RESEARCH ARTICLE

Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack

C. E. Grover · X. Zhu · K. K. Grupp ·
J. J. Jareczek · J. P. Gallagher · E. Szadkowski ·
J. G. Seijo · J. F. Wendel

Abstract Understanding the relationship between domesticated crop species and their wild relatives is paramount to germplasm maintenance and the utilization of wild relatives in breeding programs. Recently, *Gossypium ekmanianum* was resurrected as an independent species based on morphological analysis of specimens obtained from the Dominican Republic, where the original type specimen was collected. The molecular data presented here support the recognition of *G. ekmanianum* Wittmack as a distinct species that is phylogenetically close to *G. hirsutum* L. Analyses of chloroplast DNA data reveal species-specific, indel polymorphisms that unambiguously distinguish *G. ekmanianum* samples from other polyploid congeners. Furthermore, analysis of accessions that originated from the Dominican Republic demonstrate the cryptic inclusion of this sister taxon within the US National Plant Germplasm System, a germplasm collection maintained for diversity preservation and future breeding resources. The data presented here indicate that "wild" *G. hirsutum* accessions may include the closely related *G. ekmanianum*, and provide a method to easily distinguish the two.

Keywords Breeding · Cryptic species · Germplasm · *Gossypium ekmanianum* · Polyploid

C. E. Grover · K. K. Grupp · J. J. Jareczek ·
J. P. Gallagher · E. Szadkowski · J. F. Wendel (✉)
Department of Ecology, Evolution, and Organismal
Biology, Iowa State University, 251 Bessey Hall, Ames,
IA 50011, USA
e-mail: jfw@iastate.edu

X. Zhu
School of Life Sciences, Nantong University,
Nantong 226019, China

J. G. Seijo
Instituto de Botánica del Nordeste (UNNE-CONICET)
and Facultad de Ciencias Exactas y Naturales y
Agrimensura, Universidad Nacional del Nordeste, Casilla
de Correo 209, 3400 Corrientes, Argentina

Introduction

Cotton is the world's largest source of renewable plant-based fiber, with the allopolyploid species, *Gossypium hirsutum*, dominating global cotton production. *Gossypium* is remarkable in that it has been independently domesticated four times from four species, two polyploid and two diploid (Wendel and Cronn 2003). The other domesticated polyploid, *G. barbadense* L. (i.e., Pima cotton), comprises a smaller fraction of cotton plantings but is highly valued for the long, strong fibers characteristic of the species (http://www.cotton.org). Together, these two species represent about 99 % of world-wide cotton production. Given the prominence of cotton in the textile industry, considerable attention has been paid to establishing a complete, justified taxonomic and phylogenetic framework for the genus.

Gossypium is remarkably diverse, with over 50 recognized species divided into 8 diploid genome groups (Endrizzi et al. 1985) and a single, monophyletic (Grover et al. 2012) polyploid genome group. As a whole, the genus is distributed throughout the tropics and sub-tropics of Africa, Asia, Australia, and the New World (Wendel and Cronn 2003); however, the native distribution of polyploid *Gossypium* is far narrower, encompassing coastal tropical and subtropical regions and islands in the Caribbean, northern South America, and Central America, with some longer-distance range extensions into Hawaii (*G. tomentosum* Nutt. ex Seem.) and islands in the south Pacific (mostly by *G. hirsutum*) (Brubaker and Wendel 1994; Wendel and Cronn 2003). The primary littoral habitat of polyploid *Gossypium* may be unsurprising, in that this lineage established following a remarkable transoceanic dispersal of an African diploid A-genome cotton species to the New World, which subsequently combined with a native D-genome species, leading to a new clade of polyploids.

A great deal of morphological diversity is present within *G. hirsutum*, which exists in a continuum of morphologies ranging from wild to domesticated. Consequently, despite its agronomic importance, *G. hirsutum* has experienced significant taxonomic changes and reclassification of accessions. Recently, Krapovickas and Seijo (2008) resurrected *Gossypium ekmanianum* as a species distinct from *G. hirsutum*, with which it has been repeatedly merged. Briefly, *G. ekmanianum* was first proposed as a species in 1928 by Wittmack 1928, who reserved this name for specific populations of cottons endemic to the Dominican Republic that displayed "perfectly wild" morphological characteristics and which inhabited extremely arid parts of the island (Fig. 1). Subsequently, Roberty reclassified *G. ekmanianum* first as a varietal form of *G. latifolium* Murray (Roberty 1942; Fryxell 1969), and later as a variety of *G. hirsutum* (Roberty 1950; Fryxell 1969). Mauer later challenged these placements of *G. ekmanianum* and assigned it to *G. tricuspidatum* Lam. (Mauer 1954; Fryxell 1969). Most subsequent treatises of the genus excluded *G. ekmanianum* altogether, until its recent resurrection by Krapovickas and Seijo (2008).

Krapovickas and Seijo (2008) discovered several uniform populations of cotton plants growing in arid areas of the Dominican Republic, similar to the habitat noted by Wittmack. Initially, these plants were

Fig. 1 Isotype of *Gossypium ekmanianum* Wittmack, M.C.L. 1928 [family Malvaceae], collected by E. L. Ekman (#H 5792) on 1926/04/01 in the Dominican Republic

identified as "wild" *Gossypium* based on their growth habit and gross morphology; however, morphological analysis found that these populations of cottons aligned well with Wittmack's description of *G. ekmanianum* and were morphologically distinct from other polyploid cottons. Based on this information, the authors resurrected *G. ekmanianum* and proposed that these populations should be classified as a sixth species of polyploid cotton (Krapovickas and Seijo 2008). A dichotomous key was created to morphologically distinguish *G. ekmanianum*. Specifically, the character combination that distinguishes *G. ekmanianum* is as follows: (1) 2.5 cm bracteoles with 3–9 teeth; (2) 3-loculed capsules each of <2 cm; (3) lamina lobes that are less than half the length of the lamina; (4) hairy stems with 1–2 branches and stellate hairs; and (5) 1–2 cm submuticous capsule with a 2 mm apex.

In recent years, the synthesis of morphological analysis with genetic analysis has greatly aided in

Table 1 Species and accessions sampled in the present study. GenBank numbers are listed for previously published sequences, and herbaria are listed for the *G. ekmanianum* samples

Genome	Species	Accession	GenBank number/ Herbaria
A	*G. herbaceum*	subsp. *africanum*	GI: 372291463
A	*G. arboreum*	Unknown	
A	*G. arboreum*	Unknown	GI: 372291837
AD	*G. hirsutum*	Unknown	GI: 91208881
AD	*G. hirsutum*	Cascot L7	
AD	*G. hirsutum*	Coker 315	
AD	*G. hirsutum*	Maxxa	
AD	*G. hirsutum*	Paymaster 145	
AD	*G. hirsutum*	TM1	
AD	*G. hirsutum*	TX0051	
AD	*G. hirsutum*	TX0488	
AD	*G. hirsutum*	TX0799	
AD	*G. hirsutum*	TX0808	
AD	*G. hirsutum*	TX0871	
AD	*G. hirsutum*	TX0898	
AD	*G. hirsutum*	TX1075	
AD	*G. hirsutum*	TX1228	
AD	*G. hirsutum*	TX2094	
AD	*G. hirsutum*	TX2263	
AD	*G. hirsutum*	TX2264	
AD	*G. hirsutum*	TX2265	
AD	*G. hirsutum*	TX2266	
AD	*G. hirsutum*	TX2268	
AD	*G. hirsutum*	TX2270	
AD	*G. hirsutum*	TX2271	
AD	*G. hirsutum*	TX2272	
AD	*G. hirsutum*	TX2273	
AD	*G. hirsutum*	TX2274	
AD	*G. hirsutum*	TX2277	
AD	*G. hirsutum*	TX2278	
AD	*G. hirsutum*	TX488	
AD	*G. barbadense*	Unknown	GI: 119368478
AD	*G. barbadense*	GB226	
AD	*G. barbadense*	GB228	
AD	*G. barbadense*	GB249	
AD	*G. barbadense*	GB262	
AD	*G. barbadense*	GB271	
AD	*G. barbadense*	K101	
AD	*G. barbadense*	PS6	
AD	*G. tomentosum*	Unknown	GI: 372291365
AD	*G. tomentosum*	Kahoolawe	
AD	*G. mustelinum*	Unknown	GI: 372291753
AD	*G. mustelinum*	JA Lee	
AD	*G. darwinii*	PW7-2	
AD	*G. darwinii*	Unknown	GI: 372291012

Table 1 continued

Genome	Species	Accession	GenBank number/ Herbaria
AD	*G. ekmanianum*	G. Seijo 3909	JBSD, CTES, USD, SI, BM, MO, K
AD	*G. ekmanianum*	G. Seijo 3919	JBSD, CTES, USD
AD	*G. ekmanianum*	G. Seijo 3918	JBSD, CTES, USD
AD	*G. ekmanianum*	G. Seijo 3917	JBSD, CTES, K, SI
AD	*G. ekmanianum*	G. Seijo 3922	JBSD, CTES, USD, SI, MBM

List of herbarium codes: JBSD = Jardín Botánico Nacional, Dr. Rafael M. Moscoso (Dominican Republic); CTES = Instituto de Botánica del Nordeste (Argentina) = SI: Instituto de Botánica Darwinion (Argentina); BM = British Museum of Natural History (UK); MO = Missouri Botanical Garden (USA); MBM = Museu Botânico Municipal (Brazil); K = Royal Botanic Gardens Kew (UK)

questions of taxonomic circumscription and distinction, particularly in the assessment of newly discovered taxa, or taxa that remain difficult to classify based on morphological data alone. Here, we evaluate the proposed resurrection and classification of *G. ekmanianum* by using both nuclear and chloroplast sequence data. We confirm *G. ekmanianum* as a separate species. Moreover, we examined accessions of *G. hirsutum* from the Dominican Republic in the USDA National Plant Germplasm System's (NPGS) Germplasm Resources Information Network (GRIN) database (USDA, ARS, National Genetic Resources Program 2014) and report the cryptic presence of *G. ekmanianum* in the NPGS. Our findings here have implications for breeding programs, evolutionary analyses, and the preservation of biodiversity.

Materials and methods

Plant materials and sampling

Putative *G. ekmanianum* leaf tissue was obtained from two sources: (1) Dr. Seijo of the Instituto de Botánica del Nordeste (Krapovickas and Seijo 2008), and (2) the USDA National Cotton Germplasm Collection in College Station, TX. We analyzed five individual samples of desiccated leaf material from voucher specimens collected in the Dominican Republic, from which the morphological assessment of this taxon was based (Krapovickas and Seijo 2008). In addition, to uncover possible misidentified accessions of *G.*

ekmanianum, we obtained all currently available accessions of *G. hirsutum* that had been originally collected in the Dominican Republic from the USDA Cotton Germplasm Collection. The USDA accessions were grown under greenhouse conditions in the Pohl Conservatory at Iowa State University, and leaf material was harvested from one individual of each USDA accession. DNA from both sources, as well as tissue representing a *Gossypium* species from each genome group (Wendel and Cronn 2003), was extracted by using the AutoGen DNA extraction service provided at the Iowa State University (ISU) DNA Facility. A full list of sampling can be found in Table 1.

Chloroplast regions were amplified and sequenced as follows. Twenty-seven primer sets were designed by aligning the published cpDNA genomes of *G. hirsutum* and *G. barbadense*, and selecting for regions that had indel polymorphisms between these two species. These primer sets were tested for their ability to reliably distinguish different species of *Gossypium* by amplifying and sequencing a small set of diverse accessions of *Gossypium* (as per Grover et al. 2012), and evaluating the data for species-diagnostic sequence differences. Of these 27 sets, 11 primer sets (Table 2) had informative nucleotides and/or indels capable of distinguishing among species and were used to evaluate the putative *G. ekmanianum* samples, along with representatives of each diploid and polyploid genome group. In cases where amplification/sequencing was not successful, the subsequent release of whole cpDNA sequences for cotton species (Table 1) was leveraged to fill in missing data. In addition, for 4 informative loci, deep sampling of multiple accessions within *G. hirsutum* and *G. barbadense* was employed to ascertain the specific identity of the Dominican accessions of *G. hirsutum* (Table 3). Sequence data were aligned with MUSCLE (Edgar 2004) and consensus sequences were generated by CodonCode Aligner (v4.0.4; Codon Code Corporation), Bioedit v7.1.3 (Hall 1999), or Seaview v4.3.0 (Gouy et al. 2010). Distinguishing indels and nucleotide polymorphisms were tabulated (Table 2).

Previously generated sequence capture data from our laboratory (not presented herein) were mined for nuclear genes that had sufficient coverage from a representative sample of *G. ekmanianum*, as well as representatives of every other polyploid species. Briefly, targeted sequence capture of *G. barbadense* accessions was performed by using Mycroarray MY-baits custom bait libraries (http://www.mycroarray.

Table 2 Primer sequences for chloroplast indel regions

4F	5′-GCC ACC GAA TCA GCG GCT TTA TTT-3′
4R	5′-CCG GTC ACG GCA ACA TAA CAA CAT-3′
5F	5′-CGA CAT GTA TGA AGA TAC ATA TTG TG-3′
5R	5′-CAG ATG ATT GTA CAA GGA AGG CGG-3′
7F	5′-GCT TCT CCT GAT GGT TGG TTG AGT-3′
7R	5′-GGA CAG ATA CTT TCT TCA GTT CAG GG-3′
8F	5′-TGG TTT CTA AGA TCG CTT CCG GCT-3′
8R	5′-AGC TCA GAG GTT AGA GCA TCG CAT-3′
14F	5′-TGC GCT ACA TCC CTT TCA ATT GGC-3′
14R	5′-ACC CGA GTT TCG GTG AGG TTG AAT-3′
19F	5′-TTC TGG CAA TGC AAT GGA GTT GGG-3′
19R	5′-TGA TGA CTT CCA CCA CGT CAA GGT-3′
21F	5′-TCC ATC CAG ATC CCA ATT CCA-3′
21R	5′-AGG TAT AAG AGG ATT GGC CGA AC-3′T
24F	5′-GAT TCC CGA ATA TCC AAC TCA CCA-3′
24R	5′-AAG GAT CCG TAT GTA TCC GCG TCT-3′
25F	5′-GGG CAC GCA TGG AAC CCT TTA TTA-3′
25R	5′-ATC CTT TCT TGC ATG GAC CCT ACC-3′
26F	5′-AGT GAT GCT TTC GGC TAC TGG ACT-3′
26R	5′-AGG ATC CTC CGG AAC CAC AAG AAT-3′
27F	5′-TTC GAT ACG AGT AGG AGA AGC GGT-3′
27R	5′-ATC GAT TGC GTA AAG CCT GCG GTA-3′

com/) on 454 Roche Rapid Libraries prepared at the Georgia Genomics Facility (http://dna.uga.edu/). All reads were trimmed with sickle (http://github.com/najoshi/sickle) with a minimum phred quality threshold of 20 and subsequently mapped onto the recently published *G. raimondii* Ulbr. genome (Paterson et al. 2012) by using GSNAP (Wu and Nacu 2010) in combination with a *Gossypium* specific SNP index. Following mapping, reads were segregated according to parental genome with PolyCat (Page et al. 2013a) and independently reassembled by using BamBam (Page et al. 2013b). The combination of PolyCat and BamBam created a consensus sequence for each gene in the diploid species and polyploid species based on the published *G. raimondii* genome. As there were multiple accessions for both *G. hirsutum* and *G. barbadense* included in the previous experiment and each accession was individually barcoded, those libraries were treated individually such that the reconstructed sequences each represent a single gene from that accession. The sequences for each polyploid species/accession were concatenated for each gene with the appropriate model diploid progenitor (*G. arboreum* L. for A-genome

Table 3 USDA *G. hirsutum* accessions originating from the Dominican Republic

*TX2263, 501443. P-155. Dominican Republic. IRCT. IRCT 1212. Highway 2 West between km 43 and km 44, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

TX2264, 501444. P-156. Dominican Republic. IRCT. IRCT 1213. Highway 2 West between km 43 and km 44, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

*TX2265, 501445. P-157. Dominican Republic. IRCT. IRCT 1214. Highway 2 West at km 44, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

*TX2266, 501446. P-158. Dominican Republic. IRCT. IRCT 1215. Highway 2 West between km 44 and km 45, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

TX2268, 501448. P-160. Dominican Republic. IRCT. IRCT 1217. Old Highway 2 West between km 44 and km 45, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

*TX2271, 501451. P-163. Dominican Republic. IRCT. IRCT 1220. Old Highway 2 West at km 52, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

TX2272, 501452. P-164. Dominican Republic. IRCT. IRCT 1221. Old Highway 2 west between km 52 and km 53, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

*TX2273, 501453. P-165. Dominican Republic. IRCT. IRCT 1222. Old Highway 2 west at km 53, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

TX2274, 501454. P-166. Dominican Republic. IRCT. IRCT 1223. Old Highway 2 west between km 55 and km 56, Azua. Plant 2 m tall, glabrous. Lint: brown. Fuzz: brown. Pollen: yellow. Bolls: very small, round with point. Nectaries: 1. Locules: 3–4. Wild. Seed.

TX2277, 501457. P-169. Dominican Republic. Dooryard, Highway 2, Yaguate, San Cristobal. Plant 4 m tall, sparse pubescence. Leaves lobed. Lint: white. Fuzz: green. Flowers: yellow. Petal spot medium. Pollen: yellow. Bolls: medium-sized, oval, square end. Nectaries: 1. Cultivated- Seed.

Table 3 continued

TX2278, 501458. P-17 J. Dominican Republic. Field, seashore in front of Feria Intl. del Caribe, Santo Domingo. Plant 2 m tall, pilose. Leaves slightly lobed. Lint: white. Fuzz: green/brown. Flowers: yellow. Petal spot medium. Pollen: yellow. Bolls: medium-sized, oval, pointed. Nectaries: 1. Locules: 3–4. Cultivated. Seed.

Accessions denoted by an * indicate accessions that should be reclassified as *G. ekmanianum* based on the present analysis

derived polyploid sequences and *G. raimondii* for D-genome derived sequences), and the 30 longest genes containing sufficient coverage for all taxa were retained for phylogenetic analysis.

Phylogenetic analysis

Chloroplast indel regions were coded for each individual region by SeqState (v1.4.1; http://bioinfweb.info/Software/SeqState) using both simple and modified complex indel coding (Müller 2005, 2006) and removed from the alignment using Gblocks (Castresana 2000). A concatenated dataset was also generated for both simple and modified complex coding. Phylogenetic relationships were assessed by using neighbor-joining (NJ) reconstruction via ClustalW (Larkin et al. 2007) and maximum parsimony via Paup (Swofford 2003) on the indel-coded, degapped sequences. For maximum parsimony, the basic general time reversible model with gamma distribution (GTR + G) was used in a heuristic search with 1,000,000 random addition sequence replicates and TBR branch swapping. Bayesian analyses were also generated via MrBayes (Ronquist and Huelsenbeck 2003) with the following parameters: 3 runs with four chains for 10 million generations and using a burn-in fraction of 25 %. Trees were rooted with the A-genome diploid, *G. arboreum*.

The concatenated dataset containing 30 nuclear genes was used for maximum parsimony and Bayesian phylogenetic reconstructions for both the A- and D-homoeologs. The same parameters (described above) were used for the nuclear reconstruction as were employed for the chloroplast reconstruction. Phylogenetic analyses were conducted for the A- and D-homoeologs independently, using members of the respective diploid genome groups to root the resulting trees.

Table 4 Substitution and indel polymorphism data

Features unique to a species						Features that distinguish AD6 from other polyploid associations					
AD6	AD1	AD2	AD3	AD4	AD5	AD2/AD5	AD1/AD5	AD1/AD2/AD3/AD5	AD1/AD2/AD5	AD2/AD3/AD5	AD1/AD2
1 nt SNP	1 nt SNP	1 nt SNP	1 nt SNP	4 nt indel	(None)	1 nt SNP	51 nt indel	6 nt indel	62 nt PR	5 nt indel	5 nt indel
10 nt indel	1 nt SNP		2 nt indel	11 nt indel		1 nt SNP				5 nt indel	
74 nt indel	1 nt SNP		3 nt SNP			2 nt SNP					
17 nt PR	4 nt indel		6 nt indel			6 nt indel					
114 nt PR	5 nt indel		6 nt indel			6 nt indel					
144 nt PR	5 nt indel		6 nt indel								
	7 nt indel		7 nt indel								

* AD1 = *G. hirsutum*; AD2 = *G. barbadense*; AD3 = *G. tomentosum*; AD4 = *G. mustelinum*; AD5 = *G. darwinii*; AD6 = *G. ekmanianum*

SNP single nucleotide polymorphism, *PR* polymorphic region

Results and discussion

Divergence of *G. ekmanianum* in 11 chloroplast regions

Eleven chloroplast regions were sequenced from the *G. ekmanianum* Dominican Republic samples (Table 2) and representatives of each recognized polyploid species, and were characterized with respect to indel and nucleotide polymorphisms. For 7 regions, the sampling included multiple accessions of both the *G. ekmanianum* samples, as well as of *G. hirsutum*, in which this resurrected species was previously incorporated. Tabulation of SNPs and indel polymorphisms (Table 4) revealed 10 single or multi-nucleotide polymorphisms (range: 1–3 nt in length), 20 indels, and 4 polymorphic indel regions (complex regions of overlapping indels and substitutions) that distinguish the *G. ekmanianum* samples from the established polyploid species (Fig. 2). Most of these differences were shared between the *G. ekmanianum* samples and at least one polyploid species; however, 6 features (1 SNP, 2 indels, and 3 polymorphic regions) were unique to the *G. ekmanianum* samples. While these unique states could reflect intraspecific variation, all of the *G. ekmanianum* samples were uniform across these loci, with the exception of G. Seijo 3917 (Table 2), which consistently matched the *G. hirsutum* sequences. By comparison, for the multiple *G. hirsutum* and *G. barbadense* accessions we sequenced, there were few unique polymorphisms within species. Most intraspecific polymorphisms were represented by an occasional SNP; however, there was a 13 nt indel that distinguished the *G. hirsutum* accessions TX2094 and TX51 from the remaining accessions and a C/T transition that occurred in nearly half of the *G. hirsutum* accessions which likely represents an apomorphy shared by that lineage. These few differences within species may be variability that has been retained in the wild lineages; for example, TX2094 and TX51 are both wild accessions, and therefore were not a part of the domestication bottleneck that affected much of this taxon. Aside from the relatively few differences within species, the polymorphisms observed in these regions typically distinguished species rather than accessions and, thus, add weight to the inference of specific status for *G. ekmanianum*.

The ability of indel polymorphisms to aid in phylogenetic analyses has been used and discussed,

Fig. 2 Diagram depicting the polymorphisms that separate *G. ekmanianum* (AD6) from other polyploid species (AD1 = *G. hirsutum*; AD2 = *G. barbadense*; AD3 = *G. tomentosum*; AD4 = *G. mustelinum*; AD5 = *G. darwinii*). Polymorphisms shared with another species are listed on both branches and the partners designated in parentheses

AD1

1 nt SNP
1 nt SNP
1 nt SNP
4 nt indel
5 nt indel
5 nt indel
7 nt indel
5 nt indel (AD2)
6 nt indel (AD2/AD3/AD5)
51 nt indel (AD5)
62 nt PR (AD2/AD5)

AD4

4 nt indel
11 nt indel

AD3

1 nt SNP
3 nt SNP
2 nt indel
6 nt indel
6 nt indel
6 nt indel
7 nt indel
5 nt indel (AD2/AD5)
5 nt indel (AD2/AD5)
6 nt indel (AD1/AD2/AD5)

AD6

10 nt indel
1 nt SNP   144 nt PR
74 nt indel   17 nt PR
114 nt PR

1 nt SNP (AD2)
1 nt SNP (AD2)
2 nt SNP (AD2)
5 nt indel (AD2/AD3)
5 nt indel (AD2/AD3)
6 nt indel (AD2)
6 nt indel (AD2)
6 nt indel (AD1/AD2/AD3)
51 nt indel (AD1)
62 nt PR (AD1/AD2)

AD5

1 nt SNP
1 nt SNP (AD5)
1 nt SNP (AD5)
2 nt SNP (AD5)
5 nt indel (AD1)
5 nt indel (AD3/AD5)
5 nt indel (AD3/AD5)
6 nt indel (AD5)
6 nt indel (AD5)
6 nt indel (AD1/AD3/AD5)
62 nt PR (AD1/AD5)

AD2

with the general conclusion that indel polymorphisms tend to be important informative characters by either supporting or refining the inferences supported by substitution data alone (Luan et al. 2013; Müller and Borsch 2005; Müller 2006; Simmons et al. 2001). Indel polymorphisms tend to have low levels of homoplasy (Hamilton et al. 2003; Müller and Borsch 2005; Müller 2006; Simmons et al. 2001), in part due to the mutational processes involved in their generation (Müller 2006). The data presented here identify numerous diagnostic indels that distinguish *G. ekmanianum* from the remaining polyploid species, including *G. hirsutum*, which subsequent molecular analyses demonstrate is likely the closest relative of this newly resurrected species (see nuclear analyses below).

Phylogenetic placement of *G. ekmanianum*

Phylogenetic resolution of the chloroplast sequences (DNA + coded gaps) was assessed for each region individually, as well as for a concatenated alignment containing the sequences of each region. Resolution on a per-region basis was typically low (Fig. 3), which reflects the generally low variation in both nucleotide and indel polymorphisms. Due to the high number of indel polymorphisms in region 8, it gave the greatest resolution when using NJ reconstruction; however, the placement of *G. ekmanianum* in this reconstruction is not supported by parsimony nor Bayesian analyses, and it is in conflict with the placement in the concatenated tree (Fig. 3). Despite the uncertain placement of *G. ekmanianum* in the chloroplast
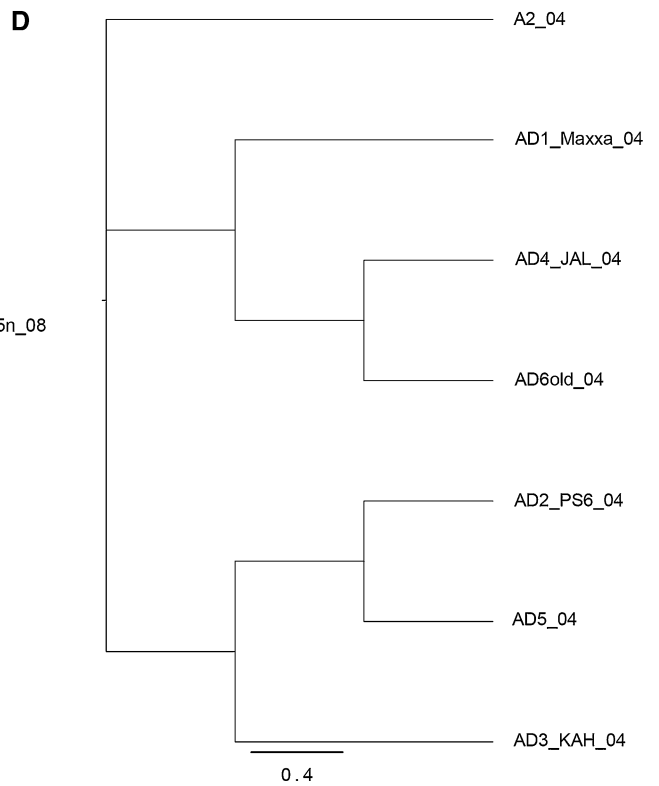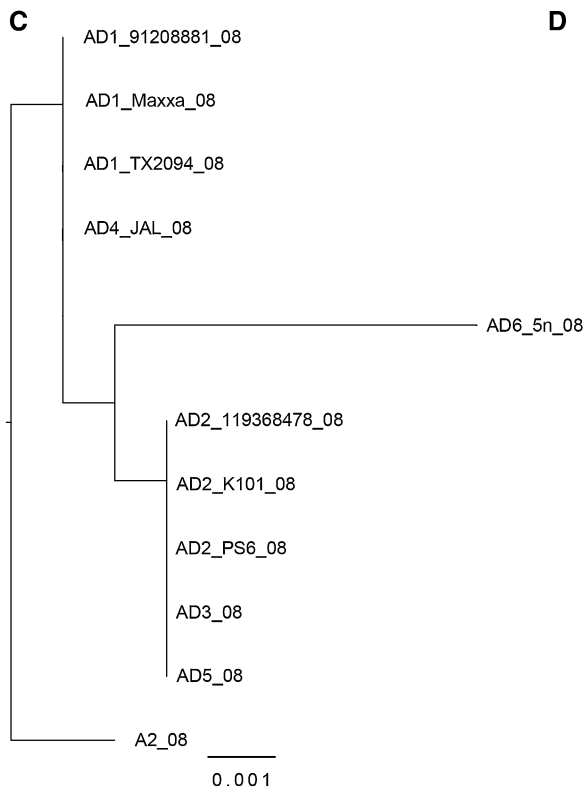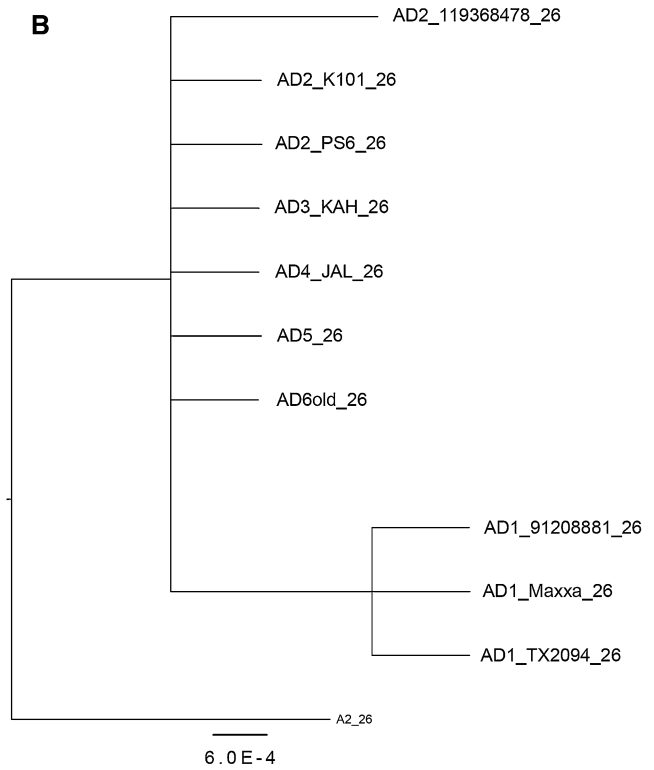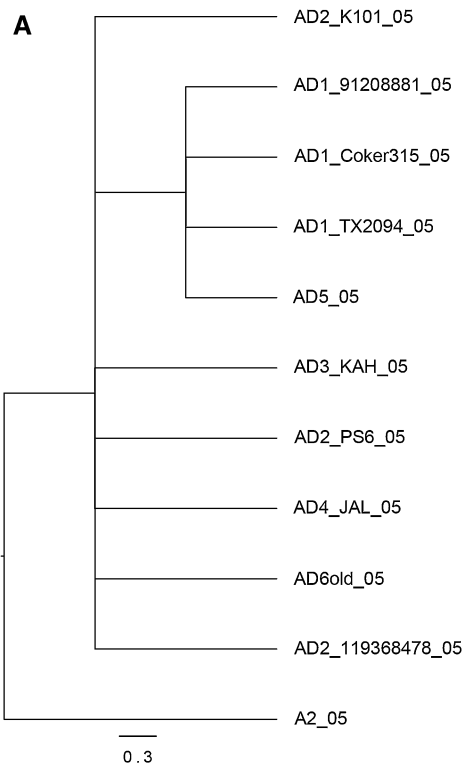
◄**Fig. 3** Phylogenetic analyses of cpDNA sequences and indels for *G. ekmanianum*. Panels **a**, **b**, and **c**, display representative trees from the individual regions for maximum parsimony, Bayesian, and Neighbor-joining (NJ) analyses, respectively. *Panel* **d** shows the maximum parsimony tree for the concatenated regions (Bayesian and NJ are similar, and therefore not displayed)

phylogenies, the separation of *G. ekmanianum* from the other polyploid species does provide molecular evidence for *G. ekmanianum* as a distinct species.

To further assess the placement of *G. ekmanianum* within *Gossypium* polyploid species, we used 30 nuclear genes for parsimony analysis (Fig. 4) and Bayesian analyses (not shown). Both parsimony and Bayesian reconstructions place the putative *G. ekmanianum* sister to the sampled *G. hirsutum* accessions (Fig. 5), which were selected to represent the full spectrum of diversity within *G. hirsutum*. This observation was true for trees generated from both A-genome and D-genome derived

homoeologs. That the putative *G. ekmanianum* sample was not included within the diversity of the G. *hirsutum* accessions sampled, nor allied with any other polyploid species, further supports recognition of *G. ekmanianum* as a separate species.

Cryptic inclusion of *G. ekmanianum* as *G. hirsutum* in established germplasm collections

The exceptional morphological diversity in *G. hirsutum*, as well as the lack of a single explicit distinguishing morphological character for *G. ekmanianum,* is likely responsible for the long-standing inclusion of *G. ekmanianum* within *G. hirsutum*. Given that *G. ekmanianum* was only recently resurrected, we considered the possibility that the publicly available USDA germplasm collection (USDA, ARS, National Genetic Resources Program (12 May 2014) for *G. hirsutum* may unknowingly contain accessions of *G. ekmanianum*. Four of the previously used chloroplast primer sets were
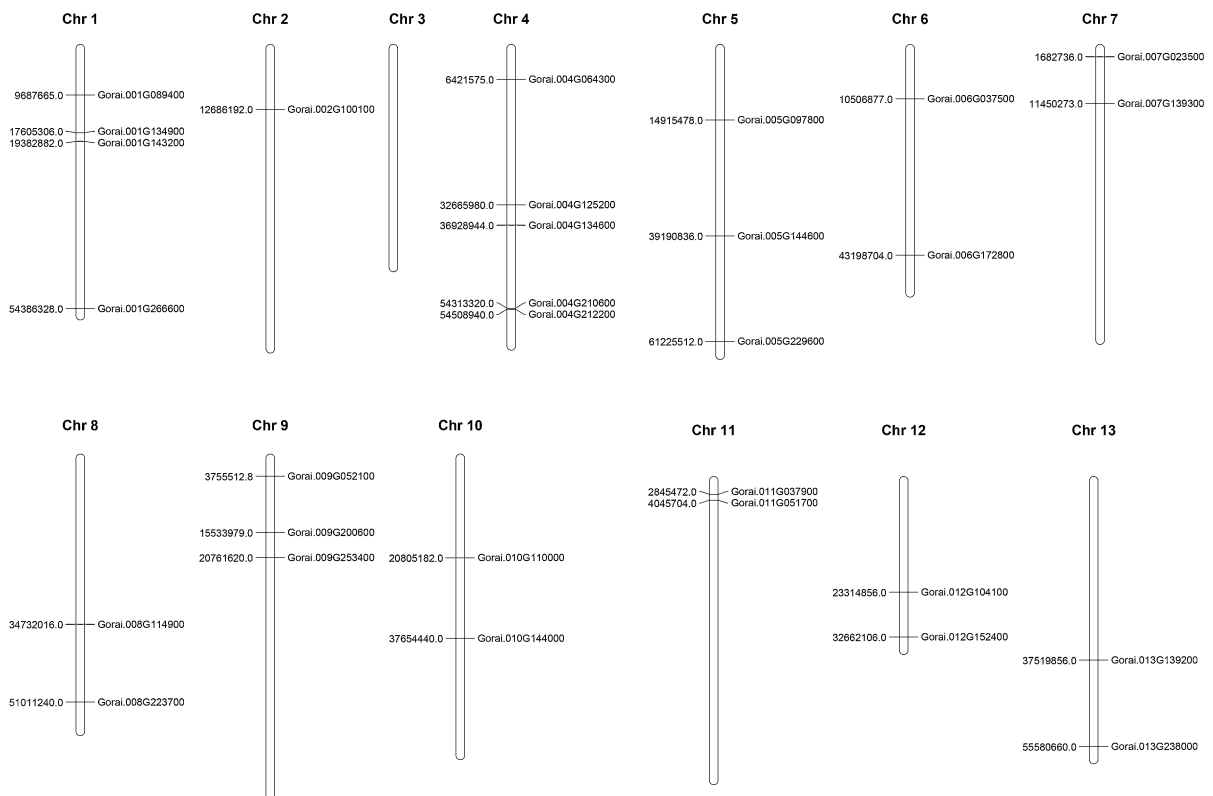
Fig. 4 Identities and genomic locations [based on the sequenced *G. raimondii* genome; (Paterson et al. 2012)] of the nuclear loci used for phylogenetic reconstructions
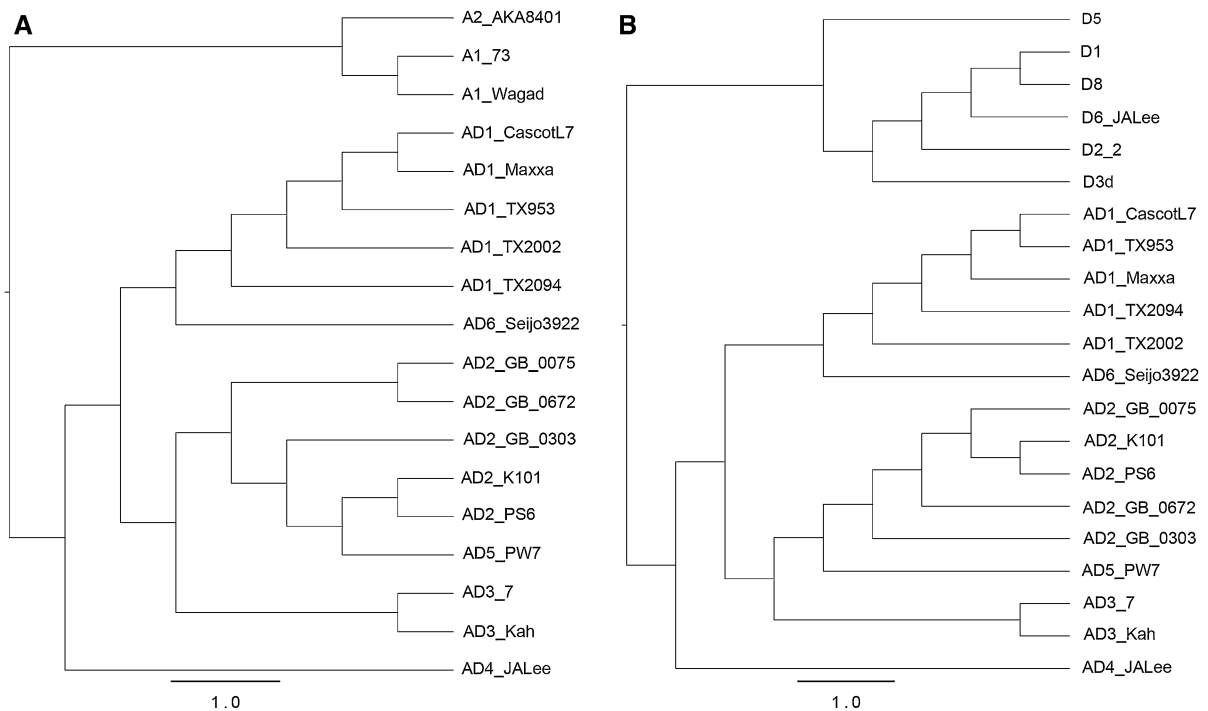
Fig. 5 Example trees for the A- and D-homoeologs (*panels* **A** and **B**, respectively) generated via maximum parsimony analyses (Bayesian results were similar; not shown)

used to screen 11 USDA accessions that were collected in the Dominican Republic (Table 2). Nearly half (5/11) of the USDA *G. hirsutum* samples displayed sequences consistent with the *G. ekmanianum* samples (Fig. 6). While this appears high, it bears noting that this panel of wild to domesticated *G. hirsutum* accessions initially chosen for comparison was selected to represent the range of diversity in the species and were not based on geographic proximity to the Dominican Republic. As *G. ekmanianum* has thus far only been identified in the Dominican Republic, this provides confidence in our assessment that the sequences displayed by these *G. ekmanianum* samples are taxon-specific. We should note, however, that there remain many more accessions of *G. hirsutum* that could be sampled to determine whether the geographic range of *G. ekmanianum* extends beyond the Dominican Republic.

Conclusions

Taxonomic knowledge regarding the relationships between domesticated crop species and their wild relatives is fundamental to both germplasm maintenance and the utilization of wild species in breeding programs. The current USDA germplasm collection for the agronomically important species, *G. hirsutum*, consists of over 6,500 accessions representing the diversity of the species, as well as representatives of four other previously recognized polyploid *Gossypium* species. The data presented here, however, indicate that this germplasm pool may cryptically contain multiple accessions of a newly resurrected sixth polyploid species. Both chloroplast DNA and nuclear genic data provide clear molecular support for *G. ekmanianum* as an independent polyploid species.

The recognition of *G. ekmanianum* as a distinct species has implications for germplasm maintenance, breeding, and evolutionary inferences, and stimulates further questions with respect to the extent of this species within germplasm collections, the native range of this species, and its overall genetic relationships. The molecular data presented here indicate a close relationship between *G. ekmanianum* and *G. hirsutum*, despite the fact that the morphological characters used to distinguish between these two taxa are less clear (Krapovickas and Seijo 2008). Clearly, further
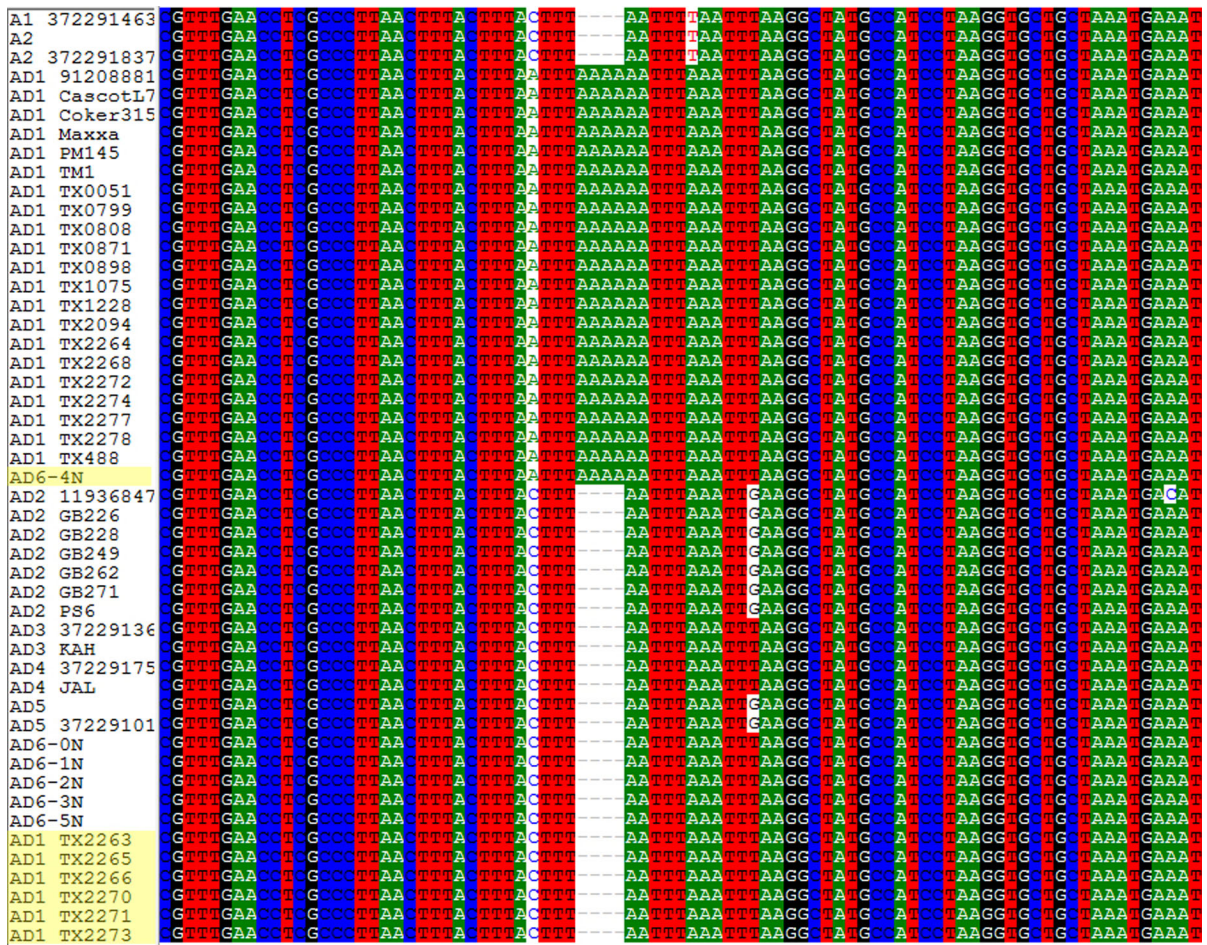
Fig. 6 Example sequence polymorphism between *G. ekmania-num* and *G. hirsutum* (both the indel and the A/C SNP 4 bases before the indel). Accessions that do not belong with their respective grouping are highlighted in *yellow*. One sample of *G. ekmanianum* obtained from G. Seijo (3917) is cytoplasmically *G. hirsutum*-like, and 6 samples of *G. hirsutum* (obtained from the USDA and originating in the Dominican Republic) show *G. ekmanianum*-like sequences

research is necessary to determine (1) the extent of this species in existing collections, and (2) the geographic range of this cryptic polyploid.

References

Brubaker C, Wendel JF (1994) Re-evaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment polymorphisms (RFLPs). Am J Bot 81:1309–1326

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi:10.1093/nar/gkh340

Endrizzi JD, Turcotte EL, Kohel RJ (1985) Genetics, cytology, and evolution of *Gossypium*. Adv Genet 23:271–375

Fryxell PA (1969) A classification of *Gossypium* L. (Malvaceae). Taxon 18:585–591. doi:10.2307/1218405

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224. doi:10.1093/molbev/msp259

Grover CE, Grupp KK, Wanzek RJ, Wendel JF (2012) Assessing the monophyly of polyploid *Gossypium* species.

Plant Syst Evol 298:1177–1183. doi:10.1007/s00606-012-0615-7

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95–98. http://jwbrown.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf

Hamilton MB, Braverman JM, Soria-Hernanz DF (2003) Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in New World species of the Lecythidaceae. Mol Biol Evol 20:1710–1721. doi:10.1093/molbev/msg190

Krapovickas A, Seijo G (2008) *Gossypium ekmanianum* (Malvaceae), algodón silvestre de la República Dominicana. Bonplandia 17:55–63

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. doi:10.1093/bioinformatics/btm404

Luan P-t, Ryder OA, Davis H, Zhang Y-p, Yu L (2013) Incorporating indels as phylogenetic characters: impact for interfamilial relationships within *Arctoidea* (Mammalia: Carnivora). Mol Phylogenet Evol 66:748–756. doi:10.1016/j.ympev.2012.10.023

Mauer FM (1954) Proiskhoshchdenie i sistematika khlopchatnika. Akademii Nauk Uzbekskoi S.S.R, Tashkent

Müller K (2005) SeqState. Appl Bioinfo 4:65–69. doi:10.2165/00822942-200504010-00008

Müller K (2006) Incorporating information from length-mutational events into phylogenetic analysis. Mol Phylogenet Evol 38:667–676. doi:10.1016/j.ympev.2005.07.011

Müller K, Borsch T (2005) Phylogenetics of *Utricularia* (Lentibulariaceae) and molecular evolution of the trnK intron in a lineage with high substitutional rates. Plant Syst Evol 250:39–67. doi:10.1007/s00606-004-0224-1

Page JT, Gingle AR, Udall JA (2013a) PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 3:517–525. doi:10.1534/g3.112.005298

Page JT, Huynh MD, Liechty ZS, Grupp K, Stelly D, Hulse AM, Ashrafi H, Van Deynze A, Wendel JF, Udall JA (2013b) Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. G3 3:1809–1818. doi:10.1534/g3.113.007229

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo M-j, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee T-h, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, Rahman Mu, Rainville LN, Rambani A, Reddy UK, Rong J-k, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MFS, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KFX, Peterson DG, Rokhsar DS, Wang X, Schmutz J (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423–427. doi:http://www.nature.com/nature/journal/v492/n7429/abs/nature11798.html#supplementary-information

Roberty G (1942) Gossypiorum revisionis tentamen. Candollea 9:19–103

Roberty G (1950) Gossypiorum revisionis tentamen. Candollea 13:9–165

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. doi:10.1093/bioinformatics/btg180

Simmons MP, Ochoterena H, Carr TG (2001) Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst Biol 50:454–462. doi:10.1080/10635150120427

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

USDA, ARS, National Genetic Resources Program (12 May 2014) Germplasm Resources Information Network—(GRIN). National Germplasm Resources Laboratory, Beltsville, Maryland. URL: http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Wittmack, L. and S. Fraenkel (1928). Systematik, Botanik und Kultur der Baumwolle, Springer Berlin, 4/1:22–176

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881. doi:10.1093/bioinformatics/btq057