

Phylogeny of the New World diploid cottons (*Gossypium* L., Malvaceae) based on sequences of three low-copy nuclear genes

I. Álvarez¹, R. Cronn², and J. F. Wendel³

¹Real Jardín Botánico de Madrid, CSIC, Madrid, Spain

²Pacific Northwest Research Station, USDA Forest Service, Corvallis, Oregon, USA

³Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA

Received October 4, 2004; accepted December 15, 2004

Published online: May 9, 2005

© Springer-Verlag 2005

Abstract. American diploid cottons (*Gossypium* L., subgenus *Houzingenia* Fryxell) form a monophyletic group of 13 species distributed mainly in western Mexico, extending into Arizona, Baja California, and with one disjunct species each in the Galapagos Islands and Peru. Prior phylogenetic analyses based on an alcohol dehydrogenase gene (*AdhA*) and nuclear ribosomal DNA indicated the need for additional data from other molecular markers to resolve phylogenetic relationships within this subgenus. Toward this end, we sequenced three nuclear genes, the anonymous locus *A1341*, an alcohol dehydrogenase gene (*AdhC*), and a cellulose synthase gene (*CesA1b*). Independent and combined analyses resolved clades that are congruent with current taxonomy and previous phylogenies. Our analyses diagnose at least two long distance dispersal events from the Mexican mainland to Baja California, following a rapid radiation of the primary lineages early in the diversification of the subgenus. Molecular data support the proposed recognition of a new species closely related to *Gossypium laxum* that was recently collected in Mexico.

Key words: *Gossypium*, *Houzingenia*, cotton, phylogeny, low-copy nuclear genes, alcohol dehydrogenase, cellulose synthase.

Introduction

New World, diploid *Gossypium* species comprise a morphological and cytogenetic (D-genome) assemblage (Cronn et al. 2002, Endrizzi et al. 1985, Wendel 1995, Wendel and Cronn 2003) that taxonomically is recognized as subgenus *Houzingenia* (Fryxell 1969, 1979, 1992). This group of plants includes 11 species distributed primarily in SW Mexico and extending northward into Arizona, in addition to two other species with disjunct distributions in Peru and the Galapagos Islands (Fryxell 1992). Although none of these species produces commercially important cotton fiber, the fact that one of the parental lineages of allotetraploid cultivated cotton (*G. hirsutum* L. and *G. barbadense* L.) belongs to this group (Cronn et al. 1999, Endrizzi et al. 1985, Small et al. 1998, Small and Wendel 2000a) gives special relevance to the understanding of their systematics and evolutionary relationships.

Numerous molecular phylogenetic analyses have demonstrated that the subgenus is monophyletic (reviewed in Wendel and Cronn 2003). However, while the circumscription of the subgenus and species boundaries within this

clade are reasonably well-understood (Fryxell 1979, 1992), phylogenetic relationships among species remain unclear, despite numerous studies (Cronn et al. 1996, DeJooode 1992, Fryxell 1971, Liu et al. 2001, Seelanan et al. 1997, Small and Wendel 2000b, Wendel 1995, Wendel and Albert 1992). Phylogenetic analysis based on chloroplast restriction site analysis and chloroplast DNA sequences (Cronn et al. 2003, DeJooode 1992, Seelanan et al. 1997, Wendel and Albert 1992) have led to a number of phylogenetic conclusions that are at odds with numerous, unlinked nuclear markers and morphological trends in the genus. Results from nuclear ribosomal sequences (Cronn et al. 1996, Seelanan et al. 1997) have provided equally controversial resolutions (e.g. Wendel et al. 1995a,b), and resolution of these closely related species appears hampered by the presence of intraindividual polymorphism, some of which appears trans-specific.

The use of low-copy nuclear genes to infer plant phylogenies is rapidly increasing, due in part to the recent accessibility of many nuclear genes (characterized in gene discovery and genome sequencing projects) and their higher resolution, as has been demonstrated for various groups (reviewed in Sang 2002, Small et al. 2004). In *Gossypium*, the characterization of numerous low-copy nuclear genes (Cronn et al. 2002, Cronn et al. 1999, Senchina et al. 2003, Small and Wendel 2000a) has yielded a wealth of nuclear gene markers for studying *Gossypium* (Cronn et al. 2002, Small et al. 2004, Small et al. 1998, Small and Wendel 2000b) and related genera (Wendel et al. 2002). To date, these markers have been applied primarily to just a few representatives of subgenus *Houzingenia*, as studies to date have focused on higher-level phylogenetic relationships (Cronn et al. 2002, 2003; Wendel et al. 2002).

The most comprehensive molecular systematic study of New World cottons to date was performed by Small and Wendel (2000b) using a member of the *Adh* gene family in *Gossypium* (Cronn et al. 1999, Small and Wendel 2000a). While two to three allozyme

loci can be resolved for *Adh* in *Gossypium* (Wendel, unpublished), the gene family is much larger, including up to seven discrete loci in some diploid cotton species (Small and Wendel 2000a). One of these loci, *AdhA*, was selected for phylogenetic applications (Small and Wendel 2000b) based on its homologous chromosome location in three genetic maps including diploid and allotetraploid cottons (Brubaker et al. 1999), and on the results of Southern hybridization analyses (Cronn et al. 1999) that indicate the existence of only one copy per genome in two species of the subgenus (*G. raimondii* Ulbrich and *G. trilobum* (DC.) Skovsted). The phylogeny based on *AdhA* (Small and Wendel 2000b) supports the taxonomically recognized subsections and is generally congruent with previous analyses in the subgenus (Cronn et al. 1996, DeJooode 1992, Seelanan et al. 1997, Wendel and Albert 1992), although relationships among sections and subsections remained unresolved.

More recently, Cronn et al. (2003) used four chloroplast genes and eight low-copy nuclear genes to reevaluate the evolutionary history of *G. gossypoides* (Ulbrich) Standley. Among the nuclear genes were two alcohol dehydrogenase genes (*AdhA*, *AdhC*), two cellulose synthase genes (*CesA1*, *CesA1b*), a fatty acid desaturase intron (*FAD2-1 intron*), and the anonymous genes *A1341*, *G1121*, and *G1262*. In this study, six of the 13 D-genome species were included. The individual molecular markers used in this study revealed levels of variation that provided modest resolution of New World species; however, results from this study indicated that combining the most informative genes might have the potential to resolve phylogenetic relationships among all 13 species in the subgenus. With this objective in mind, we sequenced the three most informative nuclear genes (*A1341*, *CesA1b*, and *AdhC*) that showed variation at different levels (Cronn et al. 2002, Seelanan et al. 1999, Senchina et al. 2003, Small et al. 1998) from representatives of all species. To these data we added previously generated sequences of the *AdhA* gene (Small and Wendel 2000b).

Insights from these four nuclear genes are compared to results obtained with nuclear ribosomal DNA, and the most recent taxonomic treatment of Fryxell (1992) is evaluated in light of these combined results.

Materials and methods

Plant materials. Sampling included the 13 American diploid cottons: *Gossypium aridum* (Rose & Standley) Skovsted, *G. armourianum* Kearny, *G. davidsonii* Kellogg, *G. gossypoides* (Ulbrich) Standley, *G. harknessii* Brandegeee, *G. klotzschianum* Andersson, *G. laxum* Phillips, *G. lobatum* Gentry, *G. raimondii* Ulbrich, *G. schwendimanii* Fryxell & Koch, *G. thurberi* Todaro, *G. trilobum*, and *G. turneri* Fryxell. Besides, we included one specimen (*Gossypium* sp.) that is suggested to be a new species related to *G. aridum* (Ulloa et al., unpublished). In some cases, more than one accession per species was sampled, based on our assessments of variability within each species (some species are narrowly distributed and relatively invariable morphologically and with respect to allozyme markers, whereas others are more widespread and/or exhibit greater variation). Six species that belong to different cytogenetic and taxonomic groups: *G. anomalum* Wawra ex Wawra & Peyritsch, *G. bickii* Prokhanov, *G. longicalyx* J. B. Hutchinson & Lee, *G. robinsonii* F. von Mueller, and *Gossypoides kirkii* (Mast.) J. B. Hutchinson and *Kokia drynarioides* (Seemann) Lewton were included as an outgroup (Table 1). At least one accession per species is identical to those used by Small and Wendel (2000b), so that direct evaluation of the *AdhA* data could be made. For the genes *A1341*, *AdhC*, and *CesA1b*, we included sequences already published for six species and the outgroup (Cronn et al. 2002), and for the remaining we used DNAs available from previous studies (Cronn et al. 2002, Cronn et al. 1996, DeJoode 1992, Seelanan et al. 1997, Wendel and Albert 1992). In a few cases, we newly isolated total DNA from plants grown in the greenhouse (at Iowa State University), using fresh leaf tissue and the Plant DNeasy kit (Qiagen) following the manufacturer's instructions. Vouchers for these plants were deposited at the Ada Hyden Herbarium (ISC) at Iowa State University, Ames.

Molecular markers. We used sequences of three independent nuclear genes (*A1341*, *AdhC*,

and *CesA1b*) as molecular markers. This selection was based on copy number of each gene (inferred to be single-copy from earlier work) and on the knowledge of their orthology across different genomes in cotton (Brubaker et al. 1999, Cronn and Wendel 1998, Small and Wendel 2000a). Additionally, we selected genes that in previous analyses (Cronn et al. 2002) showed a relatively high ratio of phylogenetically informative sites (PI) compared to other low-copy nuclear genes. The *A1341* locus (0.7 kb) is an anonymous gene that corresponds to a *PstI* mapping probe (Brubaker and Wendel 1994, Cronn and Wendel 1998, Cronn et al. 1999); this gene has a PI ratio slightly higher than the *AdhA* gene used in the previous phylogenetic analysis of the subgenus (Small and Wendel 2000b). A region of the *CesA1b* gene (Cronn et al. 2002, Cronn et al. 1999) has a similar PI ratio to plastid genes, although this low ratio is compensated by its length (1.15 kb). From the *Adh* gene family, we sequenced a portion (0.94 kb) of the *AdhC* gene that has a high PI ratio similar to some chloroplast spacers, and higher than other nuclear genes in a previous analysis (Cronn et al. 2002).

Amplification primers were those used previously (Cronn et al. 2002), namely A1341F and A1341R for the *A1341* locus, CelAF and CelAR for the *CesA1b* partial gene, and ADHx4-3 and ADH-P2 for the *AdhC* partial gene. For accessions of *G. lobatum*, a new set of internal forward and reverse primers (GTG AGG CTT CTA GGA TCA TTG G and CCA ATG ATC CTA GAA GCC TCA C, respectively), were used in a second PCR in order to obtain enough amplification product to sequence the *AdhC* partial gene. To amplify the three loci we followed protocols already described (Cronn et al. 1999, Small et al. 1998). The DNA Sequencing Facility of Iowa State University carried out direct sequencing of all amplification products.

Data analysis. Sequence alignment was performed manually using BioEdit v.5.0.9 (Hall 1999). Genomic sequences were aligned to previously published (Cronn et al. 1999) exon sequences for the corresponding gene, which aided determination of intron/exon boundaries. Alignment was straightforward in all cases, as indels were rare and uncomplicated. Data matrices are available at <http://www.eeob.iastate.edu/faculty/WendelJ/datasets.htm>

Table 1. Plant materials used, indicating geographic origin, voucher, and GenBank accession numbers for the three genes sequenced (*Al341*, *AdhC*, and *CesA1b* respectively). Sequences in italics are from previous work (Cronn et al. 2002)

Species	Geographic origin	Voucher ID	GenBank accession numbers
<i>G. anomalum</i>	Africa	JFW &TDC 305	<i>AF403074, AF419966, AF419974</i>
<i>G. aridum</i>	Mexico, Jalisco	DRD 185	AY699077; AY699104; AY699084
<i>G. aridum</i>	Mexico, Colima	DRD 168	AY699105; AY699085
<i>G. aridum</i>	Mexico, Colima	IA 1-4	AY699106; AY699086
<i>G. aridum</i>	Mexico, Guerrero	IA 14-1	AY699107; AY699087
<i>G. aridum</i>	Mexico, Sinaloa	IA 36-1	AY699109; AY699089
<i>G. armourianum</i>	Mexico, Baja California	D2-1-7	AY699078; AY699110; AY699090
<i>G. bickii</i>	Australia	JFW &TDC 557	<i>AF403077, AF419968, AF419977</i>
<i>G. davidsonii</i>	Mexico, Baja California	32	<i>AF520737, AY125059, AY125071</i>
<i>G. gossypoides</i>	Mexico, Oaxaca	D6-2	<i>AF520736, AY125058, AY125070</i>
<i>G. harknessii</i>	Mexico, Baja California	D2-2	AY699079; AY699111; AY699091
<i>G. klotzschianum</i>	Galapagos Islands	D3k-3	AY699080; AY699112; AY699092
<i>G. klotzschianum</i>	Galapagos Islands	IA 54	AY699113; AY699093
<i>G. laxum</i>	Mexico, Guerrero	L. Phillips	AY699081; AY699119; AY699094
<i>G. laxum</i>	Mexico, Guerrero	IA 26-2	AY699120; AY699095
<i>G. laxum</i>	Mexico, Guerrero	IA 25-2	AY699121; AY699096
<i>G. laxum</i>	Mexico, Guerrero	IA 40-4	AY699122; AY699097
<i>G. lobatum</i>	Mexico, Michoacan	DRD 157	AY699123; AY699100
<i>G. lobatum</i>	Mexico, Michoacan	DRD 161	AY699082; AY699115; AY699099
<i>G. lobatum</i>	Mexico, Michoacan	IA 57-6	AY699114; AY699098
<i>G. longicalyx</i>	Tanzania	TS 8	<i>AF403076, AF419967, AF419976</i>
<i>G. raimondii</i>	Peru	No accession no.	<i>AF136815, AF036568, AF139449</i>
<i>G. robinsonii</i>	Australia	AZ-50	<i>AF136817, AF036567, AF139451</i>
<i>G. schwendimanii</i>	Mexico, Michoacan	No accession no.	<i>AF520738, AY125060, AY125072</i>
<i>G. schwendimanii</i>	Mexico, Michoacan	IA 56-3	AY699116; AY699101
<i>G. thurberi</i>	Arizona	D1-17	AY699083; AY699117; AY699102
<i>G. thurberi</i>	Arizona	D1-8	AY699118; AY699103
<i>G. trilobum</i>	Mexico, Sinaloa	No accession no.	<i>AF520739, AY125061, AY125073</i>
<i>G. turneri</i>	Mexico, Sonora	D10-3	<i>AF520740, AY125062, AY125074</i>
<i>Gossypium</i> sp.	Mexico, Guerrero	IA 64-1 (US72)	AY699108; AY699088
<i>Gossypoides kirkii</i>	Madagascar	TS 3	<i>AF201877, AF169254, AF201887</i>
<i>Kokia drynarioides</i>	Hawaiian Islands	TS 6	<i>AF403078, AF419969, AF419978</i>

Phylogenetic analyses were conducted using maximum-parsimony as implemented in PAUP*4.0b10 (Swofford 1999) and a heuristic search with the TBR and ACCTRAN option for character optimization. Gaps were treated as missing data. To obtain the most parsimonious trees (m.p.t.), 100 random addition sequences were performed, saving 1000 trees per replicate. Relative support for tree branches was assessed by using

decay and bootstrap analyses. Decay values were obtained by the converse constraints approach (Bremer 1994) with the aid of the program AutoDecay (Eriksson 1998) for PAUP*. Bootstrap analyses were performed with a fast-heuristic search of 1000 replicates.

To assess congruence among datasets, all possible combinations of the three datasets obtained plus the *AdhA* dataset (Small and Wendel