PHYLOGENY OF THE GENUS *Gossypium* AND GENOME ORIGIN OF ITS POLYPLOID SPECIES INFERRED FROM VARIATION IN NUCLEAR REPETITIVE DNA SEQUENCES

A Thesis

by

YING RONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2004

Major Subject: Molecular & Environmental Plant Sciences

PHYLOGENY OF THE GENUS *Gossypium* AND GENOME ORIGIN OF ITS POLYPLOID SPECIES INFERRED FROM VARIATION IN NUCLEAR REPETITIVE DNA SEQUENCES

A Thesis

by

YING RONG

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

Hongbin Zhang (Chair of Committee)	David Stelly (Member)
David H. Byrne (Member)	John Z. Yu (Member)
Marla L. Binzel (Chair of Molecular & Environmental Plant Sciences Faculty)	Mark A. Hussey (Head of Department)

December 2004

Major Subject: Molecular & Environmental Plant Sciences

ABSTRACT

Phylogeny of the Genus *Gossypium* and Genome Origin of Its Polyploid Species Inferred from Variation in Nuclear Repetitive DNA Sequences. (December 2004)

Ying Rong, B.S., Guangxi Agricultural College

Chair of Advisory Committee: Dr. Hongbin Zhang

Knowledge of phylogenetic relationships among taxa is essential for comparative and evolutionary genomic research. Here, we report reconstruction of the phylogenetic tree of the genus *Gossypium* containing cultivated cottons of importance in agriculture by using variation of nuclear repetitive DNA sequences. Genomic DNA was isolated from 87 available accessions of 35 species representing all eight basic genome groups of the genus *Gossypium* and analyzed to infer phylogeny of the genus and genome origin of its polyploid species. Twenty-two interspersed repeated sequence clones derived from *G. hirsutum*, each representing a repeated sequence family, were hybridized to the genomic DNA of the 35 species, respectively. Southern hybridization showed that 15 of the repetitive DNA sequences could be detected in all of the eight diploid genome groups, five were A genome-specific, and two were detected in some of the non D-genome groups. A total of 642 major restriction bands of repeated sequences were used for phylogenetic analysis of the species. A phylogenetic tree of the species was constructed, based on the parsimony method and evaluated by the bootstrap approach. The tree was consistent with those previously constructed with different methods in major clades in

which the genealogical lineages of species are largely congruent with genome designations and geographical distribution; but significantly different branching among some of the species was observed. These results not only further confirm the previously phylogenetic analysis of the species and the utility of repetitive DNA sequences for phylogenetic analysis of the genus *Gossypium*, but also provide new insights into the phylogeny of the genus.

ACKNOWLEDGMENTS

I would like to thank my academic advisor, Dr. Hongbin Zhang, for his full support, encouragement, and instruction in my thesis research; and my other committee members, Dr. David Stelly, Dr. John Z. Yu, and Dr. David H. Byrne, for their valuable comments and suggestions. Appreciation is given to Dr. Edward Percival for his kindly providing most of the cotton species used in this study; Dr. Spencer J. Johnston for reviewing and correcting the thesis; and Dr. Liangtao Zhang for developing all repetitive DNA sequences used in this study. I would also like to acknowledge Dr. Mi-Kyung Lee, Dr. Taesik Uhm, Dr. Chengcang Wu, Dr. Yaning Li, Ms. Chantel Scheuring, Dr. Zhangyou Xu, Mr. Felips Santos, and Mr. Greg Hess for providing assistance on my research. Finally, I would like to express my deep appreciation to my husband, Yiming Jiang, for his assistance and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Cotton, *Gossypium hirsutum* and *G. barbadense*, is the leading textile fiber and the second most important oilseed in the world. In the USA, cotton harvested is more than all other crops except for maize, soybean, and wheat. The combined raw-product value of the U.S. cotton fiber and cotton-seed oil and meal exceed $5.5 billion annually. Annual business revenue stimulated by cotton in the U.S. economy exceeds $120 billion (NASS 1999).

The genus *Gossypium* includes about 45 diploid and 5 polyploid species that occur naturally. Relationships among these species or their selected groups have been studied using several methods including comparative morphology (Fryxell 1979, 1992), intercross fertility and cytology (Endrizzi et al. 1985), and molecular markers (Wendel and Albert 1992; Cronn et al. 1996, 2002). Diploid species ($2n = 26$) are divided into eight genome groups, designated A through G and K on the basis of chromosome size and pairing behavior in interspecific hybrids (Endrizzi et al. 1985). They distribute in Australia (C-, G-, and K-genomes), African-Arabia (A-, B-, E-, and F-genomes), and the Americas (D-genome). Five polyploid species are recognized to date, including the commercially important *G. hirsutum* (‘‘Upland cotton’’) and *G. barbadense* (‘‘Pima’’ and ‘‘Egyptian’’ cotton), and they are traditionally considered to be allotetraploids ($2n =$

This thesis follows the style of Theoretical and Applied Genetics.

52), containing A- and D-subgenomes and being endemic to the New World (Fryxell 1992).

The present phylogenetic relationships of the cotton genome groups was proposed by Wendel and Cronn (2003) according to recent molecular phylogenetic investigations, including largely cpDNA restriction site variation, and nucleotide sequence variation of a limited number of selected chloroplast genes, nuclear ribosomal DNA (5S gene and spacer, 5.8S gene and its flanking internal transcribed spacers) and low-copy nuclear genes (Wendel and Albert 1992; Cronn et al. 1996, 2002). Nevertheless, several significant questions and/or uncertainties about their phylogeny need to be further investigated.

First, uncertainties remain in the phylogenetic tree of the species with respect to several of the earliest branch points and the genome origin of allopolyploids. For example, the phylogenies inferred from different molecular data differ with respect to the resolution of the B-genome species groups. Chloroplast DNA data robustly placed the B-genome lineage sister to the combined Australian (C + G)-genome, whereas the data of nuclear locus analysis placed the B-genome lineage solidly into an African clade that includes A- and F-genome cottons (Cronn et al. 2002).

Second, the phylogenetic tree reveals that *G. raimondii* is the closest living relative of the ancestral D-genome donor and the A-genome donor is most similar to present-day *G. herbaceum*. However, the discovery that a number of A genome-specific repetitive DNAs were found in *G gossypioides* (D-genome) (Zhao et al. 1998) raised the possibility that *G gossypioides* was involved in the origin of allopolyploid cotton. Since

G. gossypioides is the sole D-genome diploid that exhibits evidence of genetic "contact" with A-genome species, it appears that *G. gossypioides* experienced nuclear introgression from an A-genome species shortly after divergence from the lineage leading to *G. raimondii*. This is incongruent with the recent phylogenetic tree, which placed *G. gossypioides* as basal within the subgenus, distant from a lineage comprising *G. raimondii* and the progenitor D-genome donor of the allopolyploids.

Finally, the current phylogenetic tree of the species was largely based on the data derived from chloroplast genome analysis, or resulted from individual genes or locus sequences of the nuclear genome, which are more likely to indicate the phylogenies of the genes or loci themselves, but not the entire plant genomes. Therefore, more lines of evidence from extensive analysis of the nuclear genomes are required for an in-depth understanding of the phylogeny of *Gossypium* and deciphering the genome origin of the allopolyploids.

Plant genomes are composed of repeated and low- or single-copy DNA sequences. Nuclear repetitive DNA sequences provide powerful tools for studies of genome relationships and construction of phylogenetic trees of the species. First, repeated sequences constitute a considerable portion of the genomes of many higher plant species (Flavell et al. 1974), accounting for most of the variation in genome size. Second, the dispersed repetitive DNA elements that represent the majority of repeated sequences in the genomes intersperse with other sequences and disperse throughout the genome, thus being well representative of the entire plant genome. Third, some repeated sequences may be only present in certain related species, but absent or undetectable in others. At

the nucleotide sequence level, they usually show extremely similar or uniform restriction patterns within a species due to their concerted evolution, but can be remarkably variable in closely related species. Finally, since each repetitive element is present in thousands of copies in a genome, a large amount of data could be collected rapidly. Variation in repeated sequence has been previously widely used to infer phylogenetic relationships among related taxa and the genome origin of polyploid plants (e.g., Dvorak and Zhang 1990; Zhang and Dvorak 1991, 1992; Zhao and Kochert 1993).

The genomes of cottons contain abundant repeat sequences (Geever et al. 1989). Recent studies showed that most of the repeated sequences are dispersed in the cotton genomes (Zhao et al. 1995, 1998). The sequences representing most, if not all, of repeated sequence families have been cloned from both the Sea Island cotton (*G. barbadense*) (Zhao et al. 1995) and the Upland cotton (*G. hirsutum*) (Zhang et al. 2002). Together, 163 repeated sequence families have been isolated, of which several subgenome-specific, dispersed repeated sequences have been characterized in detail (Zhao et al. 1995, 1998; Hanson et al. 1998; Zhang et al. 2002). The objectives of the present study were reconstruction of the phylogenetic tree of the species and deciphering of the genome origin of the allopolyploid cotton.

MATERIALS AND METHODS

Plant materials

A total of 87 accessions representing 35 species of *Gossypium* are available at the cotton germplasm bank of USDA/ARS, College Station, Texas. These species represent all genome groups of the *Gossypium* genus (Table 1). Included were five allopolyploid species, two A-genome diploid species, thirteen D-genome species, one F-genome species, two B-genome species, three E-genome species, two C-genome species, three G-genome species, and four K-genome species. Depending on availability of the seeds, from one to five accessions from each species were planted in the USDA/ARS green house at College Station, Texas, for seed production, and the identity of every accession was verified morphologically.

Nuclear DNA isolation

Young leaf tissues of each accession were collected from a single plant verified to represent its accession. Nuclear DNA was isolated with the modified CTAB method, a procedure of nuclear DNA isolation that is routinely used in our laboratory. Briefly, nuclei are first isolated in the extraction buffer (pH 7.5) containing 350 mM Sorbitol, 100 mM Tris, 5 mM EDTA, and 0.38% (w/v) bisulfate, and then lysed to release nuclear

Table 1. *Gossypium* species used for phylogenetic analysis

No.	Species name	Genome	Accession	Origin
1	*G. sturtianum*	C1	C1-4 (EP)	Australia
			C1-1 (EP)	Australia
2	*G. nandewarense*	C1-n	C1-n-5 (EP)	Australia
3	*G. costulatum*	K	C5-3	Australia
			C5-4	Australia
4	*G. nobile*	K	NWA35 (EP)	Australia
5	*G. pulchellum*	K	C8-1 (EP)	Australia
6	*G. marchantii*	K	NWA-6 (EP)	Australia
7	*G. australe*	G	C3-1 (EP)	Australia
8	*G. nelsonii*	G	C9-1	Australia
			C9-2	Australia
9	*G. bickii*	G1	G1-1	Australia
			G1-3	Australia
10	*G. thurberi*	D1	D1-1	Mexico
			D1-7 (EP)	Mexico
11	*G. trilobum*	D8	D8-7 (EP)	Mexico
			D8-8 (EP)	Mexico
			D8-9 (EP)	Mexico
			D8-10 (EP)	Mexico
12	*G. davidsonii*	D_{3d}	D_{3d}-1	Mexico
			D_{3d}-2	Mexico
13	*G. klotzchianum*	D_{3-k}	D_{3-k}-58 (EP)	Ecuador
			D_{3-k}-59 (EP)	Ecuador
14	*G. armourianum*	D_{2-1}	D_{2-1}-7 (EP)	Mexico
			D_{2-1}-9 (EP)	Mexico
15	*G. harknessii*	D_{2-2}	D_{2-2}-4	Mexico
16	*G. turneri*	D10	D10-1	Mexico
17	*G. aridum*	D4	D4-5	Mexico
18	*G. lobatum*	D7	D7-4 (EP)	Mexico
19	*G. laxum*	D9	D9-3 (EP)	Mexico
20	*G. schwendimanii*	D11	D11-1	Mexico
21	*G. gossypioides*	D6	D6-2 (EP)	Mexico
			D6-6 (EP)	Mexico
22	*G. raimondii*	D5	D5-3 (EP)	Peru
			D5-6 (EP)	Peru
			D5-8 (EP)	Peru
			0208082.05 (DS)	
23	*G. herbaceum*	A1	A1-108 (EP)	
			A1-111 (EP)	
			A1-120 (EP)	
			A1-125 (EP)	
			A1-127 (EP)	
			A1-128 (EP)	
			A1-129 (EP)	
			A1-153 (EP)	
			A1-154 (EP)	

Table 1(continued)

No.	Species name	Genome	Accession	Origin
			A1-172 (EP)	
			A1-180 (EP)	
24	*G. arboreum*	A2	A2-67A (EP)	
			0208083.10 (DS)	
			A2-142	
			A2-47	
			A2-84	
25	*G. anomalum*	B1	B1-1 (EP)	Africa
26	*G. capitis-viridis*	B3	B3-1	Portugal
27	*G. longicakyx*	F1	F1-1	Tanzania
			F1-4	Tanzania
28	*G. stocksii*	E1	E1-3	Arabia
			E1-4	Arabia
29	*G. areysianum*	E3	E3-1	Arabia
30	*G. incanum*	E4	0208081.07 (DS)	
			E4-4	
31	*G. hirsutum*	(AD)1	Wild Mexico Jack Jones (FR)	
			Clevewilt 6 (FR)	
			Auburn 56 (FR)	
			Stoneville 213 (FR)	
			Coker 201 (FR)	
			Coker 310 (FR)	
			Deltapine 16 (FR)	
			Deltapine 61 (FR)	
32	*G. barbadense*	(AD)2	Pima S6 (FR)	
			3-79 (RK)	
			(AD)2-201 (EP)	
			(AD)2-81 (EP)	
			(AD)2-372 (EP)	
			K101	
33	*G. tomentosum*	(AD)3	(AD)3-10 (EP)	USA
			(AD)3-15 (EP)	USA
			(AD)3-16 (EP)	USA
			(AD)3-17 (EP)	USA
			(AD)3-25 (EP)	USA
			0208081.05 (DS)	
34	*G. mustelinum*	(AD)4	0208082.04 (DS)	
			(AD)4-9	Brazil
			(AD)4-7	Brazil
35	*G. darwinii*	(AD)5	(AD)5-3	Ecuador
			(AD)5-7	Ecuador

Note: The plants were kindly provided by EP - Edward Percival, DS – David Stelly, RK – Russell Kohel and FR - Forest Robinson.

DNA in a nuclei lysis buffer containing 0.2 M Tris.HCI, 50 mM EDTA, 2.0 M NaCI, and 2% (w/v) CTAB. The DNA is purified with the Chloroform/Isoamyl Alcohol (24:1) mixture and collected by precipitation with Isopropanol. The concentration of isolated DNA is estimated by microfluorometry and agarose gel electrophoresis. Because the isolated DNA of several species accessions contained too much secondary compounds to be digested with restriction enzymes, fresh young leaves were collected from growing tips and used to isolate genomic DNA.

Repeated sequence probes

A total of 163 repeated sequence clones representing 163 repeated sequence families were previously isolated from the Upland cotton (*G. hirsutum*) genetics standard TM-1 (Zhang et al. 2002). All of these clones are available in our laboratory. Twenty-two dispersed repeated sequence clones were randomly selected from the 163 repeated sequences families and used as probes in the Southern analysis of the cotton nuclear DNA.

Southern blot preparation and hybridization

For each accession of the species, approximately 5 μg nuclear DNA of diploid species or 10 μg DNA of polyploid species was digested with three restriction endonucleases, *Eco*RI, *Hin*dIII and *Bam*HI, respectively, fractionated by electrophoresis on 0.8%

agarose gels, and transferred onto Hybond N+ membranes. After blotting, the membrane blots were washed in 2 x SSC (1x: 150 mM NaCl, 15 mM Na3 citric acid, pH. 7.0) and stored at 4^{o}C before use.

The random priming method was used to label repetitive DNA sequence probes with [α-^{32}P] dCTP. The labeled probes were hybridized to those Southern blots of the nuclear DNA of the *Gossypium* species at 65^{o}C in a hybridization solution containing 5 x SSC, 0.5% (w/v) SDS, 25 mM potassium phosphate buffer (pH 6.5), and 5 x Denhardt's solution overnight with gently shaking. The hybridized membranes were washed in preheated (65^{o}C) wash buffer containing 0.2 x SSC and 0.1% (w/v) SDS for three times, 15 - 30 minutes each wash, at 65^{o}C with gentle shaking. The membranes were individually wrapped with the SARAN Wrap and exposed to X-ray film (NEW BioMax, Kodak) with a sheet of intensifying screen in an autoradiography cassette at -80^{o}C for 3 - 36 hours. Finally, the X-films were developed with a Film Processor (M35A X-OMAT, KODAK) in a dark room.

Data analysis and phylogenetic tree reconstruction

Each band on the Southern blot autoradiographs was considered as a phylogenetic character and scored. Presence or absence of each band of a repetitive sequence was scored as a binary unit character, with its presence as "1" and absence as "0". The uncertainty of a band in an accession was scored as "?" for missing data. The data was analyzed by using the PAUP (Phylogenetic Analysis Using Parsimony) program version

4.0b10 (Swofford 2001). The parsimony method was used to construct the phylogenetic tree of the species with the heuristic search. The reliability of each branch of the tree was assessed by use of the bootstrap method with 100 replications. A program TREEVIEW was used for displaying and printing phylogenetic trees (Page 1996)

Inference of genome origin of polyploid cottons

The genome origin of the *Gossypium* polyploid species was inferred by calculating repeated sequence correspondence (RSC) between a diploid and a polyploid according to Zhang and Dvorak (1991, 1992).

(1) Character bands: To simplify the procedure of genome origin inference, a band detected only in one diploid genome group was defined a genome-specific band (GSB). If a GSB was also observed in a polyploid species, the band was defined a genome-marker band (GMB). Similarly, if a band was detected only in a single taxon (species), but not in others, it was considered to be species-specific band (SSB). Further, if a SSB was also detected in a polyploid species, it was defined a species-marker band (SMB). Therefore, the GMBs or GMBs could be used as diagnostic tools for identification of the genome origin of the *Gossypium* polyploids.

(2) RSC: The RSC is calculated using the following formula:

$$RSC = \sum SMBs / \sum SSBs \quad or \quad RSC = \sum GMBs / \sum GSBs$$

An RSC value reflects a relationship between a diploid species and the polyploid species under study. The value of an RSC ranges from zero to one. If none of GSBs of a diploid is encountered in a polyploid, the RSC is 0, suggesting that it is unlikely that the polyploid contains a genome from the diploid. However, if all are encountered in a polyploidy, RSC is 1, suggesting that the one or more of the genomes of the polyploid likely originated from the diploid.

RESULTS

Restriction profiles of nuclear repetitive DNA sequences

DNA of 35 *Gossypium* species (Table 1) was isolated and hybridized with 22 repeated sequence families randomly selected from 163 repeated sequences families isolated from *G. hirsutum* (Zhang et al. 2002). The results of Southern blot hybridization are summarized in Tables 2 and 3.

Of the repetitive DNA sequences analyzed, 15 were detected in all genome groups of the genus, five were A genome-specific and two were detected in a group of the genomes but the D genome. For example, repeated sequence GH1C11 was detected in all *Gossypium* genomes (Fig. 1), GH1A14 was detected in A- E-, and F- genome species, but not in the D-, and C-genome species (Fig. 2), while GH1B19, GH1A11, GH1I19, and GH1J19 were detected only in A-genome species (Fig. 3). Therefore, we classified the repetitive DNA sequences into three groups: (1) Repeated sequences present in all genomes (Fig. 1), (2) repeated sequences absent in the D-genome species, but present in some of the other diploid species (Fig. 2), and (3) repeated sequences only present in A-genome diploid species, i.e., the A genome-specific sequences (Fig. 3).

The number of bands of each repetitive DNA sequence was observed to be various, ranging from 3 to 20 bands per probe/restriction enzyme combination (Table 2; Fig. 4). A total of 642 informative restriction fragment band characters were observed for the

Table 2 Summary of the Southern blot hybridization results of repetitive DNA sequences with genomic DNA of different genome groups

Repetitive DNA	Enzyme	Bands	D	AD	A	B	E	F	C	G	K
1. GH1A14	*Bam*HI	3	-	+	+	+	+	+	-	-	-
	*Hin*dIII	11	-	+	+	+	+	+	-	-	-
	*Eco*RI	3	-	+	+	-	-	-	-	-	-
2. GH1B2	*Bam*HI	7	-	+	+	+	+	+	-	+	+
	*Hin*dIII	17	+	+	+	+	+	+	+	+	+
	*Eco*RI	20	-	+	+	+	+	+	+	-	-
3. GH1B19	*Bam*HI	10	-	+	+	-	-	-	-	-	-
	*Hin*dIII	14	-	+	+	-	-	-	-	-	-
	*Eco*RI	13	-	+	+	-	-	-	-	-	-
4. GH1A11	*Bam*HI	11	-	+	+	-	-	-	-	-	-
	*Hin*dIII	11	-	+	+	-	-	-	-	-	-
	*Eco*RI	10	-	+	+	-	-	-	-	-	-
5. GH1B4	*Bam*HI	3	-	+	+	-	+	+	-	-	-
	*Hin*dIII	9	+	+	+	+	-	+	+	-	-
	*Eco*RI	6	+	+	+	-	-	+	-	-	-
6. GH1E21	*Bam*HI	7	-	+	+	+	+	+	-	-	-
	*Hin*dIII I	10	+	+	+	+	+	+	-	-	-
	*Eco*RI I	5	+	+	+	-	+	+	-	-	-
7. GH1C10	*Bam*HI	14	+	+	+	+	+	+	+	+	+
	*Hin*dIII	19	+	+	+	+	+	+	+	+	+
	*Eco*RI I	16	-	+	+	+	+	+	-	-	-
8. GH1D3	*Bam*HI	13	-	+	+	+	+	+	+	+	+
	*Hin*dIII	13	+	+	+	+	+	+	+	+	+
	*Eco*RI	11	-	+	+	+	+	+	+	+	+
9. GH1E13	*Bam*HI	3	+	+	+	+	+	+	-	-	-
	*Hin*dIII	13	+	+	+	+	+	+	+	+	+
	*Eco*RI	5	+	+	+	+	+	+	+	+	+
10. GH1E9	*Bam*HI I	8	+	+	+	+	+	-	-	-	-
	*Hin*dIII	11	+	+	+	+	+	+	-	+	+
	*Eco*RI I	15	-	+	+	+	+	+	+	+	+
11. GH1E22	*Bam*HI	6	+	+	+	+	+	+	-	+	-
	*Hin*dIII	4	+	+	+	+	+	+	+	+	+
	*Eco*RI I	7	+	+	+	+	-	+	-	-	-

Table 2 (continued)

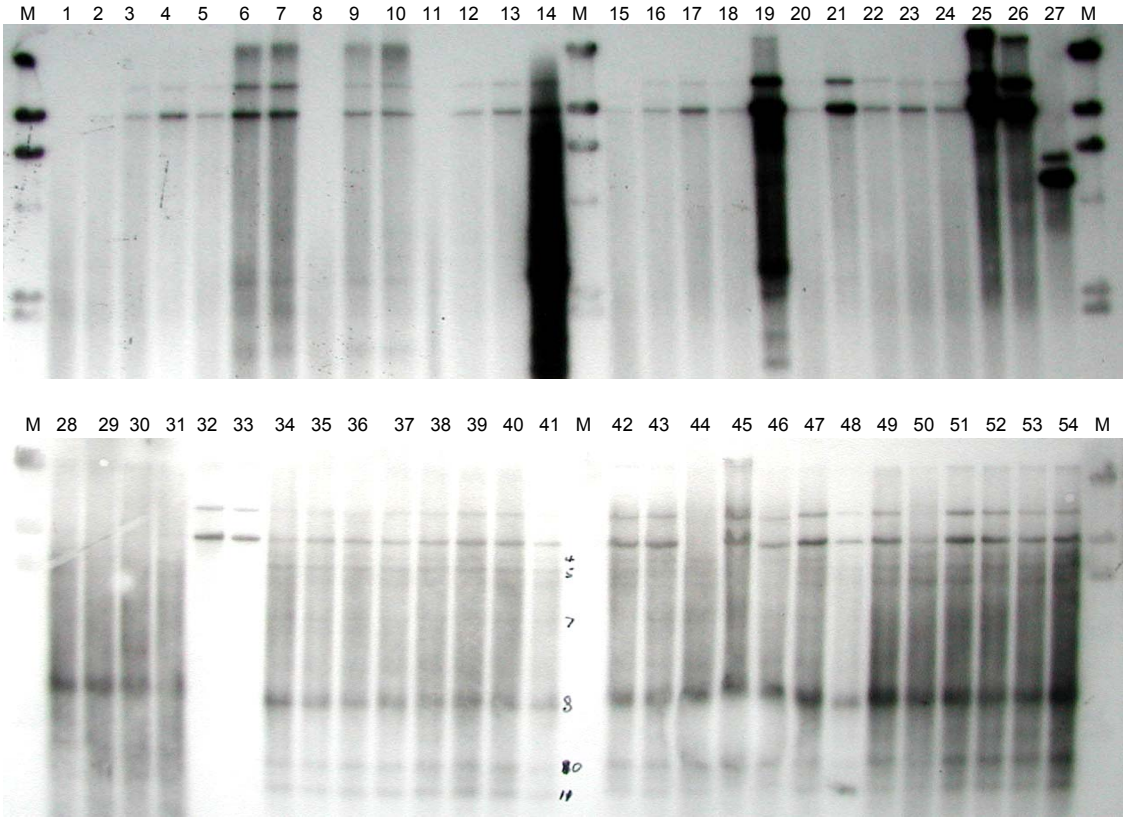
Repetitive DNA	Enzyme	Bands	D	AD	A	B	E	F	C	G	K
12. GH1E19	*Bam*HI	8	-	+	+	+	+	+	+	+	+
	*Hin*dIII	15	-	+	+	+	+	+	+	-	-
	*Eco*RI	15	-	+	+	+	+	+	+	-	-
13. GH1C11	*Bam*HI	10	+	+	+	+	+	+	+	+	+
	*Hin*dIII	15	+	+	+	+	+	+	+	+	+
	*Eco*RI I	19	+	+	+	+	+	+	+	+	+
14. GH1F8	*Bam*HI	10	+	+	+	+	+	+	+	+	+
	*Hin*dIII I	16	+	+	+	+	+	+	+	+	+
	*Eco*RI	16	+	+	+	+	+	+	+	+	+
15. GH1G12	*Bam*HI	7	+	+	+	+	+	+	+	+	+
	*Hin*dIII	10	+	+	+	+	+	+	+	+	+
	EcoRI	13	+	+	+	+	+	+	+	+	+
16. GH1F3	*Bam*HI	9	-	+	+	-	+	+	-	-	-
	*Hin*dIII	14	+	+	+	+	+	+	+	+	+
	*Eco*RI	12	-	+	+	-	+	+	-	-	-
17. GH1G14	*Bam*HI	9	+	+	+	+	+	+	+	+	+
	*Hin*dIII	17	+	+	+	+	+	+	+	+	+
	*Eco*RI I	16	+	+	+	-	+	+	-	-	-
18. GH1E8	*Bam*HI	5	-	+	+	-	-	-	-	-	-
	*Hin*dIII	6	-	+	+	-	-	-	-	-	-
	*Eco*RI	4	-	+	+	-	-	-	-	-	-
19. GH1I19	*Bam*HI	4	-	+	+	-	-	-	-	-	-
	*Hin*dIII	14	-	+	+	-	-	-	-	-	-
	*Eco*RI	10	-	+	+	-	-	-	-	-	-
20. GH1N17	*Bam*HI	5	+	+	+	+	+	+	+	+	+
	HindIII	6	+	+	+	+	+	+	+	+	+
	*Eco*RI	9	+	+	+	+	+	+	+	+	+
21. GH1J19	*Bam*HI	3	-	+	+	-	-	-	-	-	-
	*Hin*dIII	8	-	+	+	-	-	-	-	-	-
	*Eco*RI	12	-	+	+	-	-	-	-	-	-
22. GH1E14	*Bam*HI	3	+	+	+	-	+	+	+	+	-
	*Hin*dIII	8	+	+	+	-	+	+	+	+	-
	*Eco*RI	3	+	+	+	+	+	+	-	+	-
Total		642									

'+' = present bands of the repetitive DNA sequences;
'-' = absent bands of the repetitive DNA sequences.

Table 3. Contribution of the DNA bands detected by repetitive DNA sequences in diploid and polyploid species

Species name	G	TB	GSB	GMB	SSB	SMB	RSC
G. sturtianum	C	124	-	-	0	0	0
G.nandewarense	C1-n	74	-	-	0	0	0
Total	C	124	1	0	-	-	0
G. costulatum	K	71	-	-	0	0	0
G. nobile	K	65	-	-	0	0	0
G. pulchellum	K	66	-	-	0	0	0
G. marchantii	K	69	-	-	0	0	0
Total	K	96	0	0	-	-	0
G. australe	G	82	-	-	0	0	0
G. nelsonii	G	69	-	-	0	0	0
G. bickii	G1	99	-	-	0	0	0
Total	G	114	0	0	-	-	-
G. thurberi	D1	57	-	-	0	0	0
G. trilobum	D8	77	-	-	0	0	0
G. davidsonii	D3d	72	-	-	0	0	0
G. klotzchianum	D3k	74	-	-	0	0	0
G. armourianum	D21	59	-	-	0	0	0
G. harknessii	D22	53	-	-	0	0	0
G. turneri	D10	56	-	-	0	0	0
G. aridum	D4	44	-	-	0	0	0
G. lobatum	D7	32	-	-	0	0	0
G. laxum	D9	39	-	-	0	0	0
G. schwendimanii	D11	38	-	-	0	0	0
G. gossypioides	D6	49	-	-	0	0	0
G. raimondii	D5	89	-	-	7	3	0.43
Total	D	126	9	3	-	-	-
G. herbaceum	A1	431	-	-	25	17	0.68
G. arboreum	A2	413	-	-	20	11	0.55
A1+A2 Shared		370	151	136	-	-	0.90
Total	A	479	-	-	-	-	-
G. anomalum	B1	138	-	-	0	0	0
G. capitis-viridis	B3	145	-	-	0	0	0
Total	B	151	0	0	-	-	0
G. longicakyx	F1	162	-	-	7	0	0
G. stocksii	E1	89	-	-	0	0	0
G. areysianum	E3	150	-	-	0	0	0
G. incanum	E4	165	-	-	0	0	0
Total	E	181	1	0	-	-	0
G. hirsutum	AD1	314	-	-	9	-	-
G. barbadense	AD2	298	-	-	2	-	-
G. tomentosum	AD3	338	-	-	2	-	-
G. mustelinum	AD4	332	-	-	8	-	-
G. darwinii	AD5	306	-	-	13	-	-
Total	AD	426	-	-	-	-	-

Note: G- Genome; TB-Total Bands; GSB- Genome Specific Bands; GMB- Genome Marker Bands; SSB- Species Specific Bands; SMB- Species Marker Bands; RSC- Repetitive Sequences Correspondence.

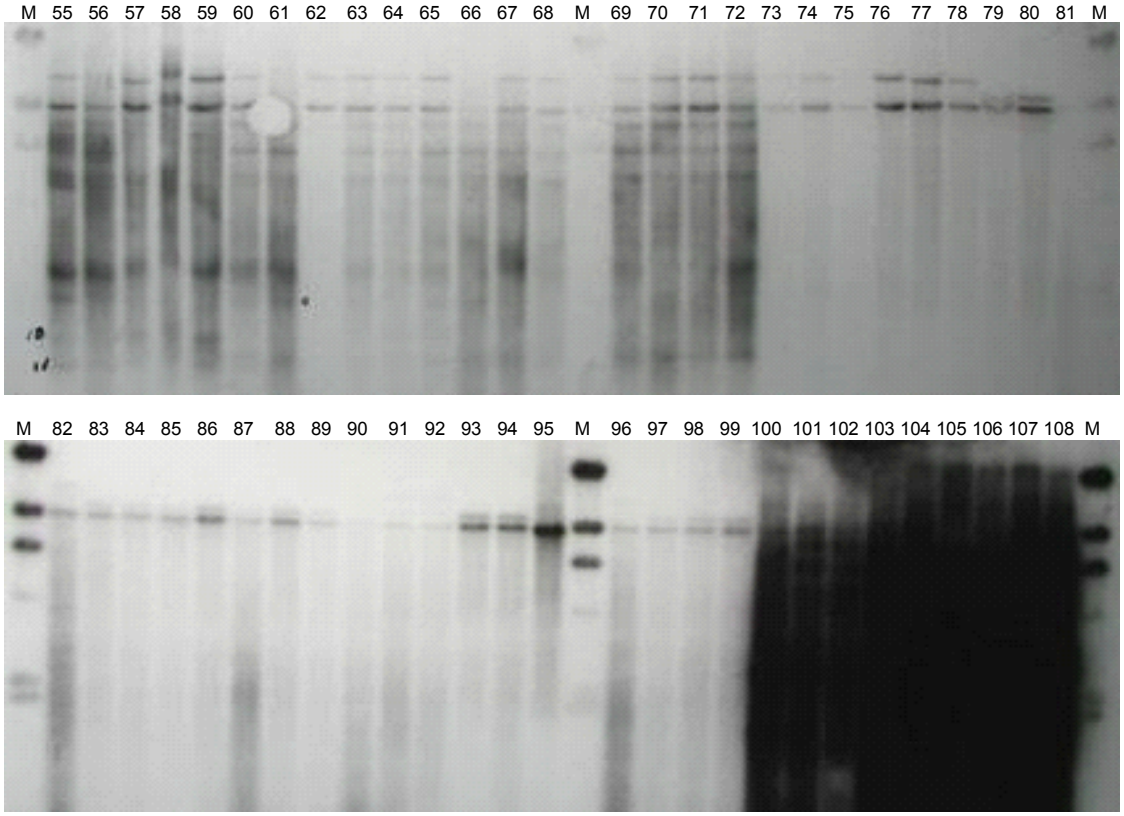


1= *G. thurberi* D1-1
2= *G. thurberi* D1-7
3= *G. trilobum* D8-7
4= *G. trilobum* D8-8
5= *G. trilobum* D8-10
6= *G. davidsonii* D3d-1
7= *G. davidsonii* D3d-2
8= *G. klotzchianum* D3k-57
9= *G. klotzchianum* D3k-58
10= *G. klotzchianum* D3K-59
11= *G. armourianum* D21-6
12= *G. armourianum* D21-7
13= *G. armourianum* D21-9
14= *G. hirsutum* TM-1
15= *G. harknessii* D22-4
16= *G. turneri* D10-1
17= *G. aridum* D4-5
18= *G. lobatum* D7-7
M= size Marker

19= *G. lobatum* 82.07
20= *G. laxum* D9-3
21= *G. laxum* 21.08
22= *G. schwendimanii* D11-1
23= *G. gossypioides* D6-2
24= *G. gossypioides* D6-6
25= *G. raimondii* D5-3
26= *G. raimondii* D5-6
27= California
28= *G. tomentosum* AD3-5
29= *G. tomentosum* AD3-7
30= *G. tomentosum* AD3-11
31= *G. tomentosum* AD3-14
32= *G. raimondii* D5-8
33= *G. raimondii* D5-8
34= *G. hirsutum* WMJJ
35= *G. hirsutum* Clevewilt
36= *G. hirsutum* Auburn56

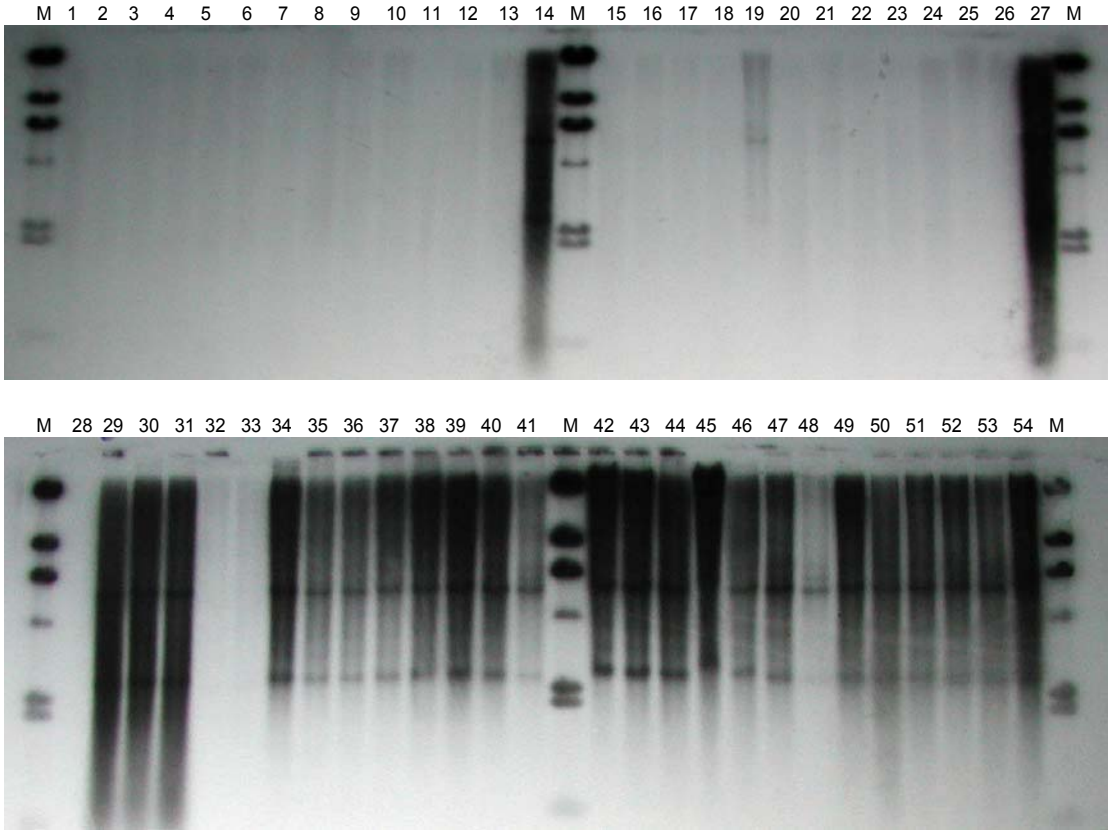
37= *G.hirsutum*Stoneville213
38= *G. hirsutum* Coker201
39= *G. hirsutum* Coker310
40= *G. hirsutum* Deltapine16
41= *G. hirsutum* Deltapine61
42= *G. barbadense* PimaS6
43= *G. barbadense* 3-79
44= *G. barbadense* K101
45= *G. barbadense* AD2-201
46= *G. barbadense* AD2-81
47= *G. barbadense* AD2-372
48= *G. tomentosum* AD3-10
49= *G. tomentosum* AD3-15
50= *G. tomentosum* AD3-16
51= *G. tomentosum* AD3-16
52= *G. tomentosum* AD3-17
53= *G. tomentosum* AD3-25
54= *G. tomentosum* 81.05

Fig. 1 Example of repetitive DNA sequences detected in all *Gossypium* genomes. Genomic DNA was digested with *Hin*dIII and probed with the GH1C11 family.



55= *G.* mustelinum 82.04 (AD4)
56= *G. mustelinum* AD4-9
57= *G. darwinii* AD5-3
58= *G. mustelinum* AD5-3
59= *G. mustelinum* AD5-7
60= *G. herbaceum* A1-108
61= *G. herbaceum* A1-111
62= *G. trilobum* D8-9
63= *G. herbaceum* A1-128
64= *G. herbaceum* A1-129
65= *G. herbaceum* A1-153
66= *G. herbaceum* A1-172
67= *G. herbaceum* A1-180
68= *G. arboretum* A2-67A
69= *G. arboretum* 83.10 (A2)
70= *G. arboretum* A2-142
71= *G. arboretum* A2-47
72= *G. arboretum* A2-84
73= *G. somalense* E2-3
74= *G. areysianum* E3-1
75= *G. incanum* 81.07 (E4)
76= *G. incanum* E4-4
77= *G. longicakyx* F1-1
78= *G. longicakyx* F1-4
79= *G. sturtianum* C1-4
80= *G. nandewarense*C1n-5
81= *G. nandewarense* C1n-6
82= *G. capitis-viridis* B3-1
83= *G. sturtianum* C1-1
84= *G. nobile* NWA-35
85= *G. pulchellum* C8-1
86= *G. marchantii* NWA-6
87= *G. australe* C3-1
88= *G. nelsonii* C9-1
89= *G. nelsonii* C9-2
90= *G. turneri* D10-2
91= *G. stocksii* E1-3
92= *G. stocksii* E1-4
93= *G. bickii* G1-1
94= *G. bickii* G1-3
95= Soybean
96= *G. anomalum* B1-1
97= *G. australe* C3-4
98= *G. costulatum* C5-3
99= *G. costulatum* C5-4
100= *G. tomentosum* AD3-26
101= *G. herbaceum* A1-120
102= *G. herbaceum* A1-127
103= *G. herbaceum* A1-154
104= *G. mustelinum* AD4-17
105= *G. tomentosum* AD3-1
106= *G. tomentosum* AD3-3
107= *G. tomentosum* AD3-4
108= *G. hirsutum* TM-1
M= size marker

Fig. 1 (continued)

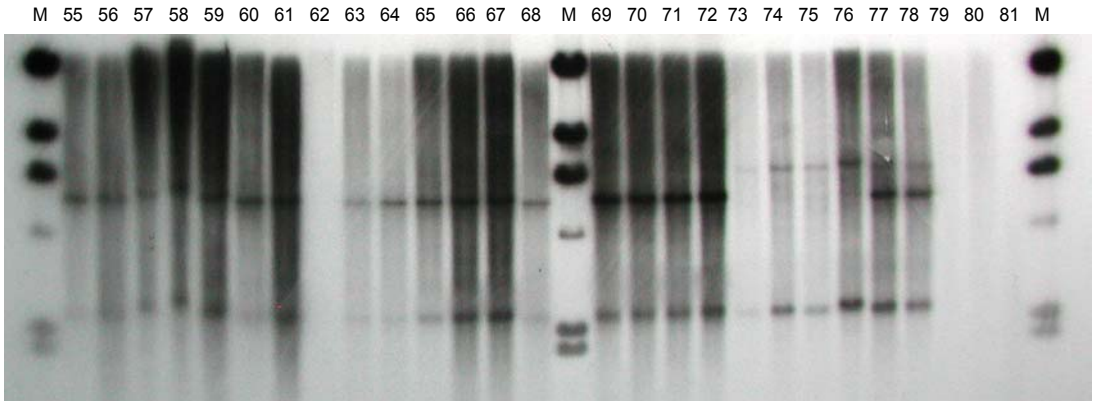


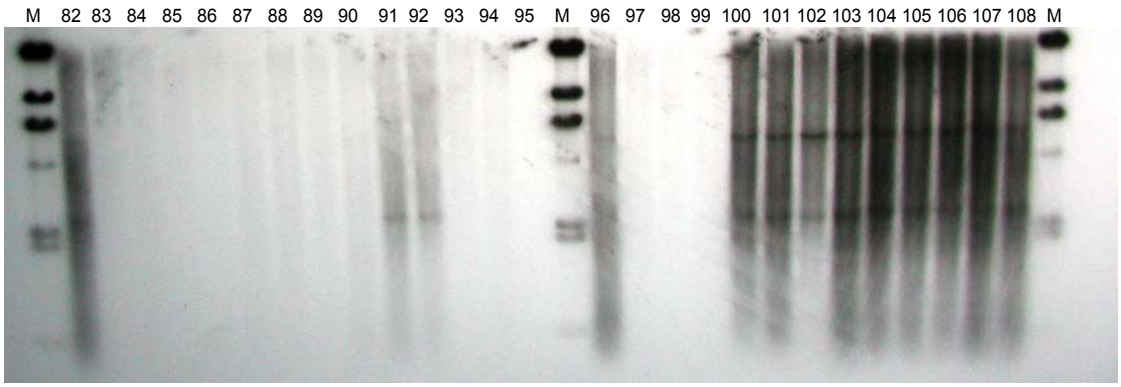
1= *G. thurberi* D1-1
2= *G. thurberi* D1-7
3= *G. trilobum* D8-7
4= *G. trilobum* D8-8
5= *G. trilobum* D8-10
6= *G. davidsonii* D3d-1
7= *G. davidsonii* D3d-2
8= *G. klotzchianum* D3k-57
9= *G. klotzchianum* D3k-58
10= *G. klotzchianum* D3K-59
11= *G. armourianum* D21-6
12= *G. armourianum* D21-7
13= *G. armourianum* D21-9
14= *G. hirsutum* TM-1
15= *G. harknessii* D22-4
16= *G. turneri* D10-1
17= *G. aridum* D4-5
18= *G. lobatum* D7-7
M= size marker

19= *G. lobatum* 82.07
20= *G. laxum* D9-3
21= *G. laxum* 21.08
22= *G. schwendimanii* D11-1
23= *G. gossypioides* D6-2
24= *G. gossypioides* D6-6
25= *G. raimondii* D5-3
26= *G. raimondii* D5-6
27= *G. tomentosum* AD3-5
28= California
29= *G. tomentosum* AD3-7
30= *G. tomentosum* AD3-11
31= *G. tomentosum* AD3-14
32= *G. raimondii* D5-8
33= *G. raimondii* D5-8
34= *G. hirsutum* WMJJ
35= *G. hirsutum* Clevewilt
36= *G. hirsutum* Auburn56

37= *G. hirsutum*Stoneville213
38= *G. hirsutum* Coker201
39= *G. hirsutum* Coker310
40= *G. hirsutum* Deltapine16
41= *G. hirsutum* Deltapine61
42= *G. barbadense* PimaS6
43= *G. barbadense* 3-79
44= *G. barbadense* K101
45= *G. barbadense* AD2-201
46= *G. barbadense* AD2-81
47= *G. barbadense* AD2-372
48= *G. tomentosum* AD3-10
49= *G. tomentosum* AD3-15
50= *G. tomentosum* AD3-16
51= *G. tomentosum* AD3-16
52= *G. tomentosum* AD3-17
53= *G. tomentosum* AD3-25
54= *G. tomentosum* 81.05

Fig. 2 Example of repetitive DNA sequences undetected in the D-genome species but present in the other diploid species. Genomic DNA was digested with *Bam*HI and probed with the GH1A14 family.



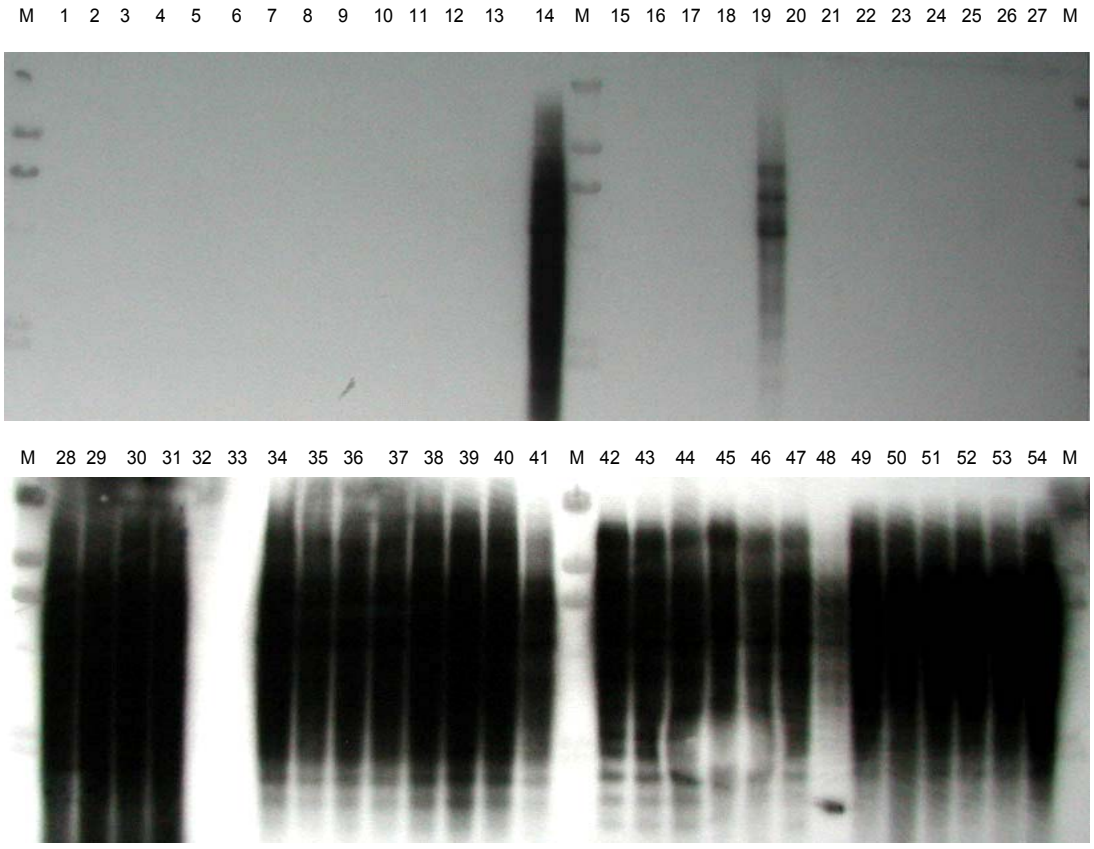


55= *G.* mustelinum 82.04 (AD4)
56= *G. mustelinum* AD4-9
57= *G. darwinii* AD5-3
58= *G. mustelinum* AD5-3
59= *G. mustelinum* AD5-7
60= *G. herbaceum* A1-108
61= *G. herbaceum* A1-111
62= *G. trilobum* D8-9
63= *G. herbaceum* A1-128
64= *G. herbaceum* A1-129
65= *G. herbaceum* A1-153
66= *G. herbaceum* A1-172
67= *G. herbaceum* A1-180
68= *G. arboretum* A2-67A
69= *G. arboretum* 83.10 (A2)
70= *G. arboretum* A2-142
71= *G. arboretum* A2-47
72= *G. arboretum* A2-84
M=size marker

73= *G. somalense* E2-3
74= *G. areysianum* E3-1
75= *G. incanum* 81.07 (E4)
76= *G. incanum* E4-4
77= *G. longicakyx* F1-1
78= *G. longicakyx* F1-4
79= *G. sturtianum* C1-4
80= *G. nandewarense*C1n-5
81= *G. nandewarense* C1n-6
82= *G. capitis-viridis* B3-1
83= *G. sturtianum* C1-1
84= *G. nobile* NWA-35
85= *G. pulchellum* C8-1
86= *G. marchantii* NWA-6
87= *G. australe* C3-1
88= *G. nelsonii* C9-1
89= *G. nelsonii* C9-2
90= *G. turneri* D10-2

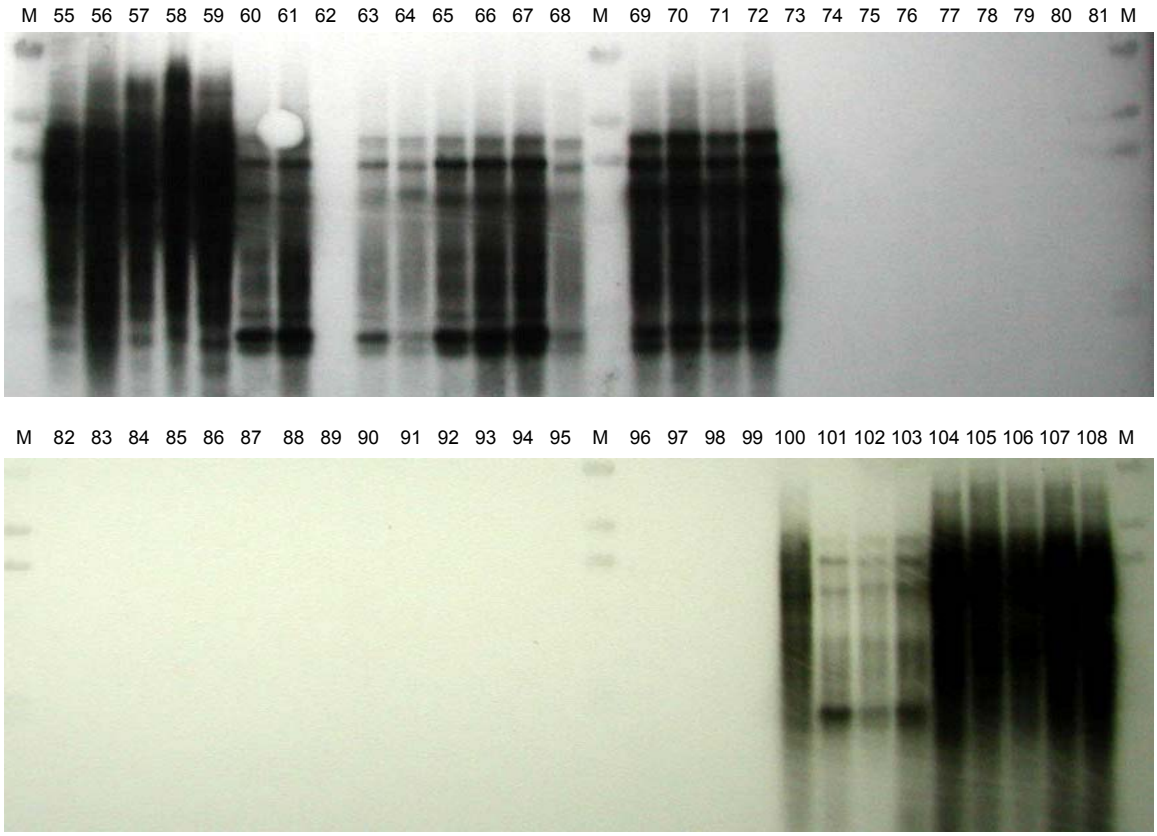
91= *G. stocksii* E1-3
92= *G. stocksii* E1-4
93= *G. bickii* G1-1
94= *G. bickii* G1-3
95= Soybean
96= *G. anomalum* B1-1
97= *G. australe* C3-4
98= *G. costulatum* C5-3
99= *G. costulatum* C5-4
100= *G. tomentosum* AD3-26
101= *G. herbaceum* A1-120
102= *G. herbaceum* A1-127
103= *G. herbaceum* A1-154
104= *G. mustelinum* AD4-17
105= *G. tomentosum* AD3-1
106= *G. tomentosum* AD3-3
107= *G. tomentosum* AD3-4
108= *G. hirsutum* TM-1

Fig. 2 (continued)



1=*G. thurberi* D1-1
2= *G. thurberi* D1-7
3= *G. trilobum* D8-7
4= *G. trilobum* D8-8
5= *G. trilobum* D8-10
6=*G. davidsonii* D3d-1
7= *G. davidsonii* D3d-2
8=*G. klotzchianum* D3k-57
9= *G. klotzchianum* D3k-58
10= *G. klotzchianum* D3K-59
11=*G. armourianum* D21-6
12= *G. armourianum* D21-7
13= *G. armourianum* D21-9
14=*G. hirsutum* TM-1
15=*G. harknessii* D22-4
16=*G. turneri* D10-1
17=*G. aridum* D4-5
18=*G. lobatum* D7-7
19= *G. lobatum* 82.07
20= *G. laxum* D9-3
21= *G. laxum* 21.08
22= *G. schwendimanii* D11-1
23= *G. gossypioides* D6-2
24= *G. gossypioides* D6-6
25= *G. raimondii* D5-3
26= *G. raimondii* D5-6
27= California
28= *G. tomentosum* AD3-5
29= *G. tomentosum* AD3-7
30= *G. tomentosum* AD3-11
31= *G. tomentosum* AD3-14
32= *G. raimondii* D5-8
33= *G. raimondii* D5-8
34= *G. hirsutum* WMJJ
35= *G. hirsutum* Clevewilt
36= *G. hirsutum* Auburn56
37= *G.hirsutum*Stoneville213
38= *G. hirsutum* Coker201
39= *G. hirsutum* Coker310
40= *G. hirsutum* Deltapine16
41= *G. hirsutum* Deltapine61
42= *G. barbadense* PimaS6
43= *G. barbadense* 3-79
44= *G. barbadense* K101
45= *G. barbadense* AD2-201
46= *G. barbadense* AD2-81
47= *G. barbadense* AD2-372
48= *G. tomentosum* AD3-10
49= *G. tomentosum* AD3-15
50= *G. tomentosum* AD3-16
51= *G. tomentosum* AD3-16
52= *G. tomentosum* AD3-17
53= *G. tomentosum* AD3-25
54= *G. tomentosum* 81.05
M=size marker

Fig. 3 Example of repetitive DNA sequences was only detected in A-genome species. Genomic DNA digested with *Hin*dIII and probed with the GH1A11 family.

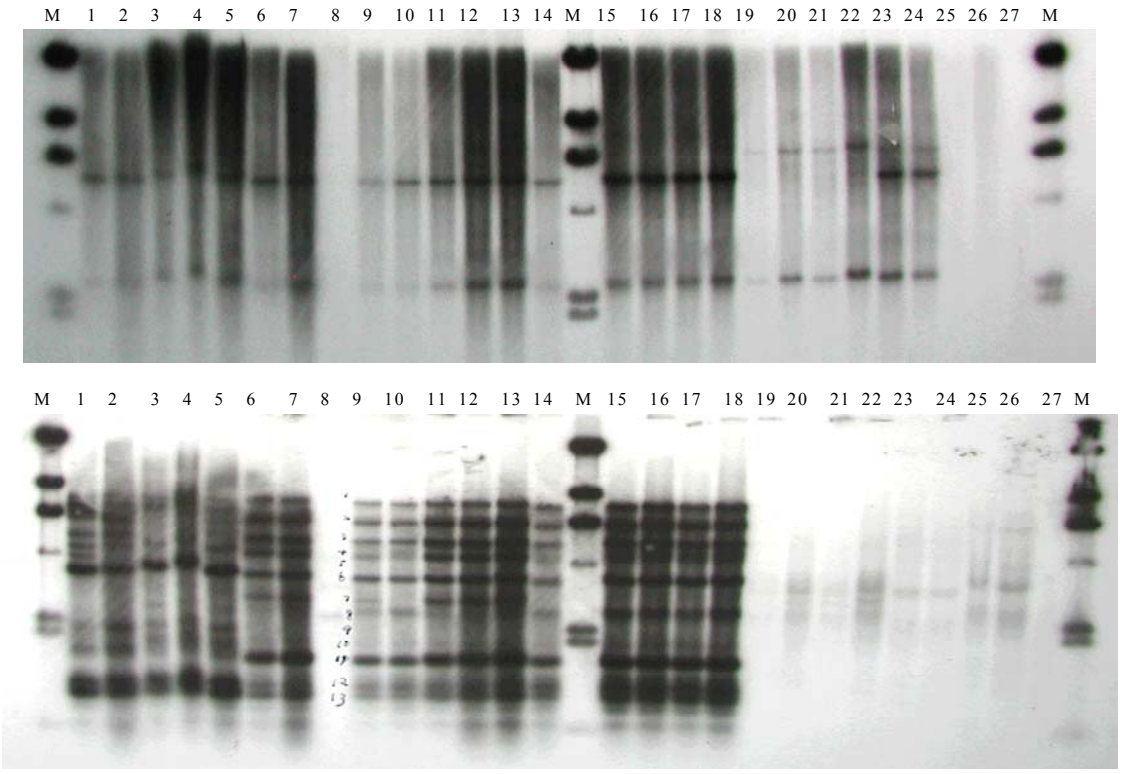


55= *G. mustelinum* 82.04 (AD4)
56= *G. mustelinum* AD4-9
57= *G. darwinii* AD5-3
58= *G. mustelinum* AD5-3
59= *G. mustelinum* AD5-7
60= *G. herbaceum* A1-108
61= *G. herbaceum* A1-111
62= *G. trilobum* D8-9
63= *G. herbaceum* A1-128
64= *G. herbaceum* A1-129
65= *G. herbaceum* A1-153
66= *G. herbaceum* A1-172
67= *G. herbaceum* A1-180
68= *G. arboretum* A2-67A
69= *G. arboretum* 83.10 (A2)
70= *G. arboretum* A2-142
71= *G. arboretum* A2-47
72= *G. arboretum* A2-84
M=size marker

73= *G. somalense* E2-3
74= *G. areysianum* E3-1
75= *G. incanum* 81.07 (E4)
76= *G. incanum* E4-4
77= *G. longicakyx* F1-1
78= *G. longicakyx* F1-4
79= *G. sturtianum* C1-4
80= *G. nandewarense*C1n-5
81= *G. nandewarense* C1n-6
82= *G. capitis-viridis* B3-1
83= *G. sturtianum* C1-1
84= *G. nobile* NWA-35
85= *G. pulchellum* C8-1
86= *G. marchantii* NWA-6
87= *G. australe* C3-1
88= *G. nelsonii* C9-1
89= *G. nelsonii* C9-2
90= *G. turneri* D10-2

91= *G. stocksii* E1-3
92= *G. stocksii* E1-4
93= *G. bickii* G1-1
94= *G. bickii* G1-3
95= Soybean
96= *G. anomalum* B1-1
97= *G. australe* C3-4
98= *G. costulatum* C5-3
99= *G. costulatum* C5-4
100= *G. tomentosum* AD3-26
101= *G. herbaceum* A1-120
102= *G. herbaceum* A1-127
103= *G. herbaceum* A1-154
104= *G. mustelinum* AD4-17
105= *G. tomentosum* AD3-1
106= *G. tomentosum* AD3-3
107= *G. tomentosum* AD3-4
108= *G. hirsutum* TM-1

Fig. 3 (continued)



1=*G. mustelinum* 82.04 (AD4)
2=*G. mustelinum* AD4-9
3=*G. darwinii* AD5-3
4=*G. mustelinum* AD5-3
5=*G. mustelinum* AD5-7
6=*G. herbaceum* A1-108
7=*G. herbaceum* A1-111
8=*G. trilobum* D8-9
9=*G. herbaceum* A1-128
10=*G. herbaceum* A1-129
11=*G. herbaceum* A1-153
12=*G. herbaceum* A1-172
13=*G. herbaceum* A1-180
14=*G. arboretum* A2-67A
15=*G. arboretum* 83.10 (A2)
16=*G. arboretum* A2-142
17=*G. arboretum* A2-47
18=*G. arboretum* A2-84
19=G. *somalense* E2-3
20=*G. areysianum* E3-1
21=*G. incanum* 81.07 (E4)
22=*G. incanum* E4-4
23=*G. longicakyx* F1-1
24=*G. longicakyx* F1-4
25=*G. sturtianum* C1-4
26=*G. nandewarense*C1n-5
27=*G. nandewarense* C1n-6
M=Size markers

Fig. 4 Example of repetitive DNA sequence restriction profile of *Gossypium* species. Top panel: Genomic DNA digested with *Bam*HI and probed with the GH1A14 family; Lower panel: Genomic DNA digested with *Hin*dIII and probed with GH1B2 family.

DNA of 35 species (Table 2). Figure 5 and Figure 6 show the distributions of repetitive DNA bands in different genomes and species, respectively. A total of 426 bands were observed in the allopolyploid species. Within the diploid species, the A-genome species produced 479 band characters, the largest number of band characters, while the K-genome species produced 96 bands, the fewest number of band characters. The D-genome species produced 126 band characters and the remaining genome species produced character bands between 119 and 181 (Table 3).

Inference of genome origin of polyploid species

In the restriction profiles of the diploid species, SSBs were observed only for the A-genome, D-genome, and F-genome species. *Gossypium raimondii* (D5*)* had seven SSBs, *G. longicakyx* (F) had seven SSBs, and *G. herbaceum* (A1) and *G. arboretum* (A2) had 25 and 20 SSBs, respectively. However, only some of the SSBs of the A-genome and D-genome species were encountered in one or more polyploid species (Table 4). Of the 7 SSBs of *G. raimondii* (D5*)*, 1 – 3 were encountered in each polyploid species, giving an RSC of 0.14 – 0.30; of the 25 SSBs of *G. herbaceum* (A1), 9 were encountered in each polyploid species, giving an RSC of 0.36; and of the 20 SSBs of *G. arboretum* (A2), 2 – 6 were encountered in each polyploid species, giving an RSC of 0.10 – 0.30. Concerning the common bands between *G. herbaceum* (A1) and *G. arboretum* (A2), a total of 151 GSBs were observed, of which 82 – 94 were encountered in each polyploid species, giving an RSC of 0.51 – 0.62.

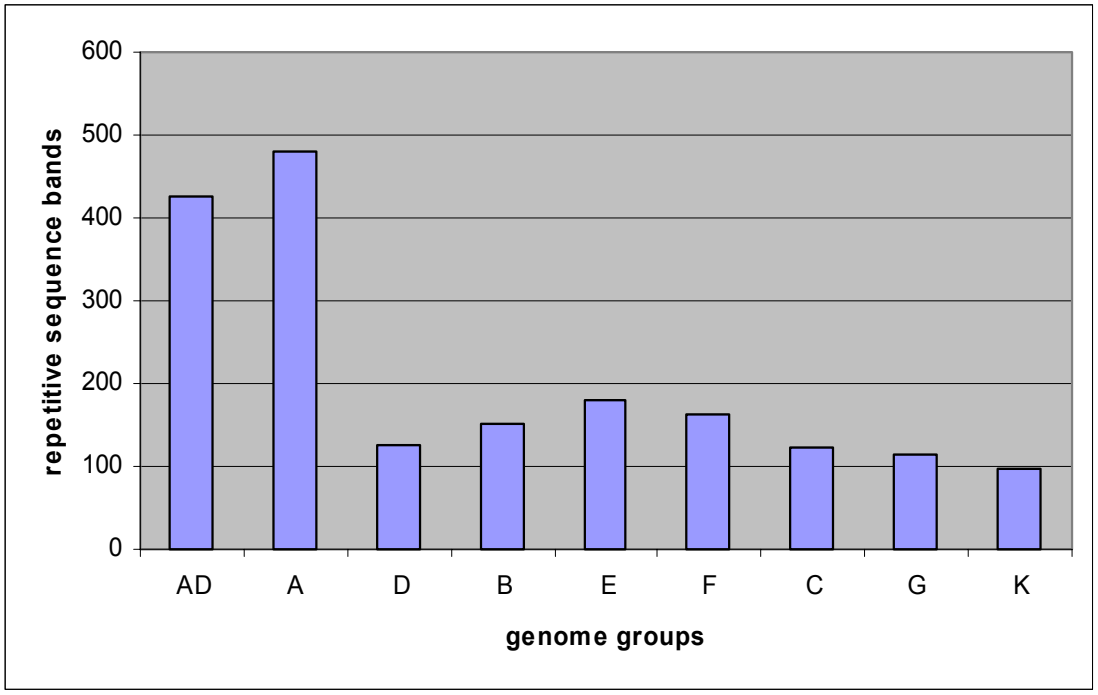


Fig. 5 Distribution of restricted repetitive DNA bands in different genomes of the genus *Gossypium*.

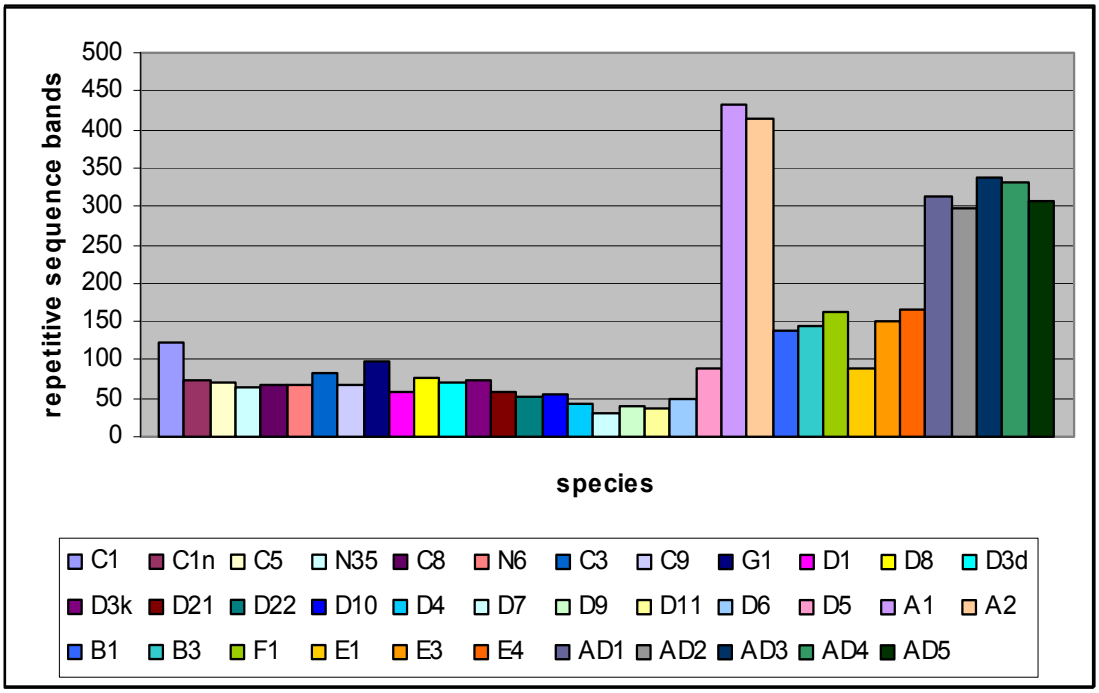


Fig. 6 Distribution of restricted repetitive DNA bands in different species of the genus *Gossypium*.

Table 4. Repetitive sequence correspondences (RSCs) between several diploid species and the polyploid species.

Species	TB	SSB /GSB	AD1 SMB	AD1 RSC	AD2 SMB	AD2 RSC	AD3 SMB	AD3 RSC	AD4 SMB	AD4 RSC	AD5 SMB	AD5 RSC
A1	431	25	9	0.36	9	0.36	9	0.36	9	0.36	9	0.36
A2	413	20	2	0.10	5	0.25	2	0.10	5	0.25	6	0.30
A1+A2	370	151	93	0.62	91	0.60	94	0.62	87	0.58	82	0.51
D5	89	7	1	0.14	3	0.43	2	0.29	2	0.29	2	0.29

Note: TB - Total Bands; GSB - Genome Specific Bands; GMB - Genome Marker Bands; SSB - Species Specific Bands; SMB - Species Marker Bands; RSC - Repetitive Sequences Correspondence; A1 - *G. herbaceum*; A2 - *G. arboretum*; and D5 - *G. raimondii.*

Reconstruction of phylogenetic tree

To reconstruct the phylogenetic tree of the species, a data matrix was constructed from the 642 informative band characters (Table 2) and analyzed by the parsimony method using the PAUP program and soybean as the outgroup. A rooted phylogenetic tree of the 35 species, including both polyploid and diploid species, was generated (Fig. 7), with a confidence of each branch ranging from 61 to 100 out of the 100 bootstrap replicates applied (Fig. 8). Furthermore, considering that five A genome-specific repetitive sequences might cause bias to the reconstruction of the phylogenetic tree, we excluded the data derived from the A genome-specific probes and reconstructed the phylogenetic tree. As a result, the same tree was obtained, suggesting the A genome-specific repeated sequence probes did not significantly affect the phylogenetic analysis of the *Gossypium* species.

The tree was consistent with the genome designations and geographical distribution of the species. A basal dichotomy divided the genus *Gossypium* into two major clades, one being composed of the New World D-genome diploid species and the other consisting of the remaining diploid species and all allopolyploid species. The former clade was grouped with a bootstrap value of 63% and the latter clade was grouped with a bootstrap value of 89%. In the latter clade, Australian C-, G-, and K-genome species were grouped into one subclade with 68% confidence and the five New World allopolyploid genome species with all African-Arabia diploid species including

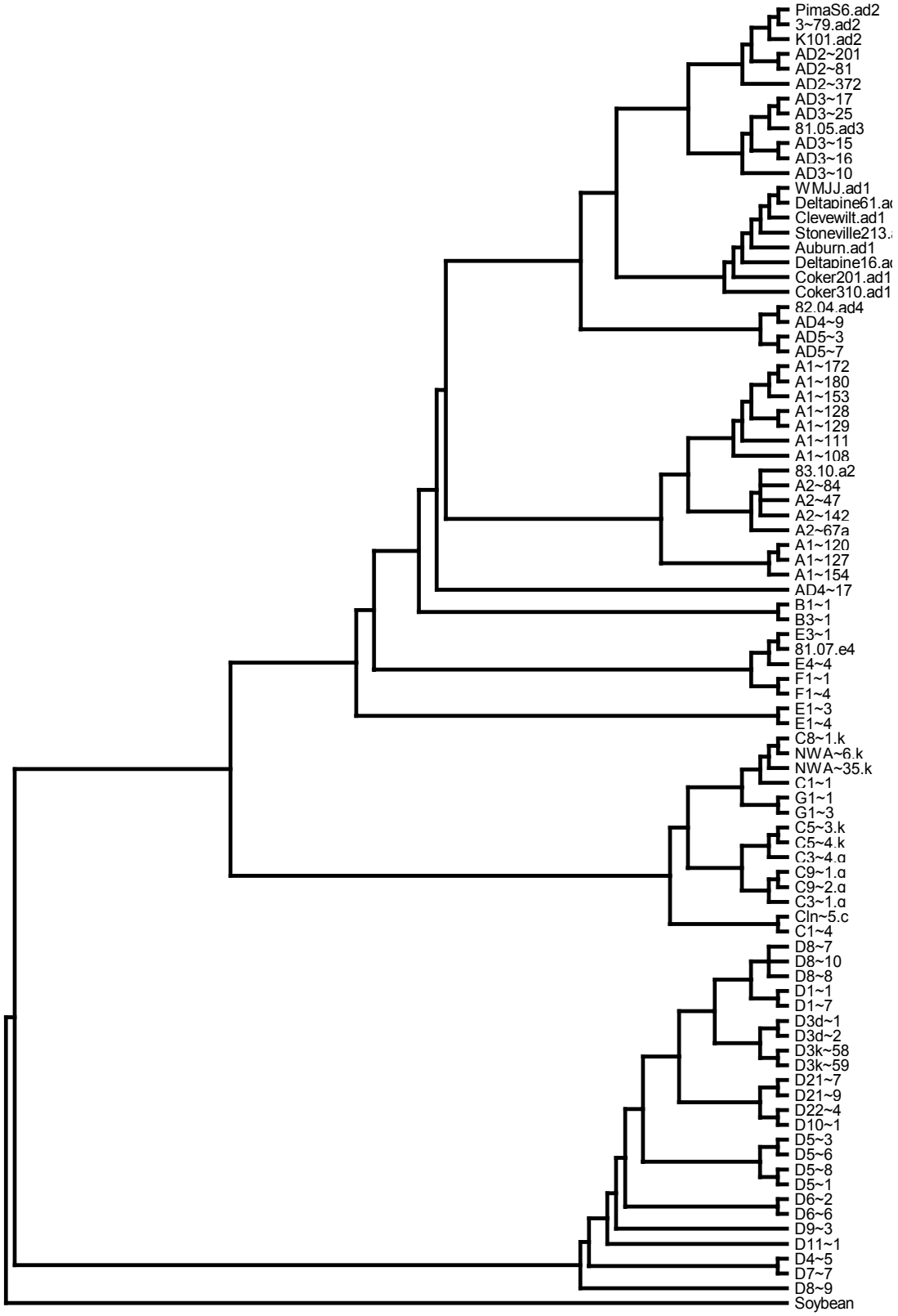


Fig. 7 Rooted phylogenetic tree resulting from analyses of 35 species of the genus *Gossypium*. Soybean was used as the outgroup.

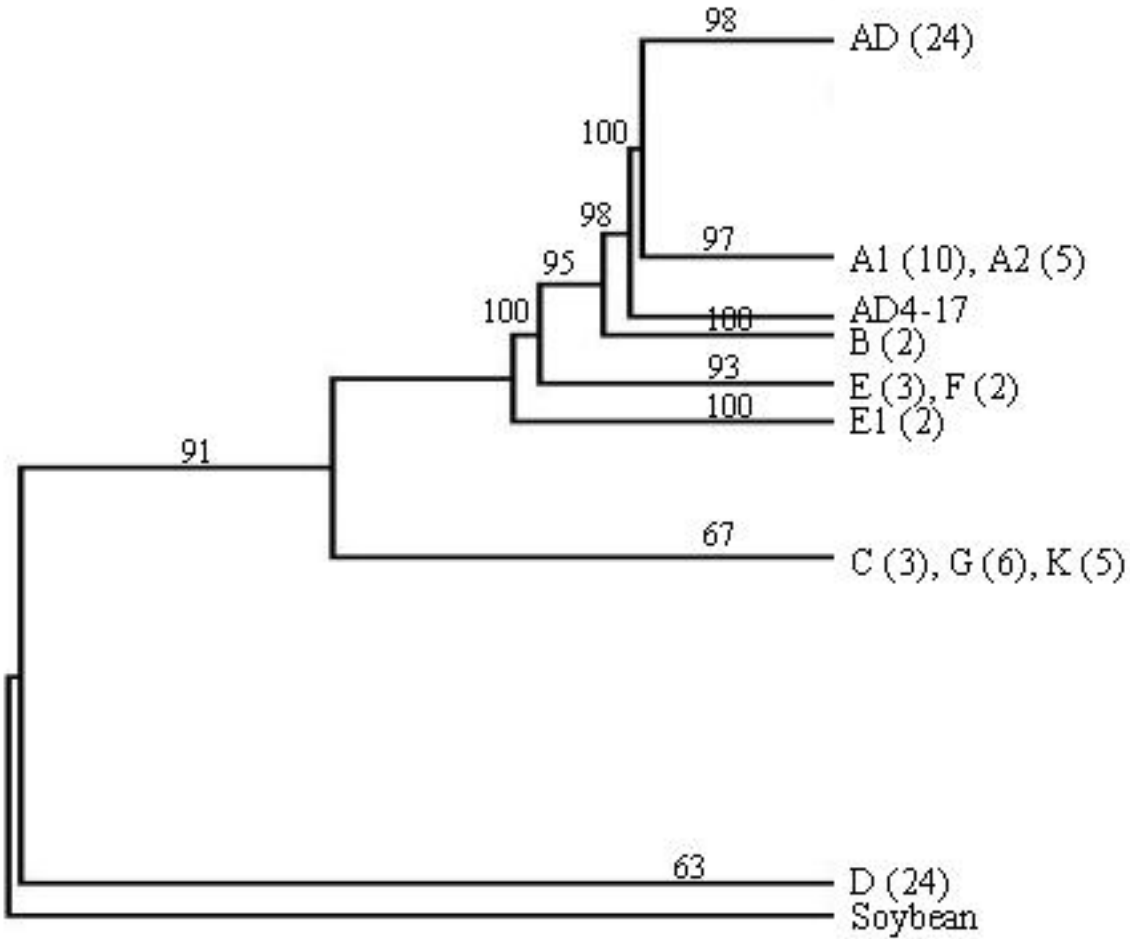


Fig. 8 Simplified phylogenetic tree of 35 species of *Gossypium*. The number above each branch is the bootstrap value in percentage. Branches without numbers had bootstrap values of less than 50. Cytogenetic groups are indicated at the right. The number inside each parenthesis is the number of the accession or species.

the A-, B-, E- and F-genome species together form the other subclade with 100% confidence. Within the subclade of polyploid species and A-, B-, E-, and F-genome species, the branch composed of the E- and F-genome species was a sister branch to the polyploid and A- and B-genome species. The A-genome species was further grouped with the allopolyploid species comprising all five allopolyploid species (Fig. 7).

Since the diploid species each contain a single genome and the tetraploid species each contain two genomes (A- and D-subgenomes), the ploidy level of which might affect the phylogenetic analysis result, we further analyzed the data and reconstructed the phylogenetic trees of diploid species and polyploid species, separately. The phylogenetic tree of the diploid species is shown in Fig. 9 and that of the polyploid species is shown in Fig. 10.

The phylogenetic tree of the diploid species was largely the same as that of combined diploid and polyploid species, but differences were observed between the trees, suggesting that the ploidy levels of the species indeed influenced the phylogenetic analysis result. In the clade consisting of 13 New World D-genome species, the branching of the species was exactly the same as that of the phylogenetic tree of the combined diploid and polyploid species. All accessions of a species, including 2 accessions of *G. thurberi* (D1), 2 accessions of *G. davidsonii* (D3d), 2 accessions of *G. klotzschianum* (D3k), 2 accessions of *G. armourianum* (D21), 2 accessions of *G. raimondii* (D5), and 2 accessions of *G. schwendimanii* (D6), were grouped into the same species branch, respectively. But, an exception was observed for *G. trilobum* (D8), one accession of the species, D8-9, occupying the basal position of the D-genome clade,

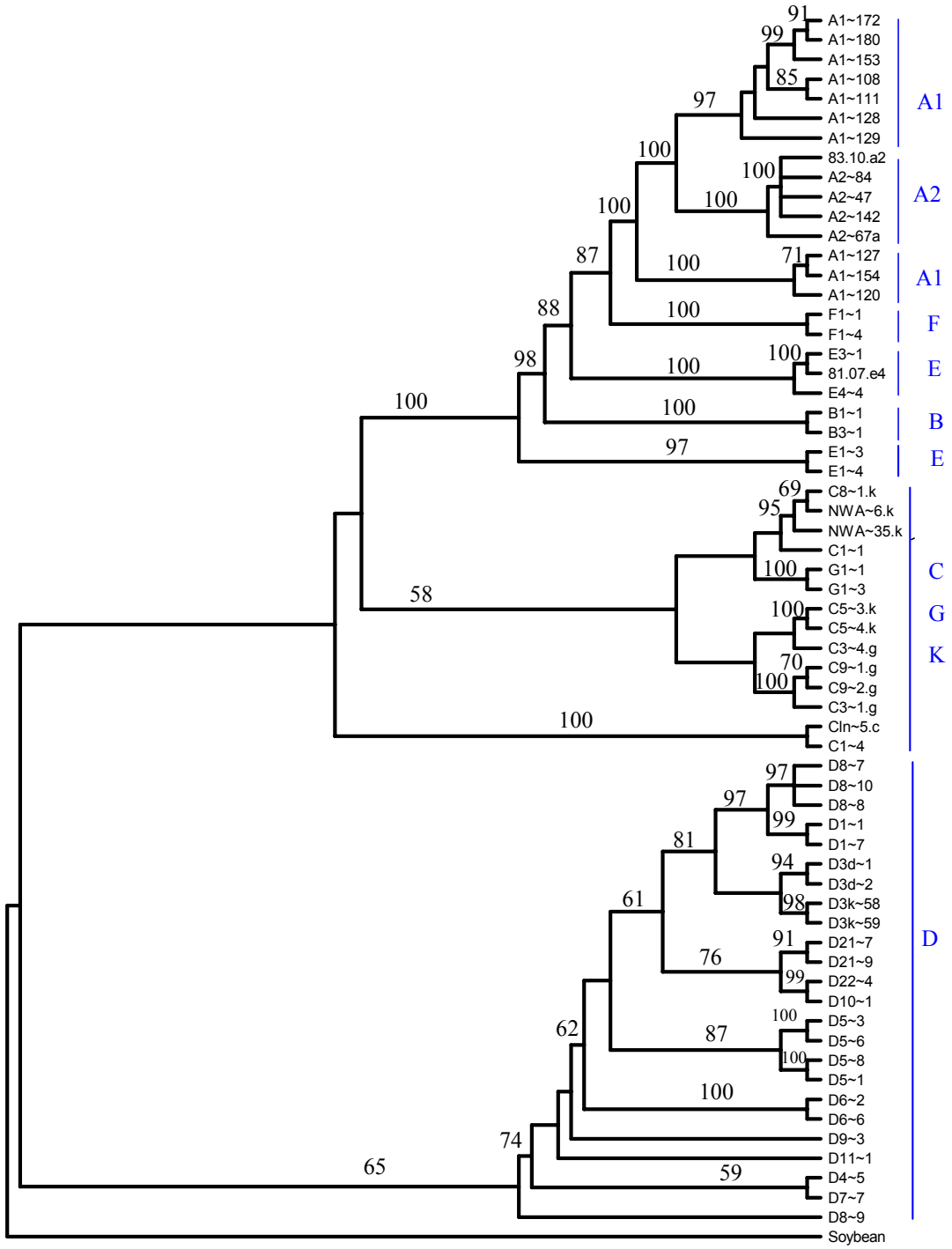


Fig 9 Rooted phylogenetic tree resulting from analyses of 30 diploid species of the genus *Gossypium*. Soybean was used as the outgroup. The number above each branch is the bootstrap value in percentage. Branches without numbers had bootstrap values of less than 50.

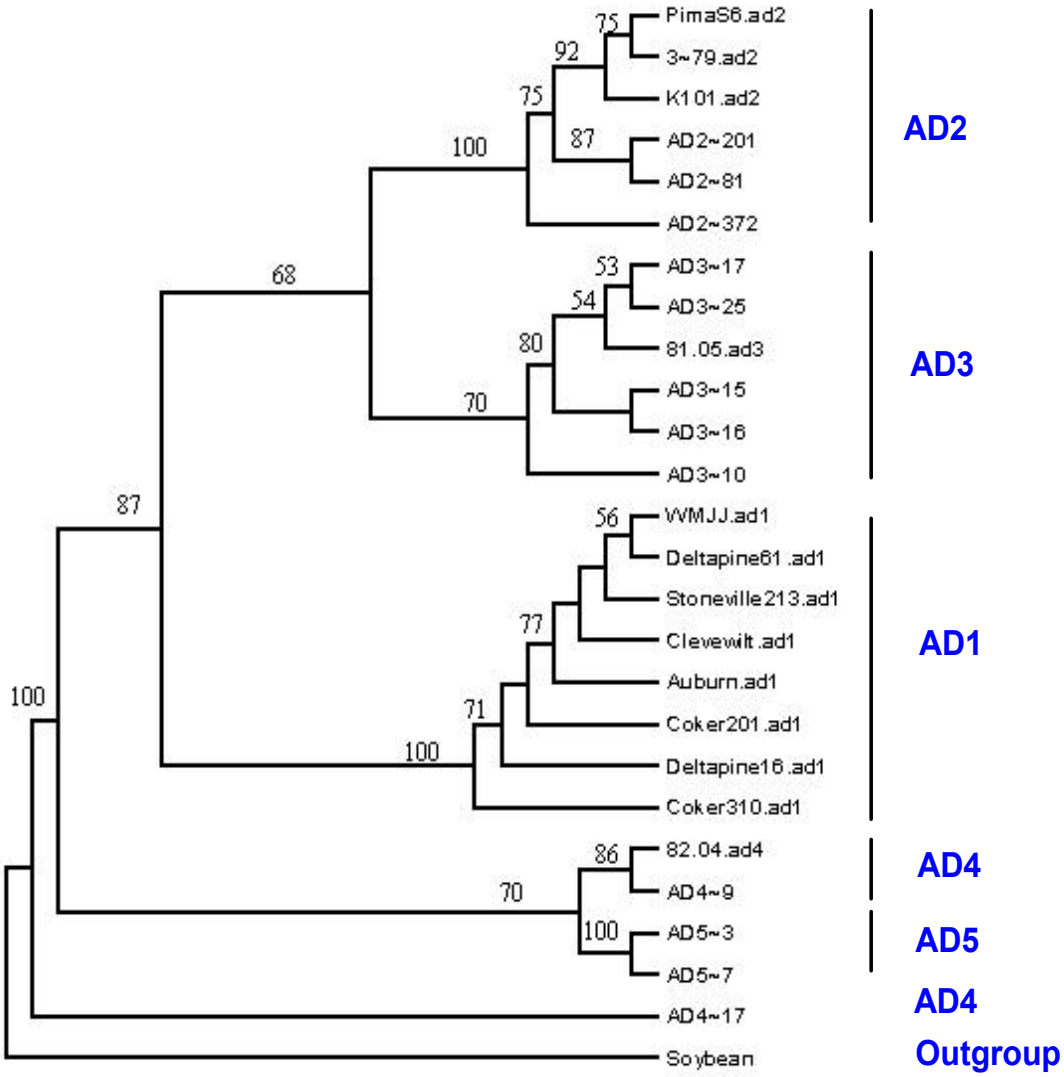


Fig. 10 Rooted phylogenetic tree resulting from analyses of 5 polyploid species of the genus *Gossypium*. Soybean was used as the outgroup. The number above each branch is the bootstrap value in percentage. Branches without numbers had bootstrap values of less than 50.

while the remaining three accessions, D8-7, D8-10, and D8-8, being placed at the tip of the D-genome clade, forming a sister branch to the *G. thurberi* (D1) branch. Within the D-genome species clade, several other sister branches were also found. They were *G. trilobum* (D8) with *G. thurberi* (D1), *G. aredum* (D4) with *G. lobatum* (D7), and *G harknessii* (D22) and *G. turneri* (D10) with *G. armourianum* (D21). The branch order of the D-genome species in the D-genome clade, from the basal node to the tip, was *G. trilobum* (D8), *G. aredum* (D4)/*G. lobatum* (D7), *G. schwendimanii* (D11), *G. laxum* (D9), *G. schwendimanii* (D6), *G. raimondii* (D5), *G. turneri* (D10)/*G harknessii* (D22)/ *G. armourianum* (D21), *G. davidsonii* (D3d)/*G. klotschianum* (D3k), and *G. trilobum* (D8)/*G. thurberi* (D1) (Fig. 9).

The clade consisting of all other genomes were grouped into two subclades, the C-, G-, and K-genome species and the A-, E-, F-, and B-genome species. Branching of the species differed from that of the combined diploid and polyploid tree. In the latter subclade, all five accessions (82.04, A2-84, A2-47, A2-142, and A2-67a) of *G. arboretum* (A2) were grouped into a single branch that was sister to the branch of 7 accessions (A1-172, A1-180, A1-153, A1-128, A1-129, A1-111, and A1-108) of *G. herbaceum* (A1). The remaining three accessions (A1-120, A1-127 and A1-154) of *G. herbaceum* (A1) formed a branch sister to the branch composed of two A-genome species (*G. herbaceum* and *G. arboretum*). The A-genome group was in the tip position of the tree clade. The sister branch to the A-genome branch was the F-genome species, followed by the E- and B-genome species branches toward the base of the clade. Within the Australian species (C-, G-, and K-genome) subclade, the phylogenetic relationships

were complicated. For example, one branch containing some accessions of the C-, G-, and K-genome species was a sister to another branch composed of the others of the K- and G-genome species (Fig. 9).

The phylogenetic tree of the five polyploid species consisted of three clades: AD4/AD5 species, AD2/AD3 species and AD1 species (Fig. 10). The tree showed that (1) almost all accessions of each polyploid species fell into one branch; (2) *G. barbadense* (AD2) and *G. tomentosum* (AD3) were the most closely related and sister to the branch of AD1; (3) *G. mustelinum* (AD4) and *G. darwinii* (AD5) formed a branch occupying the basal position of the tree (Fig. 10).

DISCUSSION AND CONCLUSIONS

Repetitive DNA sequences are abundant in the *Gossypium* genomes. Variation of repeated sequences in copy number and restriction pattern was limited within a species or a genome group of the genus, but a significant level of variation was observed among different species of the genus. These results suggest that variation of repeated sequences is suited for reconstruction of the phylogeny and deciphering of the genome origin of polyploid species of the *Gossypium* genus.

The 22 repetitive sequence probes used in this study all were from *G. hirsutum* containing A and D subgenomes. Nevertheless, five of them were found to be A genome-specific, but none was found to be D genome-specific. It was also observed that the A-genome species gave about four-fold as many bands as the D-genome species. These results indicate that the A genome seems evolving much faster than the D genome in number of repeated sequence families and at the nucleotide sequence level. In comparison, although the A genome (3.8 pg/2C) is about two-fold as large as the D genome (2.0 pg/2C), it is close to or about two-fold smaller than the other genomes, such as the K genome (7.0 pg/2C), that gave many fewer bands than the A-genome species. Therefore, the genome divergence seems to provide more appropriate explanation on why more bands were detected in the A-genome species than the other-genome species. This is consistent with previous studies. Evidence from cytogenetic and segregation data concluded that the A subgenome of allopolyploid cottons is more similar to that of the

A-genome diploid species than the D subgenome of the allopolyploid is to that of the D-genome diploid species. Data derived from amplified fragment length polymorphisms (AFLPs) showed that the number of bands shared or common between the A-genome diploids and the polyploids are much more those between the D-genome diploids and the ploypoids (Khan et al. 2000). Zhao et al. (1998) reported that 77% of the non-cross-hybridizing repetitive DNA clones isolated from *G. barbadense* (AD2) are largely restricted to the A-genome diploid species. In contrast, only 5% of them are D-genome specific or enriched.

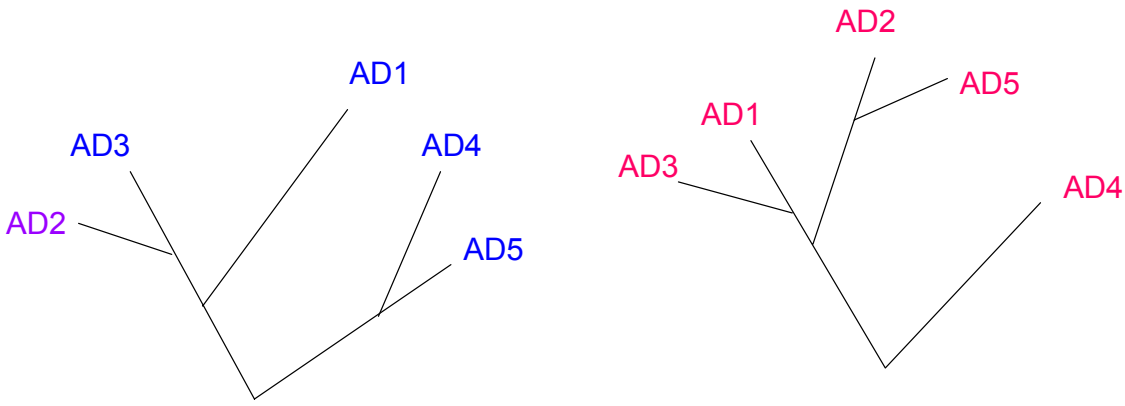
The phylogenetic tree constructed in the present study, based on variation in nuclear repetitive sequence, is largely congruent with the phylogenetic tree constructed previously (Wendel and Cronn, 2003), mainly based on cpDNA restriction site variation as well as sequence variation of nuclear ribosomal DNA, chloroplast genes, and low-copy nuclear genes, in genome designation, geographical distribution, and phylogenetic inference. Both the tree constructed in this study and that of Wendel and Cronn (2003) grouped the eight diploid genome groups into three major lineages corresponding to three continents, Australia, African-Arabia and Americas, where the carrying species naturally occur. The earliest divergence in the genus separated the New World D-genome lineage from the ancestor of all Old World taxa, making the New World and Old World diploid species into phylogenetic sister groups. Within the Old World taxa, the Australian C-, G- and K-genome species constitute a subclade sister to the subclade containing the African-Arabian A-, B-, E-, and F-genome species.

However, a few significant disagreements exist between the phylogenetic trees constructed in this study and by Wendel and Cronn (2003). The first major difference is the phylogenetic relationships among the polyploid species (Fig. 11). Wendel and Cronn (2003) classified the five polyploid species into three branches, one consisting of *G. mustelinum* (AD4), one consisting of *G. tomentosu* (AD3) and *G. hirsutum* (AD1), and the third one containing *G. barbadense* (AD2) and *G. darwinii* (AD5). However, this study shows that *G. barbadense* (AD2) with *G. tomentosu* (AD3) forms one branch, *G. mustelinum* (AD4) with *G. darwinii* (AD5) forms the second branch, and *G. hirsutum* (AD1) alone forms the third branch. The flavoid data suggested that *G. tomentosu* (AD3) is the most similar to *G. barbadense* (AD2) (Parks et al. 1975). Moreover, a high interspecific genetic identity (0.83) was found between *G. tomentosu* (AD3) and *G. barbadense* (AD2), based on DNA fingerprinting (Khan et al. 2000). These results enforce the phylogenetic tree constructed from the variation of repetitive DNA sequences.

The second difference is the position of the B-, F-, and E-genome species in the trees (Fig. 12). Wendel and Cronn (2003) showed the B-genome species is sister branches to either the A- and F-genome species, and the E-genome is basal to the A-, F-, and B-genome lineage. In our present study, the B-genome species with the E-genome species, *G. areysianum* (E3) and *G. incanum* (E4), was found to form a sister branch to the F- and A-genome species.

The third one is the phylogenetic relationship among the thirteen species of the D-genome clade. Although several branches of this lineage between the species agree with

the ones constructed by Wendel and Cronn (2003), the order of the lineage branches is very different (Fig. 13).



AD1, *G. hirsutum*
AD2, *G. barbadense*
AD3, *G. tomentosum*
AD4, *G. mustelinum*
AD5, *G. darwinii*

Fig. 11 Comparison of two phylogenetic trees of allopolyploid species of the genus *Gossypium*. (Left, a tree was constructed base on repetitive DNA sequences in this study; Right, a tree was adapted from Wendel and Cronn, 2003)

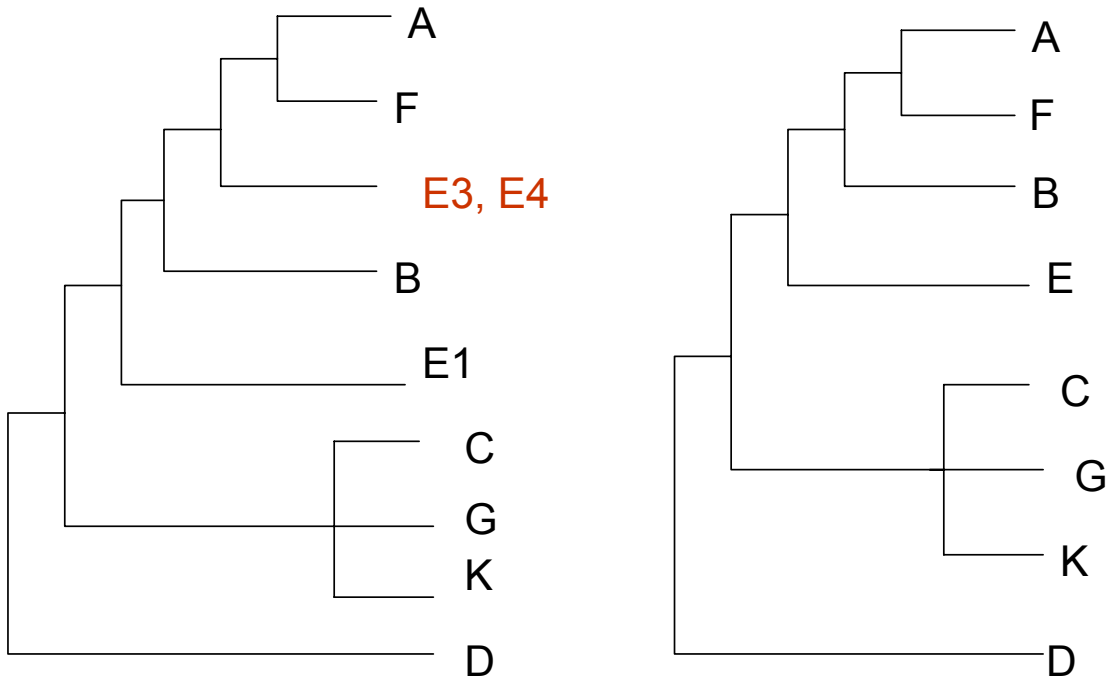
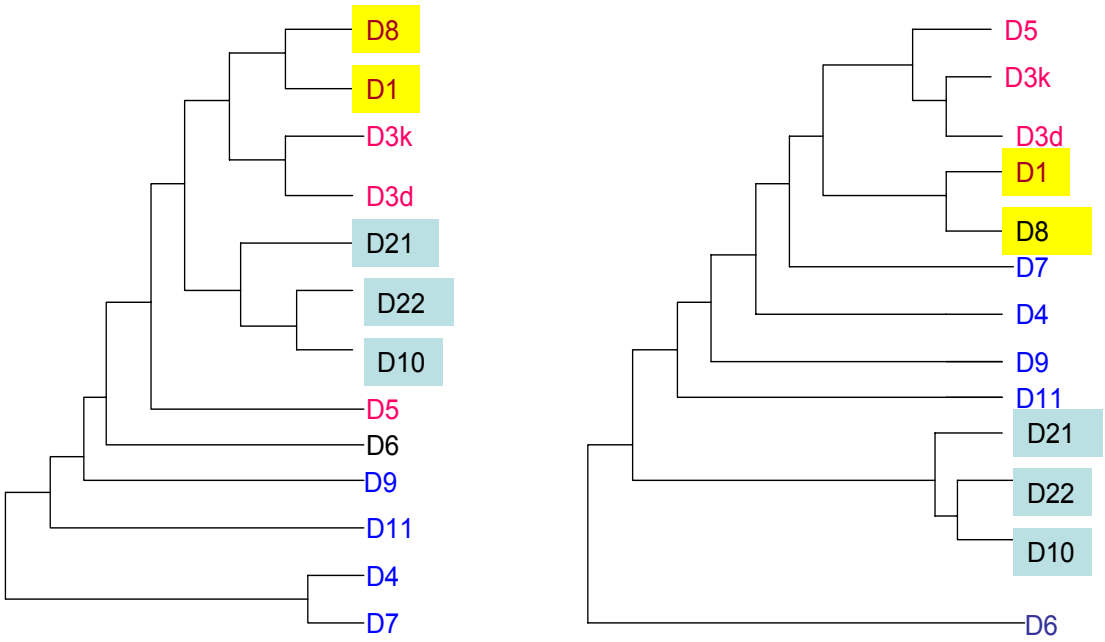


Fig. 12 Comparison of two phylogenetic trees of diploid genome groups of the genus *Gossypium*. (Left, a tree was constructed base on repetitive DNA sequences in this study; Right, a tree was adapted from Wendel and Cronn, 2003)

.



D1, *G. thurberi*
D21, *G. armourianum*
D22, *G. harknessii*
D3d, *G. davidsonii*
D3k, *G. klotzchianum*
D4, *G. aridum*
D5, *G. raimondii*

D6, *G. gossypioides*
D7, *G. lobatum*
D8, *G. trilobum*
D9, *G. laxum*
D10, *G. turneri*
D11, *G. schwendimanii*

Fig. 13 Comparison of two phylogenetic trees of diploid D-genome species of the genus *Gossypium*. (Left, a tree was constructed base on repetitive DNA sequences in this study; Right, a tree was adapted from Wendel and Cronn, 2003)

In this study, some restriction fragment bands of repetitive DNA sequences could be individually characterized as genome-specific and/or species-specific repetitive DNA sequence makers. Southern blot hybridization showed that only the A-genome and D-genome species exclusively share maker bands with allopolyploid species, confirming that only these species potentially contributed to the genomes of the polyploid species. However, analysis of RSC between the A-genome species and the polyploid species indicates that neither of the extant A-genome species, *G. herbaceum* (A1) and *G. arboretum* (A2), can be claimed the donor of the A genome of the polyploid species. Nevertheless, the ancestor represented by A1 + A2 shared significantly high RSC values with the polyploid species. This result strongly suggests that the polyploid species of *Gossypium* originated before a split between the *G. herbaceum* (A1) and *G. arboretum* (A2). Given the relatively low values of RSCs (0.51 – 0.62), the polyploid species likely originated in the early time of the A-genome species evolution. The observed significant divergence between the genomes of the D-genome diploids and the D subgenome of the polyploids further supports this inference. However, since insufficient numbers of genome- or species-specific bands were identified for the D-genome species, additional studies are needed to infer the origin of the D genome of the polyploid species. Furthermore, the additional studies may also allow addressing the questions whether the ployploid species evolved from a single or multiple polyploidization events.

REFERENCES

Cronn RC, Zhao XP, Paterson AH, Wendel JF (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

Cronn RC, Small RL, Haselkorn T, Wendel JF. (2002) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany 89: 707-725

Dvorak J, Zhang HB (1990) Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proceedings of the National Academy of Sciences of the United States of America 87: 9640-9644

Endrizzi JE, Turcotte EL, Kohel RJ. (1985) Genetics, cytogenetics, and evolution of *Gossypium*. Advances in Genetics 23: 271-375

Flavell RB, Bennett MD, Smith JB, and Smith DB (1974) Genome size and the proportion of repetitive nucleotide sequence DNA in plants. Biochem Genet 12: 257-269

Fryxell PA (1979) The natural history of the cotton tribe. Texas A&M University Press, College Station, Texas

Fryxell PA (1992) A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). Rheedea 2: 108-165

Geever RF, Katterman FRH, Endrizzi JE (1989) DNA hybridization analyses of a *Gossypium* allotetraploid and two closely related diploid species. Theor Appl Genet 77:553-559

Hanson RE, Zhao XP, Islam-Faridi MN, Paterson AH, Zwick MS, Crane CF, Mcknight TD, Stelly DM, and Price HJ (1998) Evolution of interspersed repetitive elements in *Gossypium* (Malvaceae). American Journal of Botany 85:1364-1368

Khan SA, Hussain D, Askari E, Stewart JM, Malik KA, and Zafar Y (2000) Molecular phylogeny of *Gossypium* species by DNA fingerprinting. Theor Appl Genet 101: 931-938

National Agricultural Statistics Service (NASS) (1999) Annual Crop Summary. USDA, Washington, D.C.

Page, RDM (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358

Parks CR, William DE, Dreyer DL (1975) The application of flavonoid distribution to taxonomic problems in the genus *Gossypium*. Bull Torrey Bot Club 102: 350-361

Swofford DL (2001) PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods).Version 4. Sinauer Associates, Sunderland, Mass.

Wendel JF, Albert VA. (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

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Advances in Agronomy 78: 139-186

Zhang HB, Dvorak J (1991) The genome origin of tetraploid species of *Leymus* (Poaceae: Triticeae) inferred from variation in repeated nucleotide sequences. American Journal of Botany 78(7): 871-884

Zhang HB, Dvorak J. (1992) The genome origin and evolution of hexaploid *Triticum crassum* and *Triticum syriacum* determined from variation in repeated nucleotide sequences. Genome 35: 806-814

Zhang LT, Dong JM, Decanini L, Lee MK, Ren C, Yan B, Kohel RJ, Yu J, Zhang HB, Stelly DM (2002) Development of molecular cytogenetic markers in cotton (Poster). International Plant, Animal and Microbe Genomes X Conference. January, San Diego, Calif. P285

Zhao X, Kochert G. (1993) Clusters of interspersed repeated DNA sequences in the rice genome (*Oryza*). Genome 36(5): 944-53.

Zhao XP, Wing RA, Paterson AH. (1995) Cloning and characterization of the majority of repetitive DNA in cotton (*Gossypium* L.). Genome 38: 1177-1188

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, and Paterson AH. (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Research 8: 479-492

VITA

Name: Ying Rong
Address: Department of Soil & Crop Sciences, and Institute for Plant Genomics & Biotechnology, Room 148, 2123 TAMUS, Texas A&M University, College Station, TX 77843, USA
E-mail: yingrong@neo.tamu.edu

EDUCATION

2004 Master of Science, Molecular & Environmental Plant Sciences, Texas A&M University, College Station, TX
1984 Bachelor of Science, Agronomy, Guangxi Agricultural College, People's Republic of China.

PUBLICATIONS

1. Uhm T, **Rong Y**, Xu Z, Covaleda L, Scheuring C, Zhang HB. A large-insert plant transformation-competent BIBAC library of the wild rice, *Oryza rufipogon* L. and BIBAC-based physical map of the wild rice chromosome 8 centromere. Manuscript in preparation

2. Jiang Y, Rong Y, Tao G, and Tan L. 1997. Genetic study on new-type thermo-sensitive genetic male sterile rice. Journal of Yunnan Agricultural University 12(3): 37-42

3. Jiang Y, Rong Y, Tao G, and Tan L. 1997. Breeding of Diannong S-2, a thermo-sensitive genetic male sterile line with new cytoplasm from *Japonica* rice. Southwest China Journal of Agricultural Sciences 10(3): 21-24

4. Jiang Y, Rong Y, Tao G, and Tan L. 1997. Breeding and performance of *Japonica* thermo-sensitive genetic male sterile rice line Diannong S-1 of a New Resource. Hybrid Rice 12(5): 30-31

5. Jiang Y, Rong Y and Tao G. 1997. A new type of thermo-sensitive genetic male sterility bred by hybridization. International Symposium on Two-line System Heterosis Breeding in Crops (Changsha, China) pp188-192