CSIRO PUBLISHING
Crop & Pasture Science, 2011, **62**, 859–875
http://dx.doi.org/10.1071/CP11161

Analysis of genetic diversity, population structure and linkage disequilibrium in elite cotton (*Gossypium* L.) germplasm in India

Satya Narayan Jena[A,C], *Anukool Srivastava*[A], *Uma Maheswar Singh*[A], *Sribash Roy*[A], *Nandita Banerjee*[A], *Krishan Mohan Rai*[A], *Sunil Kumar Singh*[A], *Verandra Kumar*[A], *Lal Babu Chaudhary*[A], *Joy Kumar Roy*[A,B], *Rakesh Tuli*[A,B], *and Samir V. Sawant*[A]

[A]National Botanical Research Institute (CSIR), Rana Pratap Marg, Lucknow, India.
[B]Present address: National Agri-Food Biotechnology Institute, Industrial Area, SAS Nagar, Mohali, India.
[C]Corresponding author. Email: satyanarayan@nbri.res.in

Abstract. An understanding of the level of genetic diversity is a prerequisite for designing efficient breeding programs. Fifty-one cultivars of four cotton species (*Gossypium hirsutum*, *G. barbadense*, *G. herbaceum* and *G. arboreum*) representing core collections at four major cotton research stations with a wide range of eco-geographical regions in India were examined for the level of genetic diversity, distinct subpopulations and the level of linkage disequilibrium (LD) using 1100 amplified fragment length polymorphism (AFLP) markers with 16 primer pairs combinations. The AFLP markers enabled a reliable assessment of inter- and intra-specific genetic variability with a heterogeneous genetic structure. Higher genetic diversity was noticed in *G. herbaceum*, followed by *G. arboreum*. The genetic diversity in tetraploid cotton species was found to be less than that in the diploid species. The genotypes VAGAD, RAHS14, IPS187, 221 557, Jayhellar of *G. herbaceum* and 551, DLSA17, 221 566 of *G. arboreum* were identified as the most diverse parents, useful for quantitative trait loci (QTL) analysis in diploid cotton. Similarly, LRA 5166, AS3 and MCU5 of *G. hirsutum* and B1, B3, Suvin of *G. barbadense* were most diverse to develop mapping populations for fibre quality. The internal transcribed spacer sequences were sufficient to resolve different species and subspecies of diploid cotton. Low level of genome-wide LD was detected in the entire collection ($r^2 = 0.07$) as well as within the four species ($r^2 = 0.11$–0.15). A strong agreement was noticed between the clusters constructed on the basis of morphological and genotyping data.

Additional keywords: cotton, genetic diversity, population structure, resolving power.

Received 23 June 2011, accepted 1 October 2011, published online 6 December 2011

Introduction

The genus, *Gossypium* L. comprises ~45 diploid and five tetraploid species (Rieseberg and Noyes 1998). The diploid species ($2n = 26$) are divided into eight genome groups, designated A through G and K on the basis of chromosome size and pairing behaviour in inter-specific hybrids (Endrizzi *et al*. 1985). They are globally distributed across Australia (C-, G-, and K-genomes), African-Arabia (A-, B-, E-, and F-genomes), and the Americas (D-genome). Five polyploid species are recognised to date, including the commercially important *G. hirsutum* (Upland cotton) and *G. barbadense* (Pima or Egyptian cotton). These are traditionally considered allotetraploids ($2n = 52$), containing A- and D-subgenomes and being endemic to the New World (Fryxell 1992). It is believed that AD disomic tetraploid cotton ($2n = 4x = 52$) originated ~1–2 million years ago by the hybridisation of A-genome taxon [related to the extant species *G. herbaceum* L. and *G. arboreum* L. ($2n = 2x = 26$)] with D-genome taxon [related to *G. raimondii* Ulbrich and *G. gossipioides* L. ($2n = 2x = 26$)], followed by polyploidisation (Beasley 1940, 1942; Wendel *et al*. 1992; Abdalla *et al*. 2001). The ancestral AD allotetraploid diverged in the New World to give rise to five extant AD tetraploid species including the present cultivated cottons *G. hirsutum* L. and *G. barbadense* L. (Brubaker *et al*. 1999).

India, the second largest cotton producer in the world is the ancient home of the cultivated Asiatic species of *Gossypium* L., particularly representing the origin and domestication of *G. arboreum* and *G. herbaceum* (Hutchinson *et al*. 1947). India has been growing diploid cottons since before 3000 BC. Both perennial and annual forms are found widely distributed. In addition, two tetraploid species *G. hirsutum* and *G. barbadense* were introduced by the British East India Co. and other agencies since the latter half of the 18th century. The acclimatised materials have been spread in localised pockets all over the Indian subcontinent. Some of these have also undergone isolation, introgression and continuous localised selection in various regions and are grown in smaller pockets

by local communities, besides the large-scale commercial cultivation of improved cultivars that have been developed in recent years. India has the distinction of having the largest cotton area (8.82 million ha) in the world with a record production of 24.25 million bales and a productivity of 465 kg per ha (Anonymous 2006). In India, in spite of severe competition from synthetic fibres in recent years, it is occupying the premiere position with 70 percent share in the textile industry.

An understanding of the level of genetic diversity is important for the conservation of genetic resources, identifying diverse parental lines for designing efficient breeding programs, developing mapping populations for linkage analysis and for association genetics. Diverse sets of genotypes in Indian cottons need to be identified. Such studies are important to ensure enriching the allelic diversity in gene pool (Abdalla *et al*. 2001) and for the development of DNA-based molecular markers (DeVerno and Mosseler 1997). We chose to examine genomic diversity by both amplified fragment length polymorphism (AFLP) (Vos *et al*. 1995; Mueller and Wolfenbarger 1999) and internal transcribed spacer (ITS) sequences. In addition to these markers, we have confirmed a suspected hybrid using simple sequence repeat (SSR) markers. Due to the high multiplex ratio (a large number of polymorphic loci generated in a single experiment; Rafalski *et al*. 1996) and reproducibility (Jones *et al*. 1997), AFLP is an efficient marker technique for fingerprinting and assessing genetic polymorphism (Flint-Garcia *et al*. 2003; Xiao *et al*. 2006; Bouajila *et al*. 2007; Masum Akond *et al*. 2008). Ribosomal DNA (rDNA) is a well known multi-gene family, present in tandem repeats in plants. The variation in rDNA has been used for several studies to distinguish different species, analyse genetic diversity and predict phylogenetic relationships (Pillay and Myers 1999; Sharma and Raina 2005). SSR (Jacob *et al*. 1991) are short (1–6 bp) repeat motifs that show a high level of length polymorphism due to insertion or deletion mutations of one or more repeat type (Tautz and Renz 1984). These are among the most efficient classes of molecular markers due to their hyper-variable and co-dominant nature, relatively high abundance and random distribution in the genome (Powell *et al*. 1996). Such repeats display high levels of polymorphism because of variation in repeat length and can be rapidly analysed through PCR and gel electrophoresis. SSR also allow relatively simple interpretation and genetic analysis of a single locus and can distinguish hybrids from their parents (Saghai-Maroof *et al*. 1994).

The objectives of the present study are (*i*) to estimate genetic variation within and among four cotton species represented by core collections of cotton taxa in India, by deploying both AFLP markers and ITS sequences; (*ii*) to know whether the widespread diploid species (*G. herbaceum* and *G. arboreum*) possess more genetic variation within populations than the tetraploid species; (*iii*) to establish diverse parental lines suitable for linkage mapping in diploid cotton with contrasting phenotypes and (*iv*) to evaluate genome-wide level of linkage disequilibrium (LD) in the entire collection as well as within the four cotton species. Such studies are important to identify new sources of alleles for cotton improvement.

Materials and methods

Plant materials

Fifty-one genotypes of cultivated cotton were collected from the core collections at four cotton research stations in India (Table 1). These comprised of 15 genotypes of *G. herbaceum*, 12 genotypes of *G. hirsutum*, 10 genotypes of *G. arboreum*, 13 genotypes of *G. barbadense* and one suspected hybrid (Table 1). Voucher specimens of all the accessions were deposited in the National Botanical Research Institute (NBRI) herbarium at Lucknow, India.

Phenotypic evaluation

The selection of 51 genotypes was done on the basis of morphological and yield components. The seeds of 51 genotypes were sown in nursery beds during the end of April 2009 and transplanted during the end of May 2009 in a randomised block design with three replications. Each genotype was planted in two rows with a length of 4 m at spacing of 30 × 40 cm following standard agronomical practices. Five plants were randomly selected from each replication to record the mean data of plant height, leaf size, leaf shape, leaf curvature, leaf pubescence, stem pubescence, epicalyx nature, serration of epicalyx, flower size, stigma position, petal spot, boll size, shape and presence of fuzz at the end of November 2009.

AFLP analysis

Genomic DNA was extracted from the young expanding leaves using a modified CTAB protocol (Jena *et al*. 2004). AFLP analysis was performed as described by Vos *et al*. (1995) with few modifications. Approximately 200 ng genomic DNA was digested with 10 U *Eco*RI and 10 U *Mse*I in a 20 µL reaction and incubated at 37°C for 3 h. Following heat inactivation of the enzymes, 20 µL ligation master mix containing 75 pmol each *Mse*I and *Eco*RI adapters with 20 U T4 DNA ligase in 1X T4 DNA ligase buffer was added and incubated overnight at 16°C. The digestion-ligation mixture was diluted with 80 µL sterile water. Pre-selective amplification was performed by using a tri-selective nucleotide (+3) at the 3′. AFLP amplifications were performed by adding three selective nucleotides (+3) to the end of the *Mse*I and *Eco*RI primers (Table 2). Eight +3 selective AFLP amplifications were performed using 16 primer combinations (Table 2). The *Eco*RI adaptor primers were 5′ fluorescently labelled either with FAM or JOE or NED. The *Mse*I adaptor primers were unlabelled. Each 25 µL reaction contained 5 µL diluted pre-selective solution, 1X PCR buffer, 1.5 mM $MgCl_2$, 300 µM dNTP, 4 pmol each *Eco*RI adaptor +3 primer, 25 pmol *Mse*I adaptor +3 primer, and 1 U *Taq* DNA polymerase. The amplification profile was 94°C for 2 min, 10 cycles of 94°C for 30 s, 65°C for 30 s, 72°C for 2 min, reducing the annealing temperature by 1°C per cycle, followed by 35 cycles of 94°C for 30 s, 56°C for 30 s, 72°C for 2 min, ending with 72°C for 30 min.

ITS amplification and sequencing

PCR was performed to amplify the ITS region (including ITS1, 5.8S, and ITS2) from the genomic DNA using the primers ITS-5A (5′-GGAAGTAAAAGTCGTAACAAGG-3′) and ITS-4 (5′-TCCTCCGCTTATTGATATGC-3′). These primers were

Table 1. Details of the 51 genotypes of cultivated cotton of four different species used in this study

Serial number	Genotype number	Species	Source of collection	Accession number[F]	Origin
1	AK235	*Gossypium* spp.[A]	UAS-D, Dharwad, Karnataka[B]	250308	Old World
2	579	*G. arboreum*	UAS-D, Dharwad, Karnataka	250324	Old World
3	DLSA17	*G. arboreum*	UAS-D, Dharwad, Karnataka	250392	Old World
4	576	*G. arboreum*	UAS-D, Dharwad, Karnataka	250327	Old World
5	551	*G. arboreum*	UAS-D, Dharwad, Karnataka	250326	Old World
6	575	*G. arboreum*	UAS-D, Dharwad, Karnataka	250319	Old World
7	574	*G. arboreum*	UAS-D, Dharwad, Karnataka	250323	Old World
8	577	*G. arboreum*	UAS-D, Dharwad, Karnataka	250320	Old World
9	221 568	*G. arboreum*	GAU, Banaskantha, Gujarat[C]	221568	Old World
10	221 567	*G. arboreum*	GAU, Banaskantha, Gujarat	221567	Old World
11	221 566	*G. arboreum*	GAU, Banaskantha, Gujarat	221566	Old World
12	249001	*G. barbadense*	Guwahati, Assam	249001	New World
13	249003	*G. barbadense*	Dakuapara, Assam	249003	New World
14	249005	*G. barbadense*	Kamrup, Assam	249005	New World
15	249007	*G. barbadense*	Andherijuli, Assam	249007	New World
16	249009	*G. barbadense*	Lankeswar, Assam	249009	New World
17	249011	*G. barbadense*	Sonaighuli, Assam	249011	New World
18	249012	*G. barbadense*	Lankeswar, Assam	249012	New World
19	249010	*G. barbadense*	Lankeswar, Assam	249010	New World
20	249008	*G. barbadense*	Pakharapara, Assam	249008	New World
21	249006	*G. barbadense*	Andherijuli, Assam	249006	New World
22	249004	*G. barbadense*	Dakuapara, Assam	249004	New World
23	249002	*G. barbadense*	Bhagdabari, Assam	249002	New World
24	Suvin	*G. barbadense*	TAU, Tamil Nadu[D]	250345	New World
25	VAGAD	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250311	Old World
26	GC21	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250314	Old World
27	RAHS14	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250332	Old World
28	IPS187	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250313	Old World
29	221 573	*G. herbaceum*	UAS-D, Dharwad, Karnataka	221573	Old World
30	7GP	*G. herbaceum*	UAS-D, Dharwad, Karnataka	259329	Old World
31	AH41	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250307	Old World
32	H17	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250331	Old World
33	221 557	*G. herbaceum*	GAU, Banaskantha, Gujarat	221557	Old World
34	RAHS132	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250312	Old World
35	182 LC	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250322	Old World
36	DB312	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250318	Old World
37	221 547	*G. herbaceum*	UAS-D, Dharwad, Karnataka	221547	Old World
38	RAHS 127	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250330	Old World
39	Jayhellar	*G. herbaceum*	UAS-D, Dharwad, Karnataka	250325	Old World
40	JKC703	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh[E]	250304	New World
41	JKC770	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250334	New World
42	JKC771	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250346	New World
43	JKC777	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250335	New World
44	JKC737	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250342	New World
45	JKC725	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250305	New World
46	JKC752	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250315	New World
47	JKC783	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250321	New World
48	LRA5166	*G. hirsutum*	JK Agri., Hyderabad, Andhra Pradesh	250341	New World
49	AS3	*G. hirsutum*	UAS-D, Dharwad, Karnataka	250347	New World
50	MCU5	*G. hirsutum*	UAS-D, Dharwad, Karnataka	250336	New World
51	KC2	*G. hirsutum*	UAS-D, Dharwad, Karnataka	250343	New World

[A]Genotype was not assigned taxonomically to a species.
[B]University of Agricultural Sciences, Dharward, Karnataka.
[C]Gujarat Agricultural University, Banaskantha, Gujarat.
[D]Tamil Nadu Agricultural University, Tamil Nadu.
[E]JK Agri Genetics, Hyderabad, Andhra Pradesh.
[F]Accession number assigned at National Botanical Research Institute (NBRI) herbarium, Lucknow.

designed to anneal to the 3′ terminus of the 18S (primer ITS-5A) and the 5′ terminus of the 26S rRNA genes (primer ITS-4), respectively. Using 34 cycles with a PCR profile of 40 s at 94°C, 40 s at 56–58°C and 60 s at 72°C, the ITS region was amplified, purified using PCR purification kit (Nucleospin Extract II, Macherey Nagel, GmbH & Co. KG, Duren,

Table 2. Details of amplified AFLP amplicons in respective primer pair combinations

Primer pair number	Primer pair[A]	Number of amplicons	Number (%) of polymorphic amplicons (Entire set of 51 genotypes)	PIC	RP	Number (%) of polymorphic amplicons *G. arboreum*	*G. herbaceum*	*G. barbadense*	*G. hirsutum*
1	E-acg/M-cag	71	61 (85.9)	0.49	41.8	21/67 (31.3)	37/63 (58.7)	31/66 (47.0)	13/51 (25.5)
2	E-act/M-ctc	65	45 (69.2)	0.39	27.0	18/57 (31.6)	39/62 (62.9)	23/57 (40.4)	15/46 (32.6)
3	E-aca/M-cat	146	94 (64.4)	0.48	57.3	82/109 (75.2)	82/133 (61.7)	81/143 (56.6)	52/113 (46.0)
4	E-acc/M-cta	8	6 (75.0)	0.47	43.5	4/6 (66.7)	5/7 (71.4)	4/7 (57.1)	3/6 (50.0)
5	E-aca/M-caa	95	72 (75.8)	0.45	55.8	53/82 (64.6)	66/92 (71.7)	36/87 (41.4)	43/81 (53.1)
6	E-agc/M-ctc	69	52 (75.4)	0.43	36.5	43/61 (70.5)	45/65 (69.2)	38/64 (59.4)	15/55 (27.3)
7	E-aag/M-cac	125	77 (61.6)	0.45	46.9	74/113 (65.5)	70/119 (58.8)	35/112 (31.3)	52/97 (53.6)
8	E-agg/M-cac	52	32 (61.5)	0.48	26.3	18/44 (40.9)	28/50 (56.0)	16/42 (38.1)	23/48 (47.9)
9	E-aac/M-cta	81	73 (90.1)	0.44	45.3	49/78 (62.8)	26/74 (35.1)	26/72 (36.1)	25/63 (39.7)
10	E-acg/M-cac	71	56 (78.9)	0.46	42.3	30/62 (48.4)	44/68 (64.7)	31/57 (54.4)	18/59 (30.5)
11	E-acg/M-ctt	25	14 (56.0)	0.41	11.9	12/20 (60.0)	19/25 (76.0)	11/23 (47.8)	7/21 (33.3)
12	E-acg/M-ctg	58	42 (72.4)	0.33	25.3	21/48 (43.8)	23/55 (41.8)	30/52 (57.7)	4/43 (9.3)
13	E-acg/M-cta	51	35 (68.6)	0.40	30.9	15/43 (34.9)	27/48 (56.3)	29/48 (60.4)	19/41 (46.3)
14	E-acg/M-caa	86	63 (73.3)	0.27	26.8	18/70 (25.7)	43/81 (53.1)	46/78 (59.0)	34/68 (50.0)
15	E-acg/M-cat	54	43 (79.6)	0.46	40.2	28/43 (65.1)	39/52 (75.0)	22/48 (45.8)	26/48 (54.2)
16	E-acg/M-ctc	43	32 (74.4)	0.45	39.3	18/40 (45.0)	23/41 (56.1)	21/38 (55.3)	16/37 (43.2)
Mean	–	68.8	49.8 (72.5)	0.43	37.3	31.5/58.9 (53.4)	38.5/64.7 (59.5)	30/62.1 (48.3)	22.8/54.8 (41.6)
Total	–	1100	797	–	–	504/943 (53.4)	616/1035 (59.5)	480/994 (48.3)	365/877 (41.6)

[A]E-xxx: *Eco*RI three-base extension primer; M-xxx: *Mse*I three-base extension primer.

Germany), and sequenced. The sequencing reactions were performed using the BigDye Terminator version 3.1 Cycle-Sequencing Kit (Applied Biosystems, Foster city, CA, USA). Both the strands were sequenced on an automated DNA analyser 3730xl (Applied Biosystems).

Simple sequence repeats amplification

The diploid species and the suspected hybrid were amplified following a standard PCR protocol in a touch-down PCR with 24 SSR primer pairs developed from a repeat enriched genomic library (Table S8 as Supplementary Materials to this publication and available on journal's website, Gene Bank accession number: HQ524341–HQ524715). These amplified alleles were analysed onto 3% agarose gels.

Phenotypic data analysis

The phenotypic data were converted into scores ranging from 0 to 1 with three distinct classes (0.0, 0.5 and 1.0) for further analysis. Pairwise squared Euclidean distances were calculated from the phenotypic data and cluster analysis was performed to construct a dendrogram using the Euclidean distance matrix in SPSS version 16 software (SPSS Inc., Chicago, IL, USA).

AFLP data analysis

Selective AFLP amplification products were resolved on the polymer, POP-7 (Applied Biosystems) in an ABI 3730xl DNA Analyser. Image analysis was performed using GeneMapper software, version 4.0 (Applied Biosystems) using its automatic function to read AFLP peaks. The fragment data were recorded as '1' (present) or '0' (absent). Polymorphism information content (PIC) for each AFLP primer pair was calculated according to Ghislain *et al.* (1999), i.e. $PIC = 1 - p^2 - q^2$, where p and q are the frequencies of presence and absence of the fragments. Resolving power (RP) of each AFLP primer pair was calculated according to Prevost and Wilkinson (1999), i.e. $RP = \sum l_b$, where, l_b represents informativeness of the fragments. The l_b can be represented into a 0–1 scale by the following formula: $l_b = 1 - [2 \times |0.5 - p|]$, where, p is the proportion of the 51 accessions containing the fragment.

Genetic identity among four species and the populations within the species were determined using Jaccard's unbiased measures of genetic similarity (GS), i.e. genetic similarity coefficient (Jaccard 1908) using the program PHYLIP version 3.2 (Felsenstein 1985, 1989). Cluster analysis to reveal genetic relationships was conducted using the unweighted pair group method with arithmetic averages (UPGMA). The dendrogram was constructed using the program NTSYS-pc version 2.0 (Rohlf 1998) based on the pairwise Jaccard GS. The pairwise genetic dissimilarity (GD) i.e. genetic distance matrix was calculated as GD = 1 – GS. The principal coordinate analysis (PCA) was conducted using the genetic distance matrix in Mod3D plot option of NTSYS-pc (Rohlf 1998) to visualise the genetic differences among and within the four cotton species studied.

POPGENE version 1.31 software (Yeh and Boyle 1997) was used to calculate the parameters of genetic variation under the assumption of Hardy–Weinberg equilibrium, including the percentage of polymorphic loci (Nei 1973), effective number of alleles per locus (Hartl and Clark 1989), gene diversity (H_e = expected heterozygosity; (Nei 1973), total genetic diversity (H_t), genetic diversity within populations (H_s), genetic diversity among populations (Dst), and the relative magnitude of genetic differentiation among populations (Gst = Dst/H_t; Nei and Li 1979). The phenotypic diversity quantifying the degree of AFLP polymorphism within populations was calculated using Shannon's information index ($I = -\Sigma p_i \log_{10} p_i$, where p_i is the frequency of the presence or absence of a AFLP fragment; Lewontin 1972).

Hierarchical analyses of molecular variance (AMOVA, Excoffier *et al*. 1992) based on the pairwise squared Euclidean distances among molecular phenotypes (presence or absence of fragments) were conducted using ARLEQUIN version 3.01 (Excoffier and Smouse 1994; Schneider *et al*. 2000; Excoffier 2006) to further quantify the amount of genetic variation residing at two levels (i.e. among and within populations; among species and within species). The same program was used to generate the matrix of pairwise F_{st} values, which indicate the genetic differentiation between populations, and are analogous to Gst if a locus consists of two alleles as applicable in dominant marker analyses (e.g. random amplified polymorphic DNA; Nybom and Bartish 2000). The significance levels for AMOVA were evaluated using a permutation approach (default value of 1023 replications).

To estimate the number of subpopulations in the cotton collection, population structure analysis was conducted using a model-based clustering method implemented in the program STRUCTURE version 2.2 (Pritchard *et al*. 2000) with burning length of 30 000 followed by 100 000 iterations in the admixture model. Based on the previous knowledge of the cotton collection, the population structure analysis was run assuming the number of subpopulations (K) from 2 to 7. Ten independent runs were assessed for each fixed K. We further used ΔK, an *ad hoc* quantity related to the second-order rate of change of the log probability of data with respect to the number of clusters to predict the real number of clusters (Evanno *et al*. 2005).

Evaluation of linkage disequilibrium

LD, squared allele-frequency correlations (r^2) between pairs of polymorphic loci was evaluated using the software package Tas.SEL excluding alleles with frequency under 10% (Somers *et al*. 2007). Since each AFLP analysis is based on multi-loci markers, Tas.SEL calculates a weighted average of r^2 between any two loci (Farnir *et al*. 2000) by calculating r^2 for all possible combinations, and then the allelic frequencies are used to weight them. The significance of pairwise LD (P-values) among all possible pairs of loci was also evaluated by Tas.SEL with the rapid permutations test. The loci were considered to have significant LD if $P < 0.001$ or $r^2 > 0.5$. As the AFLP markers were not known for its chromosomal localisation, intra/inter-chromosomal LD was not plotted.

ITS sequence analysis

The edited ITS region sequences were aligned using MEGA 4.0 (Tamura *et al*. 2007) with the default setting of 15 gap opening penalty and 6.6 gap extension penalty and 0.5 weighted DNA transition. The published ITS sequences of the four taxa for the present work (Gene Bank accession number U12712, U12713, U12715, and U12719) were included in the software option before analysis. Insertions or deletions (indels) introduced during the alignment of ITS sequences were treated as missing data and a subset of the indels was added to the data matrix as binary characters, in the form of presence (1) or absence (0). The inter-specific and intra-specific sequence divergence was calculated using a Kimura's two-parameter model and the tree was constructed to validate the respective cultivars in respective species. The neighbour-joining (NJ), UPGMA and parsimony methods were used to construct trees from an alignment of all 743 ITS sequences. For the NJ and UPGMA methods (Saitou and Nei 1987), we estimated Kimura's two-parameter distances (Kimura 1980) between all pairs of sequences (transition/transversion ratio = 2.00). A maximum parsimony (MP) tree using MEGA version 4.0 could not be developed due to formation of a large numbers of equally parsimonious trees. In both NJ and MP analyses, all positions containing gaps and missing data were eliminated from the dataset (Complete Deletion option). Support values of the internal branches of NJ and MP trees were evaluated through bootstrap method (500 replicates).

Statistical correlation between different matrices

The product moment correlations among similarity matrix based on AFLP data obtained in the present study, squared Euclidean distance matrix based on data of phenotypic traits, and pairwise distance matrix based on the ITS sequence were estimated using the normalised Mantel Z statistics (Mantel 1967).

Results

Level of polymorphism in four different species

Using 16 primer pairs, 1100 scorable fragments were generated and 797 of those were polymorphic among the 51 genotypes representing four cotton species (Table 2). Eighteen ambiguous and 19 redundant peaks were discarded before the subsequent analyses. The number of amplicons/primer combination ranged from 25 to 146 with an average of 69 amplicons per primer combination while the number of polymorphic amplicons ranged from 6 to 94 with an average of 50 polymorphic amplicons per primer combination (Table 2). The minimum and maximum percentages of polymorphisms were detected by the primer combinations E-acg/M-ctt and E-aac/M-cta, respectively. The PIC values of AFLP markers ranged from 0.27 to 0.49 with an average of 0.43 (Table 2). The RP of AFLP markers scored least in the primer combination, E-acg/M-ctt while maximum with primer combination, E-aca/M-cat with an average of 37.3 (Table 2).

Within the tetraploid cotton (AD-genome) species, a total of 48.3 and 41.6% polymorphic amplicons were produced by 16 primer combinations in *G. barbadense* and *G. hirsutum*, respectively (Table 2). The total number of polymorphic amplicons and average number of polymorphic amplicon were varied in two different species (Table 2). In *G. barbadense*, the minimum and maximum percentages of polymorphism were detected by the primer combinations E-aag/M-cac and E-acg/M-cta, respectively, while in *G. hirsutum*, primer combination E-acg/M-ctg scored least and E-acg/M-cat scored maximum (Table 2).

Within the diploid cotton (A-genome) species, a total of 53.4 and 59.5% polymorphic amplicons were produced by 16 primer combinations in *G. arboreum* and *G. herbaceum*, respectively (Table 2). The number of polymorphic amplicons by individual primer pair ranged from 12 to 82 with an average of 53 polymorphic amplicons per primer combination in *G. arboreum* and the same was varied from 19 to 82 with an average of 59.5 polymorphic amplicons per primer combination in *G. herbaceum* (Table 2). The minimum and maximum

percentages of polymorphism were detected by the primer combinations E-acg/M-caa and E-aca/M-cat, respectively, in *G. arboreum* while primer combination E-aac/M-cta scored least and E-acg/M-ctt scored maximum in *G. herbaceum*.

In this study, we found species-specific and ploidy-specific markers. There were five *G. arboreum* specific, and five *G. herbaceum* specific AFLP markers (Table S5 in supplementary materials available on journal's website). Similarly, there were eight *G. hirsutum*-specific and eight *G. barbadense*-specific AFLP markers (Table 3). Twenty-one markers shared between the two A-genome species (*G. arboreum* and *G. herbaceum*) are designated as 'A-specific markers' (Table S5). Twenty-nine markers shared between the tetraploid cultivars (*G. barbadense* and *G. hirsutum*) were designated as 'AD-specific markers' (Table S5).

Genetic relationship among four different species based on phenotypic traits

Genetic diversity among the 51 cotton genotypes representing four cotton species (*G. arboreum*, *G. herbaceum*, *G. barbadense* and *G. hirsutum*) was studied using the data on 22 phenotypic traits (Fig. 1). The genetic distance (squared Euclidean distance) for all possible pairs of the 51 genotypes varied with an average value of 6.1 (Table S2, refer to Supporting information section). The genotypes were grouped into four well separated clusters that are consistent with the traditional taxonomic four species (Fig. 1): 11 genotypes into *G. arboreum* cluster (A-genome), 15 genotypes into *G. herbaceum* (A-genome) cluster, 13 genotypes into *G. barbadense* cluster (AD-genome), and 12 genotypes into *G. hirsutum* cluster (AD-genome). In *G. barbadense* cluster, the genotype 'Suvin' was different from the other 12 genotypes and occupied basal position in the cluster. The three *G. arboreum* genotypes '221 566', '221 567' and '221 568' exhibited an intermediate leaf shape (a trait predominantly found in most *G. herbaceum* genotypes; data not given) unlike the rest of the *G. arboreum* genotypes. As a result, these three genotypes grouped together into a subcluster in the *G. arboreum* cluster (Fig. 1). In the *G. arboreum* cluster, the genotype 'AK235' grouped separately from the rest of the *G. arboreum* genotypes (Fig. 1) since it had large boll size, intermediate leaf shape and the presence of fold in leaf sinus.

Genetic relationship among four different species based on AFLP markers

GS coefficients among the 51 genotypes ranged from 0.20 to 0.78 with an average of 0.40 (Table S1, available as an Accessory publication to this paper). The mean GS within the geographically and historically diverse sets of the four species were more or less similar, such as 0.57 in *G. barbadense*, 0.60 in *G. arboreum*, 0.57 in *G. hirsutum* and 0.53 in *G. herbaceum*. The levels of GS between the A-genome species and AD-genome species were low. The two A-genome species ranged the GS from 0.24 to 0.58 while two AD-genome species ranged the same from 0.27 to 0.46. There was 50% GS within diploid genotypes and 51% GS in tetraploid genotypes while the GS between diploid and tetraploid genotypes was 30% (Table S3, refer to Supporting information section). The mean GS within *G. hirsutum* and *G. barbadense* was 0.57 while that in *G. herbaceum* and *G. arboreum* was 0.53 and 0.60, respectively (Table S3, refer to Supporting information section).

The data on the 797 polymorphic AFLP loci were used to make groups following UPGMA clustering (Fig. 2). The taxa fell into two well supported major clusters, consistent with the cytogenetic genome groups: 25 genotypes fell into 'AD-genome' cluster and the 26 taxa into 'A-genome' cluster. The AD-genome cluster comprised all 12 *G. hirsutum* cultivars into a single subcluster. All 13 *G. barbadense* cultivars grouped into another subcluster. The 'A-genome' cluster also comprised of two subclusters. The *G. herbaceum* subcluster contained all 15 genotypes of *G. herbaceum* and a cultivar, AK235 from southern India, (which was morphologically more similar to the *G. arboreum* species). The *G. arboreum* subcluster contained all 11 *G. arboreum* genotypes. AK235 showed mean genetic distance of 0.55 ± 0.07 with other cultivars of *G. arboreum*.

Principal coordinate analysis

To resolve three-dimensional patterns of genetic relationship among the 51 genotypes representing four cotton species, PCA was conducted using the 1100 AFLP markers. The first two

Table 3. Analysis of molecular variance among and within four *Gossypium* species as well as within and among diploid and tetraploid species

Source of variation	d.f.[A]	SSD[B]	MSD[C]	Variance components	% of total[D]	*P*-value[E]
Among tetraploid and diploid group	1	131.7	131.7	0.13	20.0	<0.001
Among species within ploidy group	2	390.9	195.5	0.06	9.2	<0.001
Among genotypes within population	50	13 224	264.5	0.46	70.8	<0.001
Within individual species						
G. hirsutum	11	3015.2	274.1	–	–	–
G. barbadense	12	2862.5	238.5	–	–	–
G. herbaceum	14	3307.2	236.2	–	–	–
G. arboreum	10	2493.1	249.3	–	–	–
Total	54	13 746.6	–	–	–	–

[A]Degree of freedom.
[B]Sum of squared deviation.
[C]Mean squared deviation.
[D]Percentage of total molecular variance.
[E]Probability of obtaining a larger component estimate; number of permutations used = 1000.

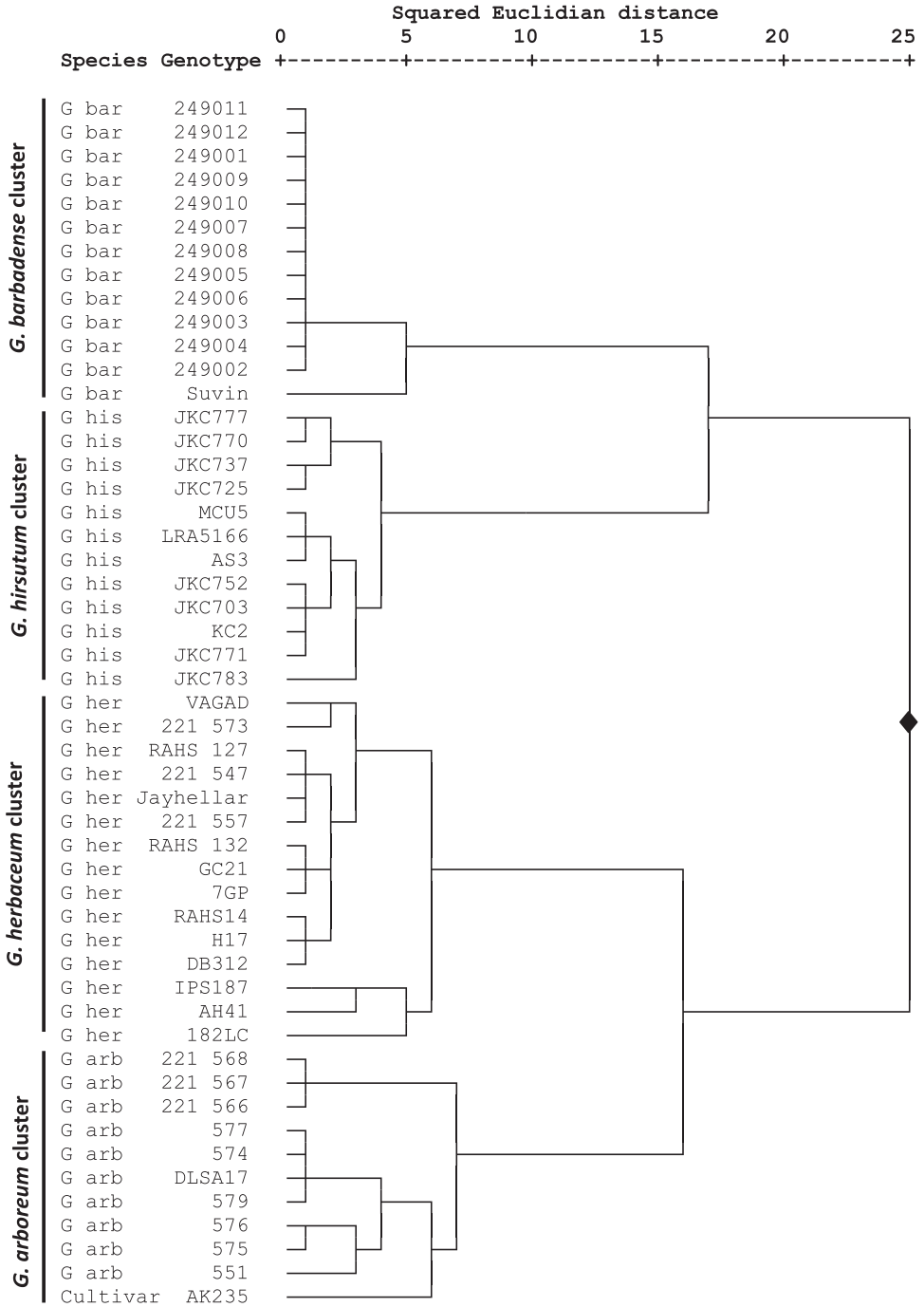

Fig. 1. Dendrogram of 51 cotton genotypes based on the data on 22 phenotypic traits; the genotype AK235 clustered with *G. arboretum*.

components (Dim-1 and Dim-2) accounted for 45.3% of the total variation. The first three components (Dim-1, Dim-2, and Dim-3) accounted for 53.3% of the total variation (Fig. 3; Table S4, refer to Supporting information section). PCA analysis of the 51 genotypes showed a clear distinction between the two AD-genome species (*G. barbadense* and *G. hirsutum*) (Fig. 3). The cultivar 'AK235' stood separately from all other accessions of *G. arboreum* and *G. herbaceum*. It was skewed in its position in the group of *G. herbaceum*, suggesting introgression between *G. herbaceum* and *G. arboreum*. The pattern of grouping of the 51 genotypes on the basis of PCA largely resembled the grouping of the genotypes obtained by UPGMA analysis (Figs 2 and 3).

Inter- and intra-genetic variation based on AFLP analysis

The level of allelic diversity within *G. arboreum*, *G. barbadense*, *G. herbaceum*, and *G. hirsutum* was estimated using mean

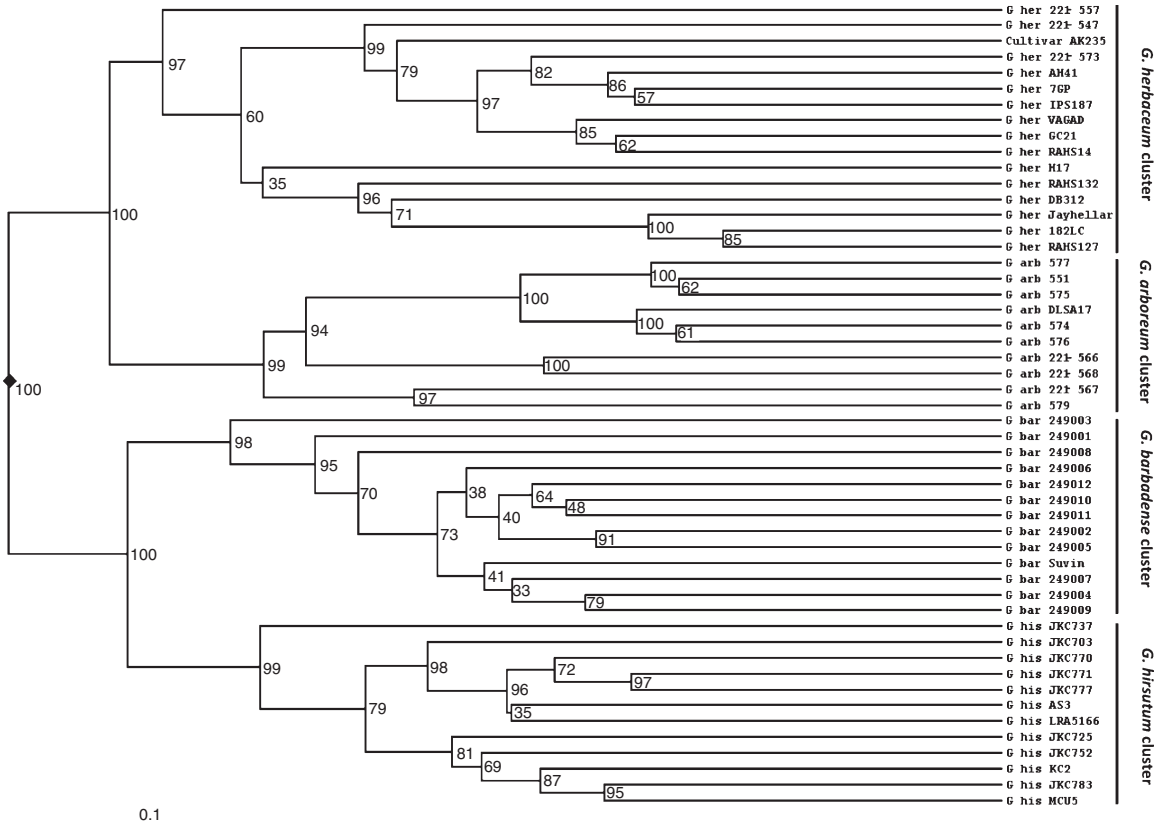

Fig. 2. Dendrogram of 51 cotton genotypes based on pairwise Jaccard similarity coefficients using data on the 797 AFLP polymorphic markers; the genotype AK235 clustered with *G. herbaceum*; the number at the node (bootstrap value) explains the robustness of the nodes.

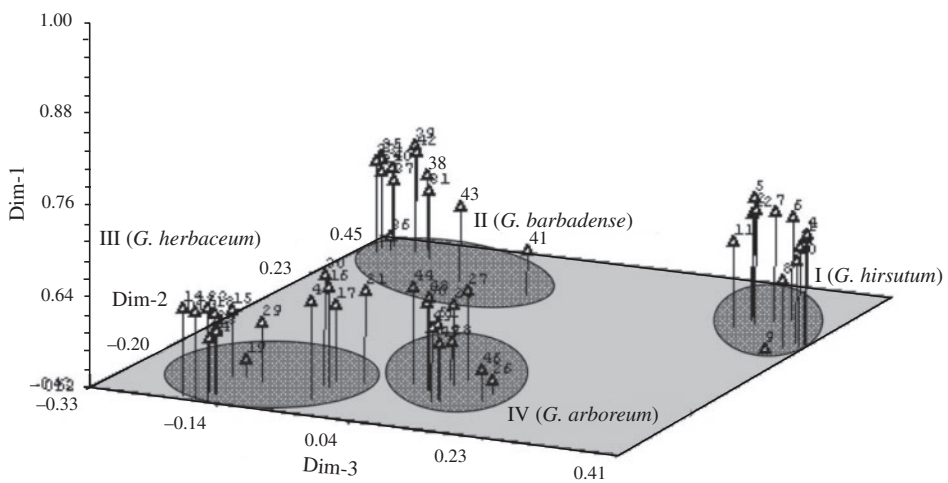

Fig. 3. Plot of 51 genotypes based on Jaccard similarity coefficient using principal coordinate analysis; the first (Dim-1), second (Dim-2), and third principal coordinates accounted for 30.96, 14.31, and 7.99% of total variation, respectively.

panmictic heterozygosity (H_t; Nei and Li 1979) and Shannon's information index (Shannon and Weaver 1949; Lewontin 1972). Mean panmictic heterozygosity, H_t was $0.23 \pm 0.0.19$ for *G. arboreum*, 0.21 ± 0.18 for *G. barbadense*, 0.25 ± 0.18 for *G. herbaceum*, and 0.21 ± 0.18 for *G. hirsutum*. Mean Shannon's information indices were 0.35 ± 0.26 for

G. arboreum, 0.33 ± 0.25 for *G. barbadense*, 0.39 ± 0.24 for *G. herbaceum*, and 0.33 ± 0.26 for *G. hirsutum*. The effective number of alleles was also estimated in the four cotton species. All four cotton species showed a low level of homozygosity (Hartl and Clark 1989) reflected as the effective number of alleles (1.34 ± 0.35 to 1.42 ± 0.35). The genotypes belonging to *G. herbaceum* had the maximum (1.42 ± 0.35) mean effective number of alleles while the mean observed number of alleles was 1.87 ± 0.34. *G. barbadense* showed the minimum (1.34 ± 0.35) mean effective number of alleles and the mean observed alleles was 1.75 ± 0.43.

AMOVA analysis using 797 polymorphic AFLP markers was conducted to partition genetic variance in the two ploidy groups (diploid and tetraploid species), among the four cotton species, and the 51 genotypes. The analysis revealed that there was significant genetic variation between the two ploidy groups ($P < 0.001$; Table 3). The genetic variance among the four species and the 51 genotypes was also significant ($P < 0.001$; Table 3). Variations between the two ploidy groups, among the four species and the 51 genotypes accounted for 20, 9.2 and 70.8% of the total molecular variance, respectively (Table 3). The inter-genotype variation differed among the four species. *G. hirsutum* had the largest mean squared deviation, whereas *G. herbaceum* had the least. The other two species *G. barbadense* and *G. arboreum* had intermediate mean squared deviation levels (Table 3).

Population structure analysis

The model-based population structure analysis assuming four subpopulations ($K = 3$) using 1100 AFLP markers in STRUCTURE program grouped the 51 genotypes into three clusters/subpopulations. All 11 *G. arboreum* and *G. herbaceum* genotypes grouped in two clusters. Similarly, the genotypes belonging to *G. hirsutum* and *G. barbadense* grouped into one single cluster (Fig. 4). Even at the assumption of higher number of subpopulations ($K = 5$–7), the number of subpopulations remained three, representing the four species (Fig. 4). By assuming two subpopulations, the structure analysis grouped the 51 genotypes into two clusters according to ploidy level: one cluster for diploid species (*G. arboreum* and *G. herbaceum*) and the second cluster for tetraploid species (*G. barbadense* and *G. hirsutum*) (Fig. 4). In most cases, the likelihood increases until the real *K* was reached, and then eased off. On the other hand, the distribution of delta *K* almost always showed a mode at the real K (Fig. 5, Table S7, refer to supporting information section).

Linkage disequilibrium analysis

LD, r^2 was estimated using 1100 AFLP markers. In the entire collection, 4.18% out of 6 044 502 possible genome-wide marker pairs were in LD at $P < 0.001$, and the strongest LD

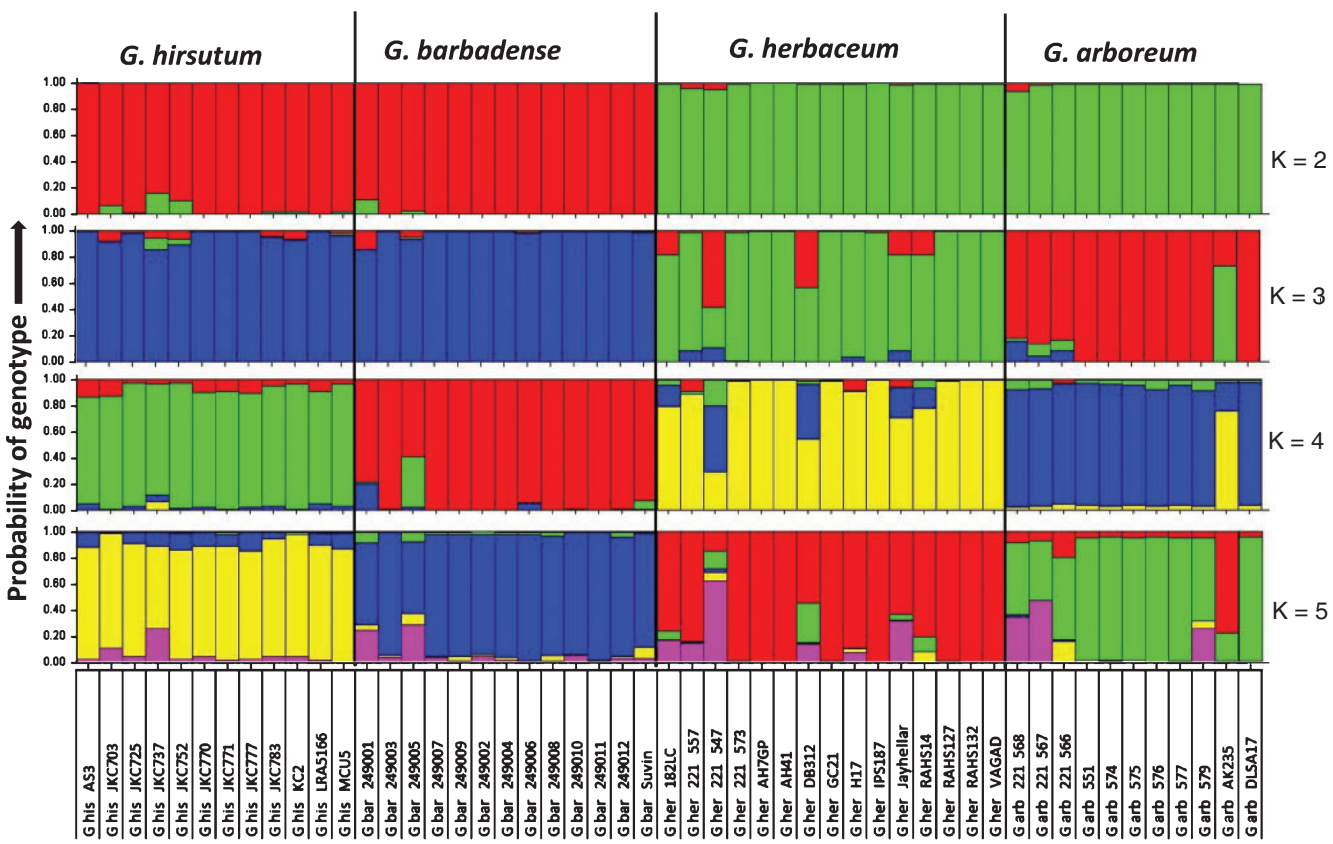

Fig. 4. Graphical plot of population structure analysis using 1100 AFLP markers in the program 'STRUCTURE; K = 2–5 represents assumption of 2–5 subpopulations and Ln P(D) represents the probability of log-likelihood of the data.

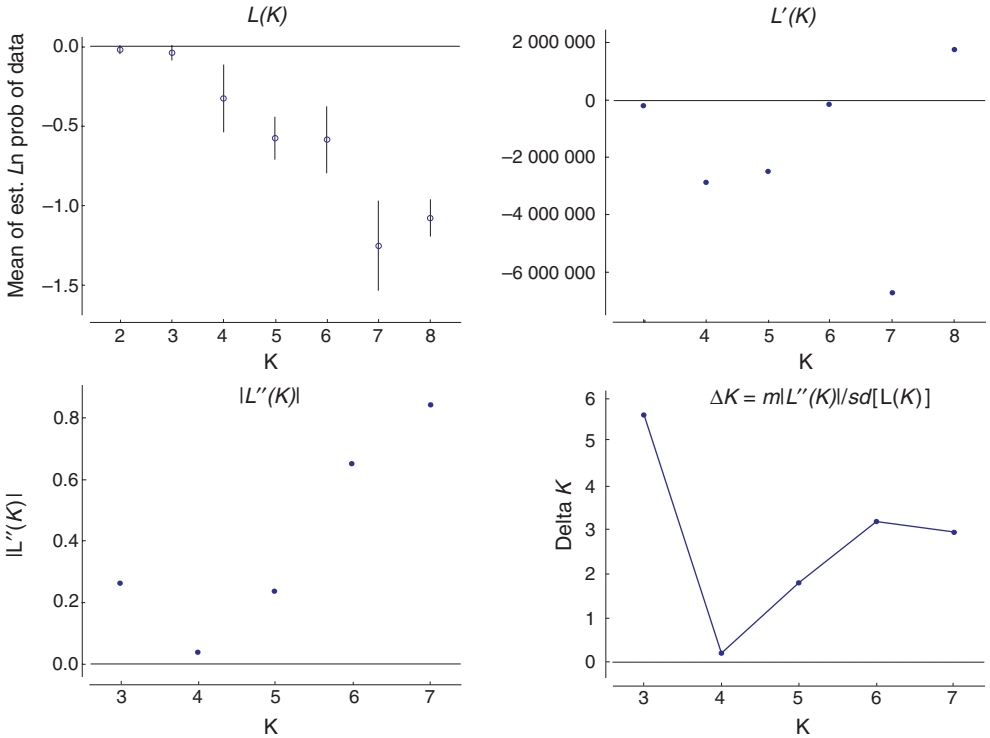

Fig. 5. Description of the four steps for the graphical method allowing detection of the true number of groups K^*. (*a*) Mean $L(K)$ (±s.d.) over three runs for each K-value. The model considered here is a hierarchical island model using all 51 individuals and 1100 AFLP loci. (*b*) Rate of change of the likelihood distribution (mean ± s.d.) calculated as: $L'(K) = L(K) - L(K-1)$. (*c*) Absolute values of the second-order rate of change of the likelihood distribution (mean ± s.d.) calculated according to the formula: $IL'(K)I = IL'(K+1) - L'(K)I$. (*d*) Delta K calculated as: $\Delta K = m|L''(K)|/sd[L(K)]$. The modal value of this distribution is the true K^* or the uppermost level of structure, here three clusters.

($r^2 = 1$) was observed for 302 marker pairs (Table 4). The level of LD (mean r^2) in the entire set was 0.07. In *G. herbaceum*, a total of 37 marker pairs were in LD at $P < 0.001$, and the strongest LD ($r^2 = 1$) was observed for 3655 marker pairs. The level of LD (mean r^2) in *G. herbaceum* was 0.11. In *G. arboreum*, no marker pairs were in LD at $P < 0.001$, and there were no marker pairs with strongest LD ($r^2 = 1$). In *G. hirsutum* and *G. barbadense*, a total of 28 and 167 marker pairs were in LD at $P < 0.001$, respectively, and the strongest LD ($r^2 = 1$) was observed for 2425 and 2863 marker pairs, respectively. The same level of LD (mean $r^2 = 0.11$) was observed in *G. hirsutum* and *G. barbadense* (Table 4).

ITS sequence analysis

The sequences of ITS regions including ITS1, 5.8S, and ITS2 were compared among the 51 genotypes. Boundaries for ITS1 and ITS2 in *Gossypium* were determined by comparison with the published sequences (Wendel *et al.* 1995*a*). The length of

Table 4. Linkage disequilibrium (LD) analysis in the entire 51 cotton genotypes as well as within individual four species

LD characteristics	Entire ($n = 51$)	*G. herabaceum* ($n = 15$)	*G. arboretum* ($n = 11$)	*G. hirsutum* ($n = 12$)	*G. barbadense* ($n = 13$)
Total number of markers	1100	688	651	644	550
Number of marker pairs	604 450	236 328	211575	207 046	150 975
Mean r^2	0.07	0.11	0.15	0.11	0.11
Number of markers pairs at $0 \leq P \leq 0.001$	25274 (4.18%)	37 (0.02%)	0	28 (0.01%)	167 (0.11%)
Number of markers pairs at $0.001 \leq P \leq 0.01$	855 (0.14%)	4332 (1.83%)	3847 (1.82%)	741 (0.36%)	873 (0.58%)
Number of markers pairs at $r^2 = 1$	302 (0.05%)	3655 (1.54%)	0	2425 (1.17%)	2863 (1.89%)

the entire ITS sequences (ITS1, 5.8S, and ITS2) in the 51 genotypes of *Gossypium* ranged from 605 (Jayhellar of *G. herbaceum*) to 731 bases (JKC725 of *G. hirsutum*). The length of ITS1 ranged from 238 (JKC752) to 293 (JK771) bases in *G. hirsutum*, 240 ('249008') to 293 ('249004' and '249009') bases in *G. barbadense*, 289 (Jayhellar) to 294 (182 LC) bases in *G. herbaceum*, and 289 ('551') to 294 (DLSA17) bases in *G. arboreum*. In case of ITS2, the sequences of 27 genotypes (out of 51) were incomplete (Fig. S1, available as Supplementary Materials on journal's website). The comparison was done for the ITS2 sequences from only 32 genotypes. The length of ITS2 ranged from 238 (JKC752) to 293 (JKC771) bases in *G. hirsutum*, 240 ('249008') to 293 ('249004' and '249009') bases in *G. barbadense*, 289 (Jayhellar) to 294 (182 LC) bases in *G. herbaceum*, and 289 ('551') to 294 (DLSA17) bases in *G. arboreum*.

We found three gaps, each of single nucleotide in ITS1 (marked as solid circles in Fig. S1). In the ITS2 sequences, we found only one gap of single nucleotide at position 600 (marked as solid circles in Fig. S1). Multiple substitutions of varying lengths were found in the diploid genotypes of *G. herbaceum* and *G. arboreum* in comparison to the tetraploid genotypes *G. hirsutum* and *G. barbadense* (Fig. S1). The GC content of the ITS region varied from 55.8 to 60.2% with an average of 57.85% (Table S6). The nucleotide frequencies were 0.205 for A, 0.214 for T, 0.296 for C, and 0.285 for G and the transition/transversion rate ratios were: $k1 = 10.397$ (purines) and $k2 = 0.29$ (pyrimidines). The overall transition/transversion bias (R) was 3.072, where $r = [A*G*k_1 + T*C*k_2]/[(A + G)*(T + C)]$. The dataset including alignment gaps and missing data comprised of 746 nucleotide positions, out of which 635 were conserved, 111 were variable and 61 were parsimony informative sites. The aligned sequences contained the conserved region of 164 bp (encoded from 307 to 470) of 5.8S rRNA.

The sequence alignment of ITS regions among the 51 genotypes was largely unambiguous, except in certain regions where several species-specific substitutions occurred. For example, *G. barbadense* was found to have two species-specific substitutions in ITS1: one was 'T' in place of 'C' (encoded at 115) and the other was 'C' in place of 'T' (encoded at 286). In *G. arboreum*, three species-specific substitutions were noticed: two in ITS1 (encoded at 72, 166) and one in ITS2 (encoded at 516). One species-specific substitution (encoded at 205) was found in *G. herbaceum*. *G. herbaceum* genotype H17 showed unique point mutations with 'T' at positions 45 and 254.

Out of four species-specific single nucleotide polymorphism (SNP) in cotton (mentioned above), both *G. arboreum* and *G. herbaceum* type alleles were identified at three positions in the genotype AK235. It showed the presence of 'T' as well as 'C' at the nucleotide position 72, both 'C' as well as 'A' at position 166, 'A' as well as 'C' at position 516. The result suggested AK235 may be a recent hybrid between *G. arboreum* and *G. herbaceum* and therefore contained both the alleles.

The ITS analysis revealed the same topology as AFLP analysis with exception of AK235 (Fig. 6). All the genotypes grouped according to the species resulting in two major clades:

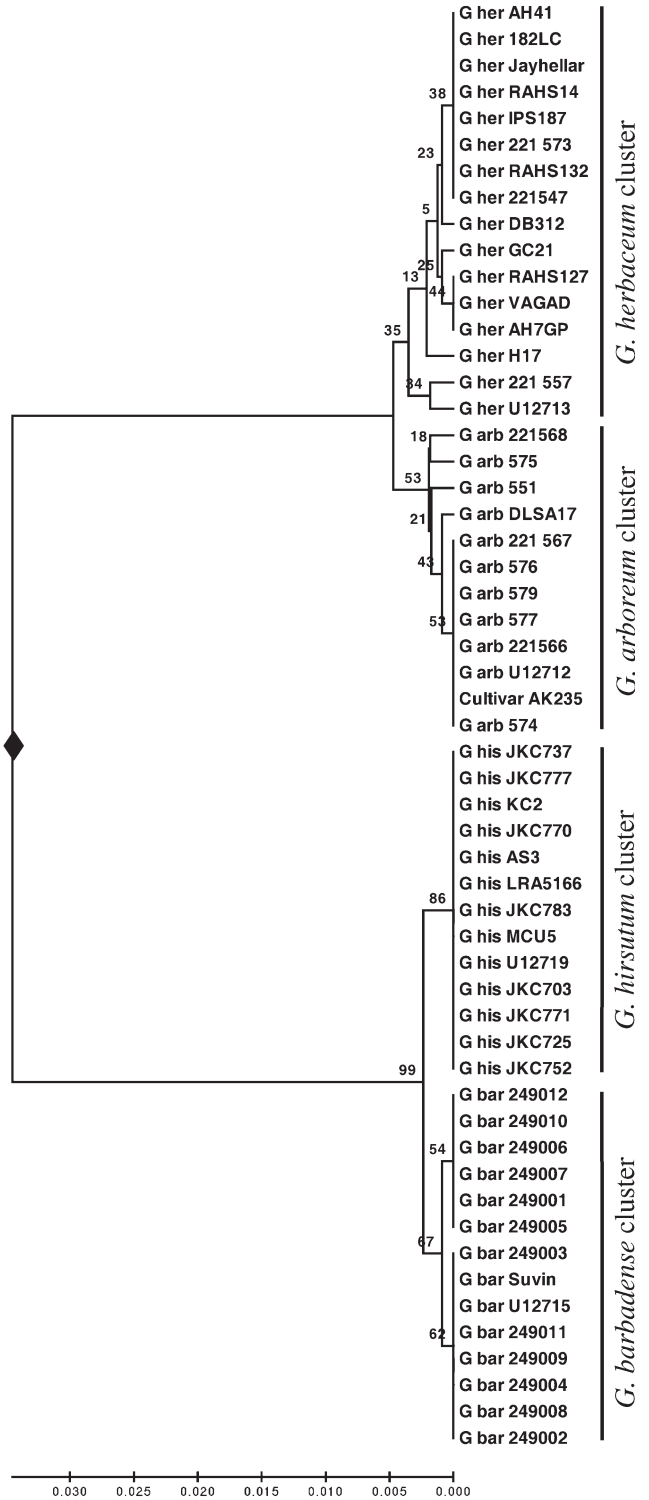

Fig. 6. Dendrogram of 51 cotton genotypes based on the ITS sequences using p-distances; numbers on the branches indicate percentage support in 500 bootstrap replication; the genotype AK235 clustered with *G. arboretum*.

one clade with diploid genotypes and the other clade with tetraploid genotypes. The diploid clade comprised of two subclusters: one with all genotypes of *G. herbaceum* and the

other with *G. arboreum*. Similarly, the tetraploid clade consisted of two subclusters, one with all genotypes of *G. hirsutum* and the other with all genotypes of *G. barbadense* (Fig. 6).

Relationship between GS/GD values based on phenotypic, AFLP, and ITS sequences

Relationships among GS/GD values for all pairs of genotypes based on AFLP markers and phenotypic traits were studied using a Mantel test. The test showed significant correlation (0.80 at $P < 0.001$), suggesting a high level of relationship between GS/GD estimates based on the AFLP marker system and the phenotype. Pairwise distances based on the ITS region and phenotypic traits were studied using a Mantel test. A significant correlation (0.88, at $P < 0.001$) suggested a high level of relationship between the pairwise distance estimates based on the ITS sequences and morphometric parameters in the cotton genotypes.

Genomic SSR amplification for hybridity

The cultivar 'AK235' grouping with *G. herbaceum* in the AFLP cluster analysis was compared with the four other cultivars of *G. herbaceum* and three other cultivars of *G. arboreum* using the genomic SSR markers. For all 12 SSR, AK235 was found to be heterozygous having both types of species-specific alleles (Fig. S2) while other 12 SSR revealed monomorphism among genotypes of *G. herbaceum* and *G. arboreum*.

Discussion

The Indian subcontinent has a long tradition of cultivation of diploid species *G. herbaceum* and *G. arboreum*, which extends back at least 4300 years (Gulatti and Turner 1928). These species have higher adaptability, particularly in the regions of low rainfall like Rajasthan, Gujarat and some part of Maharashtra. In leaf shape and pubescence, the accessions of *G. arboreum* were more similar to *G. herbaceum*, while each accession of *G. arboreum* was distinguished from the *G. herbaceum* in its large boll size, epicalyx nature, presence of fuzz, presence of fold in leaf sinus and many other characteristics. The diploid species, *G. arboreum* includes two subspecies with less prominent morphological characters with respect to epicalyx nature to the collora tube. Throughout the 19th century, there were numerous imports of new world tetraploid cultivars to India, especially *G. hirsutum* and *G. barbadense* from the North and South Americas. More recently, synthetic tetraploids derived after colchicine treatment has been incorporated into the breeding programs in India. There is an amalgam of different types of cultivated cotton in India. Therefore, it is imperative to study the level of genetic diversity and utilise the diversity for cotton improvement in our long-term goal of association and linkage mapping for various traits.

Level of polymorphism

Our results show that AFLP analysis is an efficient tool to detect significant and informative genomic polymorphism and relationship in cotton. It worked effectively across distantly related 51 genotypes selected in this study. The banding pattern obtained with AFLP was highly reproducible. A high percentage of markers were found polymorphic (72.5%) in the collection of the 51 Indian diploid and tetraploid cotton genotypes. Comparison to our study, Rana *et al*. (2005) found higher percentage of polymorphic AFLP markers (85.9%) among 24 Indian advanced breeding lines of *G. hirsutum* using six AFLP primer pairs. It could be possible as they might have used the diverse set of advanced breeding lines developed by themselves. However, in other studies the percentage of polymorphic markers was lower. Abdalla *et al*. (2001) studied the genetic diversity and phenotypic relationship among a collection of 26 USA accessions of diploid and tetraploid cottons and reported a total of 31% AFLP markers polymorphic. Hussein *et al*. (2002) reported 56% polymorphic AFLP markers among 12 cultivars of *G. barbadense* and one of *G. hirsutum* using six AFLP primer combinations. Adawy *et al*. (2006) found 50.4% polymorphic AFLP markers in 21 cotton genotypes using 16 AFLP primer combinations.

Several markers attribute like PIC, MI and RP have been used to assess the informativeness or discriminatory power of the primer pair combinations for genetic diversity studies (Varshney *et al*. 2007). PIC has been discussed extensively as a preferred parameter (Gupta and Varshney 2000; Varshney *et al*. 2007). In this study, the average PIC value of the AFLP markers was 0.43. As mentioned in the Results, only one AFLP primer combination (E-acg/M-cag) with higher PIC value (0.49) was detected in 60–80% of the accessions. The PIC value (0.49) for the E-acg/M-cag primer combination was the best for germplasm analysis on the four studied species. Li *et al*. (2008) reported a lower average PIC (0.18) in 71 glandless cotton genotypes using 10 AFLP markers. High PIC values ranging from 0.37 to 0.57 were reported in Tanzanian cotton cultivars (Lukonge *et al*. 2007).

In order to assess the discriminatory power of the AFLP primer pair combinations, Prevost and Wilkinson (1999) introduced the concept of RP. The RP values in this study ranged from 11.9 to 57.3, with the average of 37.3. Prevost and Wilkinson (1999) and Fernandez *et al*. (2002) reported a strong linear relationship between the ability of a primer combination to distinguish accessions and the RP. The primer combination E-aca/M-cat with the highest RP value (57.3) should be the most informative primer combination for distinguishing accessions of the four cotton species. We did not find any report with respect to RP of AFLP primer pair in cotton species. In comparison to cotton, the RP of Jatropha ranged from 23.11 to 46.82 with an average of 35.21 (Tatikonda *et al*. 2009). In our study, the dendrogram based on the genotyping data of AFLP analysis for the single primer (E-aca/M-cat) combination was comparable to the dendrogram based on genotypic data for all the 16 primer combinations (data not given). Thus a single marker with high resolving power can give results as good as a set of several markers with low RP.

Our results demonstrate that the biases due to fragment homoplasy can be rather high for the estimates of allele frequency and low estimates of genetic diversity (Fig. 2, Table 2). These biases depend not only on technical aspects that can be modified during the assay (number of loci scored per primer and fragment size), but also on factors usually unknown or the demographic history and structure of populations under study. Our simulations indicate that the bias associated with

size homoplasy rapidly increases with the number of true loci. Vos *et al*. (1995) recommended that, to avoid size homoplasy, the number of fragments or bands in an AFLP profile must be between 50 and 100, whereas Gort *et al*. (2006) suggested that even band numbers as low as 20 are not a guarantee of the absence of band homoplasy. These results are in line with an empirical study on the sugar beet by Hansen *et al*. (1999), who scored 456 bands in 16 AFLP lanes, giving an average of 28.5 bands per lane. They reported that 13.2% of the bands were likely to be homoplasious. In our study, the comparison between the biases associated with the estimation of population parameters with several true loci ranging from 8 to 146 indicates that ~69 experimental fragments per primer combination may represent a little upper limit to fragment homology among bands. From a literature review of 90 AFLP surveys on diverse animal, plant, fungi, and bacteria species (in the period 2004–06), it was found that the average number of fragments detected per primer combination ranges from 4 (Enjalbert *et al*. 2005) to 453 (Kalita and Malek 2006). If we take the median value of 58.8 as the more common number of detected fragments per primer combination, biases associated with the estimation of population parameters such as heterozygosity, F_{st} and allele frequency are almost negligible in most studies.

Genetic relationship

The grouping based on the AFLP analysis assigned the genotypes of four species into groups corresponding to their origin, cultivation history and/or pedigree relationships. The tetraploid Upland cotton (*G. hirsutum*) and sea-island (*G. barbadense*) cotton cultivars were nicely grouped into a major clade of AD-genome, which supports the previous report (Abdalla *et al*. 2001). In spite of the similarity at several loci in the genome, there are two subclusters under the major AD-genome clade: one contains all Upland cotton genotypes and the other comprises all sea-island cotton genotypes. In our PCA (Fig. 3) analysis, the scattered *G. hirsutum* and *G. barbadense* genotypes were grouped into two separate bunches, one above the *x*-axis and the other below the *x*-axis of the PCA. Furthermore, manual analysis of individual locus showed 60% allele sharing between the two allotetraploid cotton species. There was, however, a lower (20%) extent of allele sharing between the two tetraploid cotton and the diploid cotton species. The results suggest the need to study genotype-trait relationship in the diploid (AA) species since the A-genome appears to have evolved independently subsequent to polyploidisation. In case of the A-genome diploid species (Fig. 1), all the genotypes grouped into two subclusters of *G. herbaceum* and *G. arboreum* as also in the case of the morphometric grouping. However, the cultivar 'AK235' was an exception, since it grouped with *G. herbaceum* in AFLP analysis though it was identified as close relatives of *G. arboreum* in the morphometric features. This suggests the evolution of 'AK235' as a result of genetic exchange between *G. arboreum* and *G. herbaceum.* This genotype was obtained from Dharward Agriculture University, where the diploid *G. herbaceum* is widely cultivated. The AFLP-based resemblance of AK235 with *G. herbaceum* suggests a major recent contribution of *G. herbaceum* in its pedigree. In PCA (Fig. 3) the genotypes of *G. herbaceum* were scattered and skewed towards *G. arboreum*.

The pattern of divergence in nuclear DNA sequences was described for most diploid cotton species using nuclear rDNA (Cronn *et al*. 1996; Seelanan *et al*. 1997). In general, these studies showed that the variation among species within genome groups is limited and the divergence between species from different genome groups far exceeds the variation within genome groups.

Pairwise inter-genotype distances showed large variation. Within *G. arboreum*, the distance between the two genotypes (182 LC and RAHS127) was negligible, whereas the distance between the two genotypes in *G. herbaceum* (RAHS127 and 221 567) was substantial. Within *G. barbadense*, the distance between the two genotypes (249003 and 249008) was also substantial. This demonstrates that the differentiation between *G. herbaceum* and *G. barbadense* contributed most to the inter-variety variability in AMOVA analysis.

ITS sequence analysis and species relationship

The length of ITS ranges in the present study from 605 to 731 bases. Comparable sequence length (668 bp) was reported for ITS in *Gossypium* (Wendel *et al*. 1995*a*, 1995*b*). Our analysis based on nuclear ITS (Fig. S1) gave topology similar to that in the AFLP-based dendrogram. In ITS analysis, AK235 grouped with *G. arboreum* as in case of the morphometric grouping, rather than with *G. herbaceum* as in case of AFLP analysis. The ITS region of AK235 showed the sequence resembling *G. arboreum* rather than *G. herbaceum*. Some of the sequence features of ITS1 showed the presence of alleles for both *G. arboreum* and *G. herbaceum*, suggesting a recent genetic exchange. The nDNA (ITS) in cotton has been suggested to have evolved 2.5-fold faster than the cpDNA (Cronn *et al*. 2002). The ITS showed a high proportion of variable sites in our study also. However, not all the variable sites are informative phylogenetically. High homoplasy was noticed in the ITS regions of the A- and D-genomes in the tetraploid (AADD) cotton species in spite of sufficiently distorted base composition with high substitution rate. The variable sites revealed a nearly identical base composition as reported previously (Cronn *et al*. 2002). Although the ITS sequences of the nDNA clearly differentiated the two genomes (A- and AD-genomes), it is difficult to explain the hybrids in the two genomes.

Comparison of AFLP and ITS analysis

The dendrograms obtained from AFLP and ITS analysis generally agree with each other except in the case of AK235. Both the analyses revealed a major cluster with all tetraploid (*G. hirsutum* and *G. barbadense*) genotypes. The diploid genotypes of *G. herbaceum* and *G. arboreum* are separated into two subclusters in both the analyses. The nITS from diploid and polyploid *Gossypium* species have previously been shown to exhibit unexpected patterns of sequence evolution and apparent sequence chimera formation in the suspected diploid hybrids (Wendel *et al*. 1995*a*). It is suspected that AK235 collected in our study is a natural hybrid as revealed

by the 24 gSSR analysis. Genotypically it resembles *G. herbaceum*, but phenotypically it is closer to *G. arboreum.*

Comparison of the AFLP with SSR polymorphism shows that the AFLP markers were more efficient in resolving the genotypic diversity than the SSR markers. Being co-dominant and locus-specific, the SSR markers did not yield a large number of alleles in the closely related genotypes. However, co-dominant markers like SSR are important for applications in breeding to identify locus specific alleles. Higher polymorphism by AFLP rather than the SSR markers has been reported earlier (Wendel *et al*. 1992; Wendel and Doyle 1998; Abdalla *et al*. 2001; Iqbal *et al*. 2001; Guo *et al*. 2003; Rana *et al*. 2005; Lacape *et al*. 2007; Kantartzi *et al*. 2009).

Genetic diversity and linkage disequilibrium

Low level of genetic diversity in *G. hirsutum* has been reported previously (Wendel *et al*. 1992; Tatineni *et al*. 1996; Iqbal *et al*. 1997, 2001) and is comparable to our results. However, in the case of the diploid cotton, the diversity level is higher, which suggests that more attention is required to select the diploid cultivars for genes/QTL for various traits like drought tolerance and insect resistance. The cultivars like Vagad, RAHS 14, IPS187 221 557, Jayhellar could be used in different crosses for genetic mapping of *G. herbaceum*. Cultivars like 551, DLSA17, 221 566 are promising for use parents in an inter-specific cross for QTL mapping. Some cultivars of *G. hirsutum* like LRA5166, AS3, and MCU5 showed very little diversity, which could be used as parental lines for developing mapping populations in QTL analysis for fibre quality.

Our study revealed the influence of the population structure and the polymorphism of the assessed loci on the detected levels of LD. Many research articles on LD emphasise that population stratification with unequal distribution of alleles among the groups can cause spurious associations leading to the elevated levels of LD (Flint-Garcia *et al*. 2003; Kraakman *et al*. 2004). However, as shown in our data, in cultivated cotton, the increase of LD (higher values of r^2, and higher number of loci pairs with $r^2 > 0.05$) can also occur in a subpopulation with narrow molecular diversity as compared with a highly-structured population. In *G. arboreum* accessions with high polymorphic loci, the highest levels of LD were noticed based on r^2, whereas the extent of LD remained similar. This could happen due to non-random distribution of haplotypes at the genomic level, which may have happened in Indian diploid cotton due to strong selection pressure. Hence, evaluation of LD should be performed in a uniform set of samples showing no population structure with selected highly polymorphic markers.

Conclusions

The present study reports a large number of reliable and reproducible fingerprint profiles for the 51 accessions from different parts of India. Analysis of different marker parameters established that RP is a preferable parameter for assessing the discriminatory power of AFLP. Among the four species, higher genetic diversity was noticed in *G. herbaceum* followed by *G. arboreum*. The diversity in the accessions of tetraploid cotton examined in this study was less than that in the diploid species. It may be because the tetraploid cotton was introduced in India from America in the 18th and 19th century. The high diversity in *G. herbaceum* and *G. arboreum* may be the result of selection pressure of domestication during the long history of mixed cultivation in India. Cultivars like Vagad, RAHS 14, IPS 187, 221 557, and Jayhellar could be used in different crosses for genetic mapping of *G. herbaceum*. In addition to these, cultivars like 551, DLSA17, 221 566 are highly promising for use as the parents in inter-specific crosses for QTL mapping. Some cultivars of *G. hirsutum* like LRA5166, AS3, and MCU5 showed very little diversity, which could be used in parental lines for raising mapping populations for QTL analysis of fibre quality. Unique and rare fragments identified in different accessions would be useful in breeding programs. The ITS is useful for unambiguous species and subspecies identification. A genome-wide scan with AFLP loci in cotton allowed the detection of LD with significant values ($P < 0.001$) of r^2 >0.5 scattered over all chromosomes in the whole set of species and in a subpopulation of *G. herbaceum*, *G. arboreum*, *G. hirsutum* and *G. barbadense*.

Supporting information

This manuscript is supported by additional information given in four supporting tables: Table S1 gives details of Jaccard's similarity among the 51 genotypes of the four species of cotton; Table S2 gives PCOORD: Eigenvalue, percentage of total variation and cumulative percentage of total contribution; Table S3 gives mean Jaccard similarity among and within species and among and within ploidy level; Table S4 gives details of pairwise squared Euclidean distance matrix of the genotypes; Table S5 gives the species and ploidy-specific markers revealed by AFLP analysis; Table S6 gives details of analysis in ITS sequence in four species of cotton; Table S7 gives details of delta *K* of the Evanno statistics; Fig. S1 gives aligned sequences of ITS1, 5.8S and ITS2 for four species of cotton and Fig. S2 gives genomic profiles of four *G. herbaceum* and three *G. arboreum* and one suspected hybrid revealed by the SSR NBRI_gSSR B010.

Acknowledgements

We thank Dr Michael Baum, ICARDA and Dr A. B. Das, OUAT, Orissa for critically reading the manuscript, Dr Sreekant Somanagouda Patil, UAS, Dharward for providing some of the cotton cultivars, Dr S. K. Bag, NBRI for the statistical analysis and Prof. P. K. Gupta for valuable comments. This work was supported by the Council of Scientific and Industrial Research (CSIR), New Delhi [SIP 005].

References

Abdalla AM, Reddy OUK, El-Zik KM, Pepper AE (2001) Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. *Theoretical and Applied Genetics* **102**, 222–229. doi:10.1007/s001220051639

Adawy SS, Assem SK, Ebtissam Hussein HA, Hanaiya AE (2006) Molecular characterization and genetic relationship among cotton genotypes, 2-AFLP analysis. *Arab Journal of Biotechnology* **9**, 478–492.

Anonymous (2006) Project coordinator's report. All India Coordinated Cotton Improvement Project. CICR, Regional Station, Coimbatore.

Beasley JO (1940) The origin of American tetraploid *Gossypium* species. *American Naturalist* **74**, 285–286. doi:10.1086/280895

Beasley JO (1942) Meiotic chromosome behavior in species, species hybrids, haploids and induced polyploids of *Gossypium*. *Genetics* **27**, 25–54.

Bouajila A, Abang MM, Haouas S, Rezgui SUS, Baum M, Yahyaoui A (2007) Genetic diversity of *Rhynchosporium secalis* in Tunisia as revealed by pathotype, AFLP, and microsatellite analyses. *Mycopathologia* **163**, 281–294. doi:10.1007/s11046-007-9012-0

Brubaker CL, Bourland FM, Wendel JF (1999) The origin and domestication of cotton. In 'Cotton, origin, history, technology and production'. (Eds WC Smith, T Cothren) pp. 3–31. (John Wiley and Sons: New York)

Cronn RC, Zhao X, 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. doi:10.1007/BF02338802

Cronn RC, Small RL, Haselkorn T, Wendel JF (2002) Rapid diversification of cotton genus (*Gossypium*, Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany* **89**, 707–725. doi:10.3732/ajb.89.4.707

DeVerno LL, Mosseler A (1997) Genetic variation in red pine (*Pinus resinosa*) revealed by RAPD and RAPD–RFLP analysis. *Canadian Journal of Forest Research* **27**, 1316–1320. doi:10.1139/x97-090

Endrizzi JE, Turcotte EL, Kohel RJ (1985) Genetics, cytogenetics, and evolution of *Gossypium*. *Advances in Genetics* **23**, 271–375. doi:10.1016/S0065-2660(08)60515-5

Enjalbert J, Duan X, Leconte M, Hovmoller MS, de Vallavielle-pope C (2005) Genetic evidence of local adaptation of wheat yellow rust (*Puccina striiformis* f. sp. *tritici*) within France. *Molecular Ecology* **14**, 2065–2073. doi:10.1111/j.1365-294X.2005.02566.x

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier L (2006) Neandertal genetic diversity, a fresh look from old samples. *Current Biology* **16**, R650–R652. doi:10.1016/j.cub.2006.07.035

Excoffier L, Smouse P (1994) Using allele frequencies and geographic subdivision to reconstruct gene genealogies within a species. Molecular variance parsimony. *Genetics* **136**, 343–359.

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes, application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart N, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**, 220–227. doi:10.1101/gr.10.2.220

Felsenstein J (1985) Confidence limits on phylogenies, an approach using the bootstrap. *Evolution* **39**, 783–791. doi:10.2307/2408678

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (version 3.2). *Cladistics* **5**, 164–166.

Fernandez M, Figueiras A, Benito C (2002) The use of ISSR and RAPD markers for detecting DNA polymorphism, genotype identification and genetic diversity among barley cultivars with known origin. *Theoretical Applied and Genetics* **104**, 845–851.

Flint-Garcia SA, Thornsberry JM, Buckler ESIY (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357–374. doi:10.1146/annurev.arplant.54.031902.134907

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

Ghislain M, Zhang D, Fajardo D, Huaman Z, Hijmans RJ (1999) Marker-assisted sampling of the cultivated Andean potato *Solanum phureja* collection using RAPD markers. *Genetic Resources and Crop Evolution* **46**, 547–555. doi:10.1023/A:1008724007888

Gort G, Koopman WJM, Stein A (2006) Fragment length distributions and collision probabilities for AFLP markers. *Biometrics* **62**, 1107–1115. doi:10.1111/j.1541-0420.2006.00613.x

Gulatti AM, Turner AJ (1928) A note on the early history of cotton. Indian Central Cotton Committee, Technical Laboratory Bulletin No. 17.

Guo WZ, Wang K, Zhang TZ (2003) A and D genome evolution in *Gossypium* revealed using SSR molecular markers. *Acta Genetica Sinca* **30**, 183–188.

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* **113**, 163–185. doi:10.1023/A:1003910819967

Hansen M, Kraft T, Christiansen M, Nilsson NO (1999) Evaluation of AFLP in *Beta*. *Theoretical and Applied Genetics* **98**, 845–852. doi:10.1007/s001220051143

Hartl DL, Clark AG (1989) 'Principles of population genetics.' 2nd edn (Sinauer Associates: Sunderland, MA)

Hussein EHA, Al-Said MSh, El-Itriby HA, Madkour MA (2002) Genotyping Egyptian cotton varities (*G. barbadense*) using molecular markers. In 'Biotechnology and Sustainable Development: Voices of the South and North Conference'. Alexandria, Egypt, 16–20 March. Poster.

Hutchinson JB, Silow RA, Stephens SG (1947) 'The evolution of *Gossypium* and the differentiation of the cultivated cottons.' 1st edn. (Oxford University Press: London)

Iqbal MJ, Aziz N, Saeed NA, Zafar Y, Malik KA (1997) Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. *Theoretical and Applied Genetics* **94**, 139–144. doi:10.1007/s001220050392

Iqbal MJ, Reddy OUK, El-Zik KM, Pepper AE (2001) A genetic bottleneck in the 'evolution under domestication' of upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting. *Theoretical and Applied Genetics* **103**, 547–554. doi:10.1007/PL00002908

Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Société Vaudoise des Sciences Naturelles* **44**, 223–270.

Jacob HJ, Lindpaintner K, Lincoln SE, Kusumi K, Bunker RK, Mao YP, Ganten D, Dzau VJ, Lander ES (1991) Genetic mapping of a gene causing hypertensive rat. *Cell* **67**, 213–224. doi:10.1016/0092-8674(91)90584-L

Jena S, Sahoo P, Mohanty S, Das AB (2004) Identification of RAPD markers, *in situ* DNA content and structural chromosomal diversity in some legumes of the mangrove flora of Orissa. *Genetica* **122**, 217–226. doi:10.1007/s10709-004-2040-5

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, van de Weil C, Bredemeijer G, Vosman B, Mattes M, Daly A, Brettschneider R, Bettini P, Buiatti M, Maestri E, Malcevschi A, Marmiroli N, Aert R, Volckaert G, Rueda J, Linacero R, Vazquez A, Karp A (1997) Reproducibility testing of RAPD, AFLP, and SSR markers in plants by a network of European laboratories. *Molecular Breeding* **3**, 381–390. doi:10.1023/A:1009612517139

Kalita M, Malek W (2006) Application of the AFLP method to differentiate *Genista tinctoria* microsymbionts. *Journal of General and Applied Microbiology* **52**, 321–328. doi:10.2323/jgam.52.321

Kantartzi SK, Ulloa M, Sacks E, Stewart JM (2009) Assessing genetic diversity in *Gossypium arboreum* L. cultivars using genomic and EST-derived microsatellites. *Genetica* **136**, 141–147. doi:10.1007/s10709-008-9327-x

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Kraakman ATW, Niks RE, Berg PMMM, Stam P, Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* **168**, 435–446. doi:10.1534/genetics.104.026831

Lacape JM, Dessauw D, Rajab M, Noyer JL, Hau B (2007) Microsatellite diversity in tetraploid *Gossypium* germplasm, assembling a highly informative genotyping set of cotton SSRs. *Molecular Breeding* **19**, 45–58. doi:10.1007/s11032-006-9042-1

Lewontin RC (1972) The apportionment of human diversity. *Evolutionary Biology* **6**, 381–398.

Li Z, Wang X, Yan Z, Guiyin Z, Wu L, Jina C, Ma Z (2008) Assessment of genetic diversity in glandless cotton germplasm resources by using agronomic traits and molecular markers. *Frontiers of Agriculture in China* **2**, 245–252.

Lukonge E, Herselman L, Labuschangne M (2007) Analysis of genetic diversity in cotton (*Gossypium hirsutum* L.) varities using amplified fragment length polymorphism (AFLP) markers. In 'The World Cotton Research Conference-4'. Lubbock, Texas, USA. 10–14 Sept.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Masum Akond ASMG, Watanabe N, Furuta Y (2008) Comparative genetic diversity of *Triticum aestivum–Triticum polonicum* introgression lines with long glume and *Triticum petropavlovskyi* by AFLP-based assessment. *Genetic Resources and Crop Evolution* **55**, 133–141. doi:10.1007/s10722-007-9221-x

Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. *Trends Ecology Evolution* **14**, 389–394.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321–3323.

Nei N, Li W (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 5269–5273. doi:10.1073/pnas.76.10.5269

Nybom N, Bartish I (2000) Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. *Perspectives in Plant Ecology, Evolution and Systematics* **3**, 93–114. doi:10.1078/1433-8319-00006

Pillay M, Myers GO (1999) Genetic diversity assessed by variation in ribosomal RNA genes and AFLP markers. *Crop Science* **39**, 1881–1886. doi:10.2135/cropsci1999.3961881x

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* **2**, 225–238. doi:10.1007/BF00564200

Prevost A, Wilkinson MJ (1999) A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. *Theoretical and Applied Genetics* **98**, 107–112. doi:10.1007/s001220051046

Pritchard K, Stephens M, Donnelly PJ (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rafalski JA, Vogel JM, Morgante M, Powell W, Andre C, Tingey SV (1996) Generating and using DNA markers in plants. In 'Non-mammalian genomic analysis, a practical guide'. (Eds B Birren, E Lai) pp. 75–134. (Academic Press: London)

Rana MK, Singh VP, Bhat KV (2005) Assessment of genetic diversity in Upland cotton (*Gossypium hirsutum* L.) breeding lines by using amplified fragment length polymorphism (AFLP) markers and morphological characteristics. *Genetic Resources and Crop Evolution* **52**, 989–997. doi:10.1007/s10722-003-6113-6

Rieseberg LH, Noyes RD (1998) Genetic map-based studies of reticulate evolution in plants. *Trends in Plant Science* **3**, 254–259. doi:10.1016/S1360-1385(98)01249-7

Rohlf FJ (1998) 'NTSYS-pc, numerical taxonomy and multivariate analysis system. Version 2.0.' (Exeter Software: Setauket)

Saghai-Maroof MA, Biyashev RM, Yang GP, Zhang Q, Allard RW (1994) Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations and population dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 5466–5470. doi:10.1073/pnas.91.12.5466

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

Schneider S, Roessli D, Excoffier L (2000) 'Arlequin, for population genetics data analysis.' (Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva: Geneva)

Seelanan T, Schnabel A, Wendel JF (1997) Congruence and consensus in the cotton tribe. *Systematic Botany* **22**, 259–290. doi:10.2307/2419457

Shannon CE, Weaver W (1949) 'The mathematical theory of information.' (University of Illinois Press: Urbana, IL)

Sharma S, Raina SN (2005) Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. *Cytogenetic and Genome Research* **109**, 15–26. doi:10.1159/000082377

Somers DJ, Banks T, DePauw R, Fox S, Clarke J (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. *Genome* **50**, 557–567.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4, Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596–1599. doi:10.1093/molbev/msm092

Tatikonda L, Wani SP, Kannan S, Beerelli N, Sreedevi TK, Hoisington DA, Prathibha Devi P, Varshney RK (2009) AFLP-based molecular characterization of an elite germplasm collection of *Jatropha curcas* L., a biofuel plant. *Plant Science* **176**, 505–513. doi:10.1016/j.plantsci.2009.01.006

Tatineni V, Cantrell RG, Davis DD (1996) Genetic diversity in elite cotton germplasm determined by morphological characteristics and RAPD. *Crop Science* **36**, 186–192. doi:10.2135/cropsci1996.0011183X003600010033x

Tautz D, Renz M (1984) Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* **12**, 4127–4138. doi:10.1093/nar/12.10.4127

Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Science* **173**, 638–649. doi:10.1016/j.plantsci.2007.08.010

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP, a new technique for DNA finger printing. *Nucleic Acids Research* **23**, 4407–4414. doi:10.1093/nar/23.21.4407

Wendel JF, Doyle JJ (1998) Phylogenetic incongruence, window into genome history and molecular evolution. In 'Molecular systematics of plants II. DNA sequencing'. (Eds DE Soltis, PS Soltis, JJ Doyle) pp. 265–296. (Kluwer Academic: Boston)

Wendel JF, Brubaker CL, Percival AE (1992) Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. *American Journal of Botany* **79**, 1291–1310. doi:10.2307/2445058

Wendel JF, Schnabel A, Seelanan T (1995*a*) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences of the United States of America* **92**, 280–284. doi:10.1073/pnas.92.1.280

Wendel JF, Schnabel A, Seelanan T (1995*b*) An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Molecular Phylogenetics and Evolution* **4**, 298–313. doi:10.1006/mpev.1995.1027

Xiao M, Li Q, Guo L, Luo T, Duan WX, He WX, Wang L, Chen F (2006) AFLP analysis of genetic diversity of the endangered species *Sinopodophyllum hexandrum* in the Tibetan region of Sichuan province. *China Biochemistry Genetics* **44**, 47–60.

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany* **129**, 157.