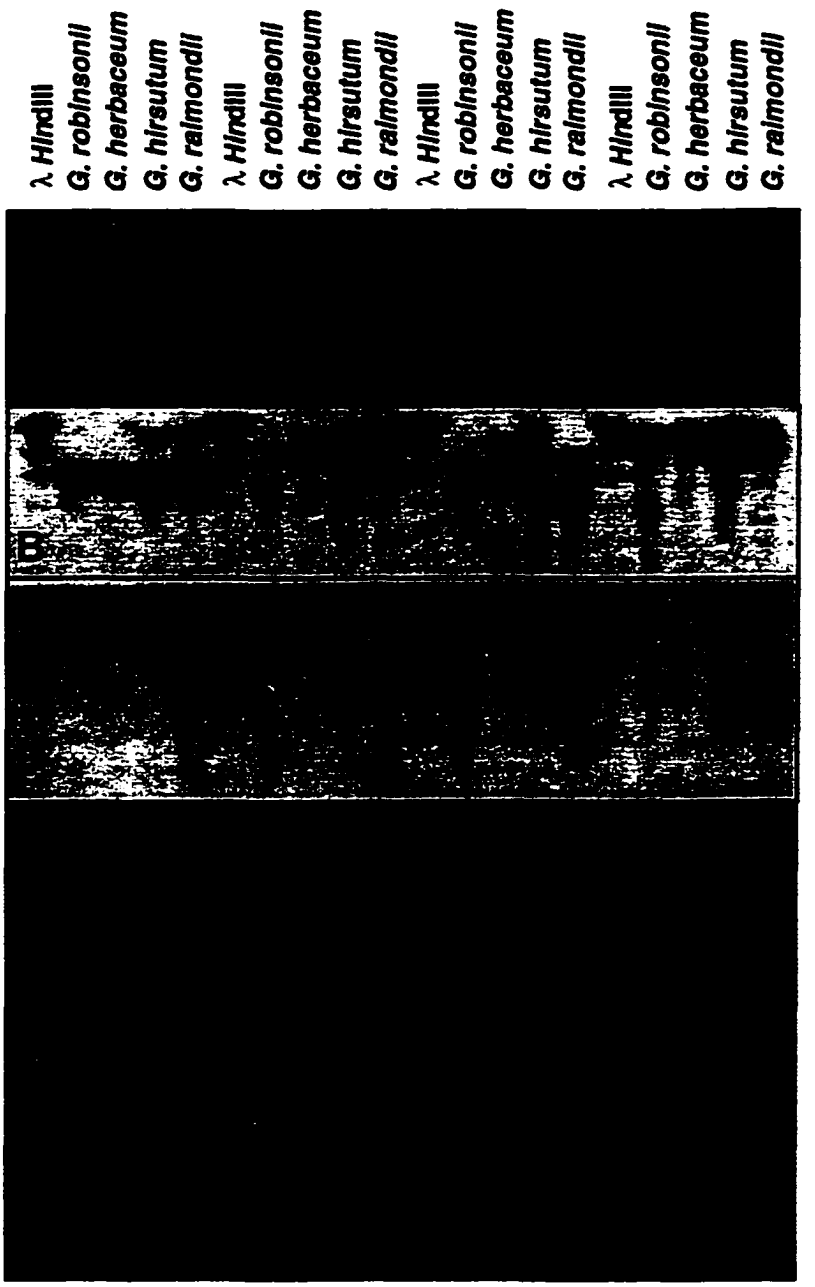


Figure 4A-E. Southern hybridization analysis of the *Gossypium Adh* genes. Each panel shows hybridization profiles for four species (diploids *G. robinsonii*, *G. herbaceum*, and *G. raimondii*; tetraploid *G. hirsutum*), each digested with four restriction enzymes (*Eco*RI, *Eco*RV, *Hin*dIII, *Xba*I). Panels A-E represent identical membranes probed with gene fragments from *AdhA*, *AdhB*, *AdhC*, *AdhD*, and *AdhE* respectively. For *AdhA*, *AdhB*, and *AdhC* the probe consisted of the intron 3/exon 4 region; for *AdhD* the probe consisted of the 3′ UTR of an *Adh1* cDNA (which is orthologous to *AdhD*, see text); and for *AdhE* the probe consisted of an exon 5 / intron 5 gene fragment.

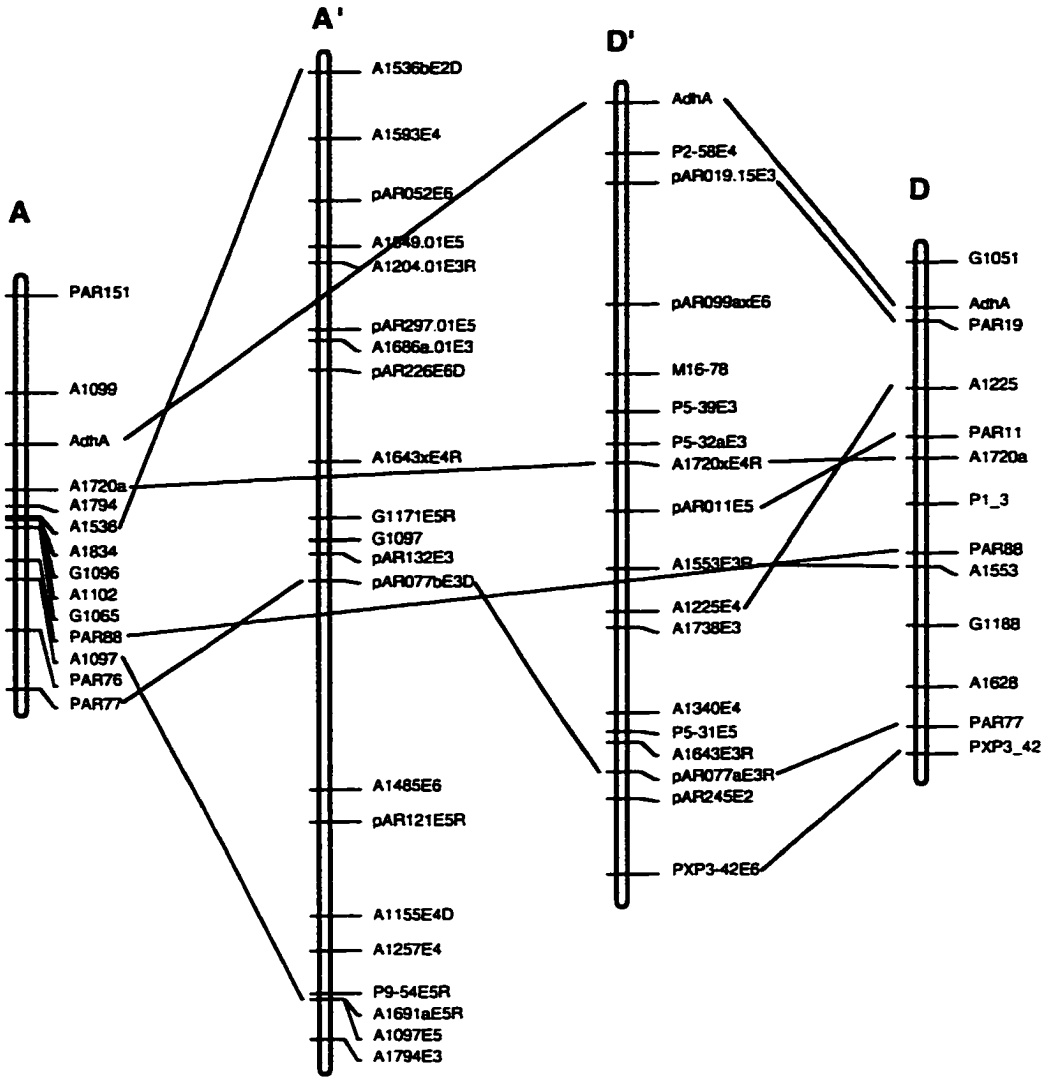


Figure 5. *AdhA* maps to homoeologous assemblage 8C of Brubaker, Paterson, and Wendel (1999) in both A- and D-genome diploid maps and in the D-subgenome of the allotetraploid map.

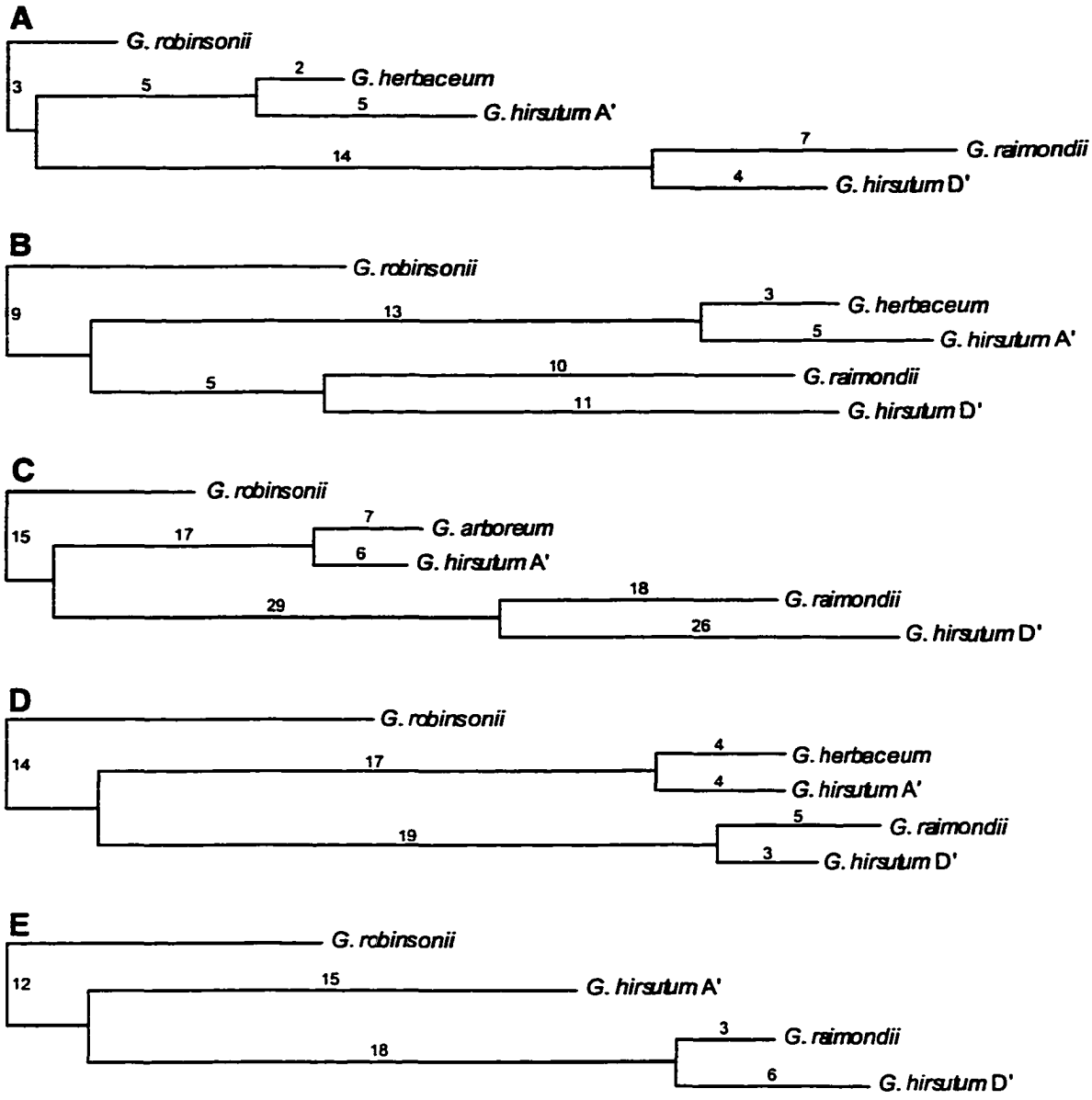


Figure 6A-E. Phylogenetic trees resulting from parsimony analysis of sequences of *AdhA*, *AdhB*, *AdhC*, *AdhD*, and *AdhE* respectively, and rooted with the *G. robinsonii* sequence. Branch lengths are given above each branch. The A- and D-subgenomic sequences of *G. hirsutum* are designated *G. hirsutum* A´ and D´ respectively. For each tree the following information is given below: tree length including autapomorphies (L), consistency index (CI), and retention index (RI). *AdhA*: L: 40, CI: 1.0, RI: 1.0; *AdhB*: L: 56, CI: 1.0, RI: 1.0; *AdhC*: L: 118, CI: 0.99, RI: 0.98; *AdhD*: L: 66, CI:0.97, RI: 0.95; *AdhE*: L: 54, CI: 0.98, RI: 0.94.

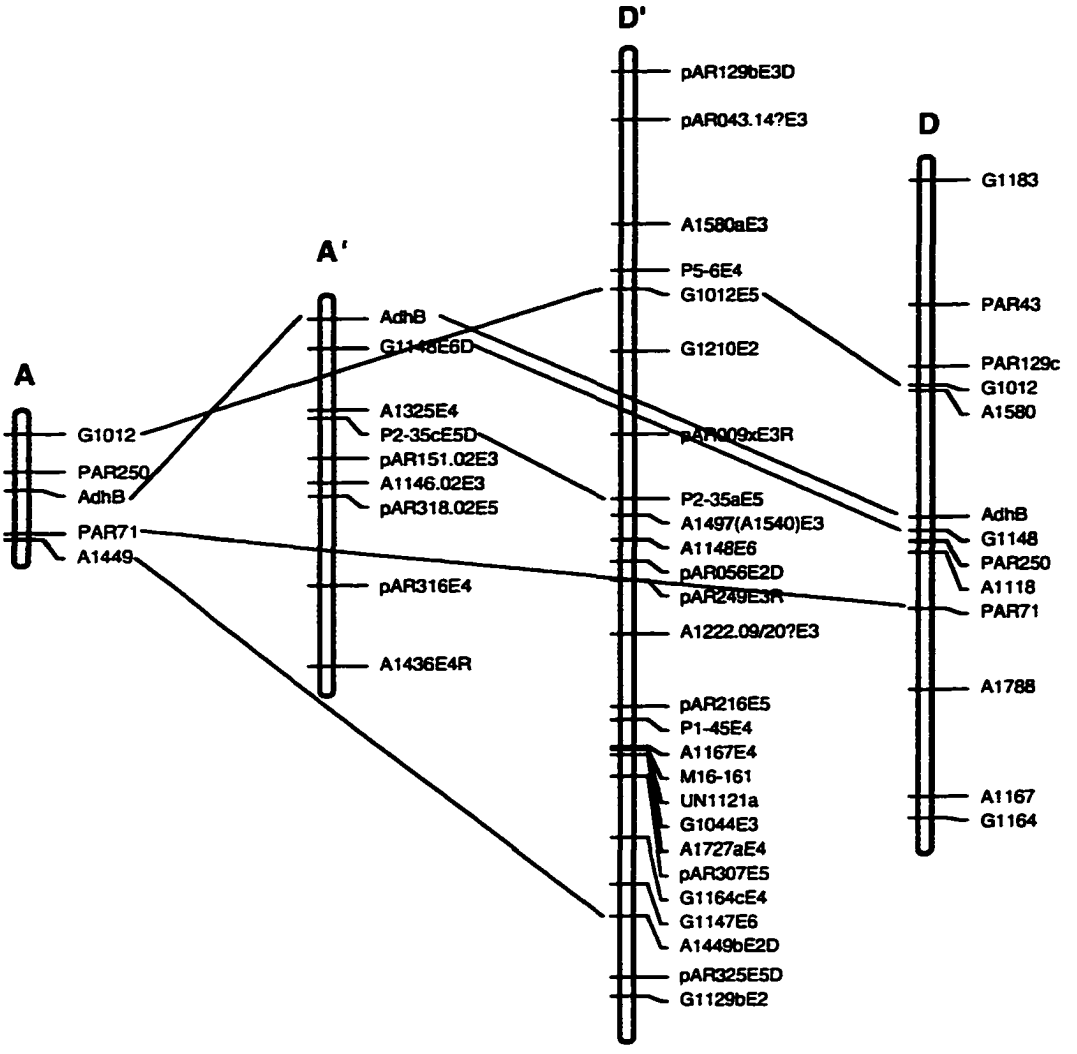


Figure 7. *AdhB* maps to homoeologous assemblage 8A of Brubaker, Paterson, and Wendel (1999) in both A- and D-genome diploid maps and in the A-subgenome of the allotetraploid map.

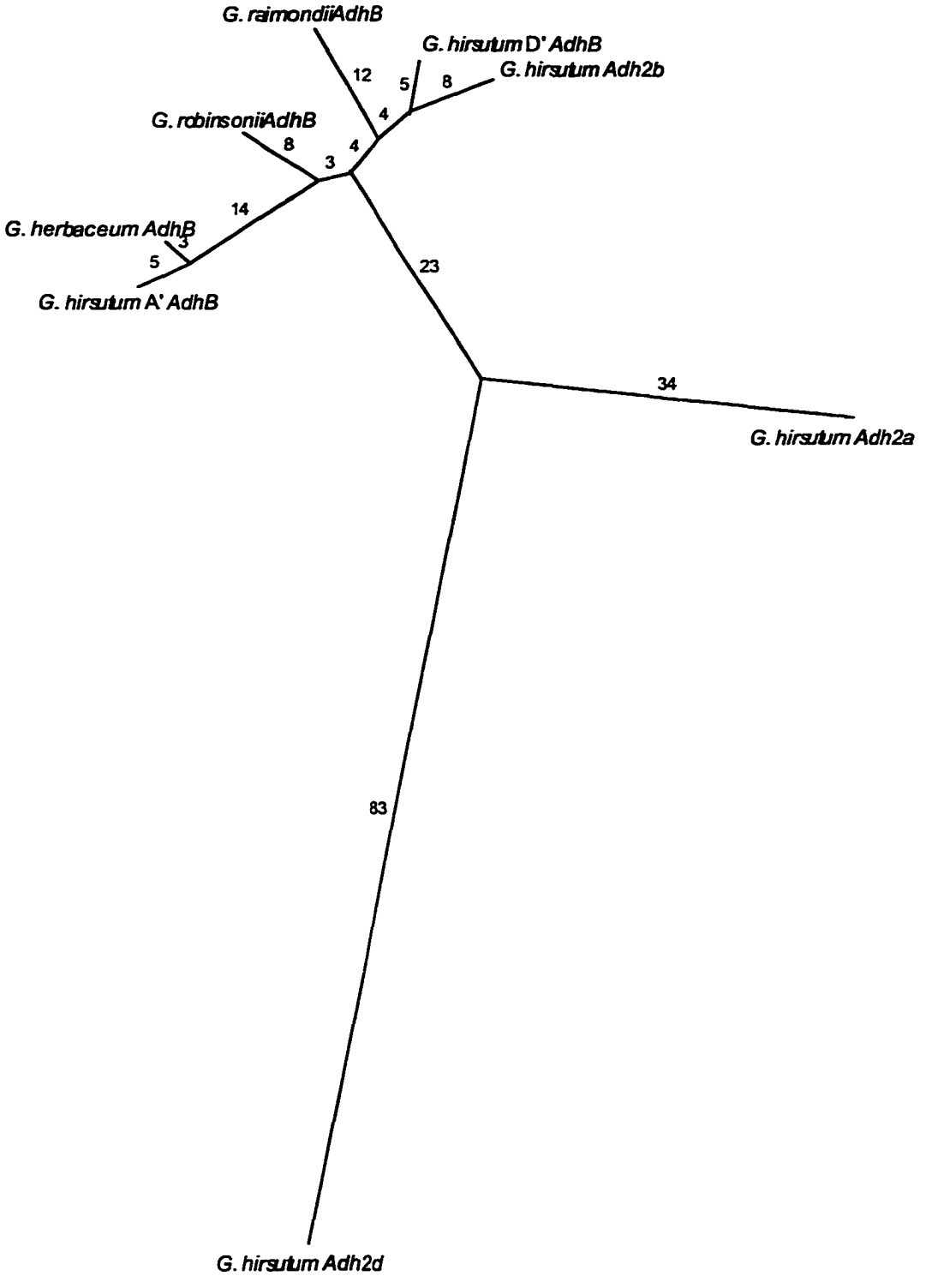


Figure 8. Phylogenetic analysis of *Gossypium AdhB* and *Adh2* (Millar and Dennis 1996) sequences, midpoint rooted, branch lengths shown above each branch. Length: 206, Consistency Index: 0.98, Retention Index: 0.91. This analysis shows *Adh2b* of Millar and Dennis (1996) as sister to the *AdhB* sequence from the D-subgenome of *G. hirsutum* suggesting that these genes are probably orthologous.

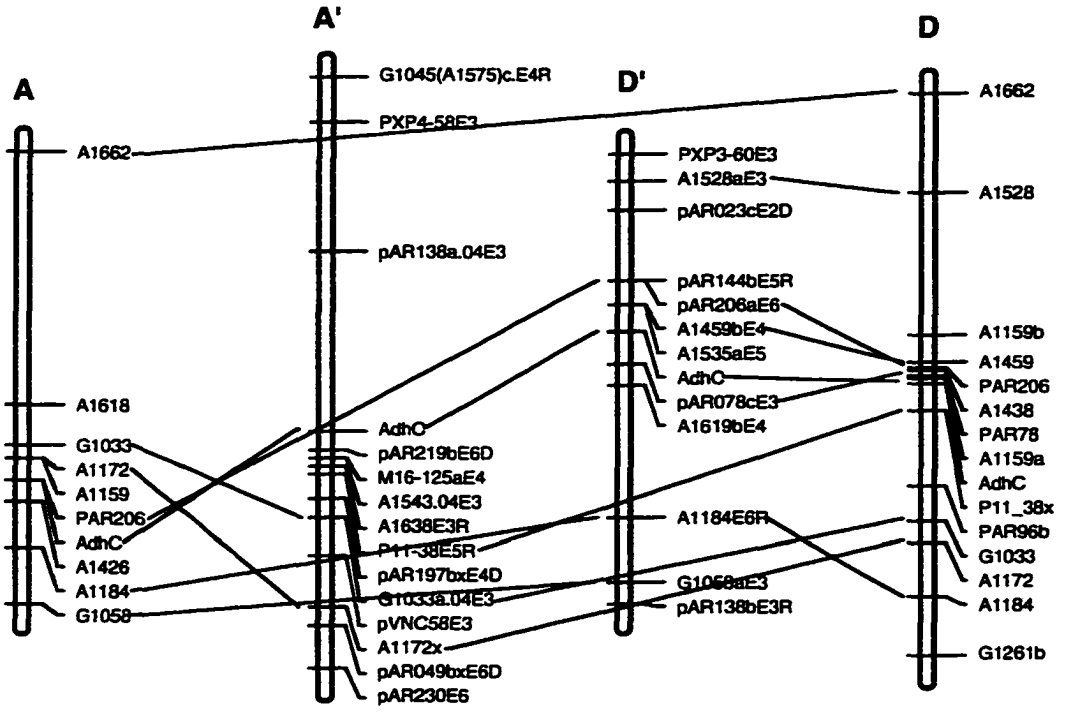


Figure 9. *AdhC* maps to homoeologous assemblage 7B of Brubaker, Paterson, and Wendel (1999) in both A- and D-genome diploid maps and in both subgenomes of the allotetraploid map.

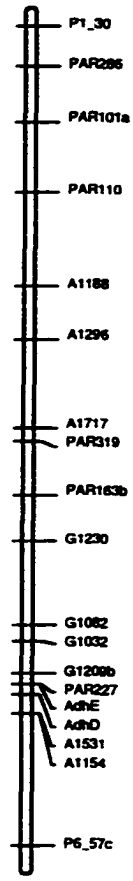


Figure 10. *AdhD* and *AdhE* are closely linked on Chromosome D7 (D-genome diploid map) in homoeologous assemblage 5 of Brubaker, Paterson, and Wendel (1999).

Figure 11. Phylogenetic analysis (neighbor-joining) of plant *Adh* genes; rooted with a *Pinus banksiana Adh* sequence.

CHAPTER 3. THE TORTOISE AND THE HARE: CHOOSING BETWEEN NONCODING PLASTOME AND NUCLEAR *ADH* SEQUENCES FOR PHYLOGENY RECONSTRUCTION IN A RECENTLY DIVERGED PLANT GROUP[1]

A paper published in the *American Journal of Botany*[1]

Randall L. Small[2], Julie A. Ryburn[2], Richard C. Cronn[2], Tosak Seelanan[2], and Jonathan F. Wendel[2]

Abstract

Phylogenetic resolution is often low within groups of recently diverged taxa due to a paucity of phylogenetically informative characters. We tested the relative utility of seven noncoding cpDNA regions and a pair of homoeologous nuclear genes for resolving recent divergences, using tetraploid cottons (*Gossypium* L.) as a model system. The five tetraploid species of *Gossypium* are a monophyletic assemblage derived from an allopolyploidization event that probably occurred within the last 0.5-2 million years. Previous analysis of cpDNA restriction site data provided only partial resolution within this clade despite a large number of enzymes employed. We sequenced three cpDNA introns (*rpl16*, *rpoC1*, *ndhA*) and four cpDNA spacers (*accD-psaI*, *trnL-trnF*, *trnT-trnL*, *atpB-rbcL*) for a total of over 7 kb of sequence per taxon, yet obtained only four informative nucleotide substitutions (0.05%) resulting in incomplete phylogenetic resolution. In addition, we sequenced a 1.65-kb region of a homoeologous pair of nuclear-encoded alcohol dehydrogenase (*Adh*) genes. In contrast with the cpDNA sequence data, the *Adh* homoeologues yielded 25 informative characters (0.76%) and provided a robust and completely resolved topology that is concordant with previous cladistic and phenetic analyses. The enhanced resolution obtained using the nuclear genes reflects an approximately three- to sixfold increase in nucleotide substitution rate relative to the plastome spacers and introns.

Key words: alcohol dehydrogenase; *Gossypium*; molecular phylogenetics; noncoding chloroplast DNA; polyploidy.

---

[1] Reprinted with permission from the American Journal of Botany, 1998, 85(9): 1301-1315.

[2] Department of Botany, Iowa State University, Ames, IA 50011.

Introduction

The ease of generating DNA sequence data has led to an explosion of molecular phylogenetic analyses in recent years (reviewed in Soltis, Soltis, and Doyle, in press). In plants, analyses of cpDNA have predominated (reviewed by Olmstead and Palmer, 1994), typically involving genes such as *rbcL*, *matK*, or *ndhF* (e.g., Chase et al., 1993; Olmstead and Palmer, 1994; Olmstead and Sweere, 1994; Steele and Vilgalys, 1994). More recently, sequencing of cpDNA noncoding regions (introns and intergenic spacers) has become popular for analyses at various taxonomic levels (e.g., Morton and Clegg, 1993; Gielly and Taberlet, 1994a, b, 1996; van Ham et al., 1994; Kita, Ueda, and Kadota, 1995; Manen and Natali, 1995; Downie, Katz-Downie, and Cho, 1996; Gielly et al., 1996; Johnson and Hattori, 1996; Jordan, Courtney, and Neigel, 1996; Kelchner and Wendel, 1996; Kelchner and Clark, 1997; Savolainen, Spichiger, and Manen, 1997; Sang, Crawford, and Stuessy, 1997). Noncoding regions have been presumed to be more useful at lower taxonomic ranks because they are less functionally constrained and are therefore freer to vary, thereby potentially providing more phylogenetically informative characters per unit of sequencing effort (Clegg et al., 1994).

One of the often-cited advantages of molecular data for phylogenetic reconstruction is the almost infinite number of characters that can be sampled. Yet, for plant groups where radiations have been relatively recent it may be extraordinarily difficult to generate sufficient phylogenetic signal because of the relatively slow accumulation of mutations, even in "rapidly evolving" noncoding DNA. The literature is replete with cladograms derived from molecular data that are well resolved internally, but that contain unresolved terminal clades of presumably closely related species (e.g., Hodges and Arnold, 1994; Bayer, Hufford, and Soltis, 1996; Soltis et al., 1996; Panero and Jansen, 1997; Sang, Crawford, and Stuessy, 1997). This phenomenon is the focus of the present paper. Specifically, we wished to address the issue of phylogenetic resolution within recent radiations by asking the following questions: (1) are mutation rates sufficiently high in noncoding cpDNA to provide phylogenetic resolution within a group of woody perennials that may be only 0.5-2 million years old? (2) do mutation rates vary among cpDNA noncoding regions, and if so, which exhibits the highest mutation rate? (3) can strictly orthologous low-copy nuclear-encoded genes be isolated, and if so, do they exhibit a higher mutation rate than noncoding cpDNA? (4) what are the relative strengths and weaknesses of the various types of molecular data for evaluating the phylogenetic relationships of recently radiated groups? As a model system for examining these questions we chose the tetraploid species of *Gossypium* L.

Gossypium includes ~ 50 species (Fryxell, 1992; Wendel, 1995; Wendel, Brubaker, and Seelanan, in press), of which the majority are diploid ($2n = 2x = 26$) and five are allotetraploids

(2n = 4x = 52). Previous studies have resulted in the phylogenetic hypothesis shown in Fig. 1. The allotetraploid species appear to be a monophyletic assemblage derived from a single polyploidization event ca. 0.5-2 million years ago (Wendel, 1989; Wendel and Albert, 1992; Seelanan, Schnabel, and Wendel, 1997), and despite extensive efforts directed at understanding relationships among tetraploid cottons, only weak resolution has been obtained (Endrizzi, Turcotte, and Kohel, 1985; Wendel, 1989; DeJoode and Wendel, 1992; Wendel and Albert, 1992; Reinisch et al., 1994; Cronn et al., 1996; Wendel, Schnabel, and Seelanan, 1995a, b; Seelanan, Schnabel, and Wendel, 1997). In addition to cpDNA and rDNA restriction site data, sequences from the nuclear ribosomal ITS regions are available for all tetraploid species (Wendel, Schnabel, and Seelanan, 1995a, b; Seelanan, Schnabel, and Wendel, 1997) and *ndhF* data are available for two of the five species (Seelanan, Schnabel, and Wendel, 1997). Given voluminous data yet little phylogenetic resolution, tetraploid *Gossypium* provide a test case for evaluating the utility of a variety of putatively quickly evolving molecular sequences for resolving the phylogeny of a recent radiation. To this end we sequenced seven cpDNA noncoding regions in each of the five tetraploid species and a representative of the diploid maternal (chloroplast donor; Wendel, 1989) lineage, *G. arboreum* L. In addition, we isolated and sequenced a region of a pair of homoeologous nuclear-encoded alcohol dehydrogenase (*Adh*) genes for these same taxa, as well as a representative of the paternal lineage, *G. raimondii* Ulbrich, and an additional outgroup, *G. robinsonii* F. Mueller.

Materials and Methods

Plant materials and DNA isolation -- The species of *Gossypium* studied include one accession from each of the five allotetraploid species, and one species from each of three diploid "genome groups." Two of these ("A" and "D" diploids) represent the lineages (maternal and paternal, respectively; Wendel, 1989) from which the allotetraploids were derived, and the third, more distantly related diploid ("C" genome) was included as an outgroup (Table 1). Previous studies support the intrageneric phylogeny shown in Fig. 1 (Wendel and Albert, 1992; Wendel, Schnabel, and Seelanan, 1995a, b; Seelanan, Schnabel, and Wendel, 1997). DNA extractions were carried out as previously described (Paterson, Brubaker, and Wendel, 1993). All sequences obtained in this study have been deposited in Genbank under the accession numbers given in Table 2.

cpDNA regions -- Many cpDNA noncoding regions (introns and intergenic spacers) have been characterized either by direct sequencing (e.g., Morton and Clegg, 1993; van Ham et al., 1994; Manen and Natali, 1995; Downie, Katz-Downie, and Cho, 1996; Gielly et al., 1996;

Johnson and Hattori, 1996; Jordan, Courtney, and Neigel, 1996; Kelchner and Wendel, 1996; Kelchner and Clark, 1997; Savolainen, Spichiger, and Manen, 1997; Sang, Crawford, and Stuessy, 1997) or by restriction site analysis of polymerase chain reaction (PCR)-amplified products (Liston, 1992; Rieseberg, Hanson, and Philbrick, 1992; Demesure, Comps, and Petit, 1996; Wolf, Murray, and Sipes, 1997; Wolfe et al., 1997). The regions we chose to study (Table 2, Fig. 2) included both cpDNA introns
and intergenic spacers and were selected based on the availability of PCR primers, and/or their size and previous reports of phylogenetic utility. These cpDNA regions all reside in the large single-copy region of the tobacco plastome (Shinozaki et al., 1986) with the exception of the *ndhA* intron, which is in the small single-copy region (Fig. 2). Phylogenetic analyses of sequence data for the cpDNA regions analyzed in this study have been previously reported from other plant groups with the exceptions of the *accD-psaI* spacer and the *ndhA* intron.

The *atpB-rbcL* spacer has been used extensively in phylogenetic and molecular evolutionary analyses (Golenberg et al., 1993; Hodges and Arnold, 1994; Manen, Savolainen, and Simon, 1994; Savolainen et al., 1994; Manen and Natali, 1995; Natali, Manen, and Ehrendorfer, 1995; Savolainen, Spichiger, and Manen, 1997). The *trnL-trnF* and *trnT-trnL* spacers were initially characterized by Taberlet et al. (1991). The *trnL-trnF* spacer has been widely exploited in molecular systematic investigations (Böhle et al., 1994; Gielly and Taberlet, 1994b; van Ham et al., 1994; Böhle, Hilger, and Martin, 1997; Sang, Crawford, and Stuessy, 1997). Curiously, the *trnT-trnL* spacer has rarely been used in systematic studies (Böhle et al., 1994; Böhle, Hilger, and Martin, 1997) despite the popularity of the other regions described in the same paper (Taberlet et al., 1991), the larger size of this region relative to the *trnL* intron and the *trnL-trnF* spacer, and the observation by Böhle et al. (1994) that this region is the most variable of the three. The *accD-psaI* spacer has been used only recently (Mendenhall, 1994; T. Barkman, University of Texas, Austin, personal communication). The PCR primers for the *accD-psaI* spacer region were originally designed by B. Milligan (New Mexico State University, Las Cruces) and were provided by T. Barkman and B. Simpson (University of Texas, Austin). The *ndhA* intron has been used in PCR-RFLP analysis (Wolf, Murray, and Sipes, 1997), and Downie, Katz-Downie, and Cho (1996) report 67.1% similarity in a comparison of the *ndhA* introns of tobacco and rice, but analyses of sequence variation among species have not previously been reported. PCR primers for the *ndhA* intron were designed based on maize, rice, and tobacco *ndhA* sequences from Genbank and were anchored in flanking exons. The *rpl16* intron has recently been used extensively for phylogenetic analyses in a variety of plant groups (Dickie, 1996; Jordan, Courtney, and Neigel, 1996; Kelchner and Wendel, 1996; Kelchner and Clark, 1997; Baum, Small, and Wendel, in press; A. Schnabel and J. Wendel, unpublished data; S. Downie, University

of Illinois, personal communication). Downie, Katz-Downie, and Cho (1996) report 64.5% similarity in a comparison of the *rpl16* introns of tobacco and rice; this is the lowest similarity reported in their comparison of cpDNA introns. The *rpoC1* intron was used by Downie, Katz-Downie, and Cho (1996) for assessing intrafamilial relationships within Apiaceae.

Nuclear-encoded alcohol dehydrogenase loci -- Alcohol dehydrogenase (*Adh*, E.C. number 1.1.1.1) is a metabolic enzyme responsible for the interconversion of ethanol and acetaldehyde, primarily in response to hypoxic conditions (Freeling and Bennett, 1985). In cotton, as in most plants, *Adh* exists as a nuclear-encoded small gene family (Millar and Dennis, 1996; Small and Wendel, unpublished data). Gene structure of *Adh* in *Gossypium* is generally conserved relative to other plant species studied (Fig. 3; Millar and Dennis, 1996; Small and Wendel, unpublished data). Because the *Gossypium* species under consideration are allotetraploids (containing A and D subgenomes; see above) each nuclear-encoded locus present in diploid species is present in two copies (homoeologues) in the tetraploid species, one per subgenome. We have PCR-amplified, cloned, and sequenced the majority of a pair of homoeologous *Adh* genes from tetraploid *Gossypium* as well as the orthologues from diploid *Gossypium* representing the parents of the allopolyploid.

An underlying assumption of any phylogenetic analysis is that the sequences included are orthologous (related by speciation), rather than paralogous (related by gene duplication). The most reliable method of demonstrating orthology for nuclear genes is comparative genetic mapping. Mapping genes to positions on homologous/homoeologous linkage groups provides strong evidence for orthology. Therefore, we have genetically mapped the sequenced *Adh* loci in both the A- and D-diploid genomes and one subgenome of the AD-allotetraploid. We found that these loci map to homologous/homoeologous linkage groups (data not shown) and so infer that they are orthologous. We term the *Gossypium* sequences reported here *AdhC* to differentiate them from the commonly used terminology *Adh1*, *Adh2*, etc., which imply homologies to *Adh* genes in other plants that are not in evidence. The *AdhC* sequences reported here are not orthologous to the *Gossypium Adh1* or *Adh2* sequences reported by Millar and Dennis (1996).

Adh sequences have been used previously in a number of phylogenetic and molecular evolutionary studies in plants (Gaut and Clegg, 1991, 1993; Goloubinoff, Pääbo, and Wilson, 1993; Hanfstingl et al., 1994; Gaut et al., 1996; Innan et al., 1996; Miyashita, Innan, and Terauchi, 1996; Morton, Gaut, and Clegg, 1996; Sang, Donoghue, and Zhang, 1997).

Amplification, cloning, and sequencing -- cpDNA regions -- PCR amplifications were performed in 50-mL reactions consisting of 1 unit *Taq* polymerase (Promega, Madison,

Wisconsin), 1X buffer (Promega), 200 mmol/L each deoxy-nucleotide triphosphate, 1.5 mmol/L $MgCl_2$, 10-20 pmol of each primer and 8-12 ng of template genomic DNA. Amplifications were carried out using the parameters described in Table 3 in an MJ Research PTC-100 thermal cycler (Watertown, Massachusetts). Amplifications were preceded by a "hotstart" consisting of 2 min at 94°C followed by 5 min at 80°C during which time the *Taq* polymerase was added to the reactions. A negative control reaction (no template DNA) was included for each set of amplifications to monitor for the possibility of contamination. All PCR primers were either obtained from other researchers or were synthesized by Integrated DNA Technologies (Coralville, Iowa). Amplification products were visualized by agarose gel electrophoresis, concentrated using Microcon-100 centrifugation separators (Amicon, Beverly, Massachusetts), and quantified fluorometrically. PCR products were either sequenced directly (*rpl16* intron, *trnL-trnF* spacer, *rpoC1* intron, *ndhA* intron) or cloned into pGEM-T (Promega) and sequenced (*atpB-rbcL* spacer, *trnT-trnL* spacer, *accD-psaI* spacer). For the cloning approach, purified PCR products were ligated into pGEM-T according to the manufacturer's instructions. Competent Top10 F′ (Invitrogen, San Diego, California) cells were transformed via electroporation and the resulting colonies were screened for plasmids with inserts by PCR using the original amplification primers. Plasmids were isolated from a single recombinant colony using an alkaline lysis/PEG precipitation protocol (Sambrook, Fritsch, and Maniatis, 1989). Cloning was performed only when PCR-amplification resulted in insufficient template for automated sequencing or when difficulties were encountered in using the amplification primers as sequencing primers. All sequencing was performed using amplification, internal, and/or vector specific primers (Table 2) at the Iowa State University DNA Sequencing and Synthesis Facility.

Adh -- PCR-amplification and cloning of *Adh* homoeologues were performed as described for the cpDNA regions except that 2.0 mmol/L $MgCl_2$ was used in PCR reactions. The primers P1 and P2 (designed by K. Schierenbeck, California State University, Fresno; Table 2) are homologous to regions in exon 2 and exon 9, respectively, of *Gossypium Adh* (Fig. 3). Initial use of these primers resulted in amplification of multiple members of the *Adh* gene family. To isolate *AdhC* sequences, the entire heterogeneous PCR product pool was cleaned and concentrated using Geneclean II (Bio 101, La Jolla, California), ligated into pGEM-T, and transformed into Top10 F′ cells. The resulting colonies were screened by PCR using the amplification primers, and colonies that contained inserts of the size corresponding to the *AdhC* sequence were identified. Because tetraploid species of *Gossypium* contain two *AdhC* loci (homoeologues), it was necessary to further screen these plasmids to isolate A and D subgenome sequences. Multiple colonies containing plasmids with appropriately sized inserts were isolated from each taxon.

Inserts from these plasmids were PCR amplified, ethanol precipitated, resuspended in a small volume of water, and then restriction digested with *Alu*I (American Allied Biochemical, Aurora, Colorado) according to the manufacturer's instructions. Visualization of the digestion products by agarose gel electrophoresis revealed subgenome-specific digestion patterns that allowed discrimination of plasmids containing either A or D subgenome *AdhC* inserts. Using this PCR/cloning approach we isolated *AdhC* sequences from the diploids *G. robinsonii*, *G. raimondii* and from both the A and D subgenomes of all five tetraploid species (Table 1). These plasmids were then isolated and sequenced as described above. We were unable, however, to isolate the corresponding *AdhC* sequence from either of the two extant A-genome diploids (*G. arboreum* or *G. herbaceum*) using this approach. We therefore employed an internal, *AdhC*-specific primer (ADHX8-2, Table 2; Fig. 3) in conjunction with P1 and amplified a ~ 1.35 kb *AdhC* fragment from *G. arboreum*. Because the primer combination is locus specific we were able to directly sequence the *G. arboreum AdhC* PCR product using the Thermosequenase cycle-sequencing kit (Amersham, Arlington Heights, Illinois).

Analyses -- Characterization of each region and sequence comparisons were facilitated by the software programs MacClade 3.05 (Sinauer, Sunderland, Massachusetts), PAUP 3.1.1 (Swofford, 1993) and MEGA 1.0 (Kumar, Tamura, and Nei, 1993). Analyses were conducted both on individual and combined data sets as follows. Individual cpDNA region data sets were analyzed separately (when warranted by the existence of sufficient variation) and then as a combined cpDNA data set. *Adh* sequences were analyzed in three separate ways: individual sequences as terminal "taxa," by subgenome, and by combining *Adh* homoeologue sequences for tetraploid taxa for an *Adh* "total evidence" analysis. For each data set a g_1 statistic (Hillis and Huelsenbeck, 1992; Hillis, Allard, and Miyamoto, 1993) was calculated using PAUP 3.1.1 to determine whether or not significant phylogenetic structure existed within the data set. For phylogenetic analyses, exhaustive searches for most-parsimonious trees were conducted with uninformative characters excluded. Due to the larger number of sequences included in the initial *Adh* analysis (each allotetraploid represented by two distinct sequences), the Branch and Bound algorithm was employed to search for maximally parsimonious trees. Relative levels of support for clades present in the most-parsimonious trees were assessed by calculating decay values, the number of extra steps required to collapse the clade (Bremer, 1988). For all phylogenetic analyses the tree lengths and consistency indices reported do not include autapomorphic characters. Rate variation among sequences was assessed using the 1D and 2D relative rate tests of Tajima (1993) as implemented in the program Tajima93 (T. Seelanan, unpublished software).

Results

cpDNA sequences -- Over 7.3 kilobase pairs (kb) of cpDNA sequence (6.4 kb of noncoding sequence) from seven different regions were determined for each of the five tetraploid species of *Gossypium* and the outgroup, *G. arboreum*. These data collectively represent ~ 5.6% (7369/130 505 bp) of the unique sequence of the tobacco plastome (i.e., counting the inverted repeat only once) and ~ 10% (6438/64 437 bp) of the unique noncoding portion of the tobacco plastome (K. Wolfe, University of Dublin, Trinity College, Ireland, personal communication). Each of the sequenced regions is characterized in Table 4. Phylogenetically informative characters were observed only in the *trnT-trnL* spacer and the *rpl16* intron. The low observed GC content (~ 30%, see Table 4) of the sequenced regions is similar to that reported for plastomes in general (Palmer, 1991).

Averaged over all cpDNA sequences the mean divergence between *G. arboreum* and the ingroup species was 0.30% and mean divergence among tetraploid *Gossypium* was 0.20%. These values, however, were not equally distributed across all regions and, in fact, divergence from *G. arboreum* ranged from 0.00 to 0.96%, while divergence among tetraploids ranged from 0.00 to 0.49% (in both cases, *rpoC1* intron and *trnT-trnL* spacer, respectively). The mean transition:transversion ratio (Ts:Tv) across all cpDNA sequences was 0.9:1, while individual values ranged from 5:1 to 0:6 (Table 4). Substitution patterns taken across all regions appear to follow the observations of Morton (1995), in that positions flanked by A or T are more likely to undergo transversions. While this pattern is evident upon inspection, the data are too few to test statistically.

Overall, 7369 characters (nucleotides) were sampled, yielding 52 variable positions (0.71%) and four potentially phylogenetically informative nucleotide substitutions (0.05%). In addition to nucleotide substitutions, we observed 15 length mutations (indels), of which four were potentially phylogenetically informative.

Phylogenetic analyses of cpDNA sequences -- Potentially phylogenetically informative characters were found in only two of the seven regions; the *trnT-trnL* spacer (four characters) and the *rpl16* intron (four characters) (see Table 4). Exhaustive searches of all possible trees were performed for each of these data sets using PAUP v. 3.1.1 (Swofford, 1993). The g_1 statistics were -1.57 and -0.23 for the *trnT-trnL* and the *rpl16* intron, respectively. For the number of taxa and characters in these data sets, only the *trnT-trnL* spacer data set is significantly more structured than random ($P < 0.01$; Hillis and Huelsenbeck, 1992). The single most-parsimonious tree resulting from analysis of the *trnT-trnL* data set is shown in Fig. 4 (length = 4; consistency

index [CI] = 1.0; retention index [RI] = 1.0). When all cpDNA data were combined into a single data set, a g_1 statistic of -1.08 was obtained which is significantly more structured than random ($P < 0.01$). Two equally most-parsimonious trees (length = 11; CI = 0.727; RI = 0.625) were found in an exhaustive search; the topology of the strict consensus tree was identical to Fig. 4. The two shortest trees differed only in the placement of *G. hirsutum* which was resolved either as sister to a *G. barbadense* + *G. darwinii* clade, or as part of an unresolved polytomy as in the strict consensus tree.

Nuclear Adh sequences -- *Adh* exists as a small gene family in *Gossypium*. We chose to analyze the locus we refer to as *AdhC*. This locus maps to homologous/homoeologous regions of the A- and D-genome diploids and AD-genome tetraploid genetic maps (data not shown); thus we are confident that we are analyzing orthologous sequences. The PCR primers P1 and P2 amplify a ~ 1.65-kb region of *Adh* from exon 2 to the 5′ end of exon 9 (Fig. 3). We obtained sequences from *G. robinsonii* (C-genome diploid outgroup; see Fig. 1), *G. raimondii* (D-genome diploid), and from both the A- and D-subgenomes of all five AD-genome tetraploid species using these primers. A *G. arboreum* (A-genome diploid) sequence was obtained using the locus-specific primer pair P1/ADHX8-2, which amplifies a region from the middle of exon 2 to the 5′ end of exon 8 (Fig. 3); the resulting PCR product was 1352 bp in length.

All *AdhC* sequences maintain the expected 5′ GT... and ...AG 3′ intron boundary sequences with the exception of a G to A transition of the first nucleotide of intron 6 of the D-subgenomes of *G. hirsutum* and *G. tomentosum*, and an A to G transition at the 3′ end of intron 3. All sequences also maintain exon integrity (presence, length, reading frame) with the following exceptions. A 67-bp deletion in the A-subgenome sequences of *G. barbadense* and *G. darwinii* begins seven nucleotides from the 3′ end of exon 4 and ends in the middle of intron 4. A large (182 bp) deletion in the *G. arboreum* sequence results in partial loss of introns 5 and 6, and all of exon 6. Finally, a G to A transition in exon 2 of the *G. arboreum* sequence results in the conversion of a tryptophan-encoding codon (TGG) to a stop codon (TAG). The relevance of the foregoing observations to *AdhC* expression was not explored.

Sequence characteristics for *AdhC* are summarized in Table 5 and are discussed below. The total aligned length of the data matrix is 1667 bp; this includes 798 bp of exon sequence and 869 bp of intron sequence. With the exception of the sequence from *G. arboreum*, the absolute sequence lengths ranged from 1579 bp to 1655 bp. GC content varied little between the A- and D-(sub)genomes, but varied greatly between exons (45.4 - 46.2%) and introns (30.1 - 32.0%). Among sequences from tetraploid taxa, transition:transversion ratios (Ts:Tv) varied between genomes, and especially between introns and exons. In the A-(sub)genome the Ts:Tv was ~

4.2:1, whereas in the D-(sub)genome the Ts:Tv was ~ 3.6:1 (Table 5). The differences between intron and exon Ts:Tv are more dramatic, ranging from 7-8:1 in exons and 1.6-3.3:1 in introns. Table 5 also reveals a marked disparity in the number of nucleotide substitutions in the two subgenomes; the number of nucleotide differences between all pairs of sequences are shown in Table 6. The D-subgenome sequences have experienced ~ 1.5 times as many nucleotide substitutions and yield almost three times as many potentially phylogenetically informative characters. This disparity is also reflected in the relative rate tests (Tajima, 1993), as summarized in Table 6. These tests indicate that, in all comparisons, *AdhC* genes from the D-(sub)genomes are accumulating substitutions at a rate that is significantly faster than are their orthologues/homoeologues in the A-(sub)genomes.

Phylogenetic analyses of Adh sequences -- Three separate analyses were conducted with the *AdhC* sequences. First, an analysis was conducted using each sequence as a terminal; secondly, sequences of each (sub)genome were analyzed separately; and finally, the data from the subgenomes were combined for each taxon for a "total evidence" analysis.

For the data set in which each sequence was treated as a terminal the g_1 statistic estimated from 10 000 random trees was -0.49, which indicates that the data are significantly more structured than random ($P < 0.01$). Phylogenetic analysis of this data set resulted in a single most-parsimonious tree (length = 97, CI = 0.93, RI = 0.98), which is shown in Fig. 5. The tree is completely resolved and divided into two primary clades — one including the D-genome diploid and D-subgenome of the allotetraploids and the second including the A-genome diploid and the A-subgenomes of the allotetraploids. Within each (sub)genomic clade the resolution is complete and the topology is identical between clades.

Analyses of the subgenome sequences individually were also performed. The g_1 statistics calculated for the A- and D-subgenome data sets were -1.55 and -1.52, respectively; both values indicate data significantly more structured than random at the $P = 0.01$ level. In both cases, phylogenetic analysis found a single most-parsimonious tree. For the A-(sub)genome the tree had a length = 8, CI = 1.0, and RI = 1.0. The D-(sub)genome tree had a length = 20, CI = 0.95, and RI = 0.95. Again, each tree was fully resolved and the resulting topologies were identical to that shown in Fig. 5.

Finally, the data for both homoeologues were combined for each taxon for an *Adh* "total evidence" analysis. For outgroup comparison, the *G. raimondii* and *G. arboreum* sequences were combined to make a "diploid progenitor" sequence and the *G. robinsonii* sequence was duplicated. This data set had a g_1 statistic of -1.39, significantly more structured than random at the $P = 0.01$ level. An exhaustive search found a single most-parsimonious tree (Fig. 6) with length = 43,

CI = 0.91, and RI = 0.91. The tree is fully resolved and well supported, as indicated by high decay values and branch lengths.

Discussion

Phylogeny of allotetraploid Gossypium -- Despite intensive study of the tetraploid species of *Gossypium*, the phylogenetic relationships among these species have remained elusive. The data presented in this paper provide a completely resolved and robustly supported phylogenetic hypothesis for tetraploid *Gossypium* (Fig. 6). Within the tetraploid clade, the Brazilian endemic *G. mustelinum* represents the sole descendant of one branch of the initial divergence, as tentatively shown by DeJoode and Wendel (1992) and predicted by Wendel, Rowley, and Stewart (1994). The remaining four taxa form a clade sister to *G. mustelinum*, and are divided into two species-pairs: *G. barbadense* + *G. darwinii*; and *G. hirsutum* + *G. tomentosum*. The relationship between *G. barbadense* and *G. darwinii* has long been established, and in fact, the two taxa have been considered conspecific (see discussion in Percy and Wendel, 1990; Wendel and Percy, 1990). The affinities of *G. hirsutum* and *G. tomentosum*, however, were unclear until the study of DeJoode and Wendel (1992), which suggested that they are sister taxa; this relationship, however, was only weakly supported by a single rDNA restriction site mutation. Subsequent analysis of ITS sequences have confirmed this observation (Wendel, Schnabel, and Seelanan, 1995a, b; Seelanan, Schnabel, and Wendel, 1997) and the *AdhC* data presented here corroborate this relationship and provide additional strong support.

Relationships hypothesized by these data additionally confirm predictions based on other sources of evidence. For example, the basal position of *G. mustelinum* predicts that it should be genetically equidistant from all other tetraploid species (Wendel, Rowley, and Stewart, 1994). This is borne out not only by the allozyme data presented by Wendel, Rowley, and Stewart (1994), but also by the *AdhC* sequence data reported in this paper; in the combined analysis (Fig. 6) there are 34, 35, 28, and 32 character-state changes between *G. mustelinum* and *G. hirsutum*, *G. tomentosum*, *G. barbadense* and *G. darwinii*, respectively (mean divergence from *G. mustelinum* = 1.0%). The *Adh* data also support the conclusion that *G. barbadense* and *G. darwinii* diverged more recently from each other than did *G. hirsutum* and *G. tomentosum*: while the branches leading to these two clades have similar lengths (10 vs. 12 steps), the number of autapomorphies each lineage has accumulated differ dramatically (9 and 10, respectively, in *G. hirsutum* and *G. tomentosum* vs. 1 and 5, respectively, in *G. barbadense* and *G. darwinii*).

Molecular evolution of noncoding cpDNA -- The impetus for the experiments described here was to explore the phylogenetic utility of various sequences rather than to provide an in-

depth analysis of patterns of molecular evolution. Nonetheless, some observations are prompted by our data. First, it has been recognized that cpDNA accumulates nucleotide substitutions more slowly than does plant nuclear DNA (Wolfe, Li, and Sharp, 1987; Wolfe, Sharp, and Li, 1989). As summarized in Tables 4 and 6, this rate difference is clearly evident in our data. In fact, the cpDNA data are astounding in their lack of informativeness, with a total of only eight phylogenetically informative characters observed among over seven thousand nucleotides surveyed. As a result of so little variation, the cpDNA provide only limited phylogenetic power.

In addition to the overall paucity of genetic variation, certain patterns observed previously are also noted here. First, the finding of Morton (1995) that transversions are more prevalent at positions flanked by A/T is supported by our data qualitatively, but sufficient data do not exist to statistically test this association. Also, previous observations that indels occur almost as frequently as nucleotide substitutions in noncoding cpDNA (Golenberg et al., 1993; Gielly and Taberlet, 1994b) are not supported by our data (Table 4). Rather, we detected over three times as many substitutions as indels in sequences from the allopolyploids (52 vs. 15, Table 4). Patterns of substitutions and indels vary between regions and in no case does the number of indels equal the number of substitutions. Of the indels that occur, two primary types are observed: insertion or deletion of a multinucleotide stretch of unique sequence; or insertion/deletion of one or a few nucleotides within a polynucleotide tract (particularly polyA/T). The former type of indel is generally easily aligned and, if cladistically informative, is usually nonhomoplasious. In our cpDNA data there were 12 such indels, of which three were phylogenetically informative and none were homoplasious. The latter type of indel (three in our data), however, appears evolutionarily labile and probably originates via slipped-strand mispairing during replication (Levinson and Gutman, 1987). These types of indels often provide homoplasious characters. For example, the single homoplasious indel character in our cpDNA data set is a deletion of a single T in a string of ten in the *rpl16* intron, which is shared by *G. hirsutum* and *G. barbadense*.

Molecular evolution of Adh -- Patterns of molecular evolution among the *AdhC* sequences will be discussed in the context of a full presentation of the evolution of the *Adh* gene family in *Gossypium*. Certain features of the data, however, are especially relevant here. In particular, the disparity of substitution rates between *AdhC* sequences of the A- and D-subgenomes is striking, consistent, and statistically significant (see Table 6). Relative rate differences may be attributed to a number of evolutionary or population genetic phenomena, including background mutational processes, generation time, lineage effects, selection, drift, and rates of recombination (Bosquet et al., 1992; Gaut et al., 1992; Gaut, Muse, and Clegg, 1993;

Clegg et al., 1994; Eyre-Walker and Gaut, 1997). Because both of the two *AdhC* homoeologues exist within the same nuclear genome, however, background mutational and population genetic phenomena should affect them equally and can therefore be ruled out as having a significant effect. Selection is one (but not the only) process that can potentially differentially affect genes in the same nucleus. Either differing levels of purifying selection on the subgenome sequences or positive (diversifying or directional) selection on the D-subgenome sequences could account for the observed rate differences. There is an almost fivefold elevation of nucleotide substitution rates in exons of the D-subgenome relative to the A-subgenome (K = 0.014 vs. 0.003, respectively; Table 5), despite the fact that intron nucleotide substitution rates are actually slightly higher in the A-subgenome sequences (K = 0.009 vs. 0.008; Table 5). Secondly, within exon sequences the synonymous nucleotide substitution rate (K_s) is over twice as high in the D-subgenome relative to the A-subgenome (K_s = 0.019 vs. 0.008; Table 5), but the nonsynonymous nucleotide substitution rate (K_a) is over six times higher (K_a = 0.013 vs. 0.002; Table 5). Finally, overall *AdhC* nucleotide substitution rates in the A-subgenome sequences are higher in the introns than in the exons (K = 0.009 vs. 0.003, respectively; Table 5) as predicted by neutral theory (Kimura, 1980); yet, in the D-subgenome sequences the nucleotide substitution rate is approximately twice as high in exons as in the introns (K = 0.014 vs. 0.008 respectively; Table 5). These data collectively suggest that selective forces may differ between homoeologues.

Relative phylogenetic utilities of molecular data -- The phylogenetic conclusions described above are based almost exclusively on the wealth of data provided by the *AdhC* sequences, despite the volume of cpDNA data generated for identical taxa. In addition to the data presented in this paper, there exist for allotetraploid *Gossypium* comparable molecular data sets for cpDNA restriction sites (Wendel, 1989; DeJoode and Wendel, 1992; Wendel and Albert, 1992), and ITS sequences (Wendel, Schnabel, and Seelanan, 1995a, b; Seelanan, Schnabel, and Wendel, 1997). Figure 7 presents a comparison of the percentage of phylogenetically informative characters for these data sets. The cpDNA data consistently exhibit lower levels of informative characters than do the nuclear-encoded loci, as expected (Wolfe, Li, and Sharp, 1987; Wolfe, Sharp, and Li, 1989; Eyre-Walker and Gaut, 1997). The percentage of phylogenetically informative characters in the cpDNA data sets varied from 0 to 0.34%, and several of the cpDNA noncoding regions yielded no informative characters. The three cpDNA data sets that did contain informative characters (*rpl16* intron, *trnT-trnL* spacer, and cpDNA restriction sites) exhibited similar levels of informativeness both in terms of percentages (0.29 - 0.34%) and absolute numbers of informative characters (3 - 4).

Among the nuclear-encoded loci there is large range of divergence values, as is expected given that each sequence type has its own unique biology. The value for ITS was partially extrapolated from the number of characters on internal branches. This was done because the ITS sequences in *G. mustelinum* have concerted to an A-genome like sequence, while the ITS sequences of the remaining tetraploids have concerted to a D-genome like sequence (Wendel, Schnabel, and Seelanan, 1995a). Despite these caveats, three results are clear from Fig. 7: (1) levels of phylogenetically informative characters are higher in nuclear-encoded sequences than in plastome data sets; (2) levels of informative characters vary among nuclear-encoded sequences; and (3) percentages of informative characters in the *AdhC* sequences are equivalent to or higher than ITS sequences. Current work is underway to examine levels of divergence among a large number of nuclear-encoded sequences in *Gossypium* (Cronn and Wendel, unpublished data).

Advantages and limitations of nuclear-encoded genes for phylogenetic analysis -- Relative rates -- It has long been recognized that nuclear-encoded sequences evolve at a faster rate than plastid-encoded sequences (e.g., Wolfe, Li, and Sharp, 1987; Wolfe, Sharp and Li, 1989; Eyre-Walker and Gaut, 1997). Despite this, in the search for the most phylogenetic information per unit of effort, nuclear-encoded sequences have been relatively ignored, with the exception of the widely used rDNA regions. The data presented here show clearly that cpDNA noncoding sequences may not be able to provide sufficient characters for robust resolution among closely related taxa, even if sampled ad infinitum. We sampled over 6 kb of cpDNA noncoding sequence (~ 10% of all unique cpDNA noncoding sequences), and yet obtained incomplete and poorly supported phylogenetic resolution. In addition, over 1000 cpDNA restriction sites were previously sampled (Wendel, 1989; DeJoode and Wendel, 1992), again with incomplete resolution. In contrast, sequences from a 1.6-kb nuclear-encoded *AdhC* gene provided complete and robust resolution among these closely related taxa. This difference in phylogenetic utility reflects simply the greatly accelerated rates of nucleotide substitution in the nuclear genome relative to the plastome, as illustrated in Fig. 7. The mean number of substitutions per site (K) in the combined cpDNA sequence data set was $K = 0.002$, while in the *AdhC* data sets $K = 0.006$ in the A-(sub)genome and $K = 0.011$ in the D-(sub)genome — a three to sixfold difference in nucleotide substitution rates. Extrapolation of these data allow the following observation. Given a total of four informative nucleotide substitutions out of a total of 6438 bp of noncoding cpDNA sequenced, and 25 informative nucleotide substitutions in the *AdhC* sequences, and assuming that levels of informative characters are constant across the chloroplast genome, over 40 kb of noncoding cpDNA would have to be sequenced to obtain an equivalent number of informative nucleotide substitutions as found in the *AdhC* sequences. This represents 62% (40

238 bp / 64 437 bp) of the unique noncoding complement of the tobacco chloroplast genome (K. Wolfe, University of Dublin, Trinity College, Ireland, personal communication).

Patterns of mutation -- In addition to levels of divergence, issues of alignability are important in selecting a genic or noncoding region for phylogenetic studies. While noncoding sequences generally accumulate nucleotide substitutions at a higher rate than coding sequences, they also appear to accumulate indels at a faster rate, occasionally equaling the rate of nucleotide substitutions (Golenberg et al., 1993; Gielly and Taberlet, 1994b). Because coding regions are constrained to maintain frame, indels occur less frequently, and when they do, they occur in multiples of three (i.e., a codon). Sequence alignment for genic regions, therefore, is usually straightforward, thereby making assessment of positional homology unambiguous. Noncoding regions, on the other hand, experience indel mutations of all lengths and at high frequency, making sequence alignment more problematic in many cases, particularly as more distantly related taxa are included (e.g., Golenberg et al., 1993; Downie, Katz-Downie, and Cho, 1996; Savolainen, Spichiger, and Manen, 1997). Additional confounding factors in assessing homology of mutations include the duplication/deletion of short repeats (or individual nucleotides in a run) via slipped-strand mispairing (Levinson and Gutman, 1987; Golenberg et al., 1993; Cummings, King, and Kellogg, 1994); the potential multiple origin of small inversions that occur in the loop of stem-loop secondary structures (Kelchner and Wendel, 1996); the higher potential for homoplasy due to a functionally reduced number of character states (due to the high AT content of noncoding cpDNA regions), and biased nucleotide substitutions in AT-rich regions (Morton, 1995). The use of coding regions can circumvent these difficulties, but at the cost of reduced levels of variation, at least in cpDNA genes. Nuclear-encoded genes, however, may offer the higher levels of variation desired, with the ease of alignment afforded by coding sequences.

Sequencing vs. restriction site data -- Jansen, Wee, and Millie (in press) have analyzed both the relative utility (in terms of number of characters) and the relative reliability (in terms of CI and RI) of gene sequencing and restriction site studies of cpDNA. They suggest that, for intrageneric comparisons, cpDNA restriction site data are preferable, both because of the greater number of informative characters and because they report that restriction site data are, in general, less homoplasious than sequence data. Their analyses, however, did not address the lower end of the divergence spectrum (as in our study), where analysis of over 1000 cpDNA restriction sites still provided only limited resolution. cpDNA restriction site data are relatively free from problems associated with sequence data such as alignability and length of sequence. Comparison of mapped restriction sites is straightforward (assuming low levels of rearrangement), but

becomes more difficult as taxonomic distance increases (Olmstead and Palmer, 1994; Jansen, Wee, and Millie, in press). Restriction site studies, however, require large amounts of clean DNA and hence, are contraindicated in situations where availability of material is limiting.

Coalescence and intraspecific variation -- Intraspecific genetic variation (i.e., allelic variation) is often observed when more than one accession of a species is sampled for molecular phylogenetic analysis. Two types of variation may be observed and their impacts on phylogenetic reconstruction are profoundly different. First, alleles within species may all be derived from a single ancestral allele present in the species — i.e., alleles coalesce within species. In this case, all intraspecific variation will be autapomorphic and therefore irrelevant for parsimony analysis. On the other hand, allelic variation may transcend species boundaries and therefore gene trees may not be equivalent to species trees simply because alleles may be older than species and multiple alleles can be maintained within a lineage (Pamilo and Nei, 1988; Hudson, 1990; Maddison, 1995; Clegg, 1997; Wendel and Doyle, in press). The probability of concordance between a species tree and a gene tree is dependent on the time (in generations) between speciation events (the greater the number of generations, the higher the probability of recovering the species tree) and population genetic factors such as effective population size and selection. Although phylogenetic analyses of nuclear-encoded genes that have sampled multiple alleles are rare (see Huttley et al., 1997; Clegg, 1997, and references therein), incomplete coalescence has been observed (Buckler and Holtsford, 1996a, b; Gaut and Clegg, 1993; Goloubinoff, Pääbo, and Wilson, 1993; Hanson et al., 1996). Problems of noncoalescence are expected to be most prevalent in species where population genetic parameters promote the maintenance of multiple alleles, for example, large population size, high migration, and outbreeding (Pamilo and Nei, 1988; Hudson, 1990; Maddison, 1995). Populations of *Gossypium* species are primarily small, isolated, and inbred. These observations, in concert with the concordance of the phylogenies estimated from the separate homoeologues and the congruence with previous analyses, suggest to us that lack of coalescence is not an issue for this locus for these taxa. Current studies are underway to assess intraspecific polymorphism and to explicitly test whether or not *Adh* loci coalesce within closely related *Gossypium* species.

Concerted evolution -- Multigene families are often subject to concerted evolution (Arnheim, 1983; Nagylaki, 1984; Walsh, 1987; Sanderson and Doyle, 1992; Elder and Turner, 1995). The ITS regions of nuclear rDNA became widely used as a source of sequence data after it became apparent that concerted evolution homogenizes sequences so that an entire array of tandemly repeated rDNA cistrons evolves as a single "locus" (Arnheim, 1983; Hillis and Dixon,

1991; Elder and Turner, 1995). Exceptions to the apparent rule of intraspecific and intraindividual sequence homogeneity are being discovered with increasing frequency, however, and the implications of these findings can be profound for phylogenetic reconstruction. Three observations that bear on the use of ITS are (1) paralogous loci are not necessarily homogenized by concerted evolution (e.g., Suh et al., 1993); (2) in polyploids, interlocus concerted evolution may serve to homogenize homoeologous rDNA loci so that only a single parental type is retained, and that this may occur differentially toward either parental type in different descendant lineages (Wendel, Schnabel, and Seelanan, 1995b; but see Waters and Schaal, 1996); and (3) rDNA pseudogenes may persist within the genome and may be preferentially sampled by PCR (Buckler and Holtsford, 1996a, b; Buckler, Ippolito, and Holtsford, 1997; Seelanan and Wendel, unpublished data). All three of the above phenomena may give rise to incongruence between the gene tree and the organismal tree, despite a well-resolved and robustly supported gene tree.

While interlocus gene conversion and recombination have been observed for low-copy nuclear-encoded gene families in plants (e.g., actins, Moniz de Sá and Drouin, 1996; heat-shock proteins, Waters, 1995; *rbcS*, Meagher, Berry-Lowe, and Rice, 1989; glutamine synthetase, Walker et al., 1995) the frequency of these events may depend on sequence conservation between paralogues (e.g., Walsh, 1987). Clearly, gene families that retain a large number of loci with strong sequence homologies are more likely to undergo interlocus concerted evolution and/or recombination than are smaller, more divergent gene families.

In our Southern hybridization experiments we used an *AdhC*-specific probe under high stringency conditions (65°C, 0.1 X SSC/0.5% SDS wash) and detected a single hybridizing band with multiple enzyme digestions for diploid taxa (data not shown) with the exception of *G. raimondii* (which showed a multibanded digestion pattern), and two hybridizing bands in the tetraploids. These Southern hybridization data, the recovery of two identical, paralogous gene trees, the genetic mapping data, and the high degree of sequence divergence between *Gossypium Adh* loci (16-25% in exons, introns are unalignable, Small and Wendel, unpublished data) provide strong evidence that homoeologues were sampled in the allotetraploids and that these sequences have been free from interlocus concerted evolution.

Conclusions -- For phylogenetic analysis to accurately reconstruct organismal history (i.e., the species tree), orthologous sequences need to be compared (Wendel and Doyle, in press). For this reason, among others, plant molecular systematics have relied primarily on cpDNA data because the chloroplast genome is nonrecombinant, generally uniparentally inherited, and "single copy." Because nuclear-encoded genes usually exist in gene families, each member of which

exists in a minimum of two copies (in diploids), and because these multiple copies may experience recombination and gene conversion, demonstration of orthology is more complex. Methods for establishing orthology (whether explicitly stated or implied) vary considerably and include criteria such as overall sequence similarity; monophyly and systematic content — i.e., reconstruction of the expected phylogeny (Gaut et al., 1996); tissue specificity (Doyle, 1991); Southern hybridization data (Matthews and Sharrock, 1996); and most convincingly, comparative genetic mapping data (Zhu et al., 1995; Cronn and Wendel, unpublished data; this paper). These data are not always available or readily obtainable, but inferences of orthology may be facilitated with only a modest investment of effort by Southern hybridization experiments conducted using locus-specific probes and multiple enzyme digestions.

By isolating and analyzing orthologous nuclear genes and a number of different cpDNA regions, we have shown that mutation rates in noncoding cpDNA do not appear high enough to provide sufficient phylogenetic information to resolve relationships of this recently radiated group of tetraploid cottons, despite sequencing over 6 kb of noncoding cpDNA. Consequently, it is difficult to draw conclusions regarding the relative utility of the various cpDNA noncoding regions used. It is clear, however, that levels of divergence vary among noncoding cpDNA sequences (as pointed out for cpDNA introns by Downie, Katz-Downie, and Cho, 1996) and our analyses tentatively identify the *rpl16* intron and the *trnT-trnL* intergenic spacer as among the fastest evolving cpDNA regions (Table 4); this agrees with Downie, Katz-Downie, and Cho, (1996) who suggested that *rpl16* should be the fastest evolving cpDNA intron.

As an alternative source of phylogenetic evidence, orthologous, low-copy, nuclear-encoded loci such as *AdhC* in *Gossypium*, may be isolated, and may exhibit mutation rates up to six times higher than cpDNA noncoding sequences (Fig. 7). The use of nuclear-encoded genes for phylogeny reconstruction has both advantages and limitations. Primary among the advantages are the higher mutation rates and the ability to analyze large regions of sequence with interspersed coding and noncoding regions. The limitations, however, need to be considered. Demonstration of orthology among sequences is imperative and requires additional experimental effort. In addition, cognizance of issues such as coalescence and concerted evolution are required even when strict orthologues are recovered. Our study provides reason for both encouragement and caution in the continuing quest for additional and more informative tools for phylogenetic analysis in plants.

Acknowledgments

The authors thank S. Downie, B. Gaut, K. Schierenbeck, T. Barkman, B. Simpson and R. Wallace for providing PCR primers; R. Jansen and an anonymous reviewer for comments on an earlier draft of the manuscript; and the National Science Foundation for financial support (to JFW).

Literature Cited

ARNHEIM, N. 1983. Concerted evolution of multigene families. *In* M. Nei and R. K. Koehn [eds.], Evolution of genes and proteins, 38-61. Sinauer, Sunderland, MA.

BAUM, D. A., R. SMALL, AND J. F. WENDEL. In press. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology*.

BAYER, R. J., L. HUFFORD, AND D. E. SOLTIS. 1996. Phylogenetic relationships in Sarraceniaceae based on rbcL and ITS sequences. *Systematic Botany* 21: 121-134.

BÖHLE, U.-R., H. HILGER, R. CERFF, AND W. F. MARTIN. 1994. Non-coding chloroplast DNA for plant molecular systematics at the infrageneric level. *In* B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle [eds.], Molecular ecology and evolution: approaches and applications. 391-403. Birkhäuser Verlag, Basel.

BÖHLE, U.-R., H. H. HILGER, AND W. F. MARTIN. 1997. Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). *Proceedings of the National Academy of Sciences, USA* 93: 11740-11745.

BOSQUET, J., S. H. STRAUSS, A. H. DOERKSEN, AND R. A. PRICE. 1992. Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. *Proceedings of the National Academy of Sciences, USA* 89: 7844-7848.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 795-803.

BUCKLER, E. S., AND T. P. HOLTSFORD. 1996a. *Zea* systematics: ribosomal ITS evidence. *Molecular Biology and Evolution* 13: 612-622.

—, AND —. 1996b. *Zea* ribosomal repeat evolution and substitution patterns. *Molecular Biology and Evolution* 13: 623-632.

—, A. IPPOLITO, AND T. P. HOLTSFORD. 1997. The evolution of ribosomal DNA: divergent paralogs and phylogenetic implications. *Genetics* 145: 821-832.

CHASE, M. W., ET AL. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Garden* 80: 528-580.

CLEGG, M. T. 1997. Plant genetic diversity and the struggle to measure selection. *Journal of Heredity* 88: 1-7.

—, B. S. GAUT, G. H. LEARN, JR., AND B. R. MORTON. 1994. Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences, USA* 91: 6795-6801.

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. *Journal of Molecular Evolution* 42: 685-705.

CUMMINGS, M. P., L. M. KING, AND E. A. KELLOGG. 1994. Slipped-strand mispairing in a plastid gene: *rpo*C2 in grasses (Poaceae). *Molecular Biology and Evolution* 11: 1-8.

DEJOODE, D. R., AND J. F. WENDEL. 1992. Genetic diversity and origin of the Hawaiian islands cotton, Gossypium tomentosum. *American Journal of Botany* 79: 1311-1319.

DEMESURE, B., B. COMPS, AND R. J. PETIT. 1996. Chloroplast DNA phylogeography of the common beech (*Fagus sylvatica* L.) in Europe. *Evolution* 50: 2515-2520.

DICKIE, S. L. 1996. Phylogeny and evolution in the Subfamily Opuntioideae (Cactaceae): insights *rpl*16 intron sequence variation. Master's thesis, Iowa State University, Ames, IA.

DON, R. H., P. T. COX, B. J. WAINWRIGHT, K. BAKER, AND J. S. MATTICK. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research* 19: 4008.

DOWNIE, S. R., D. S. KATZ-DOWNIE, AND K.-J. CHO. 1996. Phylogenetic analysis of Apiaceae subfamily Apioideae using nucleotide sequences from the chloroplast *rpo*C1 intron. *Molecular Phylogenetics and Evolution* 6: 1-18.

DOYLE, J. J. 1991. Evolution of higher-plant glutamine synthetase genes: tissue specificity as a criterion for predicting orthology. *Molecular Biology and Evolution* 8: 366-377.

ELDER, J. F., JR., AND B. J. TURNER. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology* 70: 297-320.

ENDRIZZI, J. E., E. L. TURCOTTE, AND R. J. KOHEL. 1985. Genetics, cytology, and evolution of *Gossypium*. *Advances in Genetics* 23: 271-375.

EYRE-WALKER, A., AND B. S. GAUT. 1997. Correlated rates of synonymous site evolution across plant genomes. *Molecular Biology and Evolution* 14: 455-460.

FREELING, M., AND D. C. BENNETT. 1985. Maize *Adh1*. *Annual Review of Genetics* 19: 297-323.

FRYXELL, P. A. 1992. A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). *Rheedea* 2: 108-165.

GAUT, B. S., AND M. T. CLEGG. 1991. Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. *Proceedings of the National Academy of Science, USA* 88: 2060-2064.

—, AND —. 1993. Molecular evolution of the *Adh1* locus in the genus *Zea*. *Proceedings of the National Academy of Science, USA* 90: 5095-5099.

—, S. V. MUSE, W. D. CLARK, AND M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. *Journal of Molecular Evolution* 35: 292-303.

—, —, AND M. T. CLEGG. 1993. Relative rates of nucleotide substitution in the chloroplast genome. *Molecular Phylogenetics and Evolution* 2: 89-96.

—, 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*. *Proceedings of the National Academy of Sciences, USA* 93: 10274-10279.

GIELLY, L., AND P. TABERLET. 1994a. Chloroplast DNA polymorphism at the intrageneric level and plant phylogenies. *Comptes Rendus des Seances, Academie des Sciences (Paris); Serie III Sciences de la vie/Life sciences* 317: 685-692.

—, AND —. 1994b. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbc*L sequences. *Molecular Biology and Evolution* 11: 769-777.

—, AND —. 1996. A phylogeny of the European gentians inferred from chloroplast *trn*L (UAA) intron sequences. *Botanical Journal of the Linnean Society* 120: 57-75.

—, Y.-M. YUAN, P. KÜPFER, AND P. TABERLET. 1996. Phylogenetic use of noncoding regions in the genus *Gentiana* L.: chloroplast *trn*L (UAA) intron versus nuclear ribosomal internal transcribed spacer sequences. *Molecular Phylogenetics and Evolution* 5: 460-466.

GOLENBERG, E. M., M. T. CLEGG, M. L. DURBIN, J. DOEBLEY, AND D. P. MA. 1993. Evolution of a noncoding region of the chloroplast genome. *Molecular Phylogenetics and Evolution* 2: 52-64.

GOLOUBINOFF, P., S. PÄÄBO, AND A. C. WILSON. 1993. Evolution of maize inferred from sequence diversity of an *Adh2* gene segment from archaeological specimens. *Proceedings of the National Academy of Sciences, USA* 90: 1997-2001.

HANFSTINGL, U., A. BERRY, E. A. KELLOGG, J. T. COSTA III, W. RÜDIGER, AND F. M. AUSUBEL. 1994. Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? *Genetics* 138: 811-828.

HANSON, M. A., B. S. GAUT, A. O. STEC, S. I. FUERSTENBERG, M. M. GOODMAN, E. H. COE, AND J. F. DOEBLEY. 1996. Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. *Genetics* 143: 1395-1407.

HILLIS, D. M., AND M. T. DIXON. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology* 66: 411-453.

—, AND J. P. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity* 83: 189-195.

—, M. W. ALLARD, AND M. M. MIYAMOTO. 1993. Analysis of DNA sequence data: phylogenetic inference. *Methods in Enzymology* 224: 456-487.

HODGES, S. A., AND M. L. ARNOLD. 1994. Columbines: a geographically widespread species flock. *Proceedings of the National Academy of Sciences, USA* 91: 5129-5132.

HUDSON, R. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7: 1-44.

HUTTLEY, G. A., M. L. DURBIN, D. E. GLOVER, AND M. T. CLEGG. 1997. Nucleotide polymorphism in the chalcone synthase-A locus and evolution of the chalcone synthase multigene family of common morning glory *Ipomoea purpurea*. *Molecular Ecology* 6: 549-558.

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.

JANSEN, R. K., J. L. WEE, AND D. MILLIE. In press. Comparative utility of chloroplast DNA restriction site and DNA sequence data for phylogenetic studies in plants. *In* D. Soltis, P. Soltis, and J. Doyle [eds.], Molecular systematics of plants, 2d ed. Chapman and Hall, New York, NY.

JOHNSON, D. A., AND J. HATTORI. 1996. Analysis of a hotspot for deletion formation within the intron of the chloroplast *trn*I gene. *Genome* 39: 999-1005.

JORDAN, W. C., M. W. COURTNEY, AND J. E. NEIGEL. 1996. Low levels of intraspecific genetic variation at a rapidly evolving chloroplast DNA locus in North American duckweeds (Lemnaceae). *American Journal of Botany* 83: 430-439.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. *In* H. N. Munro [ed.] Mammalian Protein Metabolism, 21-132. Academic Press, New York, NY.

KELCHNER, S. A., AND J. F. WENDEL. 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Current Genetics* 30: 259-262.

—, AND L. G. CLARK. 1997. Molecular evolution and phylogenetic utility of the chloroplast *rpl16* intron in *Chusquea* and the Bambusoideae (Poaceae). *Molecular Phylogenetics and Evolution* 8: 385-397.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

KITA, Y., K. UEDA, AND Y. KADOTA. 1995. Molecular phylogeny and evolution of the Asian *Aconitum* subgenus *Aconitum* (Ranunculaceae). *Journal of Plant Research* 108: 429-442.

KUMAR, S., K. TAMURA, AND M. NEI. 1993. MEGA: Molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State University, University Park, PA.

LEVINSON, G., AND G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203-221.

LISTON, A. 1992. Variation in the chloroplast genes *rpo*C1 and *rpo*C2 of the genus *Astragalus* (Fabaceae): evidence from restriction site mapping of a PCR-amplified fragment. *American Journal of Botany* 79: 953-961.

MADDISON, W. 1995. Phylogenetic histories within and among species. *In* P. C. Hoch and A. G. Stephenson [eds.], Experimental and molecular approaches to plant biosystematics, *Monographs in Systematic Botany from the Missouri Botanical Garden*, vol. 53, 273-287.

MANEN, J.-F., AND A. NATALI. 1995. Comparison of the evolution of ribulose-1, 5-bisphosphate carboxylase (*rbcL*) and *atpB-rbcL* noncoding spacer sequences in a recent plant group, the tribe Rubieae (Rubiaceae). *Journal of Molecular Evolution* 41: 920-927.

—, V. SAVOLAINEN, AND P. SIMON. 1994. The *atpB* and *rbcL* promoters in plastid DNAs of a wide dicot range. *Journal of Molecular Evolution* 38: 577-582.

MATTHEWS, S., AND R. A. SHARROCK. 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. *Molecular Biology and Evolution* 13: 1141-1150.

MEAGHER, R. B., S. BERRY-LOWE, AND K. RICE. 1989. Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. *Genetics* 123: 845-863.

MENDENHALL, M. 1994. Phylogeny of *Baptisia* and *Thermopsis* (Leguminosae) as inferred from chloroplast DNA and nuclear ribosomal DNA sequences, secondary chemistry, and morphology. Ph.D. dissertation, University of Texas, Austin, TX.

MILLAR, A. A., AND E. S. DENNIS. 1996. The alcohol dehydrogenase genes of cotton. *Plant Molecular Biology* 31: 897-904.

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*. *Molecular Biology and Evolution* 13: 433-436.

MONIZ DE SÁ, M., AND G. DROUIN. 1996. Phylogeny and substitution rates of angiosperm actin genes. *Molecular Biology and Evolution* 13:1198-1212.

MORTON, B. R. 1995. Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proceedings of the National Academy of Sciences, USA* 92: 9717-9721.

—, AND M. T. CLEGG. 1993. A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbcL* in the grass family (Poaceae). *Current Genetics* 24: 357-365.

—, B. S. GAUT, AND M. T. CLEGG. 1996. Evolution of alcohol dehydrogenase genes in the palm and grass families. *Proceedings of the National Academy of Sciences, USA* 93: 11735-11739.

NAGYLAKI, T. 1984. Evolution of multigene families under interchromosomal gene conversion. *Proceedings of the National Academy of Sciences, USA* 81: 3796-3800.

NATALI, A., J.-F. MANEN, AND F. EHRENDORFER. 1995. Phylogeny of the Rubiaceae-Rubioideae, in particular the tribe Rubieae: evidence from a non-coding chloroplast DNA sequence. *Annals of the Missouri Botanical Garden* 82: 428-439.

NEI, M., AND T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3: 418-426.

OLMSTEAD, R. G., AND J. D. PALMER. 1994. Chloroplast DNA systematics: a review of methods and data analysis. *American Journal of Botany* 81: 1205-1224.

—, AND J. A. SWEERE. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* 43: 467-481.

PALMER, J. D. 1991. Plastid chromosomes: structure and evolution. *Cell Culture and Somatic Cell Genetics of Plants* 7A: 5-53.

PAMILO, P., AND M. NEI. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568-583.

PANERO, J. L. AND R. K. JANSEN. 1997. Chloroplast DNA restriction site study of *Verbesina* (Asteraceae: Heliantheae). *American Journal of Botany* 84: 382-392.

PATERSON, A. H., C. L. BRUBAKER, AND J. F. WENDEL. 1993. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Molecular Biology Reporter* 11: 122-127.

PERCY, R. G., AND J. F. WENDEL. 1990. Allozyme evidence for the origin and diversification of *Gossypium barbadense* L. *Theoretical and Applied Genetics* 79: 529-542.

REINISCH, A. J., J. DONG, C. L. BRUBAKER, D. M. STELLY, J. F. WENDEL, AND A. H. PATERSON. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. *Genetics* 138: 829-847.

RIESEBERG, L. H., M. A. HANSON, AND C. T. PHILBRICK. 1992. Androdioecy is derived from dioecy in Datiscaceae: evidence from restriction site mapping of PCR-amplified chloroplast DNA fragments. *Systematic Botany* 17: 324-336.

SAMBROOK, J., E. F. FRITSCH, AND T. MANIATIS. 1989. Molecular cloning, a laboratory manual 2d ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

SANDERSON, M. J., AND J. J. DOYLE. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. *Systematic Biology* 41: 4-17.

SANG, T., D. J. CRAWFORD, AND T. F. STUESSY. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *American Journal of Botany* 84:1120-1136.

—, M. J. DONOGHUE, AND D. ZHANG. 1997. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. *Molecular Biology and Evolution* 14: 994-1007.

SAVOLAINEN, V., J. F. MANEN, E. DOUZERY, AND R. SPICHIGER. 1994. Molecular phylogeny of families related to Celastrales based on rbcL 5' flanking sequences. *Molecular Phylogenetics and Evolution* 3: 27-37.

—, R. SPICHIGER, AND J.-F. MANEN. 1997. Polyphyletism of Celastrales deduced from a chloroplast noncoding DNA region. *Molecular Phylogenetics and Evolution* 7: 145-157.

SEELANAN, T., A. SCHNABEL, AND J. F. WENDEL. 1997. Congruence and consensus in the cotton tribe (Malvaceae). *Systematic Botany* 22: 259-290.

SHINOZAKI, K., ET AL. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its organization and expression. *EMBO Journal* 5: 2043-2049.

SOLTIS, D. E., P. S. SOLTIS AND J. J. DOYLE. In press. Molecular systematics of plants, 2d edition. Chapman and Hall, New York, NY.

SOLTIS, P. S., D. E. SOLTIS, S. G. WELLER, A. K. SAKAI, AND W. L. WAGNER. 1996. Molecular phylogenetic analysis of the Hawaiian endemics *Schiedea* and *Alsinidendron* (Caryophyllaceae). *Systematic Botany* 21: 365-379.

STEELE, K. P., AND R. VILGALYS. 1994. Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the plastid gene *mat*K. *Systematic Botany* 19: 126-142.

SUH, Y., L. B. THIEN, H. E. REEVE, AND E. A. ZIMMER. 1993. Molecular evolution and phylogenetic implications of internal transcribed spacer sequences of ribosomal DNA in Winteraceae. *American Journal of Botany* 80: 1042-1055.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1.1. Computer program distributed by the Illinois Natural History Survey, Champaign, IL.

TABERLET, P., L. GIELLY, G. PAUTOU, AND J. BOUVET. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17: 1105-1109.

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

VAN HAM, R. C. H. J., H. HART, T. H. M. MES, AND J. M. SANDBRINK. 1994. Molecular evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species. *Current Genetics* 25: 558-566.

WALKER, E. L., N. F. WEEDEN, C. B. TAYLOR, P. GREEN, AND G. M. CORUZZI. 1995. Molecular evolution of duplicate copies of genes encoding cytosolic glutamine synthetase in *Pisum sativum*. *Plant Molecular Biology* 29: 1111-1125.

WALSH, J. B. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? *Genetics* 117: 544-557.

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

—, AND B. A. SCHAAL. 1996. Biased gene conversion is not occurring among rDNA repeats in the *Brassica* triangle. *Genome* 39: 150-154.

WENDEL, J. F. 1989. New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Sciences, USA* 86: 4132-4136.

—. 1995. Cotton. *In* J. Smartt and N. W. Simmonds [eds.], Evolution of crop plants, 2d edition, 358-366. Longman Scientific & Technical, Essex.

—, AND J. J. DOYLE. In press. Phylogenetic incongruence: window into genome history and molecular evolution. *In* D. Soltis, P. Soltis and J. Doyle [eds.], Molecular systematics of plants, 2d edition. Chapman and Hall, New York, NY.

—, AND R. G. PERCY. 1990. Allozyme diversity and introgression in the Galapagos Islands endemic *Gossypium darwinii* and its relationship to continental *G. barbadense*. *Biochemical Systematics and Ecology* 18: 517-528.

—, AND V. A. ALBERT. 1992. Phylogenetics of the cotton genus (*Gossypium*): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Systematic Botany* 17: 115-143.

—, C. L. BRUBAKER, AND T. SEELANAN. In press. The origin and evolution of *Gossypium*. *In* J. Stewart, D. Oosterhuis, and J. Heitholt [eds.], Cotton physiology, Book II. Cotton Foundation, Memphis, TN.

—, A. SCHNABEL, AND T. SEELANAN. 1995a. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Molecular Phylogenetics and Evolution* 4: 298-313.

—, —, AND —. 1995b. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences USA* 92: 280-284.

—, R. ROWLEY, AND J. MCD. STEWART. 1994. Genetic diversity in and phylogenetic relationships of the Brazilian endemic cotton, *Gossypium mustelinum* (*Malvaceae*). *Plant Systematics and Evolution* 192: 49-59.

WOLF, P. G., R. A. MURRAY, AND S. D. SIPES. 1997. Species-independent, geographical structuring of chloroplast DNA haplotypes in a montane herb *Ipomopsis* (Polemoniaceae). *Molecular Ecology* 6: 283-291.

WOLFE, A. D., W. J. ELISENS, L. E. WATSON, AND C. W. DEPAMPHILIS. 1997. Using restriction-site variation of PCR-amplified cpDNA genes for phylogenetic analysis of Tribe Cheloneae (Scrophulariaceae). *American Journal of Botany* 84: 555-564.

WOLFE, K. H., W.-H. LI, AND P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054-9058.

—, P. M. SHARP, AND W.-H. LI. 1989. Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution* 29: 208-211.

ZHU, T., L. SHI, J. J. DOYLE, AND P. KEIM. 1995. A single nuclear locus phylogeny of soybean based on DNA sequence. *Theoretical and Applied Genetics* 90: 991-999.

Table 1. Plant materials. All voucher specimens are deposited at ISC. Voucher abbreviations are as follows: TS = Tosak Seelanan, JFW & TDC = J. F. Wendel and T. D. Couch

Taxon	Accession	Voucher
C-genome diploid		
Gossypium robinsonii F. Mueller	AZ-50	TS 12
D-genome diploid		
Gossypium raimondii Ulbrich	#436	JFW & TDC 127
A-genome diploid		
Gossypium arboreum L.	A_2-74	JFW & TDC 312
AD-genome tetraploids		
Gossypium hirsutum L.	"Palmeri"	JFW & TDC 632
Gossypium barbadense L.	K101	JFW & TDC 612
Gossypium tomentosum Nuttall ex Seemann	WT936	JFW & TDC 621
Gossypium mustelinum Miers ex Watt	W400	JFW & TDC 622
Gossypium darwinii Watt	WB1215	JFW & TDC 620

Table 2. Regions studied, PCR primer sequences, and Genbank accession numbers.

Region	Primer sequences (written 5′ to 3′).	References	Genbank accession numbers
atpB-rbcL spacer	*atpB*: GTG GAA ACC CCG GGA CGA GAA GTA GT *rbcL*: ACT TGC TTT AGT TTC TGT TTG TGG TGA	Hodges and Arnold, 1994	AF031445 - AF031450
trnL-trnF spacer	E: GGT TCA AGT CCC TCT ATC CC F: ATT TGA ACT GGT GAC ACG AG	Taberlet et al., 1991	AF031439 - AF031444
trnT-trnL spacer	A: CAT TAC AAA TGC GAT GCT CT B: TCT ACC GAT TTC GCC ATA TC *trnT*-I: CTG ACT CCA TTT TTA TTT TC	Taberlet et al., 1991 this paper	AF031433 - AF031438
accD-psaI spacer	*accD*-769F: GGA AGT TTG AGC TTT ATG CAA ATG G *psaI*-75R: AGA AGC CAT TGC AAT TGC CGG AAA *accD*I: GGG CTT TGA CTT TGT GAC	T. Barkman and B. Simpson, University of Texas, Austin, personal communication	AF031580 - AF031585
ndhA intron	*ndhA*-F: GGW CTT CTY ATG KCR GGA TAT RGM TC *ndhA*-R: CTG YGC TTC MAC TAT ATC AAC TGT AC *ndhA*-I: ATT CTG CTT TCG GAT CTG	this paper	AF031574 - AF031579
rpl16 intron	F71: GCT ATG CTT AGT GTG TGA CTC GTT G R1661: CGT ACC CAT ATT TTT CCA CCA CGA C R1516: CCC TTC ATT CTT CCT CTA TGT TG	Jordan, Courtney, and Neigel, 1996; Kelchner and Wendel, 1996	AF031451 - AF031456
rpoC1 intron	5′*rpoC1* exon: GGT CTT CCT AGY TAY ATH GC *rpoC1* exon 2: ATT TCA TAT TCG AAY AAN CC	Downie, Katz-Downie, and Cho, 1996	AF031457 - AF031462
AdhC	P1: CTG CKG TKG CAT GGG ARG CAG GGA AGC C P2: GCA CAG CCA CAC CCC AAC CCT G ADHX6-2: TCA ATA CCA ATG ATC CTA GAA ADHX4-1: TCA TGT TCT CCC TAT CTT CAC ADHX8-2: GAA ACC ATG GCC TGG GTG	K. Schierenbeck, California State University, Fresno, personal communication (P1, P2); this paper (ADHX6-2, 4-1, and 8-2)	AF036567 - AF036579

Table 3. PCR amplification conditions.

Region	Denaturation temperature/Time	Annealing temperature/Time	Extension temperature/Time
atpB-rbcL spacer	94°C / 1 min	55° - 50° C / 1 min[a]	72°C / 3 min
trnL-trnF spacer	94°C / 1 min	48° C / 1 min	72°C / 3 min
trnT-trnL spacer	94°C / 1 min	48° C / 1 min	72°C / 3 min
accD-psaI spacer	94°C / 1 min	65° C / 1 min 20 s	72°C / 3 min
ndhA intron	94°C / 1 min 30 s	42° C / 1 min 30 s	72°C / 2 min
rpl16 intron	95°C / 1 min	50° C / 1 min[b]	65°C / 4 min
rpoC1 intron	94°C / 1 min	42° C / 1 min	72°C / 3 min
AdhC	94°C / 1 min	50° C / 1 min	72°C / 2 min

[a]Touchdown PCR (Don et al., 1991); initial annealing temperature of 55°C, followed by a 0.5° reduction in annealing temperature every cycle for ten cycles, followed by an additional 20 cycles with a 50°C annealing temperature.

[b]Following a 50°C annealing step for 1 min the temperature was ramped to 65°C by 1°/8 s.

Table 4. Characterization of cpDNA sequences (coding and noncoding).

Regions analyzed	Aligned length (bp)[a]	GC content	Divergence from A diploid outgroup[b]	Divergence within tetraploids[c]	Ts:Tv[d]	Substitutions[e]	Indels[f]
Intergenic spacers							
atpB-rbcL	976 (18)	28.3%	0.20%	0.20%	5:1	6(0)	5(0)
trnL-trnF	437 (42)	33.7%	0.12%	0.24%	1:2	3(0)	0
trnT-trnL	1394 (22)	22.9%	0.96%	0.49%	0.8:1	23(2)	6(2)
accD-psaI	1146 (390)	29.3%	0.40%	0.30%	2:1	12(0)	0
Introns							
ndhA	1140 (82)	31.9%	0.12%	0.04%	1:1	2(0)	0
rpl16	1173 (24)	30.4%	0.34%	0.28%	0:6	6(2)	4(2)
rpoC1	1103 (353)	37.0%	0.00%	0.00%	----	0	0
Total	7369 (931)	30.0%	0.30%	0.20%	0.9:1	52(4)	15(4)

[a]Length of coding sequence in parenthesis.

[b]Calculated as the mean nucleotide percentage difference between sequences from the outgroup (*G. arboreum*) and all ingroup species (gaps treated as missing data).

[c]Calculated as the mean nucleotide percent difference among all pairwise comparisons of sequences from tetraploid species (gaps treated as missing data).

[d]Ratio of transitions to transversions.

[e]Number of nucleotide substitutions; number of potentially phylogenetically informative substitutions in parenthesis.

[f]Number of indels; number of potentially phylogenetically informative indels in parenthesis.

Table 5. Characterization of *Adh* sequences.

			Tetraploid taxa and diploid outgroups								
Region analyzed	Aligned length (bp)	GC content	Divergence from A/D diploid outgroup[a]	Divergence from C diploid outgroup[b]	Divergence within tetraploids[c]	K_s[d]	K_a[d]	K[d]	Ts:Tv[e]	Substitutions[f]	Indels[g]
A (sub)genome											
Exons	798	46.2%	1.0%	1.1%	0.3%	0.009	0.004	0.005	8:1	11(2)	2(1)
Introns	847	32.0%	1.0%	3.2%	0.9%	—	—	0.009	3.3:1	20(5)	2(0)
Total	1645	39.0%	1.0%	2.1%	0.6%	—	—	0.007	4.2:1	31(7)	4(1)
D (sub)genome											
Exons	798	45.4%	1.9%	2.3%	1.4%	0.028	0.013	0.016	7:1	33(12)	1(0)
Introns	865	30.1%	2.6%	5.5%	0.7%	—	—	0.014	1.6:1	30(6)	3(1)
Total	1663	37.5%	2.3%	3.9%	1.1%	—	—	0.015	3.6:1	63(18)	4(1)

[a]Calculated as the mean nucleotide percentage difference between the relevant subgenome outgroup (A - *G. arboreum* or D - *G. raimondii*) and the corresponding sequences from the tetraploid species (gaps treated as missing data).

[b]Calculated as the mean nucleotide percentage difference between the C-genome diploid outgroup (*G. robinsonii*) and sequences from the tetraploid species (gaps treated as missing data).

[c]Calculated as the mean nucleotide percentage difference among all pairwise comparisons of sequences from tetraploid species (gaps treated as missing data).

[d]Nucleotide substitutions among tetraploid taxa and their relevant diploid outgroup. Number of synonymous substitutions per synonymous site (K_s), nonsynonymous substitutions per nonsynonymous site (K_a), and substitutions per site (K) calculated with the Jukes and Cantor (1969) correction for multiple hits. K_s and K_a calculated according to the method of Nei and Gojobori (1986).

[e]Ratio of transitions to transversions among sequences from tetraploid taxa and the relevant subgenome outgroup.

[f]Number of nucleotide substitutions among sequences from tetraploid taxa and the relevant subgenome outgroup; number of potentially phylogenetically informative substitutions in parenthesis.

[g]Number of indels among sequences from tetraploid taxa and the relevant subgenome outgroup; number of potentially phylogenetically informative indels in parenthesis.

[h]Nucleotide substitutions among tetraploid taxa only. K_s, K_a, and K calculated with the Jukes and Cantor (1969) correction for multiple hits. K_s and K_a calculated according to the method of Nei and Gojobori (1986).

Table 5. continued.

Tetraploid taxa only		
K_s^h	K_a^h	K^h
0.008	0.002	0.003
---	---	0.009
---	---	0.006
0.019	0.013	0.014
---	---	0.008
---	---	0.011

Table 6. Results of Tajima (1993) 2D relative rate tests for *Adh* sequences (below diagonal) and number of nucleotide differences between *Adh* sequences (above diagonal). Significantly different rates are denoted by asterisks as follows: * $0.05 > P > 0.01$; ** $0.01 > P > 0.005$; *** $P < 0.005$. A′ and D′ refer to the A and D subgenomes of the tetraploids. In all cases *G. robinsonii* was used as the reference taxon.

Species	1	2	3	4	5	6	7	8	9	10	11	12
1 *G. raimondii* (D)	---	44	36	41	33	38	28	70	62	68	64	64
2 *G. hirsutum* D′		---	22	9	23	24	34	77	69	77	73	71
3 *G. barbadense* D′			---	19	17	4	28	71	63	71	67	65
4 *G. tomentosum* D′				---	20	21	31	76	68	76	72	70
5 *G. mustelinum* D′					---	19	29	68	60	68	64	62
6 *G. darwinii* D′						---	30	73	65	73	69	67
7 *G. arboreum* (A)	**	***	**	***	*	**	---	8	4	8	5	4
8 *G. hirsutum* A′	**	***	***	***	**	***		---	8	12	12	10
9 *G. barbadense* A′	**	***	***	***	***	***			---	13	9	2
10 *G. tomentosum* A′	**	***	***	***	*	***				---	14	15
11 *G. mustelinum* A′	**	***	***	***	**	***					---	11
12 *G. darwinii* A′	*	***	***	***	**	***						---

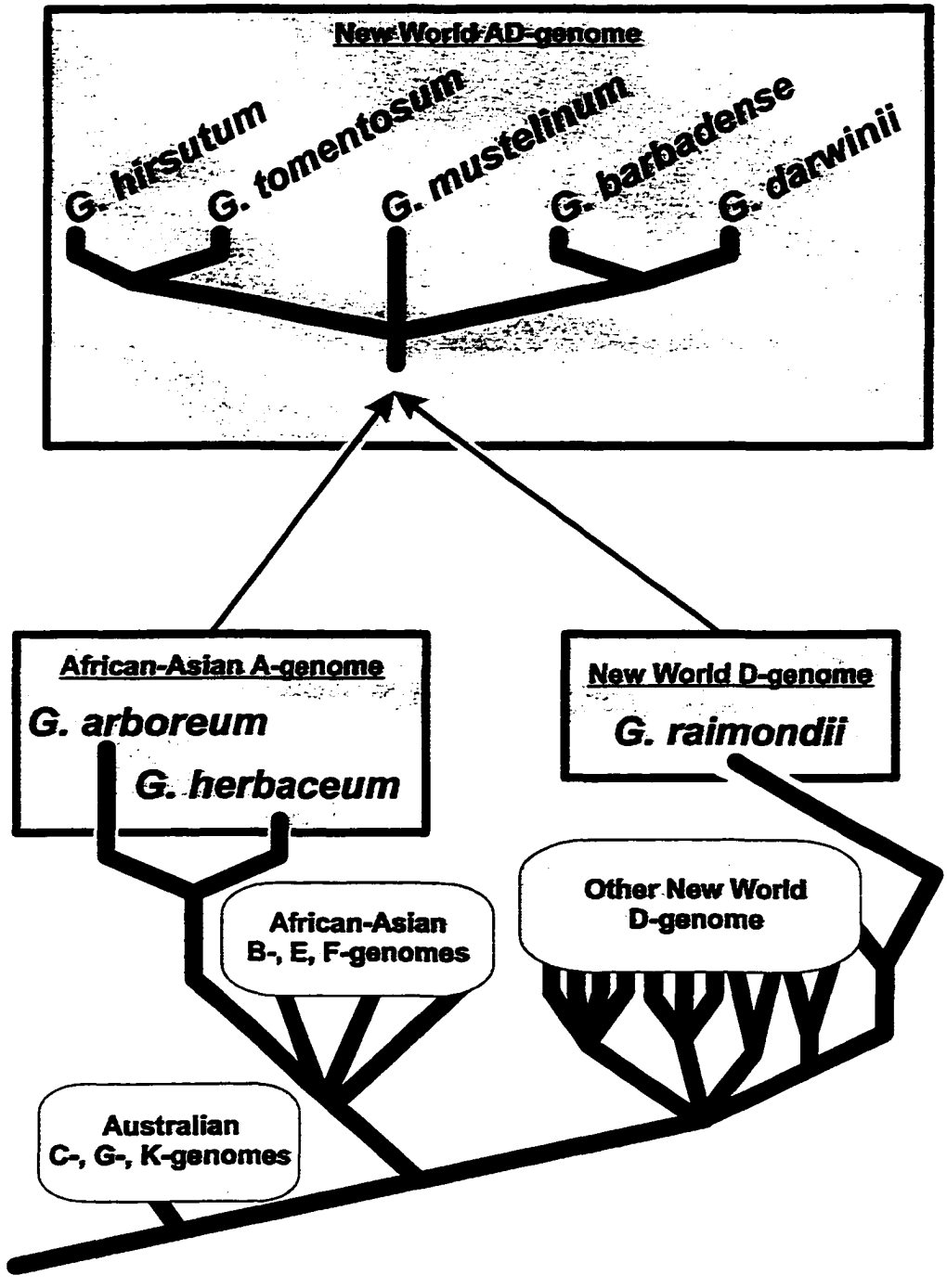


Fig. 1. Phylogenetic hypothesis for intrageneric relationships in *Gossypium*, including the origin of the allotetraploid species. The maternal diploid parent is represented by the extant A-genome species, *G. arboreum* and *G. herbaceum*, while the paternal diploid parent is represented by the extant D-genome species, *G. raimondii*.

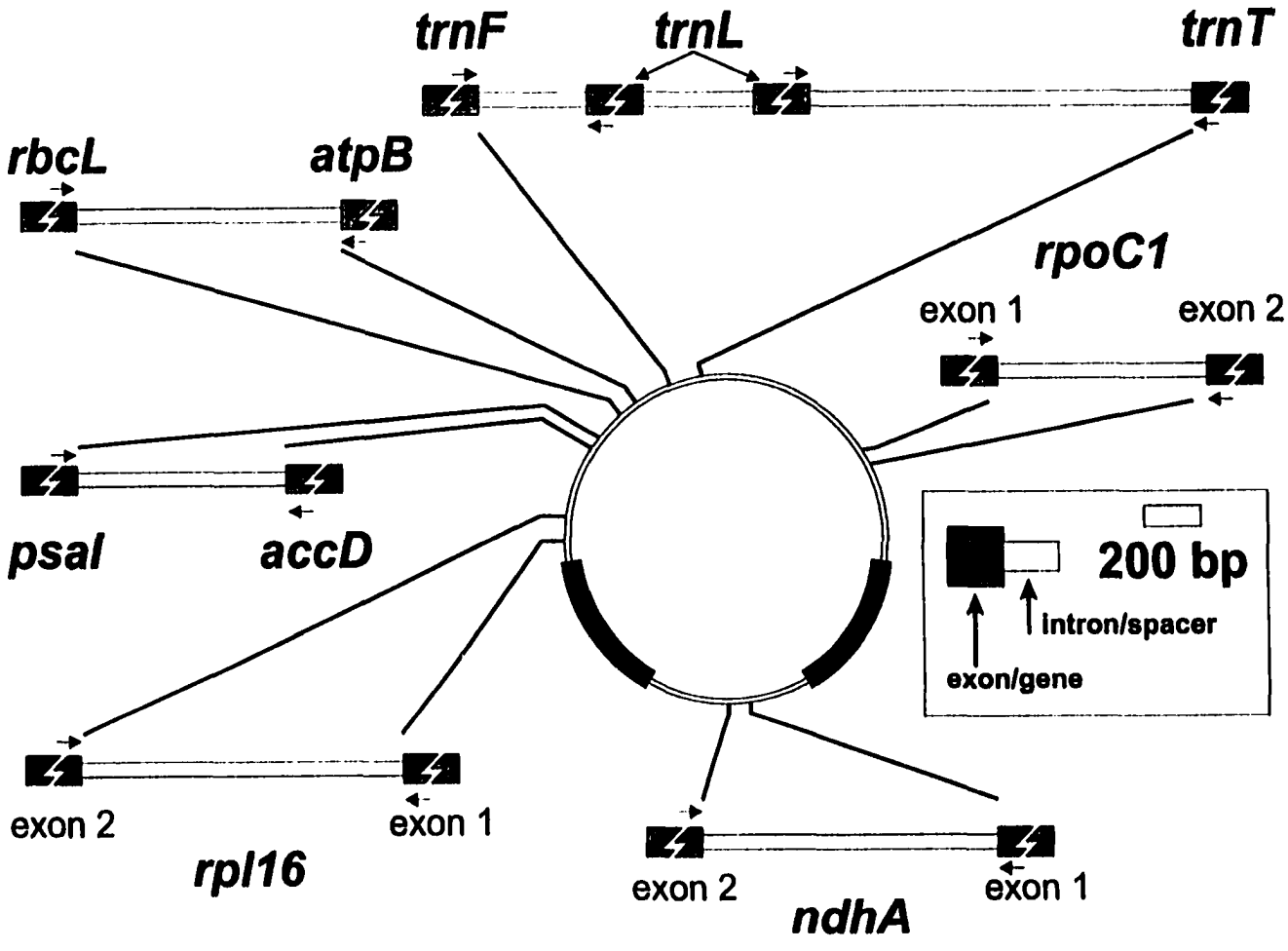


Fig. 2. Chloroplast DNA noncoding regions sampled. The circle represents the chloroplast genome, with shaded regions representing the inverted repeats. Sequenced regions are shown as mapped in the tobacco chloroplast genome (Shinozaki et al., 1986). For each region exons are represented by shaded boxes and are not drawn to scale; introns and spacers are represented by open boxes and are drawn approximately to scale.

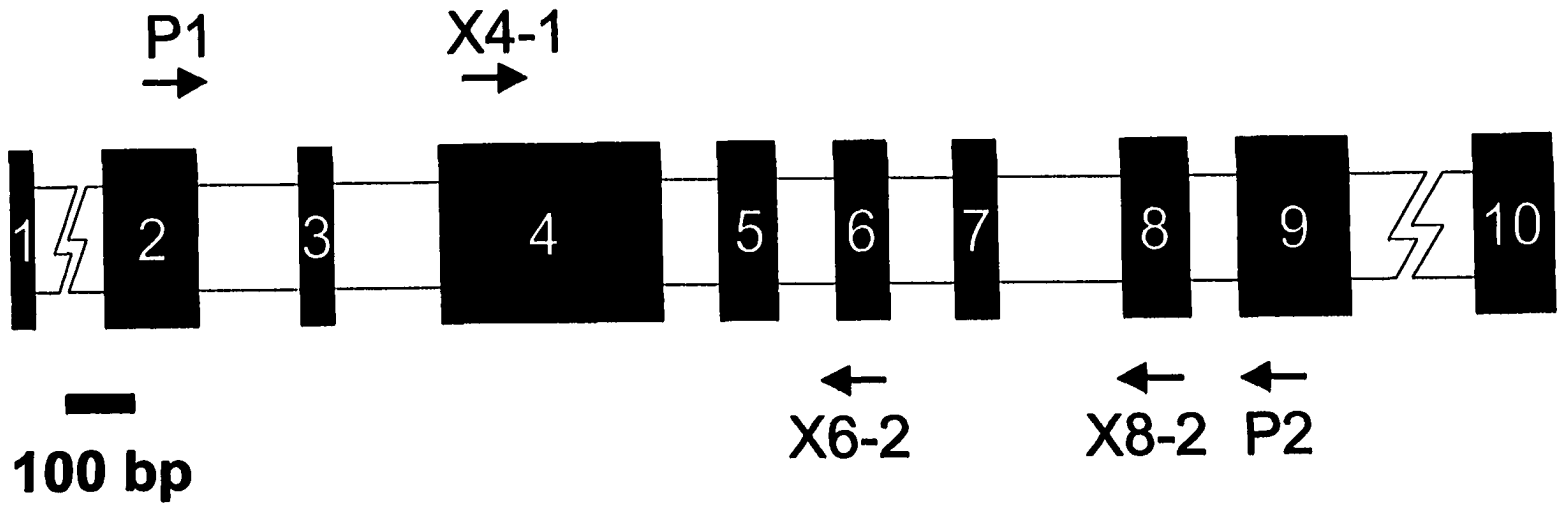


Fig. 3. Schematic representation of the *AdhC* genic region. Exons are represented by numbered and shaded boxes; introns are represented by open boxes. All regions are drawn to scale except introns 1 and 9 for which data are unavailable. Relative positions of the forward PCR primers are shown above the gene, reverse primers below.

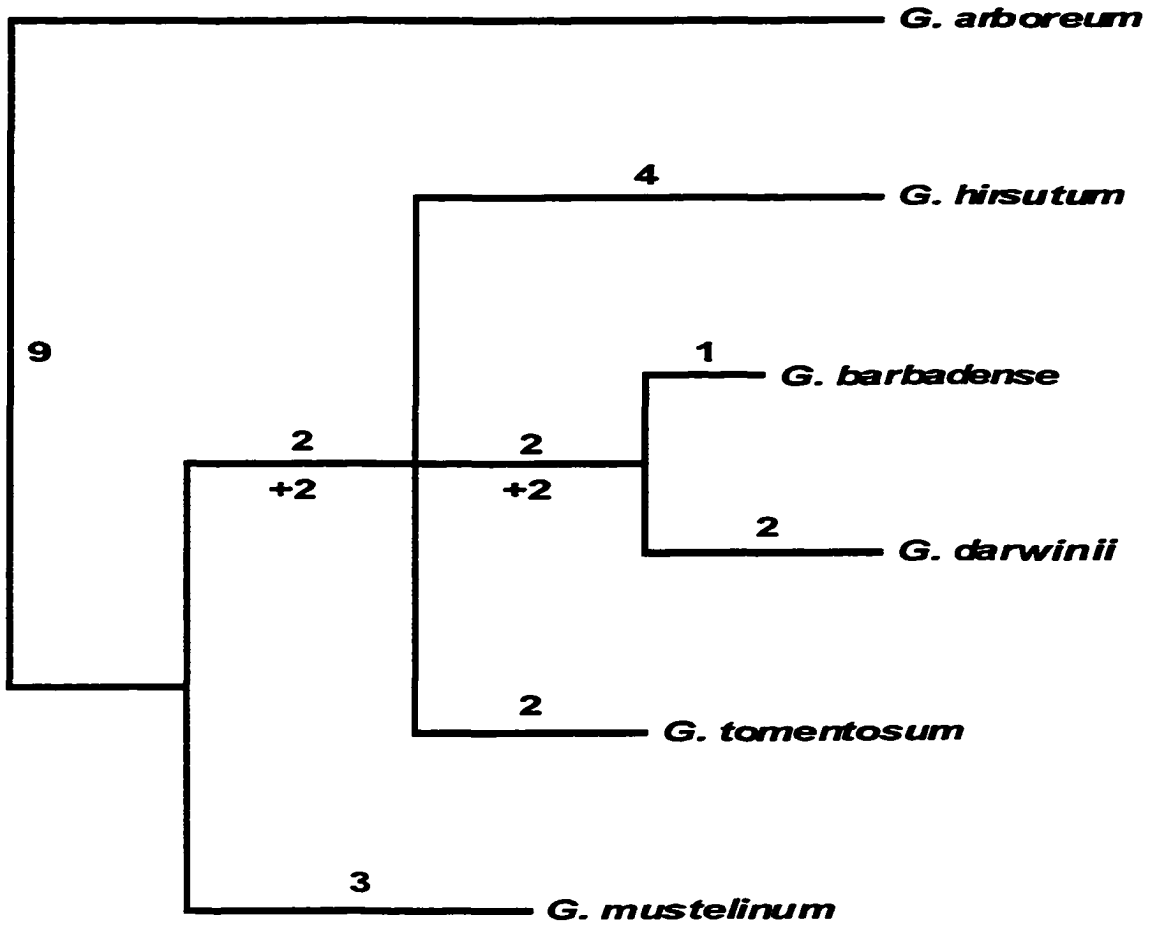


Fig. 4. Single most-parsimonious tree (length = 25, CI = 1.0, RI = 1.0) from analysis of the *trnT-trnL* spacer region. Branch lengths are shown above, and decay values below each branch.

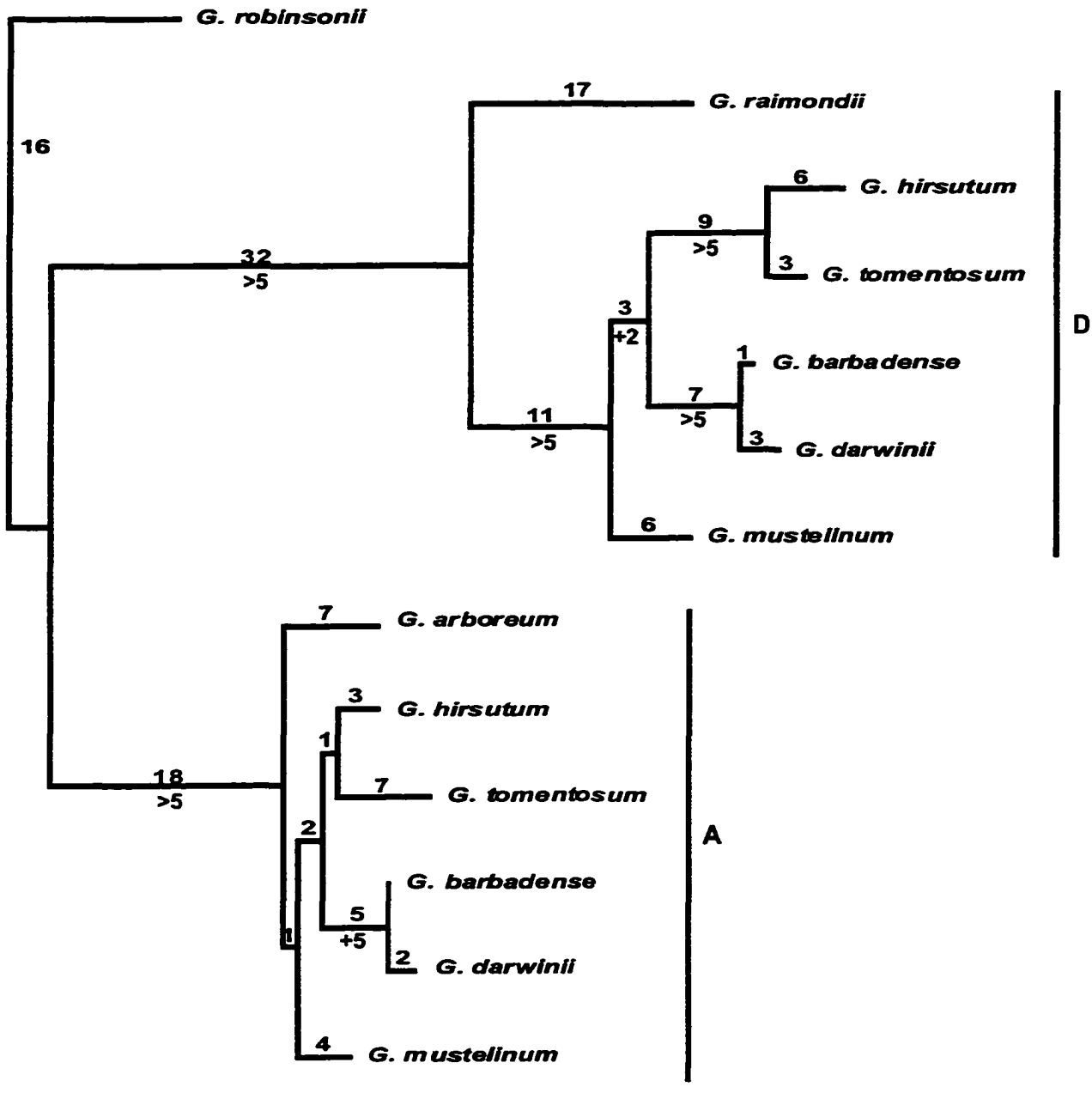


Fig. 5. Single most-parsimonious tree (length = 97, CI = 0.93, RI = 0.98) from analysis of individual *AdhC* sequences. Branch lengths are shown above, and decay values below each branch. Nodes without decay values shown collapse in the strict consensus tree of trees one step longer than the most parsimonious.

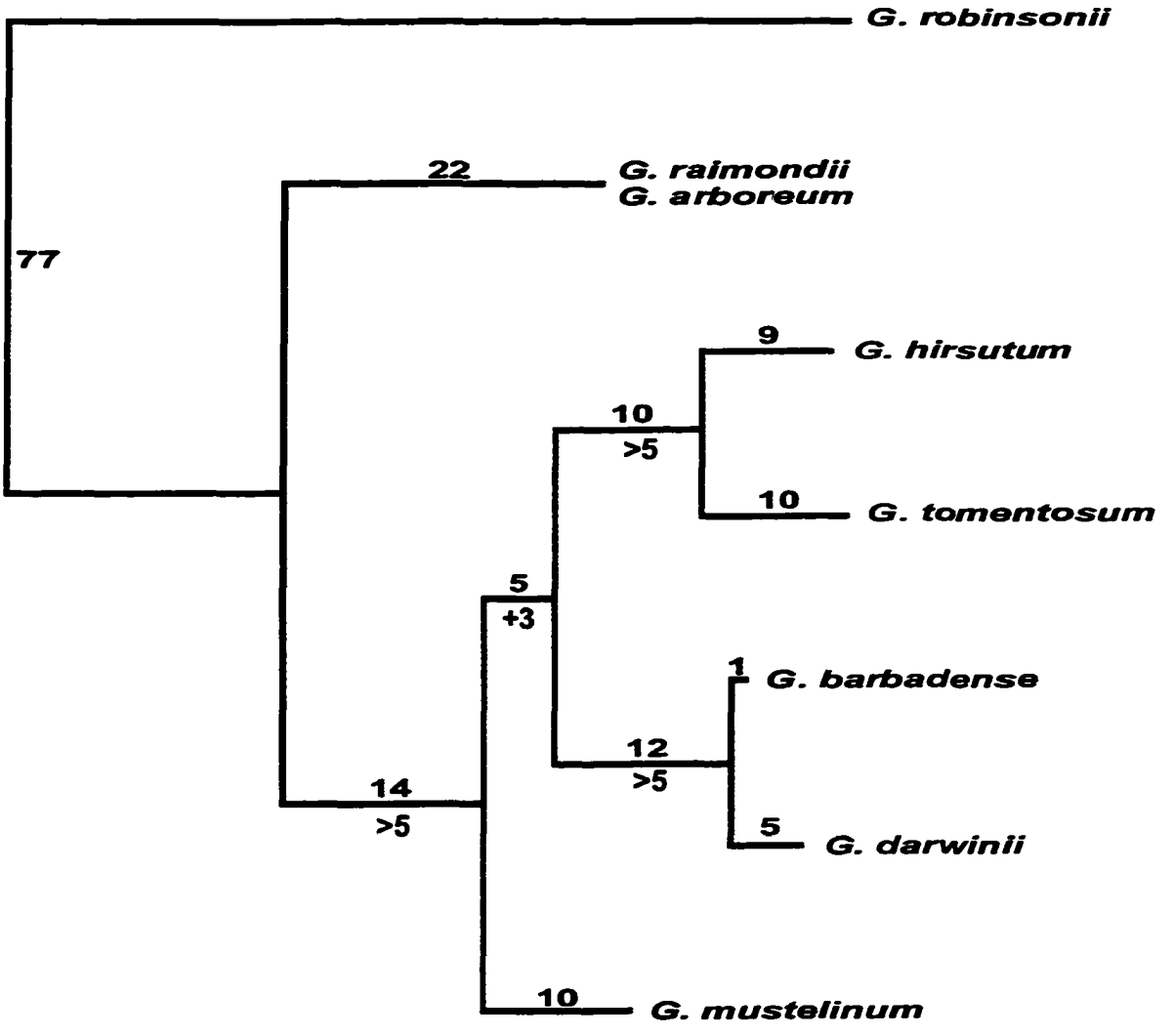


Fig. 6. Single most-parsimonious tree (length = 43, CI = 0.91, RI = 0.91) from analysis of combined *AdhC* data. Branch lengths are shown above, and decay values below each branch.

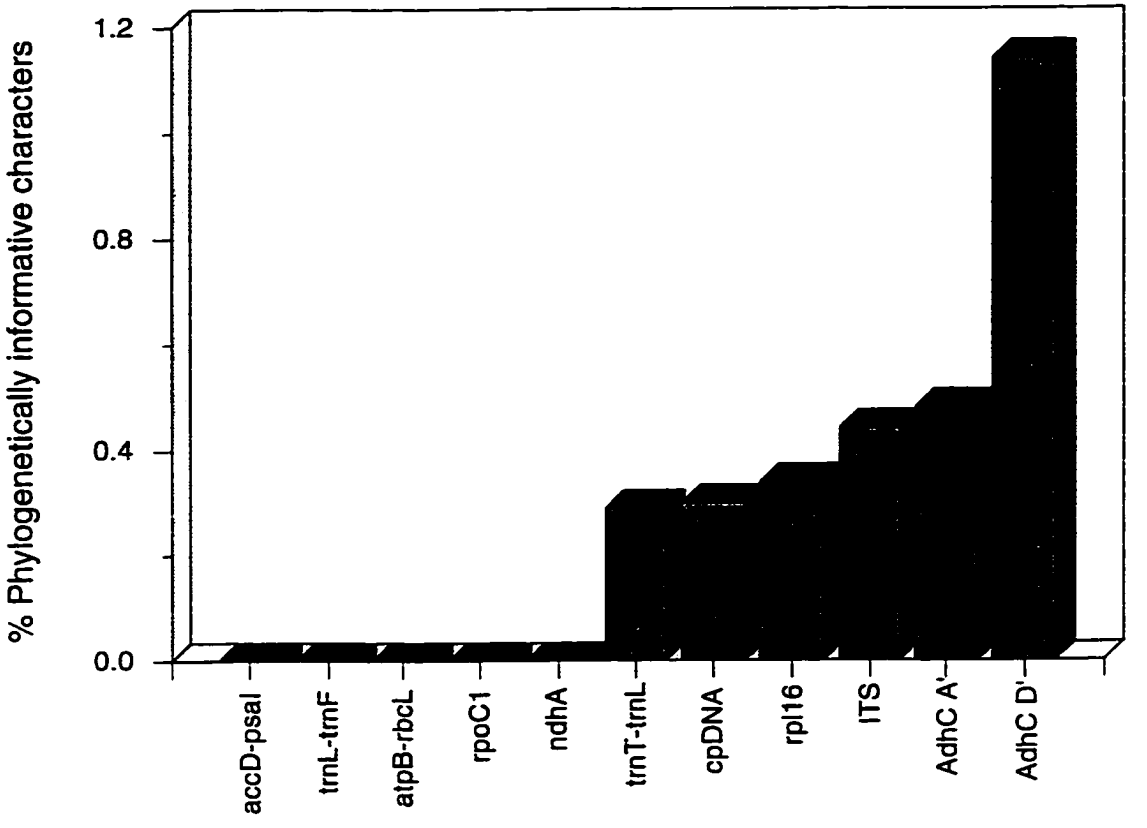


Fig. 7. Percentages of phylogenetically informative characters for several molecular data sets applied to tetraploid *Gossypium*. Number of informative ITS characters were partially extrapolated (see text).

CHAPTER 4. LOW LEVELS OF NUCLEOTIDE DIVERSITY AT HOMOEOLOGOUS *ADH* LOCI IN ALLOTETRAPLOID COTTON (*GOSSYPIUM* L.)

A paper published in *Molecular Biology and Evolution*[1]

Randall L. Small[2], Julie A. Ryburn[2], and Jonathan F. Wendel[2]

Abstract

Levels of genetic diversity within and among populations and species are shaped by both external (population-level) and internal (genomic and genic) evolutionary forces. To address the effect of internal pressures we estimated nucleotide diversity for a pair of homoeologous *Adh* loci in an allotetraploid species, *Gossypium hirsutum*. These data represent the first such estimates for a pair of homoeologous nuclear loci in plants. Estimates of nucleotide diversity for *AdhA* in *Gossypium* are lower than for any plant nuclear gene yet described. This low diversity appears to reflect primarily a history of repeated, severe genetic bottlenecks associated with both speciation and recent domestication, supplemented by an unusually slow nucleotide substitution rate and an autogamous breeding system. While not statistically supportable, the sum of the observations also suggest differential evolutionary dynamics at each of the homoeologous loci.

Key words: alcohol dehydrogenase, cotton, *Gossypium*, homoeology, nucleotide diversity, polyploidy

Introduction

Levels and patterns of genetic diversity vary greatly within and among populations and species. This variation reflects the interplay of myriad historical factors and evolutionary forces, involving external forces such as natural selection, population size and history, gene flow, and breeding system, as well as internal genomic and genetic factors such as recombination, mutation rate, and gene conversion (Aquadro and Begun 1993; Tajima 1993b; Moriyama and Powell 1996; Clegg 1997; Clegg, Cummings, and Durbin 1997; Amos and Harwood 1998). Recent studies have revealed varying patterns of nucleotide diversity within plant species (Gaut and Clegg 1993a, 1993b; Hanfstingl et al. 1994; Hanson et al. 1996; Innan et al. 1996; Miyashita, Innan, and Terauchi 1996; Huttley et al. 1997; Kawabe et al. 1997; Terauchi,

---

[1] Reprinted with permission from Molecular Biology and Evolution, 1999, 16(4): 491-501.

[2] Department of Botany, Iowa State University, Ames, IA 50011.

Terachi, and Miyashita 1997; Bergelson et al. 1998; Cummings and Clegg 1998; Eyre-Walker et al. 1998; Liu, Zhang, and Charlesworth 1998). While these and other studies have yielded a number of insights into the factors that shape naturally occurring variation, in any particular case the evolutionary or historical forces responsible for the diversity patterns observed may be difficult to discern. This is especially true for comparisons between species, where numerous potentially confounding life-history features and population histories may influence both the amount and apportionment of diversity. Allopolyploid species, which contain duplicated genes in the same nucleus, may be particularly useful in isolating potentially relevant internal genetic and genomic factors from external population-level processes. In allopolyploids some external processes (e.g., selection) can differentially effect duplicated genes, while others (e.g., genetic drift, breeding system) are expected to effect duplicated genes equivalently. Thus, examining molecular evolution at duplicated genes in a polyploid allows at least some population-level effects to be ruled out as having contributed to differential evolution.

The cotton genus (*Gossypium*) provides a model system for studying molecular evolution of genes duplicated by allopolyploidy. The five tetraploid *Gossypium* species ($n = 26$) are a monophyletic assemblage derived from a single allopolyploidization event that occurred approximately 1-2 MYA (Wendel 1989; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998). A robust phylogenetic framework has been developed for both the diploid and allopolyploid members of the genus (Fig. 1; Wendel and Albert 1992; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998). Diploid *Gossypium* species (all $n = 13$) have been divided into genomic groups (A-K) based on differences in chromosome size and pairing behavior in interspecific hybrids (Endrizzi, Turcotte, and Kohel 1985; Stewart 1995). The two diploid species that gave rise to the allotetraploids were from the A-genome and D-genome groups, and are best represented by the extant species *G. herbaceum* L. and *G. raimondii* Ulbr., respectively (Endrizzi, Turcotte, and Kohel 1985; Wendel, Schnabel, and Seelanan 1995; Small et al. 1998). Tetraploid species are therefore termed the AD-genome group, and their two constituent genomes are referred to as the A- and D-subgenomes.

Two of the allotetraploid species, *G. hirsutum* L. and *G. barbadense* L., were independently domesticated within the last 5000 years for their seed fiber (reviewed in Wendel 1995). The genetic consequences of domestication of *G. hirsutum*, the species that presently dominates world cotton commerce, have been explored in depth at the isozyme and RFLP levels (Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994). Among the conclusions of these studies is that genetic diversity in *G. hirsutum* is very low, and is especially restricted in the gene pool represented by modern annualized cultivars. *Gossypium hirsutum* probably was first domesticated in the Yucatan peninsula. The only extant form of *G. hirsutum* that arguably is

wild, race "yucatanense," is found here as a common component of the indigenous beach strand vegetation (Stephens 1958; Sauer 1967), where it exists as a sprawling, perennial shrub. Evidence suggests that following initial domestication, the original perennial cultivated forms became widely dispersed throughout the Yucatan peninsula. Later, localized derivatives developed, the most important of which was the annualized race "latifolium," which is suggested to have spread to Guatemala and southern Mexico where further agronomic development took place, leading to cultivated forms that spread via human-mediated diffusion throughout Mesoamerica. Molecular marker evidence (Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994) shows that most of the gene pool of the modern, annual forms of *G. hirsutum*, including the Upland cotton cultivars that are common in the cotton belt of the United States, traces to Mexican stocks that had been transported there from Guatemala and southern Mexico (Niles and Feaster 1984). This history of sequential genetic bottlenecks and rapid population expansion is thought to be responsible for the constrained levels of genetic diversity, as often observed in crop plant gene pools (Doebley 1989, 1992).

The goal of the present study was to quantify nucleotide diversity at a pair of homoeologous *Adh* loci in the allotetraploid species, *G. hirsutum*. With these data we asked the following questions: (1) Are estimates of nucleotide diversity equivalent at the two homoeologous loci that resulted from the allopolyploidization event? (2) How do estimates of nucleotide diversity for *Adh* in cotton compare with nucleotide diversity estimates from other species? (3) Are nucleotide diversity data consistent with previous information on the history of domestication for *G. hirsutum*? (4) How do direct measurements of genetic diversity based on DNA sequence data compare with indirect estimates derived from isozyme and RFLP data?

Materials and Methods

Plant materials

Based on variation at 205 anonymous nuclear loci detected by low-copy RFLP probes, Brubaker and Wendel (1994) defined 18 genetic/geographic groups of wild or feral *G. hirsutum* that represent the diversity encompassed by the species. In addition, they sampled several cultivars ("Upland cotton") from the cotton belt of the United States and several accessions of *G. barbadense* ("Pima cotton"; "Egyptian cotton"). For this study we surveyed one accession from each of the 18 groups identified by Brubaker and Wendel (1994) and three Upland cultivars of *G. hirsutum*. We also included five accessions of *G. barbadense* for comparison (Table 1). Because one of our objectives was to compare diversity detected using various molecular tools, the same accessions were used as in earlier studies employing isozymes (Wendel, Brubaker, and

Percival 1992) and RFLPs (Brubaker and Wendel 1994). Genomic DNAs used in the present study were the same as those described in Brubaker and Wendel (1994).

PCR amplification, cloning, and sequencing

The *Adh* gene family in *Gossypium* consists of a minimum of five genetic loci in the diploid species and five pairs of homoeologous loci in the allotetraploids (RLS and JFW, unpublished data). *Adh* genes encode alcohol dehydrogenase (alcohol:NAD^+ oxidoreductase, E.C. 1.1.1.1), metabolic enzymes involved primarily in anaerobic respiration. We have termed the locus discussed here *AdhA*. *AdhA* exists as a single-copy locus in the diploids *G. herbaceum* (A-genome) and *G. raimondii* (D-genome) and at a pair of homoeologous loci in the allotetraploids (Fig. 2). As expected from the organismal history, allotetraploid cotton contains a pair of homoeologous loci corresponding to the copies donated by the A- and D-genome ancestors at the time of allopolyploid formation. While most *Gossypium Adh* genes have the classical 10 exon/9 intron *Adh* structure (Millar and Dennis 1996; RLS and JFW, unpublished data), *AdhA* has lost introns 4 and 7 (Fig. 3). Intron loss has been observed in other plant *Adh* loci (Chang and Meyerowitz 1986; Trick et al. 1988; Charlesworth, Liu, and Zhang 1998), as well as in other plant genes (e.g., Drouin and Moniz de Sá 1997; Frugoli et al. 1998), and is presumably accomplished via gene conversion or recombination between an intact gene and a reverse-transcribed cDNA or processed pseudogene (Drouin and Moniz de Sá 1997; Frugoli et al. 1998). A more complete analysis of the structure and evolution of the *Gossypium Adh* gene family will be presented elsewhere.

To isolate *AdhA* sequences we designed *AdhA*-specific PCR primers homologous to regions in exons 2 and 8 (Fig. 3; *Adh*x2-1: CTT CAC TGC TTT ATG TCA CAC T; *Adh*x8-1: GGA CGC TCC CTG TAC TCC) and amplified a ~ 1 kb fragment of *AdhA*. PCR reactions were performed as described (Small et al. 1998). Because *AdhA* exists as a pair of homoeologous loci, the resulting PCR product contained a mixture of sequences from both the A- and D-subgenomes. To separate these products into their respective subgenomic sequences, we recovered the PCR products with Geneclean II (Bio 101), ligated the PCR products into pGEM-T (Promega), and transformed competent *E. coli* Top10 F′ cells (Invitrogen). Resulting colonies were screened for inserts by resuspending bacterial colonies in 10 µl of water, boiling for 10 minutes, centrifuging for 30 seconds, and using 2.5 µl of the supernatant as a template in a 10 µl PCR reaction using the original amplification primers and reaction conditions. PCR products that were the correct size (indicating presence of an *AdhA* insert) were ethanol precipitated, resuspended in 10 µl of water, and subjected to restriction digestion with *Taq*$^{\alpha}$I (New England Biolabs), an enzyme that has one recognition site in sequences from the A-subgenome and two sites in sequences from the

D-subgenome. This procedure allowed us to distinguish the subgenomic origin of each individual clone. To eliminate sequencing artifacts caused by misincorporation during PCR, for each accession we isolated and pooled ten plasmids from each subgenome and sequenced this pool using the amplification primers as sequencing primers. For 39 of 48 templates (81%), this procedure resulted in a monomorphic sequencing ladder; i.e., no apparent heterozygotes were detected. The remaining templates showed polymorphism at one or more sites. To evaluate whether these polymorphisms reflected true heterozygosity or misincorporation, we repeated the amplification, cloning, and sequencing steps. For all sequences that were initially based on clones, we were unable to reproduce the polymorphisms detected earlier, and in several cases new polymorphisms appeared. We concluded that in all but one case, polymorphisms did not reflect true heterozygosity but instead arose due to PCR error.

After the initial results were obtained, a second strategy was employed to isolate subgenome-specific *AdhA* sequences. This approach involved the use of two pairs of homoeologue-specific PCR primers that would amplify *AdhA* from only one subgenome at a time (A-subgenome-specific primers — *AdhA*x2i3-A: AAG GTA TTA CTG TAC GAT AA; *AdhA*x9i8-A: CCT GTA ATT CAA GAA GAA G; D-subgenome-specific primers — *AdhA*x2i3-D: AAG GTA TTA CTG TTC GAT AT; *AdhA*x9i8-D: CCT GTA ATT CAA GAA GCA T). These primers generated homogeneous PCR pools that could be directly sequenced, obviating the laborious cloning and restriction digestion steps. In addition, direct sequencing of PCR products (as opposed to sequencing cloned PCR products) greatly reduces the likelihood of detecting misincorporated nucleotides since these are expected to be present in low concentrations relative to the correct products. Therefore, we reamplified *AdhA* (using the homoeologue-specific primers) and sequenced the PCR products directly from those accessions that had shown polymorphism. Sequences obtained using this approach were monomorphic, indicating that we had eliminated PCR artifacts.

All DNA sequencing was performed using the Thermo Sequenase ^{33}P-radiolabeled terminator cycle sequencing kit (Amersham). Sequencing reactions were electrophoresed on 5-6% Long Ranger sequencing gels (FMC). Because so little polymorphism was detected, templates were sequenced on one strand only. After the entire data set had been collected, each sequence was rechecked at all polymorphic sites to confirm the original reads. The sequences reported here have been submitted to GenBank under accession numbers AF085064-AF085085, AF085812-AF085821, and AF090146-AF090168.

Statistical analyses

Gossypium hirsutum and *G. barbadense* are genomic allotetraploids and display disomic inheritance. For analytical purposes we assumed that our approach detected both alleles, and we therefore represented each locus in each accession by the two alleles present notwithstanding the high level of homozygosity observed (cf. Gaut and Clegg 1993b). Our experimental design of pooling ten plasmids per homoeologue for each accession was designed to eliminate *Taq* error, as well as to ensure cloning and sequencing of both alleles. Assuming equal representation of both alleles in the PCR product pool and equivalent success of cloning each allele, the probability of not including both alleles in the plasmid pool is quite small (0.5^{10} or P=0.001). In our experience, even if an allele is represented only once in the plasmid pool it would be detected in the sequencing ladder. Finally, in other studies of *AdhA* in diploid *Gossypium* species, we used identical PCR primers and readily amplified both alleles of heterozygous individuals. The foregoing observations suggest that our approach is expected to detect both alleles at a locus and that the monomorphic sequencing ladders we obtained were the result of homozygosity.

For each subgenome of both *G. hirsutum* and *G. barbadense* we calculated two measures of nucleotide diversity per base pair: π (Nei 1987, pp. 256-257) and θ_w (Watterson 1975). The former measure quantifies the mean percentage of nucleotide differences among all pairwise comparisons for a set of sequences, whereas the latter is simply an index of the number of segregating (polymorphic) sites. Under neutral expectations, θ_w is equal to π (Tajima 1989; 1993b). A 95% confidence interval around θ_w was calculated for *AdhA* from both subgenomes of *G. hirsutum*, using methods described by Kreitman and Hudson (1991). Tests of neutral evolution were performed as described by Tajima (1989), Fu and Li (1993), and Hudson, Kreitman, and Aguadé (1987). Recombination was explored using the algorithm of Hudson and Kaplan (1985). Many of the above calculations were expedited by the software program DnaSP v. 2.52 (Rozas and Rozas 1997). Estimates of genetic diversity (mean number of alleles per locus — A; mean panmictic [expected] heterozygosity — $H_T = 1 - \Sigma[p_i]^2$ where p_i represents allele frequencies; cf. Brubaker and Wendel 1994) were calculated using our sequence data as well as previously published isozyme (Wendel, Brubaker, and Percival 1992) and RFLP (Brubaker and Wendel 1994) data for a comparable set of accessions (identical accessions for RFLP data; three missing accessions for isozyme data).

Given the phylogenetic framework of the genus *Gossypium* and estimates of the timing of several major branching points in the phylogeny (Fig. 1; Wendel and Albert 1992; Seelanan, Schnabel, and Wendel 1997), we were able to estimate an absolute mutation rate for *AdhA*. Specifically, using unpublished *AdhA* sequences of *G. robinsonii* (C-genome outgroup), *G. herbaceum* (A-genome diploid), and *G. raimondii* (D-genome diploid), we generated, using exon

data only, a synonymous site Jukes-Cantor (JC) distance matrix using MEGA v. 1.0 (Kumar, Tamura, and Nei 1993). The timing of the two branch-point estimates shown in Fig. 1 were derived from analyses of chloroplast *ndhF* sequences (Seelanan, Schnabel, and Wendel 1997). Using these divergence time points and the JC distances, we estimated the absolute synonymous mutation rate as the JC distance divided by twice the time since divergence.

Results

DNA polymorphism

We determined approximately 1 kb of sequence from both the A- and D-subgenomic homoeologues of *AdhA* for 22 accessions (44 alleles per subgenome) of *G. hirsutum* and five accessions (10 alleles per subgenome) of *G. barbadense*. Thus, approximately 108 kb of effective sequence data were generated (27 accessions X 2 alleles/homoeologue X 2 homoeologues). Sequence data for each allele consists of 662 bp of coding sequence and 336 bp of intron sequence; this represents a mean of 500.33 nonsynonymous sites and 482.67 silent sites (synonymous or intron; excluding gaps).

All sequences appeared homozygous, with the exception of one *G. hirsutum* cultivar (Paymaster H86048) which was heterozygous for alleles 1D and 2D (see Fig. 4). The distribution of nucleotides at all polymorphic sites for both homoeologues is shown in Fig. 4. In the A-subgenome of *G. hirsutum* we observed only one polymorphic site (position 571), which included approximately equal representation among accessions of the alternative nucleotides G and A. This transitional and silent substitution was at a third codon position. In the A-subgenome of *G. barbadense* no polymorphic sites were observed. In the D-subgenome of *G. hirsutum* there were three polymorphic sites, all within introns. Two of these three sites (positions 84 and 942) reflected transitional mutations, while the third polymorphism resulted from a [G,T] transversion (position 684). For all three polymorphic sites, the minority state occurred in either 5 or 6 of the 22 accessions sampled. One polymorphic site was revealed in the D-subgenome of *G. barbadense* (position 511, a third codon position transition).

No nucleotides at either *AdhA* homoeologue distinguish all *G. barbadense* alleles from those of *G. hirsutum*. For the A-subgenome locus, all five *G. barbadense* accessions are homozygous for an allele shared by eight of the 22 *G. hirsutum* accessions. Similarly, for *AdhA* from the D-subgenome, four of the five *G. barbadense* accessions are homozygous for an allele shared by 11 of the 22 *G. hirsutum* accessions (Fig. 4).

Estimates of nucleotide diversity (π, θ_w; gaps treated as missing data) are shown for each data set in Table 2. These estimates show that nucleotide diversity is approximately twice as high for *AdhA* from the D-subgenome as it is for *AdhA* from the A-subgenome.

The two homoeologues of *AdhA* differed by a minimum of 20 nucleotide substitutions within both *G. hirsutum* and *G. barbadense*, representing 12 transitions and 8 transversions (Fig. 4). Thus, the two *AdhA* homoeologues exhibit approximately 2% sequence divergence based on non-gapped positions. In addition, the two homoeologues are differentiated by four gaps, all of which are intron nucleotides present in the A-subgenomic homoeologues that are absent from the corresponding locus in the D-subgenome. All available data indicate that the *AdhA* homoeologues have evolved independently subsequent to polyploid formation; i.e., there is no evidence of intersubgenomic gene conversion or recombination. This inference is supported by the 20 nucleotide substitutions and four gaps that distinguish the homoeologues, the majority of which are also shared with the respective diploid progenitors.

Tests of neutral evolution, recombination, and rates of nucleotide substitution

Several statistical tests were used to test the hypothesis that *AdhA* sequences have been evolving in accordance with expectations under neutral theory. The tests of Tajima (1989) and Fu and Li (1993) compare different estimates of θ ($4N_e\mu$), each of which makes certain assumptions about how sequences evolve (Simonsen, Churchill, and Aquadro 1995; Wayne and Simonsen 1998). These tests were conducted on each of the four data sets (two subgenomes in two species), and the results are shown in Table 2. None of these tests returned significant *P*-values. This is not surprising, given the small number of variable positions and the relatively low statistical power of these tests (Simonsen, Churchill, and Aquadro 1995; Wayne and Simonsen 1998). The HKA test (Hudson, Kreitman, and Aguadé 1987) compares levels of polymorphism between genes or regions both within and between species, the assumption being that levels of neutral polymorphism should be correlated with rates of evolution across genomes. While the original intent of this test was to compare an unknown region to a region that is presumed to be evolving neutrally, we adapted it to test the assumption that the two homoeologues are evolving equivalently. Intraspecific polymorphism at the *G. hirsutum AdhA* A-subgenome homoeologue was compared to the *AdhA* D-subgenome homoeologue; the same regions from *G. barbadense* provided the interspecific comparison. The HKA test result was not significant (P=0.75). The Hudson and Kaplan (1985) estimate of the minimum number of recombination events was zero, *viz.*, all sites were compatible with a history devoid of inter-allelic recombination. A network depicting allele relationships and their corresponding taxonomic and geographic distribution is shown in Fig. 5.

An absolute rate of nucleotide substitution was estimated for *AdhA* using two separate calibration points (Fig. 1) derived from analyses of chloroplast DNA sequence data (Seelanan, Schnabel, and Wendel 1997). Using the divergence of the [A+D]-genome clade from the C-

genome clade (synonymous site JC distance = 0.035; divergence time of 12 Myr), a rate of 1.47 x 10^{-9} synonymous substitutions/synonymous site/year was obtained. Using the split between the A- and D-genomes (synonymous site JC distance = 0.045; divergence time of 11 Myr), the substitution rate was estimated to be 2.05 x 10^{-9} synonymous substitutions/synonymous site/year.

Comparisons of measures of genetic diversity

One difference between DNA sequencing and indirect methods of assessing genetic variation (e.g., isozymes or RFLPs) is that all polymorphisms will be detected in the region sequenced, not just those that result in restriction site mutations (cf. RFLP analysis) or electrophoretically detectable charge or conformational changes (cf. isozyme analysis). Thus, one might expect levels of allelic diversity and heterozygosity to be higher for nucleotide sequence data than for other data sets; this expectation was met in the present study (Table 3). Previous studies have assayed isozyme and RFLP diversity in *G. hirsutum* and *G. barbadense* (Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994). We recalculated genetic diversity statistics for the isozyme and RFLP data by pruning the data sets to include only those accessions sampled here (Table 3). In general, allelic diversity was higher for *AdhA* sequence data than for isozymes or RFLPs, as expected. In addition, *expected* heterozygosity (= mean panmictic heterozygosity) was also higher for the sequence data, but *observed* heterozygosity at *AdhA* was zero in all cases except for the D-subgenome of *G. hirsutum*. The single heterozygote observed was for a cultivar (Paymaster H86048); heterozygosity in this accession may reflect the results of a breeding program or germplasm maintenance.

Discussion

Nucleotide diversity in allopolyploid *Gossypium*

A primary conclusion of the present study is that nucleotide diversity for *AdhA* in *G. hirsutum* and *G. barbadense* is very low. Estimates reported here are lower than previously reported values not only for plant *Adh* sequences (see Table 3 of Cummings and Clegg 1998; Liu, Zhang, and Charlesworth 1998), but for other plant nuclear genes as well (*C1* in maize — Hanson et al. 1996; *ChiA* in *Arabidopsis* — Kawabe et al. 1997; *ChsA* in *Ipomoea* — Huttley et al. 1997; *Pgi* in *Dioscorea* — Terauchi, Terachi, and Miyashita 1997). Nucleotide diversity per base pair for nuclear genes in other plant species range from a low of θ_w=0.001 at *Pgi* in *Dioscorea* (Terauchi, Terachi, and Miyashita 1997) to a high of θ_w=0.025 at *Adh1* in *Zea mays* ssp. *parviglumus* (Eyre-Walker et al. 1998); our values of θ_w ranged from 0.000 (*G. barbadense* A-subgenome) to 0.0007 (*G. hirsutum* D-subgenome).

Potential explanations for such low levels of nucleotide diversity include one or more recent genetic bottlenecks, a low mutation rate, a self-pollinating reproductive biology, and a selective sweep. We present evidence that the first three of these factors have been important in shaping the population genetic structure of cotton and are sufficient to explain our observations thus obviating the need to invoke additional mechanisms such as selective sweeps. Operating together, the historical and life-history features of *G. hirsutum* have had a net effect of severely constraining levels of genetic diversity, as discussed in the following paragraphs.

Genetic bottlenecks — Accumulated evidence indicates that *G. hirsutum* and *G. barbadense* are allotetraploids that are derived from a single polyploidization event that occurred ~ 1-2 MYA (Wendel 1989; Seelanan, Schnabel, and Wendel 1997; Small et al. 1998). Because the two parental diploid genomes are confined to different continents (A-genome: Africa-Asia; D-genome: New World, primarily Mexico), polyploidization appears to have been precipitated by trans-oceanic dispersal of an A-genome propagule to the New World, followed by hybridization and allopolyploidization with the native D-genome donor. It seems likely that only one to a few A-genome propagules made this trans-oceanic voyage, and similarly probable that only one to very few individuals were involved in the initial hybridization event from which allopolyploid *Gossypium* emerged. Thus, the process by which the lineage formed is characterized by a severe genetic bottleneck; presumably one or a few hybrid individual(s) constituted the entire gene pool from which the extant tetraploids are derived. Subsequent diversification of the nascent allopolyploid into the five modern tetraploid species implicates additional genetic bottlenecks associated with these more recent speciation events. Finally, more recent bottlenecks undoubtedly occurred as a consequence of the domestication of *G. hirsutum*, perhaps 4000-5000 years ago (Wendel, Brubaker, and Percival 1992; Brubaker and Wendel 1994; Wendel 1995).

Thus, the agronomic development of modern *G. hirsutum* varieties has been characterized by sequential genetic bottlenecks followed by rapid range expansions. A similar history with an approximately equivalent time-scale has been described for *G. barbadense* (Percy and Wendel 1990). For both species, these episodic bottlenecks undoubtedly contributed to a winnowing of nucleotide diversity, which may not have been especially extensive even in the wild progenitors. This winnowing process has occurred over a period of time that is exceptionally brief on an evolutionary timescale, especially in light of the timeframe necessary for the introduction of genetic diversity through mutation.

Breeding system — An additional constraint on levels of nucleotide diversity levels in *G. hirsutum* and *G. barbadense* stems from their reproductive biology; both species are self-

compatible and produce a high proportion of their seed through self-pollination (Wendel 1995). Self-pollination is known to reduce effective population size, which in turn reduces expected levels of genetic diversity (Pollak 1987; Liu, Zhang, and Charlesworth 1998). In addition to reducing levels of genetic diversity, selfing is expected to reduce observed heterozygosity relative to expected heterozygosity as allelic variation manifests itself as alternative homozygotes, rather than heterozygotes. Thus, the breeding system is consistent with our observation of a near-complete absence of observed heterozygosity at *AdhA*, where the sole exception was for a *G. hirsutum* cultivar that may have acquired its heterozygosity (either intentionally or unintentionally) through a breeding program or during seed increase for germplasm maintenance.

Low mutation rate — The absolute synonymous substitution rate calculated for *AdhA* in *Gossypium* is 1.47×10^{-9} to 2.05×10^{-9} substitutions/site/year. Wolfe, Li, and Sharp (1987) estimated that synonymous substitution rates at plant nuclear genes range from $5\text{-}30 \times 10^{-9}$, and average $5.1\text{-}7.1 \times 10^{-9}$ (Wolfe, Sharp, and Li 1989). Gaut (1998) has estimated a synonymous rate of 6.03×10^{-9} in a comparison of nine nuclear genes in rice and maize. The lowest published synonymous rate for a plant nuclear gene is 2.61×10^{-9} for *AdhA* in palms (Gaut et al. 1996). The synonymous substitution rate for *AdhA* in *Gossypium* is therefore 2.5 to 4 times lower than average rates, and is lower than any previously published rates. This estimate is, in fact, within the range ($1.0\text{-}3.0 \times 10^{-9}$) of synonymous substitution rates in chloroplast genes (Wolfe, Li, and Sharp 1987). Given this slow mutation rate, there has been little time for the genesis of allelic diversity since species formation (perhaps 1-2 million years), and even less time since *G. hirsutum* and *G. barbadense* were domesticated (perhaps 4000-5000 years). Even under a scenario of complete retention of genetic diversity, i.e., no loss due to sampling or genetic bottlenecks (as discussed above), the expectation is that for *AdhA*, with approximately 500 silent sites and mutation rates as estimated above, only one or two nucleotides, on average, are expected to become polymorphic in each million years. Hence, the observation of only 1 and 3 polymorphic sites in the A- and D-subgenomic homoeologues, respectively, is consistent with expectations based on our understanding of mutation rates and the history and biology of the species. It therefore seems unnecessary to invoke additional mechanisms such as selective sweeps.

Lack of coalescence — One of the noteworthy observations of this study is that *AdhA* alleles do not coalesce within species. In both the A- and D-subgenomes, the predominant allele found in *G. barbadense* also occurs at high frequency in *G. hirsutum*. This result is consistent with the low mutation rates and phylogenetic history discussed above, or alternatively, with an hypothesis

of large-scale introgression of *G. hirsutum* alleles into *G. barbadense* (Brubaker, Koontz, and Wendel 1993). With respect to the former, molecular phylogenetic data have led to the suggestion that post-polyploidization, there was a rapid radiation into the extant clades represented by the five modern species (Fig. 1; Wendel 1989; Small et al. 1998). Under such circumstances (low variation and rapid radiation) it is expected that alleles would be shared across species boundaries, unless a high mutation rate and a high fixation rate was operating.

Comparison within and between homoeologous locus-pairs

One of the initial goals of this study was to test the hypothesis that homoeologous loci exhibit equivalent evolutionary dynamics. Given a single origin of the tetraploid *Gossypium* species, levels and patterns of genetic diversity should be equivalent for homoeologous loci, assuming an absence of selection, differential recombination, or other forces that might differentially affect members of a homoeologous locus-pair. All population-level factors other than selection (e.g., effective population size, genetic drift, breeding system) are equivalent.

A previous study (Small et al. 1998) has shown that for another alcohol dehydrogenase locus in *Gossypium* (*AdhC*), evolutionary rates at the two homoeologues differ significantly, with the locus from the D-genome diploid and D-subgenome of the tetraploids evolving at a faster rate. Neutral theory predicts that evolutionary rate and genetic diversity should be positively correlated — this is, in fact, the basis of the HKA test (Hudson, Kreitman, and Aguadé 1987). We applied this test to the *AdhA* data presented here, not to detect departure from neutrality, but to detect differences in evolutionary dynamics between homoeologues; the result was not significant. Likewise, the 95% confidence intervals calculated for θ_w largely overlap. Finally, application of Tajima's (1993a) 1D relative rate test comparing *AdhA* sequences from the A-genome diploid and the A-subgenome of *G. hirsutum* to the D-diploid and D-subgenome of *G. hirsutum* returned only one significant departure from rate homogeneity (*G. herbaceum* vs. *G. raimondii*), despite a qualitatively obvious rate difference (Fig. 6). Although none of the statistical tests supports an inference of rate inequality among the *AdhA* homoeologues, allelic diversity is twice as high in the D-subgenome, and nucleotide diversity is two to three times higher in the D-subgenome, results that are directionally consistent with the previously reported *AdhC* data (Small et al. 1998). These observations may or may not be consequential; data from other homoeologous pairs are needed to evaluate the possibility that the subgenomes of *G. hirsutum* are subject to different evolutionary pressures.

Comparisons among homoeologous locus-pairs may also provide insight into processes of genomic evolution. The evolutionary dynamics appear different for *AdhA* and *AdhC* in tetraploid *Gossypium*. For example, in the tetraploid species of *Gossypium*, percent sequence

divergence for *AdhA* (A-subgenome vs. D-subgenome) averages 2%, but is over 4% for *AdhC* (Small et al. 1998). As described above, increased evolutionary rate should be correlated with increased genetic diversity, which would predict a higher level of diversity at *AdhC* than at *AdhA*. We are currently conducting studies to test this hypothesis. Finally, previous studies have shown that a correlation exists between chromosomal position (and associated rates of recombination) and levels of genetic diversity at a locus. In general, the more distal a locus is from the centromere, the higher the recombination rate and genetic diversity will be (Begun and Aquadro 1992; Aquadro and Begun 1993; Dvorak, Luo, and Yang 1998). Genetic mapping data allow us to speculate that *Gossypium Adh* loci may show the opposite trend. The slowly-evolving, low-diversity locus *AdhA* resides at the distal end of a linkage group, while the quickly-evolving locus *AdhC* maps near the middle of a linkage group (RLS and JFW, unpublished data). While we are currently unable to correlate these genetic mapping data with a physical map and therefore pinpoint distances from the centromere or telomere, these preliminary data may provide an exception to the general relationship between genetic diversity and chromosomal position.

Acknowledgments

We thank C. Brubaker for his previous work on genetic diversity and mapping in tetraploid *Gossypium*, which greatly facilitated this investigation. We also thank B. Gaut for helpful suggestions and discussion. We thank C. Brubaker, R. Cronn, J. Doebley, B. Gaut, and P. Soltis for comments on an earlier draft of this manuscript. This research was supported by funding from the National Science Foundation (to JFW).

Literature Cited

AMOS, W., AND J. HARWOOD. 1998. Factors affecting levels of genetic diversity in natural populations. Phil. Trans. R. Soc. Lond. B **353**:177-186.

AQUADRO, C. F., AND D. J. BEGUN. 1993. Evidence for and implications of genetic hitchhiking in the *Drosophila* genome. Pp. 159-178 *in* N. Takahata and A. G. Clark, eds. Mechanisms of molecular evolution. Sinauer, Sunderland, Mass.

BEGUN, D. J., AND C. F. AQUADRO. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356**:519-520.

BERGELSON, J., E. STAHL, S. DUDEK, AND M. KREITMAN. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148**:1311-1323.

BRUBAKER, C. L., J. A. KOONTZ, AND J. F. WENDEL. 1993. Bidirectional cytoplasmic and nuclear introgression in the New World cottons, *Gossypium barbadense* and *G. hirsutum* (Malvaceae). Amer. J. Bot. **80**:1203-1208.

BRUBAKER, C. L., AND J. F. WENDEL. 1994. Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Amer. J. Bot. **81**:1309-1326.

CHANG, C., AND E. 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.-L. LIU, AND L. ZHANG. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol. Biol. Evol. **15**:552-559.

CLEGG, M. T. 1997. Plant genetic diversity and the struggle to measure selection. J. Heredity **88**:1-7.

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.

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.

DOEBLEY, J. F. 1989. Isozymic evidence and the origin of crop plants. Pp. 165-186 *in* D. E. Soltis and P. S. Soltis, eds. Isozymes in plant biology. Dioscorides Press, Portland, OR.

DOEBLEY, J. F. 1992. Molecular systematics and crop evolution. Pp. 202-222 *in* P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds. Molecular systematics of plants. Chapman and Hall, NY.

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.

DVORÁK, J., M.-C. LUO, AND Z.-L. YANG. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing Aegilops species. Genetics **148**:423-434.

ENDRIZZI, J. D., E. L. TURCOTTE, AND R. J. KOHEL. 1985. Genetics, cytology, and evolution of *Gossypium*. Adv. Genet. **23**: 271-375.

EYRE-WALKER, A., R. L. GAUT, H. HILTON, D. L. FELDMAN, AND B. S. GAUT. 1998. Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95**:4441-4446.

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.

FU, Y.-X., AND W.-H. LI. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693-709.

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. 1993a. Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA **90**:5095-5099.

GAUT, B. S., AND M. T. CLEGG. 1993b. Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics **135**:1091-1097.

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.

HANFSTINGL, U., A. BERRY, E. A. KELLOGG, J. T. COSTA III, W. RÜDIGER, AND F. M. AUSUBEL. 1994. Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138**:811-828.

HANSON, M. A., B. S. GAUT, A. O. STEC, S. I. FUERSTENBERG, M. M. GOODMAN, E. H. COE, AND J. F. DOEBLEY. 1996. Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics **143**:1395-1407.

HUDSON, R. R., AND N. L. KAPLAN. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**:147-164.

HUDSON, R. R., M. KREITMAN, AND M. AGUADÉ. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153-159.

HUTTLEY, G. A., M. L. DURBIN, D. E. GLOVER, AND M. T. CLEGG. 1997. Nucleotide polymorphism in the chalcone synthase-A locus and evolution of the chalcone synthase multigene family of common morning glory *Ipomoea purpurea*. Mol. Ecol. **6**:549-558.

INNAN, H., F. TAJIMA, R. TERAUCHI, AND N. MIYASHITA. 1996. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics **143**:1761-1770.

KAWABE, A., H. INNAN, R. TERAUCHI, AND N. T. MIYASHITA. 1997. Nucleotide polymorphism in the acidic chitinase locus (*Chi*A) region of the wild plant *Arabidopsis thaliana*. Mol. Biol. Evol. **14**:1303-1315.

KREITMAN, M., AND R. R. HUDSON. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127**:565-582.

KUMAR, S., K. TAMURA, AND M. NEI. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version. 1.0. The Pennsylvania State University, University Park, PA 16802.

LIU, F., L. ZHANG, AND D. CHARLESWORTH. 1998. Genetic diversity in *Leavenworthia* populations with different inbreeding levels. Proc. R. Soc. Lond. B **265**:293-301.

MILLAR, A. A., AND E. S. DENNIS. 1996. The alcohol dehydrogenase genes of cotton. Pl. Mol. Biol. **31**:897-904.

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.

MORIYAMA, E. N., AND J. R. POWELL. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**:261-277.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, NY.

NILES, G. A., AND C.V. FEASTER. 1984. Breeding. Pp. 201-231 *in* R. J. Kohel and C. F. Lewis, eds. Cotton. Amer. Soc. Agron., Madison, WI.

PERCY, R. G., AND J. F. WENDEL. 1990. Allozyme evidence for the origin and diversification of *Gossypium barbadense* L. Theor. Appl. Genet. **79**:529-542.

POLLAK, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117**:353-360.

ROZAS, J., AND R. ROZAS. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. **13**:307-311.

SAUER, J. 1967. Geographic reconnaissance of seashore vegetation along the Mexican Gulf Coast. Coastal Studies Ser. No. 21. Louisiana St. Univ. Press, Baton Rouge, LA.

SEELANAN, T., A. SCHNABEL, AND J. F. WENDEL. 1997. Congruence and consensus in the cotton tribe. Syst. Bot. **22**:259-290.

SIMONSEN, K. L., G. A. CHURCHILL, AND C. F. AQUADRO. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413-429.

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 phylogeny reconstruction in a recently diverged plant group. Amer. J. Bot. **85**:1301-1315.

STEPHENS, S. G. 1958. Salt water tolerance of seeds of *Gossypium* species as a possible factor in seed dispersal. Amer. Nat. **92**:83-92.

STEWART, J. MCD. 1995. Potential for crop improvement with exotic germplasm and genetic engineering. Pp. 313-327 *in* G. A. Constable and N. W. Forrester, eds. Challenging the future: proceedings of the world cotton research conference-1. CSIRO, Melbourne.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.

TAJIMA, F. 1993a. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599-607.

TAJIMA, F. 1993b. Measurement of DNA polymorphism. Pp. 37-59 in N. Takahata and A. G. Clark, eds. Mechanisms of molecular evolution. Sinauer, Sunderland, Mass.

TERAUCHI, R., T. TERACHI, AND N. T. MIYASHITA. 1997. DNA polymorphism at the *Pgi* locus of a wild yam, *Dioscorea tokoro*. Genetics **147**:1899-1914.

TRICK, M., E. S. DENNIS, K. J. R. EDWARDS, AND W. J. PEACOCK. 1988. Molecular analysis of the alcohol dehydrogenase gene family of barley. Pl. Mol. Biol. **11**:147-160.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7**:256-276.

WAYNE, M. L., AND K. L. SIMONSEN. 1998. Statistical tests of neutrality in the age of weak selection. TREE **13**:236-240.

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* N. Simmonds and J. Smartt, eds. Evolution of crop plants. 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., C. L. BRUBAKER, AND A. E. PERCIVAL. 1992. Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. Amer. J. Bot. **79**:1291-1310.

WENDEL, J. F., A. SCHNABEL, AND T. SEELANAN. 1995. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol. Phyl. 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.

WOLFE, K. H., P. M. SHARP, AND W.-H. LI. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. **29**:208-211.

Table 1. Plant materials used in this study.

Species	Accession	Geographic Origin
Gossypium hirsutum	pfx	Oaxaca, Mexico
	TX-1	Guerrero, Mexico
	TX-6	Pueblo, Mexico
	TX-21	Chiapas, Mexico
	TX-44	Chiapas, Mexico
	TX-51	Chiapas, Mexico
	TX-93	Jutiapa, Guatemala
	TX-94	Zacapa, Guatemala
	TX-98	Chiquimula, Guatemala
	TX-111	Jutiapa, Guatemala
	TX-116	Santa Rosa, Guatemala
	TX-119	El Salvador
	TX-166	Zacapa, Guatemala
	TX-188	Baja Verapaz, Guatemala
	TX-192	Oaxaca, Mexico
	TX-367	Santa Rosa, Guatemala
	TX-481	Yucatan, Mexico
	TX-706	Honduras
	TX-766	Belize
	Paymaster H86048	cultivar
	Deltapine 50	cultivar
	BR115	cultivar
G. barbadense	Pima S5	cultivar
	B106	Dominican Republic
	B250	Belize
	B444	Colombia
	B559	Venezuela

Table 2. Estimates (x 10^3) of nucleotide diversity per base pair (π, θ_w), and tests of neutral evolution.

	π	θ_w (θ_L, θ_U)[a]	D[b]	D[c]	F[c]
G. hirsutum A-subgenome	0.50	0.24 (0.0006, 1.52)	1.47	0.55	0.94
G. hirsutum D-subgenome	1.23	0.74 (0.16, 2.38)	1.42	0.91	1.24
G. barbadense A-subgenome	0.00	0.00	—	—	—
G. barbadense D-subgenome	0.36	0.36 (0.009, 2.73)	0.62	0.74	0.67

[a]Lower (θ_L) and upper (θ_U) bounds (x 10^3) of the 95% confidence intervals in parentheses.

[b]Test statistic of Tajima (1989); no results are statistically significant.

[c]Test statistics of Fu and Li (1993); no results are statistically significant.

Table 3. Genetic diversity statistics for isozyme, RFLP and *AdhA* sequence data.

	A[a]	H_T[b]	Obs. Het.[c]
G. hirsutum			
Isozymes	1.4	0.126	0.006
RFLPs	1.6	0.144	0.004
AdhA, A-subgenome	2	0.463	0.000
AdhA, D-subgenome	4	0.656	0.045
AdhA loci, mean	3	0.556	0.023
G. barbadense			
Isozymes	1.2	0.074	0.000
RFLPs	1.2	0.062	0.008
AdhA, A-subgenome	1	0.000	0.000
AdhA, D-subgenome	2	0.320	0.000
AdhA loci, mean	1.5	0.160	0.000

[a]Mean number of alleles per locus.
[b]Mean panmictic (expected) heterozygosity.
[c]Observed heterozygosity (# of heterozygous accessions/total number of accessions sampled).

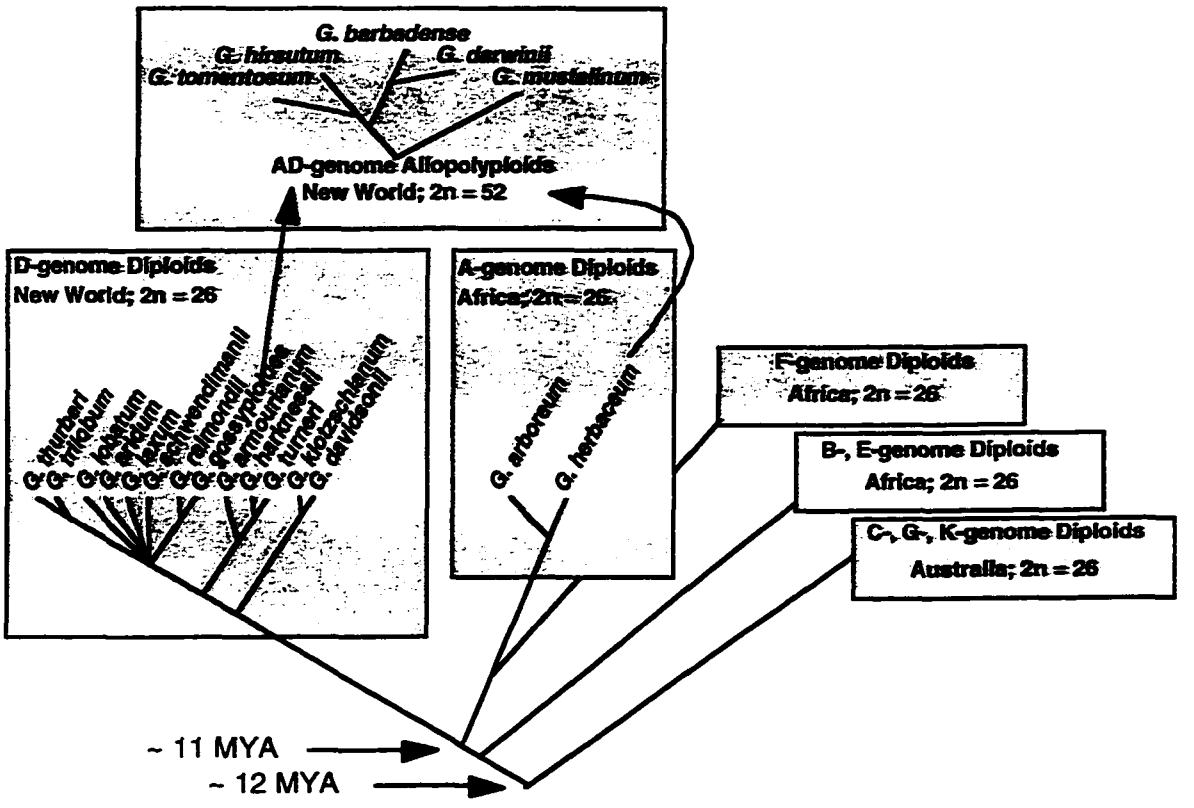


Figure 1. Composite phylogenetic hypothesis for the genus *Gossypium* based on analyses of multiple molecular data sets. Individual genome groups are boxed, and their geographic origin and chromosome numbers are identified. This hypothesis depicts both the divergent evolution of the diploid species (A through K genomes; $2n = 26$) and the reticulate origin of the allotetraploids (AD-genome; $2n = 52$) via hybridization and polyploidization between A- and D-genome diploids (best represented by *G. herbaceum* and *G. raimondii*, respectively). Arrows indicate estimated time of divergence for two nodes: initial divergence within the genus at ca. 12 MYA and the divergence between the A- and D-genome groups at ca. 11 MYA.

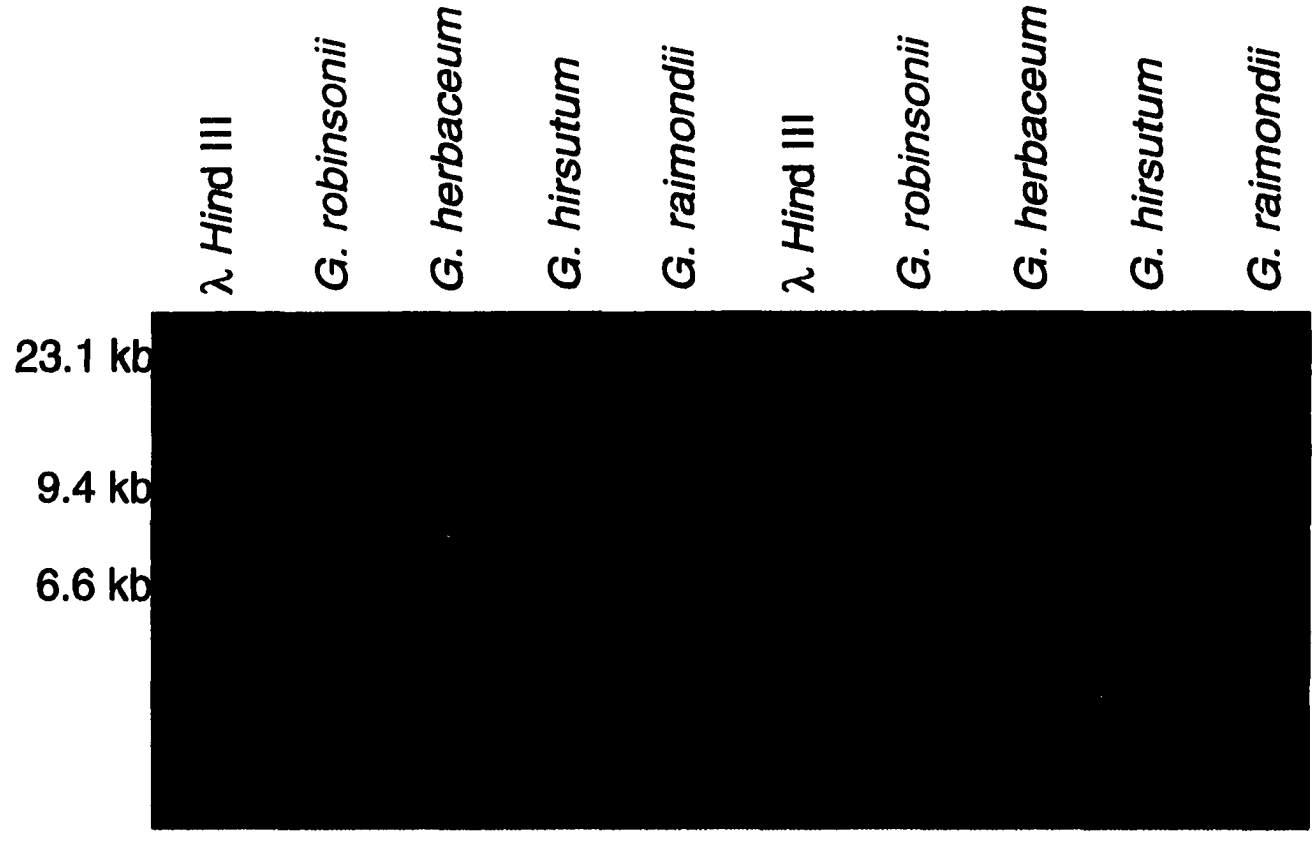


Figure 2. Southern blot of an *AdhA*-specific probe hybridized to *Hin*dIII and *Xba*I digested genomic DNAs of three diploid (*G. robinsonii*, C-genome; *G. herbaceum*, A-genome; *G. raimondii*, D-genome) and one tetraploid (*G. hirsutum*) cotton species. In both sets of digests the probe reveals only a single band per diploid genome, indicating that *AdhA* is single copy.

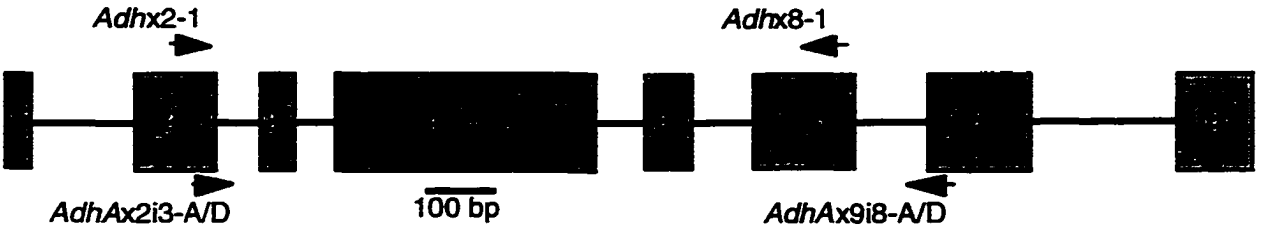


Figure 3. Diagrammatic representation of the *Gossypium AdhA* locus. Exons are shown as shaded boxes, introns as the line connecting the exons. Genomic sequence data are available only for exons 2-8; the lengths of exons 1, 9, and 10 and introns 1 and 9 are extrapolated from other *Gossypium Adh* sequences. PCR amplification primers are shown in their approximate positions. *AdhA*-specific primer pair *Adh*x2-1 and *Adh*x8-1 is shown above the diagram while homoeologue-specific primer pairs (*AdhA*x2i3-A + *AdhA*x8i9-A; *AdhA*x2i3-D + *AdhA* x9i8-D) are shown below the diagram. A 100-bp scale bar is included for reference.

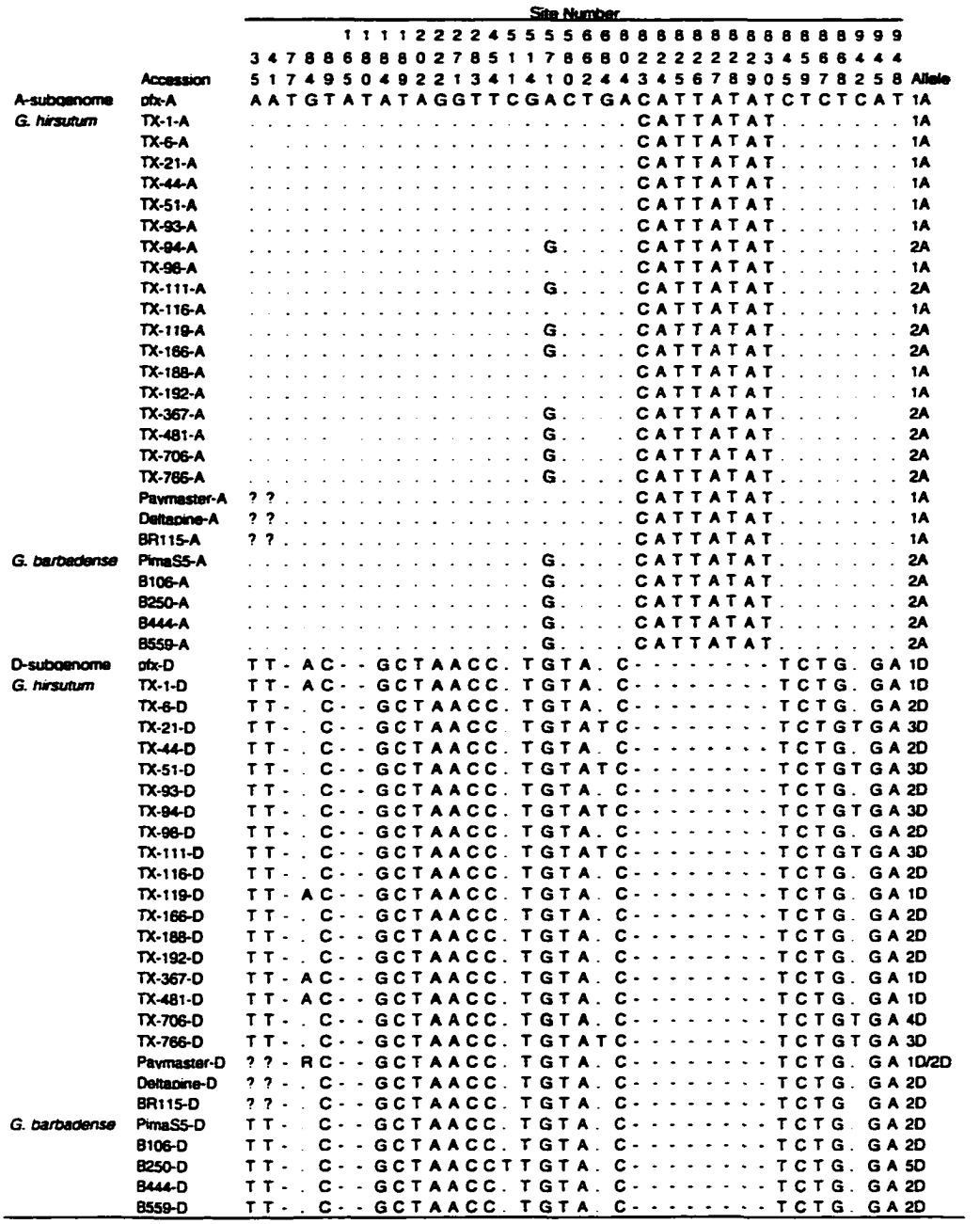


Figure 4. Polymorphic nucleotide positions in the *AdhA* data set. At each polymorphic nucleotide site (numbers shown above the sequences), the nucleotide state observed in each accession is given relative to the *G. hirsutum* pfx-A sequence. A period denotes identity, while question marks are used for missing data. Alignment gaps are indicated by dashes. Allelic designations are given in the final column.

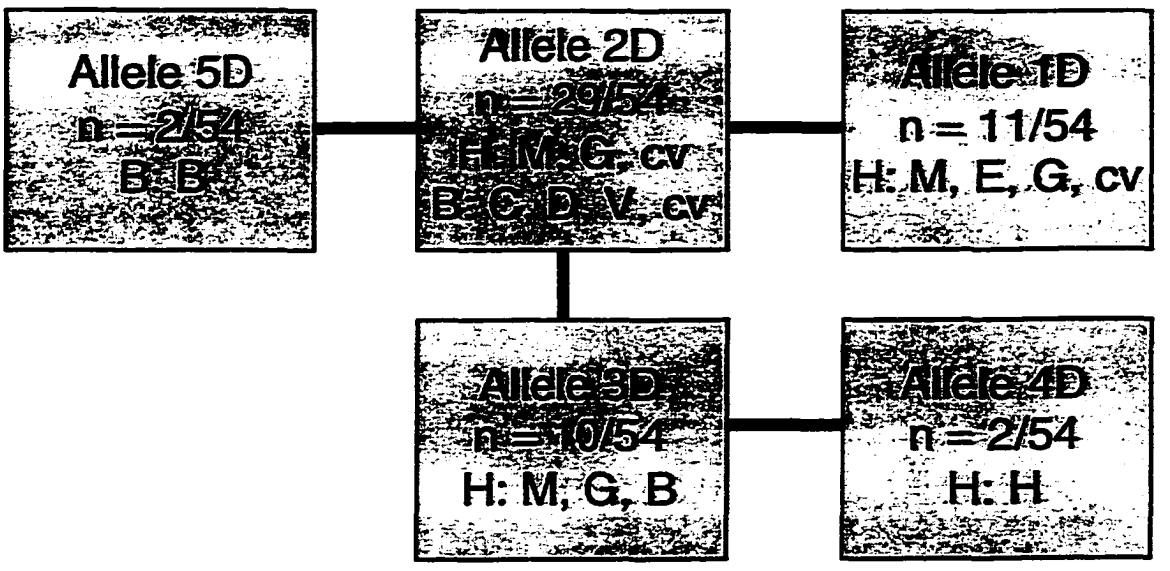


Figure 5. Allele network depicting relationships among alleles observed at the D-subgenome homoeologue of *G. hirsutum* and *G. barbadense*; each allele differs from alleles to which it is connected by a single nucleotide substitution. Allele designations follow Fig. 4. For each allele the number of times it was observed out of a total of 54 alleles (44 *G. hirsutum* alleles and 10 *G. barbadense* alleles) is given. The taxonomic and geographic distributions of alleles are as follows: H — allele detected in *G. hirsutum*; B — allele detected in *G. barbadense*; B — Belize; C — Colombia; D — Dominican Republic; E — El Salvador; G — Guatemala; H — Honduras; M — Mexico; V — Venezuela; and cv— cultivar.

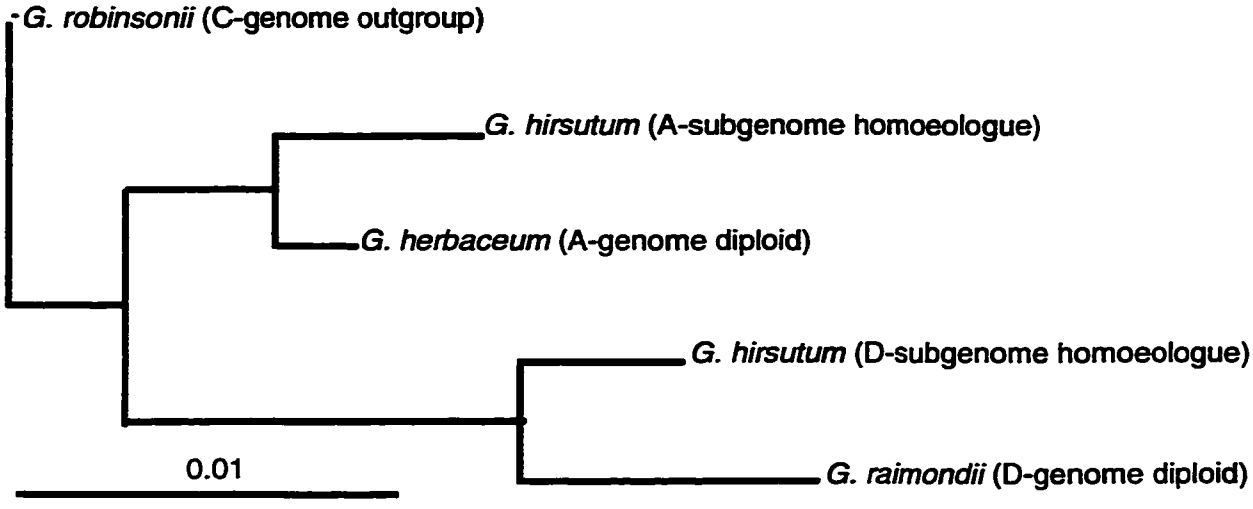


Figure 6. Neighbor-joining tree of *AdhA* sequences from diploid and tetraploid *Gossypium*, rooted with *G. robinsonii* (the topology resulting from maximum parsimony analysis is identical). Homoeologues from allotetraploid *G. hirsutum* (accession pfx) cluster with their respective diploid progenitors. Branch length leading to the *G. hirsutum* D-homoeologue and *G. raimondii* (D-diploid) is ca. twice as long as the branch leading to the *G. hirsutum* A-homoeologue and *G. herbaceum* (A-diploid). Tajima's (1993a) 1D relative rate test returns a statistically significant rate difference only in the comparison of *G. raimondii* and *G. herbaceum*.

CHAPTER 5. GENERAL CONCLUSIONS

An understanding of the evolutionary dynamics of nuclear-encoded gene families is important for a number of reasons, both theoretical and practical. From the theoretical point of view data are becoming available to test long-standing hypotheses about the structure and origin of nuclear gene families. A prime example of this is the widely held hypothesis that pairs of nuclear genes such as "*Adh1*" and "*Adh2*" are the result of an ancient gene duplication and therefore all "*Adh1*" genes are more closely related to each other than any are to "*Adh2*" genes (Gottlieb 1982). Similarly, nuclear-encoded gene families in angiosperms are often considered to be relatively stable in terms of the number of loci they include (Clegg et al. 1997). Data presented in this dissertation challenge both of these hypotheses.

From a more practical point of view, nuclear-encoded gene families represent a largely untapped reservoir of sequences that can be co-opted for studies of phylogeny, molecular evolution, and genetic diversity. In plants to date, data to address these issues have derived primarily from chloroplast DNA (cpDNA) and nuclear-encoded ribosomal DNA (rDNA) structure and sequences. Both cpDNA and rDNA suffer from limitations of either the number of sequences available for a given question, or their rate of evolution is inappropriate for a given question. Nuclear-encoded genes, however, are almost functionally infinite in terms of the number of loci that can be sampled and they display evolutionary rates that span a wide range, so that a sequence with an appropriate level of variation can be employed. Again, data presented in this dissertation have shown that data from nuclear-encoded sequences can be more informative than data from cpDNA or rDNA.

The goals of the study presented here were to use two model systems – the genus *Gossypium* and the alcohol dehydrogenase gene family – to explore two areas of plant evolutionary biology. The first goal was to describe and characterize the evolution of the alcohol dehydrogenase gene family in both diploid and allotetraploid *Gossypium* species. The second goal was to use the information derived from these foundational studies to address problems in phylogenetics, genetic diversity and molecular evolution using the tools developed in the preliminary study. The three original research papers included in this dissertation represent the fruition of those goals.

The first of these papers [Chapter Two: Organization and evolution of the alcohol dehydrogenase gene family in diploid and tetraploid cotton (*Gossypium* L.)] provides a description of the *Adh* gene family in *Gossypium* species. The data presented include genomic sequences, estimates of the number of loci encoded in the gene family, and structures of the isolated genes. In addition, evolutionary rates were estimated and compared both across loci and

between lineages for the same loci. The results of this study show that the *Gossypium Adh* gene family is the largest yet described in angiosperms and is at least as large as that of *Pinus banksiana* (jack pine), which has the largest published *Adh* gene family in plants (Perry and Furnier 1996). These data run counter to suggestions that the *Adh* gene family has a small and stable number of loci (usually 2-3; Gaut et al. 1996) and may reflect the fact that estimates of the number of genes in a given gene family are often derived from small sample sizes in terms of species assayed. Also, it seems likely that many genes may go undetected due to methodological limitations of the approaches employed for their detection.

In addition, we have shown that gene structure is variable among *Gossypium Adh* loci. Most plant *Adh* genes have a 10 exon/9 intron structure, although a few examples of intron loss have been published (e.g., *Arabidopsis thaliana*, Chang and Meyerowitz 1986). The majority of the *Gossypium Adh* genes we isolated had a 10 exon/9 intron structure as well, but *AdhA* was missing two of the nine introns, and interestingly, these are two of the same introns missing in *Arabidopsis* and some other members of the Brassicaceae.

Finally, we observed evolutionary rate variation, both among loci and among lineages at a given locus. Synonymous substitution rates varied ca. two-fold among loci, while nonsynonymous substitution rates varied almost ten-fold. In addition to among-locus variation we observed statistically significant among-lineage variation at two *Adh* loci (*AdhA* and *AdhC*). In both cases the lineage with the faster rate was the D-(sub)genome lineage, suggesting that some common evolutionary pressure (or lack thereof) is promoting rate acceleration in this lineage relative to the A- and C-genome species.

The second research paper (Chapter Three: The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group) describes an application of the *Adh* data to a practical problem in plant systematics. Specifically, examples abound of studies at low taxonomic levels where there is insufficient genetic variation among species to produce a robust hypothesis of relationships. One common approach to resolve such situations is to employ noncoding regions of cpDNA (e.g., introns or intergenic spacers) because such regions are likely to be under less functional constraint than genes and should therefore accumulate nucleotide substitutions at a faster rate. An alternative approach is to employ a nuclear-encoded sequence, as nuclear-encoded genes have been shown to have a faster evolutionary rate than cpDNA sequences (Wolfe et al. 1987; Gaut 1998). Using the group of closely related allotetraploid *Gossypium* species we tested the phylogenetic utility of seven different cpDNA noncoding regions as well as a nuclear-encoded gene, *AdhC*. Analysis of the cpDNA sequences resulted in little phylogenetic resolution, and weak support for the resolution obtained. The *AdhC* data, on the other hand, provided complete

and robust resolution of the relationships among the species, despite the fact that we sampled over 4 times as much cpDNA data. This increased resolution is due to a nucleotide substitution rate that is three to six times faster in *AdhC* than in the cpDNA sequences. This analysis clearly shows that nuclear-encoded sequences can be useful for phylogenetic analysis and that they may provide greater resolution than many previously used data sets.

The third research paper [Chapter 4: Low levels of nucleotide diversity at homocologous *Adh* loci in allotetraploid cotton (*Gossypium* L.)] describes our exploitation of the *Adh* data to explore a question of genetic diversity in allopolyploid cotton. Because allopolyploid cottons are hypothesized to have formed only once we can postulate that the subgenomes of tetraploid cotton ought to have originated with equivalent levels of genetic diversity. Thus, all things being equal, each subgenome ought to harbor equivalent levels of genetic diversity in extant species as well. Alternatively, evolutionary pressures acting differentially on the subgenomes may allow one to acquire diversity at a faster rate than the other. To attempt to distinguish among these alternatives we sequenced *AdhA* from both subgenomes of 22 accessions of the tetraploid *G. hirsutum*, as well as five accessions of a closely related tetraploid, *G. barbadense*. These accessions were chosen to represent the genetic and geographic diversity of the species (Wendel et al. 1992; Brubaker and Wendel 1994). The results show that, as indicated earlier, genetic diversity is low in *G. hirsutum*. Despite the low levels of genetic diversity observed, however, the D-subgenome of both *G. hirsutum* and *G. barbadense* harbored greater allelic and nucleotide diversity than the A-subgenome, suggesting that differential evolutionary pressures are acting on these two subgenomes.

Taken together, these analyses have broadened our understanding of the evolutionary dynamics of nuclear-encoded gene families, and have provided evidence that nuclear-encoded genes can be useful in studies of phylogeny and genetic diversity. These studies point the way for at least two logical extensions of the present work. The generality of the results obtained here can only be determined by additional sampling of both gene families and organismal systems.

One direction to pursue is to perform similar types of analyses in *Gossypium* with other low-copy nuclear-encoded gene families. Such analyses could provide important information on the generality of the inferences from the *Adh* work. For example, the *Adh* studies indicate that the *Adh* gene family in *Gossypium* is larger than reported for other angiosperms; are other *Gossypium* gene families larger than the "average" angiosperm gene family or is the *Adh* gene family unusual? In a similar vein, relative rate differences have been detected within and among loci and lineages in the *Gossypium Adh* gene family; do other gene families also display rate heterogeneity within and among loci and lineages? Evidence has also been provided that nucleotide diversity is unequally apportioned between the subgenomes of allotetraploid cotton at

Adh loci; is this true across all loci, and if so, what mechanistic explanations can be provided for such a bias? We have not detected any evidence of intersubgenomic interactions at *Adh* loci in the allotetraploid species, despite evidence presented for repetitive DNA that such interactions are not unusual (Wendel et al. 1995; Hanson et al. 1998); does this observation hold for other low-copy nuclear encoded genes?

An alternative approach is to apply the tools developed in *Gossypium* to phylogenetic or molecular evolutionary studies of other organismal groups. The tools and inferences derived from the *Gossypium Adh* studies may be best suited for studies of other members of the Malvaceae. One group that particularly stands out as a candidate is the genus *Hibiscus*, specifically section *Furcaria*. There are a number of parallels between *Hibiscus* sect. *Furcaria* and *Gossypium*: both have a long history of cytogenetic investigation, they have similar geographic distributions, and both contain polyploids (up to dodecaploids in *Hibiscus* sect. *Furcaria*). Despite these similarities, and the potential ornamental value of *Hibiscus* spp., little is known regarding the systematics of *Hibiscus* sect. *Furcaria*. Thus, this group provides a test case in which the *Adh* tools developed in *Gossypium* could be applied to studies of an unknown group.

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). Amer. J. Bot. **81**:1309-1326.

CHANG, C., AND E. MEYEROWITZ. 1986. Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA **83**:1408-1412.

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.

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

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.

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

HANSON, R. E., X.-P. ZHAO, M. N. ISLAM-FARIDI, A. H. PATERSON, M. S. ZWICK, C. F. CRANE, T. D. MCKNIGHT, D. M. STELLY, AND H. J. PRICE. 1998. Evolution of interspersed repetitive elements in *Gossypium* (Malvaceae). Amer. J. Bot. **85**:1364-1368.

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.

WENDEL. J. F., C. L. BRUBAKER, AND A. E. PERCIVAL. 1992. Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. Amer. J. Bot. **79**:1291-1310.

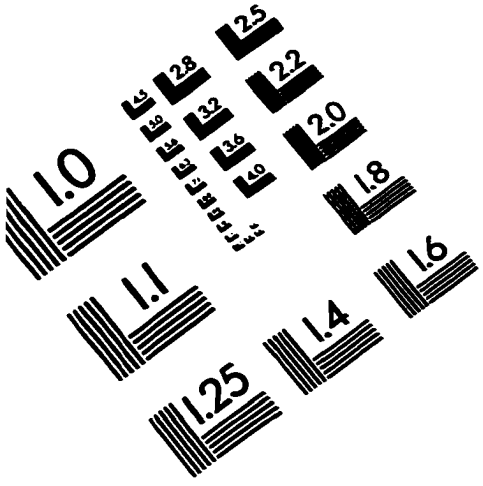
WENDEL, J. F., A. SCHNABEL, AND T. SEELANAN. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA **92**:280-284.

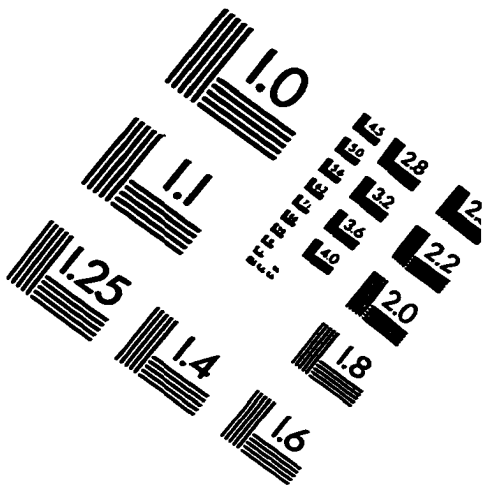
WOLFE, K. H., W.-H. LI, AND P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA. **84**:9054-9058.

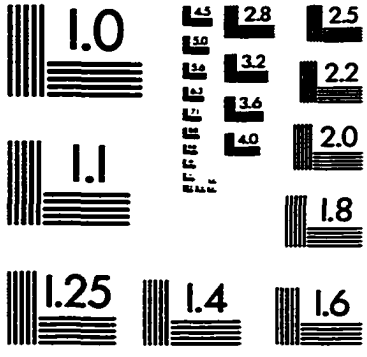
ACKNOWLEDGMENTS

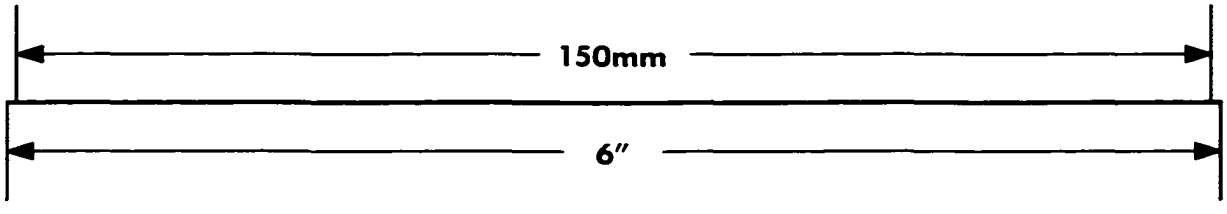
I would like to acknowledge the help and support of a number of people, without whom this dissertation would never have been completed. First, I am indebted to those professors who have believed in me and encouraged me (in somewhat chronological order): David Knoecklein, Gary Wallace, Darrell Moore, Karen Renzaglia, Lee Pike, John Warden, Jim Hickey, Mike Vincent, T.K. Wilson, and Jonathan Wendel. For guidance as well as freedom I thank my POS committee: Jim Colbert, Fred Janzen, Randy Shoemaker, Rob Wallace, and Jonathan Wendel. For the gift of mentorship I am deeply grateful to Jonathan Wendel. Of course, one does not learn in a vacuum and I'd like to thank the many graduate student and post-doc colleagues who helped me along the way; most of all, however, my two comrades-in-arms, Tosak Seelanan and Rich Cronn. For the gift of friendship, dedication, and some damned hard work I thank the best undergraduate "lab slave" to ever walk this (or any other) campus; thanks Jules. Above and beyond all else, to my wife Sarah, and my children Jesse and Laurel, thank you for your patience and your love that give my life meaning.

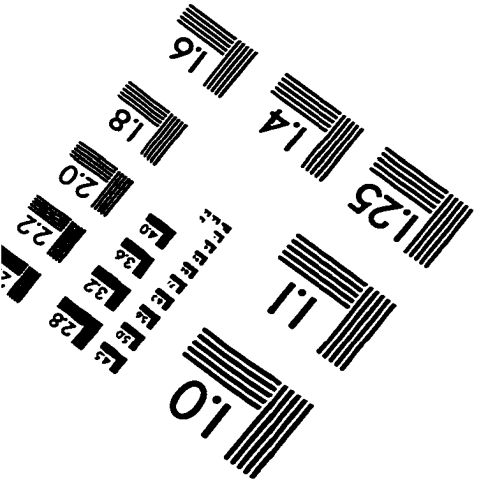
IMAGE EVALUATION
TEST TARGET (QA-3)

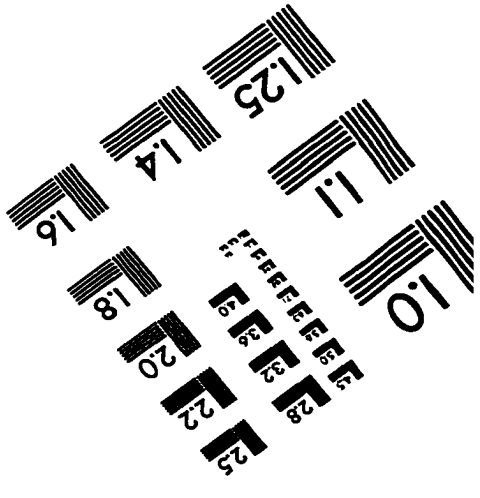












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989