INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
18–1856/2019/22–4–669–676
DOI: 10.17957/IJAB/15.1114
http://www.fspublishers.org

Full Length Article

Genetic Diversity and Population Structure Analysis in Upland Cotton Germplasm

Imtiaz Ali[1], Naqib Ullah Khan[1*], Samrin Gul[1], Shahid Ullah Khan[2], Zarina Bibi[3], Kashif Aslam[4], Ghulam Shabir[4], Hafiz Abdul Haq[5], Sher Aslam Khan[6], Ijaz Hussain[6], Sheraz Ahmed[1] and Ajmalud Din[1]
[1]*Department of Plant Breeding and Genetics, The University of Agriculture, Peshawar, Pakistan*
[2]*Institute of Biotechnology and Genetic Engineering (IBGE), The University of Agriculture, Peshawar, Pakistan*
[3]*Department of Soil Science, Faculty of Agriculture, Gomal University, Dera.Ismail Khan, Pakistan*
[4]*Institute of Molecular Biology and Biotechnology, Bahauddin Zakarya University, Multan, Pakistan*
[5]*PCCC Central Cotton Research Institute (CCRI), Multan, Pakistan*
[6]*Department of Agricultural Sciences, University of Haripur, Haripur, Pakistan*
[*]For correspondence: nukmarwat@yahoo.com

Abstract

Elite plant breeding programs could likely benefit from the unexploited standing genetic variation of obsolete genotypes without the yield drag typically associated with wild accessions. Twenty-eight upland cotton genotypes were characterized using 100 SSR (simple sequence repeat) markers. However, only 22 (out of 100) SSR markers were polymorphic and justified further analysis. Major allele frequency ranged from 0.29 (MGHES-20) to 0.93 (MGHES-15) with an average of 0.54. The average gene diversity was 0.57 varying from 0.13 (MGHES-15) to 0.78 (MGHES-20). Polymorphism information content (PIC) values ranged from 0.12 (MGHES-15) to 0.75 (BNL-3280) with an average of 0.53. Phylogenetic analysis supported the subgroups identified by STRUCTURE. Average genetic distance between genotypes was 0.57 indicating low levels of genetic diversity in upland cotton germplasm pool. Results from both phylogenetic tree and population structure analyses were in agreement with pedigree information; however, there were few exceptions like NIBGE-115, NN-3 and NIBGE-2472 which showed a little admixture in structural analysis and not in the phylogenetic tree. Further, core sets of different sizes representing different levels of allelic richness in upland cotton were identified. Establishment of genetic diversity and population structure from this study could be useful for genetic and genomic analysis and systematic utilization of the standing genetic variation in upland cotton. © 2019 Friends Science Publishers

Keywords: Genetic diversity; Phylogenetic tree; Population structure; SSR markers; Polymorphism information content (PIC); *Gossypium hirsutum* L

Introduction

Gossypium hirsutum L., also known as upland cotton, represents 90% of the global cotton fiber production. *Gossypium barbadense* L. known as Pima and Egyptian cotton is valued for its higher and longer fibers and contributes some share to the global cotton production. The other three tetraploid species (*G. mustelinum* L., *G. darwinii* L.and *G. tomentosum* L.) are wild and are not grown commercially (Lacape *et al.*, 2007; D'Eeckenbrugge and Lacape, 2014). Hybridization between A-genome (old world cotton) and D-genome (new world cotton) diploids and subsequent polyploidization about 1.5 million years ago created the five AD allotetraploid lineages belonging to the primary gene pool that are indigenous to America and Hawaii (Adams *et al.*, 2004). These new world allotetraploid cottons includes the commercially important species *i.e.*, *G. hirsutum* L. and *G. barbadense* L. which are extensively cultivated worldwide (Abdurakhmonov, 2007; Campbell *et al.*, 2010). One of the most important events in US cotton breeding history was the introduction of Mexican highland stocks in the early 19th century, which laid the foundation of current upland germplasm. However, within species, the *G. hirsutum* shows great phenotypic diversity than the other three cultivated cotton species (Abdurakhmonov *et al.*, 2012; Ali *et al.*, 2017).

With a heightened risk of genetic vulnerability to disease epidemics and climate change, the elite breeding programs could benefit from the unexploited standing genetic variation of old cultivars without yield losses which is typically associated with wild accessions. It is also noted that even within the domesticated upland cotton, unfavorable agronomic effects were observed when un-adapted germplasm from a different area is used in a breeding program (Rana *et al.*, 2005). The most effective

To cite this paper: Ali, I., N.U. Khan, S. Gul, S.U. Khan, Z. Bibi, K. Aslam, G. Shabir, H.A. Haq, S.A. Khan, I. Hussain, S. Ahmed and A. Din, 2019. Genetic diversity and population structure analysis in upland cotton germplasm. *Intl. J. Agric. Biol.*, 22: 669–676

utilization of the genetic diversity of *Gossypium* requires modern genomic technologies that help to reveal the molecular basis of agronomically important genetic variations (Dahab *et al*., 2013). By characterizing genetic diversity between and within groups, breeding efforts can be greatly improved through better parental selection for generating segregating populations.

Genetic diversity information is also helpful to identify heterotic groups, understand population structure, and identify a core set of lines for studying genetic analysis. Thus, assessment of genetic diversity and population structure are important in the upland cotton. Genetic diversity among cultivated plants is of high value in biodiversity due to their wide range of contributions to the world economy and principal position in the production of important crops (Ditta *et al*., 2018). Genetic diversity estimates have been made using pedigree and morphological data, biochemical markers and DNA-based molecular markers (Yu *et al*., 2012). In the pedigree-based studies, estimate of genetic relatedness between two accessions depends on the availability of breeding records and validity of certain assumptions. In the absence of such information, pedigree-based methods cannot be used to accurately estimate genetic diversity. This is especially true of ancestral lines or introductions, for which detailed breeding records are not available. Genetic diversity of cotton cultivars can be effectively evaluated by molecular markers and the study provides useful information on the selection of parental genotypes in the development of cotton cultivars and hybrids (Wu *et al*., 2006; Ullah *et al*., 2012).

Molecular markers, on the other hand, are more reliable and informative which can directly measure the allelic diversity and provide robust estimates of genetic distances. The most effective utilization of the genetic diversity of *Gossypium* requires modern genomic technologies that help to reveal the molecular basis of agronomically important genetic variations (Dahab *et al*., 2013). A multitude of DNA-based marker systems including restriction fragment length polymorphism (RFLP) (Becelaere *et al*., 2005), random amplified polymorphic DNA (RAPD) (Rahman *et al*., 2008; Ali *et al*., 2018), amplified fragment length polymorphism (AFLP) (Abdalla *et al*., 2001), simple sequence repeat (SSR) (Zhang *et al*., 2011), and inter-simple sequence repeat (ISSR) (Liu and Wendel, 2001) markers were used for measuring genetic diversity in cotton. The recent development of plentiful cotton SSR markers has had a more positive effect on the molecular characterization of the cotton germplasm released from specific cotton breeding programs across the world (Lacape *et al*., 2007; Zhang *et al*., 2008). The SSRs have proven to be a best marker system due to their co-dominant nature, reproducibility, multiallelic and PCR-based (Preetha and Raveendren, 2008; Yonca *et al.*, 2011). Therefore, a comprehensive study involving a broad collection of germplasm and more efficient genotyping platforms is still needed to quantify overall genetic diversity in upland cotton for its effective utilization in breeding, genetic, and genomic studies. The objective of this study was to estimate the genetic diversity through phylogenetic tree and population structure analysis by using SSR markers in upland cotton germplasm.

Materials and Methods

For the identification of genetic diversity in 28 upland cotton genotypes at different locations of Khyber Pakhtunkhwa and Punjab, Simple Sequence Repeat (SSR) analysis was performed (Table 1). A total of 100 SSR markers located on different chromosomes were surveyed for DNA polymorphism. DNA was extracted from all the cotton genotypes and polymerase chain reaction (PCR) and gel electrophoresis were carried out.

DNA Extraction

The delinted seeds of the 28 upland cotton genotypes were sown in disposable glasses filled with sand in glass house of National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan. After germination and when the plants reached to 3–4 leaves stage the young leaves from each genotype were cut and stored in a freezer for DNA extraction.

In CTAB method, the DNA was extracted from 2–3 days old seedlings leaves (Iqbal *et al*., 1997). Water bath was turned and set at 65°C to heat 2 x CTAB with 1% 2-mercapthanol. Pestle and mortar were autoclaved first and then pre-cooled with liquid nitrogen. Four to five young leaves were cut and grinded to a very fine powder in liquid nitrogen. This powder was then transferred to a 50 mL falcon tube. Fifteen mL of hot (65°C) 2 × CTAB was added to the tube, mixed gently and incubate at 65°C for half an hour. After half an hour, 15 mL of chloroform/ isoamylalcohol (24:1) was added and mix gently to form an emulsion. Mixture was centrifuged for 10 min at 9000 rpm. Supernatant solution was transferred to a new 50 mL falcon tube, whereas, the remaining chloroform phase was discarded. This step was repeated twice as to ensure the complete digestion of various cell components and phenolic compounds. To precipitate the DNA, 0.6 volumes of chilled 2-propanol was added to the supernatant and then centrifuged at 9000 rpm for five min. The supernatant was discarded. The pellet was washed thrice with 70% ethanol and air-dried. The pellets were re-suspended in 0.5 mL 0.1 × TE buffer. The suspension was transferred into an eppendorf tube (1.5 mL) and then 5 µL of RNAs was added to digest all the RNAs incubating for one hour at 37°C. After it, equal volume of chloroform/isoamylalcohol (24:1) was added and mixed gently and, centrifuged for 10 min at 13000 rpm in a microcentrifuge. The supernatant was transferred to a new eppendorf tube and 1/10th volume of 3*M* NaCl solution was added to supernatant and mixed gently. DNA was precipitated with chilled absolute ethanol

Table 1: Pedigree of 28 upland cotton genotypes used in the studies

Genotypes	Parentage	Breeding center	Released / under Approval
IR-NIBGE-901	PGMB-33/FH-90	NIBGE, Faisalabad, Pakistan	2011
IR-NIBGE-1524-4	PGMB-33/NIBGE-2	-do-	2010
IR-NIBGE-3	PGMB-33/FH-100	-do-	2012
IR-NIBGE-4	PGMB-33/CIM-448	-do-	2011
IR-NIBGE-5	PGMB-33/CIM496	-do-	Under approval
IR 3300-24	PGMB-33/BH-160	-do-	Under approval
IR 3300-13	PGMB-33/BH-160	-do-	Under approval
NIBGE-115	S-12/LRA-5166	-do-	2012
NN-3	S-12/LRA-5166	-do-	Under approval
NIBGE-2472	S-12/LRA-5166	-do-	Germplasm
NIBGE-2	LRA/S-12	-do-	2006
IR-2379	PGMB-33/FH-100	-do-	Germplasm
IR-NIBGE-3701-38	PGMB-33/CIM-448	-do-	2010
IR 1526	PGMB-33/NIBGE-2	-do-	Germplasm
NIBGE-314	S-12/LRA	-do-	Under approval
NIBGE-5	S-12/LRA	-do-	Germplasm
NIBGE-4	S-12/ CIM-448	-do-	Germplasm
IR NIBGE-2620	IR-901/Rajhans	-do-	Germplasm
NIBGE 758-8	S-12/ CIM-448	-do-	Germplasm
IR NIBGE-3701-33-6	PGMB-33/CIM-448	-do-	2010
SLH-284	-	CRS, Sahiwal, Pakistan	Under approval
CIM-446	CP 15/2 × S 12	CCRI, Multan, Pakistan	1998
CIM-473	CIM-402 × LRA 5166	-do-	2002
CIM-496	CIM-425 × 755-6/93	-do-	2005
CIM-499	CIM-433 × 755-6/93	-do-	2003
CIM-506	CIM-360 × CP 15/2	-do-	2004
CIM-554	2579-04/97 × W-1103	-do-	2009
CIM-707	CIM-243 × 738-6/93	-do-	2004

(2 volumes), spinned at 13000 rpm for 10 min, supernatant was discarded and pellets were washed with 70% ethanol. Pellets were air dried, re-suspended in 0.1 × TE buffer and quantified.

A total of 20 μL volume was used for polymerase chain reactions (PCR) using 15 ng of cotton DNA, 10 X buffer, 25 m*M* $MgCl_{2,}$ Primer-F 30 ng/μl, Primer-R 30 ng/µl, Taq polymerase 5 U/μl and deoxy-nucleotide triphosphates 2.5 mM. The amplificantoin profile consisted of initial period of denarturation at 94°C for 5 min, followed by cycle (step-1) of 94°C for 30 s, 50°C for 30 s annealing, 72°C extension for 1 min. The PCR amplifications were followed by incubation at 72°C for 10 min. DNA quantification was carried using the NanoDrop® ND-1000. Quality of DNA was observed by running 50 ng DNA in 0.8% agarose gel. The DNA samples giving smear in the gel were rejected. Moreover, the quantity of DNA was also confirmed by comparing with Quantification Standards Phage λ DNA (Gibco BRL) in 0.8% gel. Dilutions of 15 ng/μl were made from stock solutions. The dilutions were also checked by comparing them with the DNA quantification standards in agarose gel. PCR reaction was carried out in eppendorf master cycler gradient, Germany. To confirm that observed bands amplified from genomic DNA, and not primer artifact, genomic DNA was omitted from control reaction. No amplification products were detected without genomic DNA in any PCR.

Genetic Markers

For present study, the 100 SSR markers were provided by Plant Genomics and Molecular Breeding (PGMB) Laboratory, NIBGE, Faisalabad, Pakistan. These markers were selected on the basis of their reproducible nature, PCR based, highly polymorphic, small quantity of genomic DNA requirement, easy interpretation in genotyping and easy automation. However, only 22 (out of 100) SSR markers were polymorphic and justified further analysis.

Agarose Gel Electrophoresis

After PCR amplification, the concentration of amplicons was determined on 1.2% agarose gel stained with ethidium bromide. Then loading concentrations for agarose-based gel electrophoresis (PAGE) was made according to the brightness of bands on 2% agarose gel. All the PCR products were loaded into the wells by using of pipette. The gel was loaded at room temperature of 15 degrees while immersed in 1 x Tris/Boric acid / EDTA (TBA) buffer. Gels were run at 80 volts. Under these conditions the PCR products usually was separated after 80 min. Voltage gradient can be raised as high as 16 volts/cm to shorten time and improve band resolution. After the run was complete, the gel was moved into a large UV illuminator and photographed.

Scoring of Data

Amplification profiles of different cotton varieties was compared with each other and bands of DNA fragments was scored as present (1) or absent (0). The data was used to estimate the similarity based on number of shared amplification products (Nei and Li, 1979). A phylogenetic tree based on similarity coefficient was generated using the Un-weighted Pair Group Method of Arithmetic means (UPGMA).

Power Marker

The power Marker v. 3.25 was used for measuring the genetic diversity among various cotton genotypes.

Polymorphism Information Content

Polymorphism information content (PIC) value provides the information about the polymorphism of a primer.

STRUCTURE Analysis

Main aim of association mapping was to find out the markers which have association with QTLs controlling yield and yield related traits. Population structure is a central part in association mapping analysis because it can lessen type-1 error between molecular markers and traits of interest in self-pollinated species (Yu and Buckler, 2006). False positive is the major issue in association analysis. There are different approaches to reduce the effect of false positive. For population structure analysis, the most frequently used methods are implemented in the software STRUCTURE v. 2.3.1 developed by Pritchard *et al*. (2000a). The number of populations denoted by K while Delta-K values determines the sub-populations for K-ranging. Ten runs were conducted for each value of number of populations (*K*), with *K* ranging from 2 to 12. The length burn-in and number of replications were 10,000 each. Accessions were assigned to subgroup if the probability of membership was greater than 70% (Liu *et al*., 2003). If membership was <70 %, then the accessions were assigned to the mixed subgroup.

Results

Twenty-eight upland cotton genotypes were characterized using 100 SSR markers. Out of these, only 22 markers were polymorphic, 65 markers were monomorphic and 13 SSRs were not amplified (Table 2). However, these 100 SSR markers were identified and distributed on 17 chromosomes of the said cotton genotypes. Major allele frequency was ranging from 0.29 (MGHES-20) to 0.93 (MGHES-15) with an average of 0.54. The average gene diversity was 0.57 varying from 0.13 (MGHES-15) to 0.78 (MGHES-20). Polymorphism information content (PIC) values ranged from 0.12 (MGHES-15) to 0.75 (BNL-3280) with an average of 0.53. The number of alleles, average number of alleles, gene diversity and PIC values are provided in Table 2.

Table 2: Genetic diversity among 28 upland cotton genotypes based on 22 (out of 100) microsatellite markers

SSR Markers	Major Allele Frequency	Allele number	Gene Diversity	PIC
MGHES-3	0.6429	4.0000	0.5434	0.5063
MGHES-15	0.9286	2.0000	0.1327	0.1239
BNL-1667	0.3571	5.0000	0.7219	0.6715
MGHES-22	0.5000	2.0000	0.5000	0.3750
BNL-1673	0.7500	3.0000	0.4005	0.3586
BNL-4108	0.3214	6.0000	0.7449	0.7011
BNL-3254	0.3929	4.0000	0.7117	0.6606
MGHES-18	0.3929	3.0000	0.6607	0.5868
MGHES-20	0.2857	5.0000	0.7781	0.7420
MGHES-53	0.4643	4.0000	0.5944	0.5101
MGHES-55	0.7143	5.0000	0.4668	0.4431
MGHES-60	0.5357	6.0000	0.6556	0.6207
MGHES-63	0.7143	3.0000	0.4464	0.4014
MGHES67	0.4286	5.0000	0.6862	0.6303
MGHES-75	0.6786	4.0000	0.4974	0.4580
BNL-3886	0.8214	2.0000	0.2934	0.2503
BNL-1066	0.3929	4.0000	0.6709	0.6048
BNL-2449	0.7143	5.0000	0.4668	0.4431
BNL-3280	0.4286	10.0000	0.7679	0.7483
BNL-2495	0.5000	4.0000	0.6301	0.5666
BNL-3255	0.4643	7.0000	0.7015	0.6630
MGHES-14	0.6071	2.0000	0.4770	0.3633
Means	0.5471	4.3182	0.5704	0.5195

Table 3: Genetic distances of different groups from phylogenetic analysis

Groups	Range	Average	Genotypes
G1	0.046 ------ 0.1419	0.0912	6
G2	0.092 ------- 0.1494	0.1149	6
G3	0.0805 ----- 0.3103	0.1509	6
G4	0.0575 ----- 0.1954	0.1369	6
G5	0.1379 ------ 0.2299	0.1589	4
Overall	0.0460 - -----0.3563	0.1932	28

Phylogenetic Analysis

Neighbor-joining based phylogenetic tree based on genotypic data for upland cotton genotypes was constructed through software 'POWER MARKER'. The genetic distance among genotypes was estimated through un-weighted pair group method of arithmetic means. On average, for genetic distance the genotypes ranged from 0.0460 to 0.3563. Genotypes IR-NIBGE-1524-4 and CIM-446 owned maximum genetic distance (0.3563) (Table 3). However, minimum genetic distance (0.0460) was observed among genotypes IR-2379, IR-1526 and IR-2379 and NIBGE-2.

Based on phylogenetic tree, the cotton genotypes were divided mainly into two groups and further sub-divided into five groups (Fig. 1). In group first, six genotypes were observed with genetic distance ranging from 0.046 to 0.0912 with mean genetic distance of 0.0912. Maximum genetic distance (0.1494) was observed among genotypes *i.e*., NIBGE-5 and NIBGE-2, NIBGE-5 and IR-NIBGE-3701-38, while minimum genetic distance (0.046) was observed between genotypes *viz*., NIBGE-2 and IR-2379,

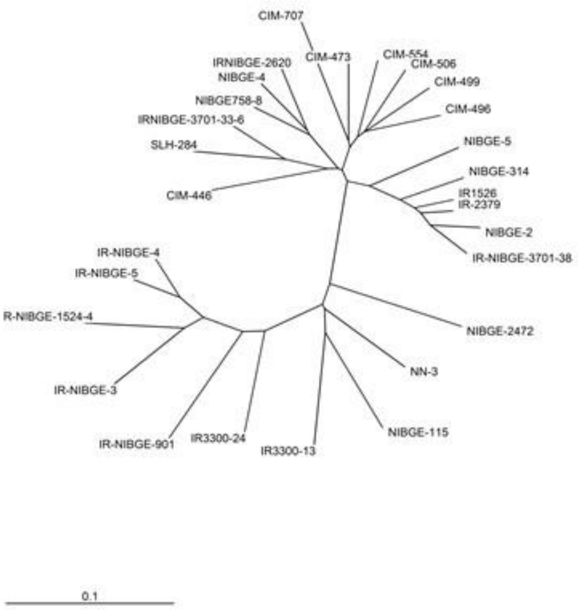

Fig. 1: Neighbor-joining clustering of the 28 cotton genotypes based on 22 (out of 100) microsatellite markers

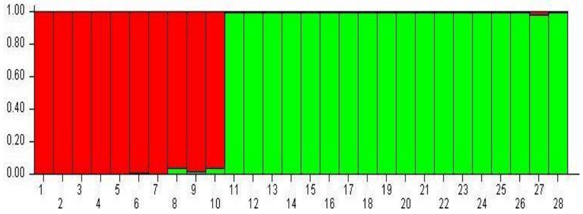

Fig. 2: Q-plot showing clustering of 28 upland cotton genotypes based on analysis of genotypic data using STRUCTURE. Each genotype is represented by a vertical bar. The colored subsections within each vertical bar indicate membership coefficient (Q) of the genotype to different clusters. Identified subgroups are group 1 (red color) and group 2 (green color)

Legends 1 = IR-NIBGE-901; 2 = IR-NIBGE-1524-4; 3 = IR-NIBGE-3; 4 = IR-NIBGE-4; 5 = IR-NIBGE-5; 6 = IR 3300-24; 7 = IR 3300-13; 8 = NIBGE-115; 9 = NN-3; 10 = NIBGE-2472; 11 = NIBGE-2; 12 = IR-2379; 13 = IR-NIBGE-3701-38; 14 = IR 1526; 15 = NIBGE-314; 16 = NIBGE-5; 17 = NIBGE-4; 18 = IR NIBGE-2620; 19 = NIBGE 758-8; 20 = IR NIBGE-3701-33-6; 21 = SLH-284; 22 = CIM-446; 23 = CIM-473; 24 = CIM-496; 25 = CIM-499; 26 = CIM-506; 27 = CIM-554; 28 = CIM-707

IR-2379 and IR-1526. In group second, six genotypes were found with genetic distance of 0.092 to 0.1494 with an average genetic distance of 0.115. Maximum genetic distance (0.1494) was observed among genotypes CIM-707, CIM-499 and CIM-496. Minimum genetic distance (0.092) was observed among various groups of genotypes CIM-554 and CIM-506, CIM-506 and CIM-499, CIM-506 and CIM-496, CIM-499 and CIM-496. The third group consists of six genotypes ranging from 0.0805 to 0.3103 with mean genetic distance of 0.151. Maximum genetic distance (0.3103) was recorded among genotypes NIBGE-4, IR-NIBGE-3701-33-6 and CIM-446. However, Minimum genetic distance (0.805) was observed among genotypes NIBGE-758, IR-NIBGE-2620 and NIBGE-4. Group fourth was having six genotypes ranged from 0.0575 to 0.1954 with mean genetic distance of 0.1368. Maximum genetic distance (0.1954) was recorded between IR-NIBGE-1524-4 and IR-3300-4, while minimum (0.0575) was observed between genotypes IR-NIBGE-5 and IR-NIBGE-4. In group five, four genotypes were observed with genetic distance ranging of 0.1379 to 0.1589 with average value of 0.1589. Maximum genetic distance was observed between genotypes NIBGE-2472 and NIBGE-115 (0.2299) while minimum genetic distance (0.1379) was recorded between genotypes NIBGE-115 and IR-3300-13 (Table 3).

Population Structure

The selected 87 (minus not amplified) SSR markers were used to estimate population structure of 28 cotton cultivars. An admixture model assumes that individuals may also have inherited a fraction of their genome from its ancestors in a different subpopulation, and thus having a mixed ancestry. The main objective of the population structure analysis was to avoid false positives during marker trait associations. In software options, a burn in length of 30,000 iterations and run length of 30,000 durations were used to test the K value in the range of 1–28. Structure analysis revealed that all the cotton genotypes were divided into two main sub-groups *i.e*., G1 (1–10) and G2 (11–28) (Fig. 2). Based on K value, all the studied cotton genotypes were also divided into two major groups (Fig. 3).

Discussion

In present study, the estimated PIC value (0.53) was in the range of average PIC values reported previously for cotton SSRs *i.e*., 0.122 to 0.80 (Abdurakhmonov *et al*., 2008; Zhang *et al*., 2011). In these upland cotton genotypes, lesser alleles per locus with low PIC values were noted. Lower number of alleles per locus and low PIC values in upland cotton, as also observed in the current study, further substantiated the previous reports on narrow genetic base in cotton (Campbell *et al*., 2009). Some other studies also reported narrow genetic base in upland cotton genotypes (Zhang *et al*., 2005; Ahmad *et al*., 2007; Ali *et al*., 2017, 2018). According to Khan *et al*. (2009), Indo-Pak and Pakistani cotton cultivars released since 1914–2005 had narrow genetic base. Rahman *et al*. (2008) also reported narrow genetic diversity among elite cotton cultivars in Pakistan. However, before the incidence of cotton leaf curl virus (CLCuV) and its epidemic expression in 1991/92, the genetic diversity was high among the cotton genotypes. It has been estimated that CLCuV epidemic caused a loss of 4.98 million bales of cotton with an estimated value of US$7.4 billion (Khan *et al*., 2009). After this heavy loss in cotton, resistance against CLCuV got much attention and breeders were concerned about the need of CLCuV resistant

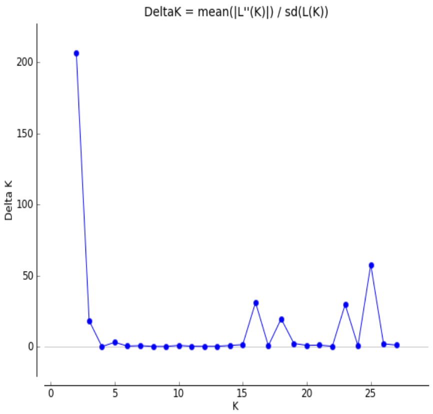

Fig. 3: Estimating number of sub-populations using delta *K* values for *K* ranging from 2 to 28 using method proposed by Evanno *et al*. (2005)

cultivars in cotton. As a result, a major shift in cotton breeding priorities was seen. Majority of the breeding programs focused on development of CLCuV resistant cultivars, however, fewer resistant sources were available in this regard. The germplasm used in the current study has been developed after that epidemic era. Therefore, the cotton genotypes developed after CLCuV epidemic had narrow genetic base which also authenticated by the present findings. Extensive use of closely related cultivars in cotton breeding has resulted in narrowing the genetic base. the present study manifested narrow genetic base for most of the genotypes released after 2000. This might be due to rigorous reuse of the available cotton genotypes in the breeding programs (Haidar *et al*., 2012).

In several past studies conducted on different cotton genotypes, similar kind of low genetic diversity was also reported in upland cotton (Zhang *et al*., 2005; Shaheen *et al*., 2010; Fang *et al*., 2013). However, the said estimates of genetic diversity might be higher because the monomorphic data of SSRs loci was excluded in the present study. Most of the cotton genotypes in mixed group were located between major clusters in the neighbor-joining tree. Results further revealed that there was a good agreement between present study and pedigree information. However, for some genotypes there were discrepancies between pedigree information and marker-based relationships. The discrepancies might be due to the mutation in the available cotton genotypes or may be using same lines again and again in breeding programs by local breeders. Past observations also reported discrepancies between cotton pedigree information and genetic relationships based on SSR markers (Iqbal *et al*., 2001; Ahmad *et al*., 2007; Fang *et al*., 2013).

According to phylogenetic tree, the genetic diversity within each group was lower. However, the present observations suggested that cultivars developed through different breeding programs might be suitable to specific geographic locations. This also could explain that in spite of narrow genetic base in upland cotton, breeders were able to develop improved cotton cultivars through breeding procedure. Thus, the present results could help breeders to identify and select appropriate genotypes for enhancement of seed cotton yield through different breeding programs and conservation of genetic diversity. Irrespective of originating institutions, division of upland cotton into two major groups was observed for some genotypes in phylogentic tree. This might be due the same breeding material being used in different institutions. Past studies revealed that genetic diversity in 48 rice accessions had divided them into two main groups and then into sub-groups irrespective to their originating stations (Aslam and Arif, 2014; Ali *et al*., 2018). Buyyarapu *et al*. (2011) also observed phylogenetic tree with four major sub-clusters for 23 species, while three species branched out individually in upland cotton genotypes.

According to Pritchard *et al*. (2000b), the software "STRUCTURE 2.3.1" was used to determine the population structure of all the cotton genotypes before marker trait association analysis. The ideal number of clusters (K) were found through method of Evanno *et al*. (2005) using online program "Structure Harvester" (Yu *et al*., 2006). This value reached a plateau when minimum number of groups that best describes the population structure has been reached (Pritchard *et al*., 2000a, b; Evanno *et al*., 2005). Population structure analysis identified two different sub-populations in upland cotton genotypes which were studied in this experiment across different locations. Twenty-eight genotypes were assigned to mixed group indicating little admixture among the studied genotypes. This admixture is possibly a result of closely related germplasm being shared among different breeding programs. Admixture was also observed between the two *Gossypium* species. Such admixture between *G. hirsutum* and *G. barbadense* is expected since introgressions from *G. barbadense* has been used for cultivar development. An admixture model assumes that individuals may also have inherited a fraction of their genome from its ancestors in a different subpopulation, thus having a mixed ancestry. Another reason could be the frequent appearance of a few lines with favorable agronomic traits of upland cotton in multiple breeding programs. The clustering of individuals into subpopulations is based on the genotypic data consisting of unlinked markers in upland cotton (Guo *et al*., 2007; Li *et al*., 2008; Khan *et al*., 2010; Paterson *et al*., 2010; Ali *et al*., 2018). Phylogenetic analysis supported the subgroups identified by software 'STRUCTURE 2.3.1' (Tyagi *et al*., 2014).

Conclusion

The present study confirmed that pedigree information was same as revealed by both phylogenetic and population structure analysis in upland cotton germplasm and identified the same groups of the genotypes. Average genetic distance between the genotypes was indicating low levels of genetic diversity in upland cotton germplasm. Present study revealed that establishment of genetic diversity and population structure analysis could be useful for genetic and genomic analysis and systematic utilization of the currently available genetic variation in upland cotton.

Acknowledgements

This research was funded by the Higher Education Commission (HEC), Islamabad - Pakistan. We also pay thanks to the University of Agriculture, Peshawar - Pakistan for administrative support, and the Department of Plant Breeding and Genetics for various assistances throughout the project. The views and opinions expressed in this research manuscript are solely those of the authors and do not represent those of the University of Agriculture, Peshawar - Pakistan and the funding agency.

References

Abdalla, A.M., O.U.K. Reddy, K.M. El-Zik and A.E. Pepper, 2001. Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. *Theor. Appl. Genet*., 102: 222–229

Abdurakhmonov, I.Y., 2007. Exploiting genetic diversity. *In: Proceedings of World Cotton Research Conference*, p: 2153. Lubbock, Texas, USA

Abdurakhmonov, I.Y., Z.T. Buriev, S.E. Shermatov, A.A. Abdullaev, K. Urmonov, F. Kushanov, S.S. Egamberdiev, U. Shapulatov, A. Abdukarimov, S. Saha, J.N. Jenkins, R.J. Kohel, J.Z. Yu, A.E. Pepper, S.P. Kumpatla and M. Ulloa, 2012. Genetic Diversity in G*ossypium* genus. *In: Genetic Diversity in Plants,* pp: 313–338. Caliskan, M. (Ed.). In-TechOpen, London Bridge St., London, UK

Abdurakhmonov, I.Y., R.J. Kohel, J.Z. Yu, A.E. Pepper, A.A. Abdullaev, F.N. Kushanov, L.B. Salakhutdinov, Z.T. Buriev, S. Saha, B.E. Scheffler, J.N. Jenkins and A. Abdukarimov, 2008. Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics*, 92: 478–487

Adams, K.L., R. Percifield and J.F. Wendel, 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics,* 168: 2217–2226

Ahmad, S., T. Zhang, N. Islam, T. Shaheen and M. Rahman, 2007. Identifying genetic variation in *Gossypium* based on single nucleotide polymorphism. *Pak. J. Bot*., 39: 1245–1250

Ali, I., N.U. Khan, M. Rahman, R. Gul, Z. Bibi, S. Gul, S. Ahmed, S. Ali, N. Ali, K. Afridi and H.A. Haq, 2018. Genotype by environment and biplot analyses for yield and fiber traits in upland cotton. *Intl. J. Agric. Biol*., 20: 1979–1990

Ali, I., N.U. Khan, F. Mohammad, M.A. Iqbal, A. Abbas, F. Ullah, Z. Bibi, S. Ali, I.A. Khalil, S. Ahmad and M. Rahman, 2017. Genotype by environment and GGE-biplot analysis for seed cotton yield in upland cotton. *Pak. J. Bot*., 49: 2273–2283

Aslam, K. and M. Arif, 2014. SSR analysis of chromosomes 3 and 7 of rice accessions with grain length. *Pak. J. Bot*., 46: 1363–1372

Becelaere, V.G., E.L. Lubbers, A.H. Paterson and P.W. Chee, 2005. Pedigree *vs.* DNA marker-based genetic similarity estimates in cotton. *Crop Sci*., 45: 2281–2287

Buyyarapu, R., R.V. Kantety, J.Z. Yu, S. Saha and G.C. Sharma, 2011. Development of new candidate gene and EST-based molecular markers for *Gossypium* species. *Intl. J. Plant Genom*., 2011: 1–9

Campbell, B.T., S. Saha, R. Percy, J. Frelichowski, J.N. Jenkins, W. Park, C.D. Mayee, V. Gotmare, D. Dessauw, M. Gband, X. Du, Y. Jia, G. Constable, S. Dillon, I.Y. Abdurakhmonov, A. Abdukarimov, S.M. Rizaeva, A.A. Abdullaev, P.A.V. Barrose, J.G. Padua, L.V. Hoffman and L. Podolnaya, 2010. Status of global cotton germplasm resources. *Crop Sci*., 50: 1161–1179

Campbell, B.T., V.E. Williams and W. Park, 2009. Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica*, 169: 285–301

Dahab, A.A., M. Saeed, B.B. Mohamed, M.A. Ashraf, A.N. Puspito, K.S.BajwaA.A. Shahid and T. Husnain, 2013. Genetic diversity assessment of cotton (*Gossypium hirsutum* L.) genotypes from Pakistan using simple sequence repeat markers. *Aust. J. Crop Sci*.,7: 261–267

D'Eeckenbrugge, G. and J.M. Lacape, 2014. Distribution and differentiation of wild, feral, and cultivated populations of perennial upland cotton (*Gossypium hirsutum* L.) in Mesoamerica and the Caribbean. *PLoS One*, 9: e107458

Ditta, A., Z. Zhou, X. Cai, M. Shehzad, X. Wang, K. Okubazghi and K. Wang, 2018. Genome-wide mining and characterization of SSR markers for gene mapping and gene diversity in *Gossypium barbadense* L. and *Gossypium darwinii* G. *watt* accessions. *Agronomy*, 8: 181–196

Evanno, G., S. Regnaut and J. Goudet, 2005. Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol*., 14: 2611–2620

Fang, D.D., L.L. Hinze, R.G. Percy, P. Li, D. Deng and G. Thyssen, 2013. A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in upland cotton (*G. hirsutum* L.) cultivars from major cotton-growing countries. *Euphytica*, 191: 391–401

Guo, W.Z., C.P. Cai, C.B. Wang and Z.G. Han, 2007. A microsatellite-based, gene-rich linkage map reveals genome structure, function, and evolution in cotton. *Genetics*,176: 527–541

Haidar, S., M. Aslam, M. Hassan, H.M. Hassan and A. Ditta, 2012. Genetic diversity among upland cotton genotypes for different economic traits and response to cotton leaf curl virus (CLCV) disease. *Pak. J. Bot*.,44: 1779–1784

Iqbal, M.J., O.U.K. Reddy, K.M. Elzik and A.E. Pepper, 2001. A genetic bottleneck in the evolution under domestication of upland cotton (*G. hirsutum* L.) examined using DNA fingerprinting. *Theor. Appl. Genet*., 103: 547–554

Iqbal, M.J., N. Aziz, N.A. Saeed, Y. Zafar and K.A. Malik, 1997. Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. *Theor. Appl. Genet*.,94: 139–144

Khan, A.I., F.A. Awan, B. Sadia, R.M. Rana and I.A. Khan, 2010. Genetic diversity studies among colored cotton genotypes by using RAPD markers. *Pak. J. Bot*., 42: 71–77

Khan, A.I., Y. Fu and I. Khan, 2009. Genetic diversity of Pakistani cotton cultivars as revealed by simple sequence repeat markers. *Commun. Biometr. CropSci.*, 4: 21–30

Lacape, J.M., D. Dessauw, M. Rajab, J.L. Noyer and B. Hau, 2007. Microsatellite diversity in tetraploid *Gossypium* germplasm. Assembling a highly informative genotyping set of cotton SSRs. *Mol. Breed*., 19: 45–58

Li, Z., X. Wang, Y. Zhang, G. Zhang, L. Wu, J. Chi and Z. Ma, 2008. Assessment of genetic diversity in glandless cotton germplasm resources by using agronomic traits and molecular markers. *Front. Agric. Chin.*, 2: 245–252

Liu, B. and J.F. Wendel, 2001. Inter-simple sequence repeat (ISSR) polymorphisms as a genetic marker system in cotton. *Mol. Ecol. Notes*, 1: 205–208

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler and J. Doebley, 2003.Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics,* 165:2117–2128

Nei, M. and W.H. Li, 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci*., 76: 5269–5273

Paterson, A.H., J. Rong, M. Rahman and Y. Zafar, 2010. Sequencing and utilization of the *Gossypium* genomes. *Trop. Plant Biol*., 3: 71–74

Preetha, S. and T.S. Raveendren, 2008. Molecular marker technology in cotton. *Biotechnol. Mol. Biol.*, 3: 32–45

Pritchard, J.K., M. Stephens and P. Donnelly, 2000a. Inference of population structure using multilocus genotype data. *Genetics,* 155: 945–959

Pritchard, J.K., M. Stephens, N.A. Rosenberg and P. Donnelly, 2000b. Association mapping in structured populations. *Amer. J. Human Genet*., 67: 170–181

Rahman, M., T. Yasmin, N. Tabbasam, I. Ullah, M. Asif and Y. Zafar, 2008. Studying the extent of genetic diversity among *Gossypium arboreum* L. genotypes/cultivars using DNA fingerprinting. *Genet. Resour. Crop Evol*., 55: 331–339

Rana, M.K., V.P. Singh and K.V. Bhat, 2005. Assessment of genetic diversity in upland cotton (*Gossypium hirsutum* L.) breeding lines by using amplified fragment length polymorphism (AFLP) markers and morphological characteristics. *Genet. Resour. Crop Evol.,* 52: 989–997

Shaheen, T., Y. Zafar and M. Rahman, 2010. Detection of single nucleotide polymorphisms in the conserved ESTs regions of *G. arboreum*. *Electr. J. Biotechnol*., 13: 3–5

Tyagi, P., M.A. Gore, D.T. Bowman, B.T. Campbell, J.A. Udall and V. Kuraparthy, 2014. Genetic diversity and population structure in the US upland cotton (*G. hirsutum* L.). *Theor. Appl. Genet*., 127: 283–295

Ullah, I., A. Iram, M.Z. Iqbal, M. Nawaz, S.M. Hasni and S. Jamil, 2012. Genetic diversity analysis of *Bt* cotton genotypes in Pakistan using simple sequence repeat markers. *Genet. Mol. Res*., 11: 597–605

Wu, Y., A.C. Machado, R.G. White, D.J. Llewellyn and E.S. Dennis, 2006. Expression profiling identifies genes expressed early during lint fiber initiation in cotton. *Plant Cell Physiol*., 47: 107–127

Yonca, S., B. Col and B. Burun, 2011. Genetic diversity and identification of some Turkish cotton genotypes (*Gossypium hirsutum* L.) by RAPD-PCR analysis. *Turk. J. Biol.,* 36: 143–150

Yu, J. and E.S. Buckler, 2006. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol*., 17: 155–160

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi and M. Yamasaki, 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet*., 38: 203–208

Yu, J.Z., D.D. Fang, R.J. Kohel, M. Ulloa, L.L. Hinze, R.G. Percy, J. Zhang, P. Chee, B.E. Scheffler and D.C. Jones, 2012. Development of a core set of SSR markers for the characterization of *Gossypium* germplasm. *Euphytica*, 187: 203–213

Zhang, Y., X.F. Wang, Z.K. Li, G.Y. Zhang and Z.Y. Ma, 2011. Assessing genetic diversity of cotton cultivars using genomic and newly developed expressed sequence tag-derived microsatellite markers. *Genet. Mol. Res*., 10: 1462–1470

Zhang, J., Y. Lu, R.G. Cantrell and E. Hughs, 2005. Molecular marker diversity and field performance in commercial cotton cultivars evaluated in the Southwestern U.S.A. *Crop Sci*., 45: 1483–1490

Zhang, Y., Z. Lin, Q. Xia, M. Zhang and X. Zhang, 2008. Characteristics and analysis of simple sequence repeats in the cotton genome based on a linkage map constructed from a BC1 population between *Gossypium hirsutum* and *G. barbadense*. *Genome*, 51: 534–546

[Received 14 Dec 2018; Accepted 07 May 2019; Published (online) 20 Aug 2019]