


RESEARCH

Open Access



# Genetic diversity and population structure of *Gossypium arboreum* L. collected in China

JIA Yinhua<sup>\*</sup> , PAN Zhaoe, HE Shoupu, GONG Wenfang, GENG Xiaoli, PANG Baoyin, WANG Liru and DU Xiongming<sup>\*</sup>

## Abstract

**Background:** *Gossypium arboreum* is a diploid species cultivated in the Old World. It possesses favorable characters that are valuable for developing superior cotton cultivars.

**Method:** A set of 197 *Gossypium arboreum* accessions were genotyped using 80 genome-wide SSR markers to establish patterns of the genetic diversity and population structure. These accessions were collected from three major *G. arboreum* growing areas in China. A total of 255 alleles across 80 markers were identified in the genetic diversity analysis.

**Results:** Three subgroups were found using the population structure analysis, corresponding to the Yangtze River Valley, North China, and Southwest China zones of *G. arboreum* growing areas in China. Average genetic distance and Polymorphic information content value of *G. arboreum* population were 0.34 and 0.47, respectively, indicating high genetic diversity in the *G. arboreum* germplasm pool. The Phylogenetic analysis results concurred with the subgroups identified by Structure analysis with a few exceptions. Variations among and within three groups were observed to be 13.61% and 86.39%, respectively.

**Conclusion:** The information regarding genetic diversity and population structure from this study is useful for genetic and genomic analysis and systematic utilization of economically important traits in *G. arboreum*.

**Keywords:** *Gossypium arboreum* L., Population structure, Genetic diversity, Genetic differentiation, Simple sequence repeat (SSR) markers

## Background

Cotton is the most important natural fiber crop in the world. It includes approximately 45 diploid ( $2n = 2x = 26$ ) and 5 allotetraploid ( $2n = 4x = 52$ ) species distributed mostly in tropical and subtropical regions throughout the world (Fryxell 1992; Wendel and Albert 1992; Wendel and Cronn 2003). Tetraploid species, including *Gossypium hirsutum* and *Gossypium barbadense*, arose in the New World from inter-specific hybridization between an A genome and a D genome diploid species which are believed to have originated from ancestors

similar to modern *G. herbaceum* race *africanum* and *G. raimondii*, respectively (Stephens 1944; Seelanan et al. 1997; Brubaker et al. 1999; Liu et al. 2001; Chen et al. 2007; Sunilkumar et al. 2006). Diploid species ( $2n = 26$ ) are classified into eight genomic groups (A-G and K), occurring naturally in Africa, Asia, America, and Australia (Wendel and Cronn 2003). Worldwide, four species are cultivated: two of these cultivated species are diploids ( $2n = 2x = 26$ ) and two are allotetraploids ( $2n = 4x = 52$ ).

*Gossypium arboreum* is a diploid species cultivated in the Old World. It was first domesticated near the Indus Valley before 6000 BC (Hutchinson 1954; Fryxell 1979, 1992; Moulherat et al. 2002). The primitive *G. arboreum* was perennial, and was once considered to have evolved from the wild *G. herbaceum* in Africa (Hutchinson

<sup>\*</sup> Correspondence: [jiayinhua\\_0@sina.com](mailto:jiayinhua_0@sina.com); [dujeffrey8848@hotmail.com](mailto:dujeffrey8848@hotmail.com)  
Institute of Cotton Research of Chinese Academy of Agricultural Sciences, State Key Laboratory of Cotton Biology, Key Laboratory of Biology and Genetic Improvement of Cotton, Ministry of Agriculture, Anyang 455000, Henan, China



1954). More recently, Wendel et al. have presented evidence that *G. arboreum* was independently domesticated from a different wild plant that gave rise to *G. herbaceum* (Wendel et al. 1989; Renny-Byfield et al. 2016). *G. arboreum* lost photoperiod sensitivity when it spread from the West India to the North and East India (Hutchinson 1954). The annual types of *G. arboreum* facilitated extension to larger areas and evolved tolerance to diseases, pest and frost (Silow 1944). Furthermore, environmental conditions associated with geographic distribution and domestication resulted in the development of considerable variation, which has been classified into six races, *soundanense*, *indicum*, *burmanicum*, *cernuum*, *bengalense*, and *sinense* in different regions (Silow 1944; Brubaker et al. 1999).

*Gossypium arboreum* was introduced into China from various routes, and was domesticated as a local crop between the 7th and the 13th centuries (Watt 1907; Guo et al. 2006). It was thought that two primary routes of importation were overland from Bengal-Assam to the Yellow River, and by sea from Indo-China to the Yangtze River Valley (Silow 1944). In the south of the Five Ridges area, Hainan Island and Yunan, *G. arboreum* was only grown as garden plants until an extremely early-fruited type which were developed from Indian and Indo-Chinese varieties. After the new technology of weaving was brought to the Yangtze River Valley in the thirteenth century, various landraces were developed and widely cultivated in the area of the middle and lower Yangtze River Valley, then spread to Northern China encouraged by Imperial edict in the fifteenth century (Watt 1907; Silow 1944; Guo et al. 2006). The three major growing regions of *G. arboreum* including the Southwest region, the Yangtze River Valley, and the Northern region were gradually formed with the breeding of local varieties (Guo et al. 2006). Then, the most important type, race *sinense*, was developed in China, until it was completely replaced by Upland cotton (*Gossypium hirsutum* L.) in the 1950s (Huang 1996; Guo et al. 2006).

As the cultivated 'Old World' diploid cotton, *Gossypium arboreum* experienced from natural and artificial selection due to environmental stress, and evolved to possess favorable characters that the tetraploid cultivars lack, such as drought tolerance, disease resistance, and insect pest resistance which makes it well adapted to biotic and abiotic stresses (Kantartzi et al. 2009; Mehetre et al. 2003), spinnable fiber with various colors and high strength that are good for weaving (Park et al. 2005; Mehetre et al. 2003). These *G. arboreum* landraces with adaptive features are important genetic resources for the improvement of tetraploid cotton, and can help to develop cultivars with invaluable genes for early maturity, stress tolerance, and high fiber strength in cotton-breeding programs (Xiang 1988; Rahman et al.

2002; Mehetre et al. 2003; Liu et al. 2006). Understanding the genetic relationships among the landraces of *G. arboreum* would facilitate efficient use for developing superior cotton cultivars with favorable agronomic traits.

## Methods

### Plant material

One hundred and ninety-seven accessions of *Gossypium arboreum* were collected from 19 provinces in China, and were preserved in the Gene Bank of Institute of Cotton Research of Chinese Academy of Agricultural Sciences. These accessions were cultivated in the main cotton growing areas of China including the North region, the Yangtze River Valley and the Southwest region. Their accession numbers and passport data are listed in Additional file 1: Table S1. A panel of 24 accessions were selected to screen the polymorphic microsatellites for the analysis of diversity and structure of the natural population.

### Genotyping with SSR markers

DNA from young and fully expanded leaves of each accession was extracted as described by Paterson and Smith (1999). SSR primers information was obtained from the Cottongen (Cottongen, <https://www.cotton-gen.org/>). PCR is conducted in 10 µL volumes, which included 1.0 µL 10× Buffer (consisting of 20 mmol·L<sup>-1</sup> MgSO<sub>4</sub>, 100 mmol·L<sup>-1</sup> KCl, 80 mmol·L<sup>-1</sup> (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 100 mmol·L<sup>-1</sup> Tris-HCl, pH 9.0, 0.5% NP-40), 50 ng template DNA, 0.5 mmol·L<sup>-1</sup> dNTP, 0.4 units of *Taq* DNA polymerase, 0.5 µmol·L<sup>-1</sup> forward and reverse primers. The amplification program of PCR included a 3 min pre-denaturation step at 95 °C, 30 cycles of 94 °C for 45 s, 57 °C for 45 s, 72 °C for 1 min, and 7 min extension at 72 °C. All reactions were completed using a PTC-100TM thermocycler. The PCR products were stored at 4 °C before being running on the 8% non-denatured PAGE gel (Sambrook et al. 1989). The gel was stained using the method of Zhang et al. (2000), and was photographed using SYNGENE gel system.

### Data collection and analysis

The most intensely amplified band for each SSR locus was scored using a standard 50 base pairs (bp) DNA marker (Takara Biotech, Dalian, China) as reference. Presence of amplified fragments was scored as 1, and the absence was labelled as 0 for the SSR locus. Missing data was represented as "-9". Diversity was calculated based on the genotype data for 80 polymorphic SSRs in 197 individuals. SpaGeDi software was used to calculate allele frequencies (Hardy and Vekemans 2002). The polymorphic information content (PIC) was also estimated using the Powermarker software package version 3.25 (Liu and Muse 2005). Powermarker software

package version 3.25 was used to calculate the genetic distance (GD). Principal coordinate analysis (PCA) was done with NTSYS-pc software version 2.1 in using Dcenter and Eigen functions (Rohlf 2000). Analysis of molecular variance (AMOVA) among and within groups was performed using Arlequin ver 3.5 software (Excoffier and Lischer 2010).

### Population structure analysis

Population structure was estimated using Structure version 2.3.4 (Pritchard et al. 2000) based on co-dominant genotypic data. The number of populations tested was assumed as K from 1 to 10. The length of running time was 100 000 and replication after burning was 10 000 for the STRUCTURE with the admixture model. The second graphs for Pn and  $\Delta K$  (Delta K) were built to find a proper number of K values using the method of Evanno et al. (2005).

## Results

### SSR marker analysis

A total of 116 SSR primer pairs were selected to detect the genotypes of all accessions. Among the SSRs, 24 primers were found to be monomorphic, and 12 primers could not be scored clearly. These 36 SSRs were deleted, leaving 80 SSR primer pairs for analysis. Accessions that missed more than 5% SSR data were also removed. Finally, 197 accessions and 80 SSR primer pairs were used for further analysis. In these accessions, a total of 255 SSR alleles were detected with an average of 3.2 alleles per SSR marker (from 2 to 6 alleles) (Table 1). The number of effective allele varied from 1.1 to 4.8 with an average of 2.3 effective alleles per locus. A summary of marker statistics for all the accessions is presented in Additional file 2: Table S2.

### Population structure

Population structure of the 197 accessions was performed with the software Structure version 2.3.4. The log-likelihood increased with the value of K, but no evidence showed that the number of subpopulations could be identified from the plot of probability for K (Fig. 1a). Then, the plot of  $\Delta K$  was built using the method described by Evanno et al. (2005) (Fig. 1b). A strong signal for the number of clusters was successfully identified to be three based on  $\Delta K$  value. Among all accessions, 128 accessions could be assigned to three different subgroups based on 60% membership threshold, and the

remaining 69 accessions were considered to have admixed parentage. The subgroups were showed as three different colored bar plots that reflected the single ancestral genetic background (Fig. 2). Detailed membership probabilities of all accessions were described in Additional file 1: Table S1.

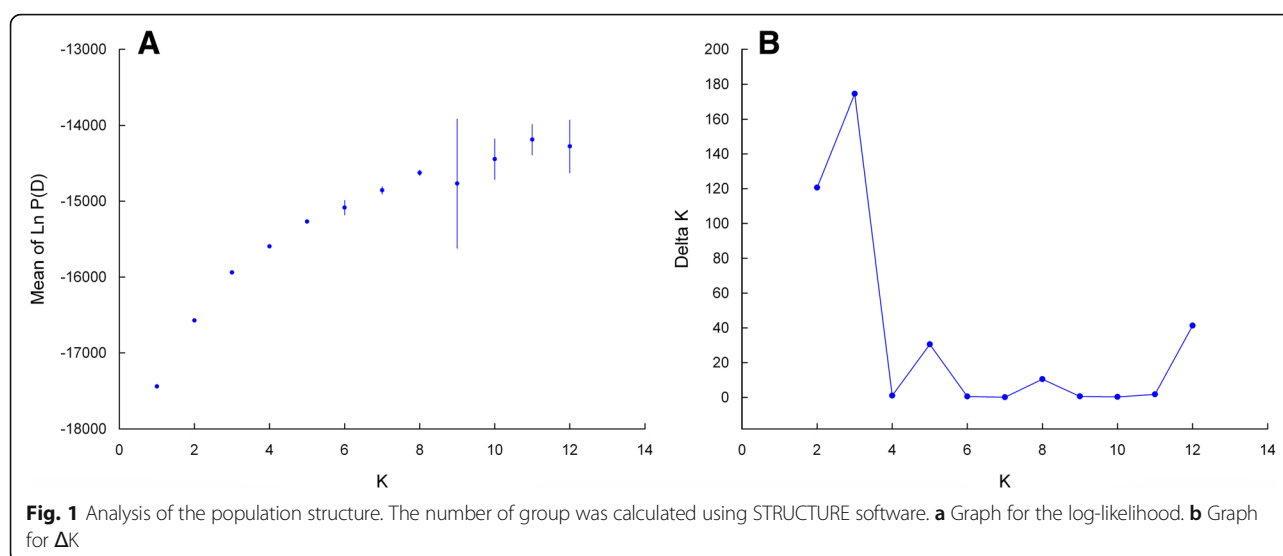
The whole group of *G. arboreum* were separated into three subgroups. These three subgroups consisted of 42, 40 and 46 accessions, and were labeled as Group 1, 2 and 3, respectively. These subgroups were found to correspond to the three traditional *G. arboreum* growing zones in China namely the Yangtze River Valley, North China and Southwest China. Most of the accessions in Group 1 (Additional file 1: Table S1) were from the Yangtze River Valley, except seven accessions from Southwest China and one from North China, which meant that they have been selected and evolved to adapt to the local environment. Accessions in Group 2 (Additional file 1: Table S1) were mainly from North China excluding nine accessions from the Yangtze River Valley and one from Southwest China. Almost all the accessions in Group 3 (Additional file 1: Table S1) were from Southwest China excluding three accessions from the Yangtze River Valley and one from North China. Accessions that have admixed parentage could be found in all the three zones meaning that they had mixed genetic background.

### Genetic diversity

Genetic diversity for all accessions was analyzed using Powermarker software package. The overall PIC value for SSRs ranged from 0.17 to 0.79 with an average of 0.47 (Table 1). The average genetic distance was 0.32 and ranged from 0.02 to 0.55 (Table 1). The highest genetic distance of 0.55 was between accession No.1 named Guichi Xiaozimian Baizi from the Yangtze River Valley and accession No.151 named Donglan Changjing Zhongmian 1 from Southwest China. The lowest genetic distance of 0.02 was between accession No.155 named Changrong Zhongmian from the Yangtze River Valley and accession No.157 named Wangmo Sanglang Da Mianhua from Southwest China. Among the three groups, Group 2 and Group 3 had the highest genetic distance of 0.205 indicating that the accessions from North and Southwest China are genetically far from each other (Table 2). However, Groups 1 and 3 had the lowest genetic distance of 0.196 reflecting the proximity of descent between the two groups (Table 2). The mixed group

**Table 1** Summary of SSR polymorphisms

No. of polymorphic SSRs				Polymorphic information content (PIC)		Genetic distance	
Total alleles	Average allele per marker	Effective allele per marker	Range	Range	Average	Range	Average
255	3.2	2.3	1.1–4.8	0.17–0.79	0.47	0.02–0.55	0.32



was similar and low distance with the three groups, certifying the parentage from the multi-regions. Within the groups, Group 3 had the highest average genetic distance of 0.308, and Group 1 had the lowest average genetic distance of 0.253.

A phylogenetic tree was constructed based on the distance matrix using the Neighbor Joining (N-J) algorithm. In the dendrogram, three major clusters were found (Fig. 3). The dendrogram was also compared with the results of Structure analysis. Three groups identified in Structure were found to be in agreement with the clusters observed in the phylogenetic tree shown with different colored lines (Fig. 3). Most of lines that were grouped in one cluster of phylogenetic tree were found to be from the same group of Structure analysis. Although a few lines were incongruent in the clusters and structure, most of accessions in the clustering pattern were grouped close to their pedigree parents.

Further study of the genetic relationships between the accessions was carried out using Principal coordinate analysis (PCA) (Fig. 4). Accessions from each subgroup identified in Structure were spread out over three-dimensional plane with some overlapping between accessions from different subgroups. The proportions of first two axes of PCA were up to 68.1% of the variation, which indicates that level of genetic diversity between

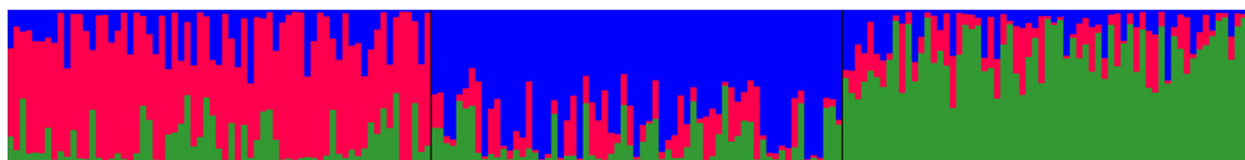
the subgroups was enough high for the identification of possible genes in *G. arboreum* germplasm.

#### Analysis of molecular variance

Analysis of molecular variance (AMOVA) was performed with Arlequin ver 3.5 software. Significant variation between the groups was observed which contributed 13.61% of the total variation ( $P < 0.0001$ ) (Table 3). A larger amount of variation within the groups was found to be 86.39% ( $P < 0.0001$ ) (Table 3). Differentiation estimate ( $F_{ST}$ ) from the genetic structure analyses was 0.137 highly significant at  $P < 0.0001$ , which was concurred with the analysis of molecular variance. Based on pairwise  $F_{ST}$  values, the highest genetic differentiation was observed between Group 1 and Group 2, which revealed that accessions of Group 1 are distinct from accessions of Group 2 ( $P < 0.0001$ ) (Table 4). The lowest genetic differentiation was between Group 2 and Group 3, suggesting that accessions of these two groups are closer to each other.

#### Discussion

In the present study, 80 SSR primer pairs were used to evaluate the diversity of *Gossypium arboreum* accessions. These primer pairs produced 255 alleles in the population of 197 accessions with an average of 3.2 alleles per



**Fig. 2** Population structure: The bar plot of Q-matrix estimates for the accessions: Groups are represented in different colors (Red for Group 1, Blue for Group 2, Green for Group 3)



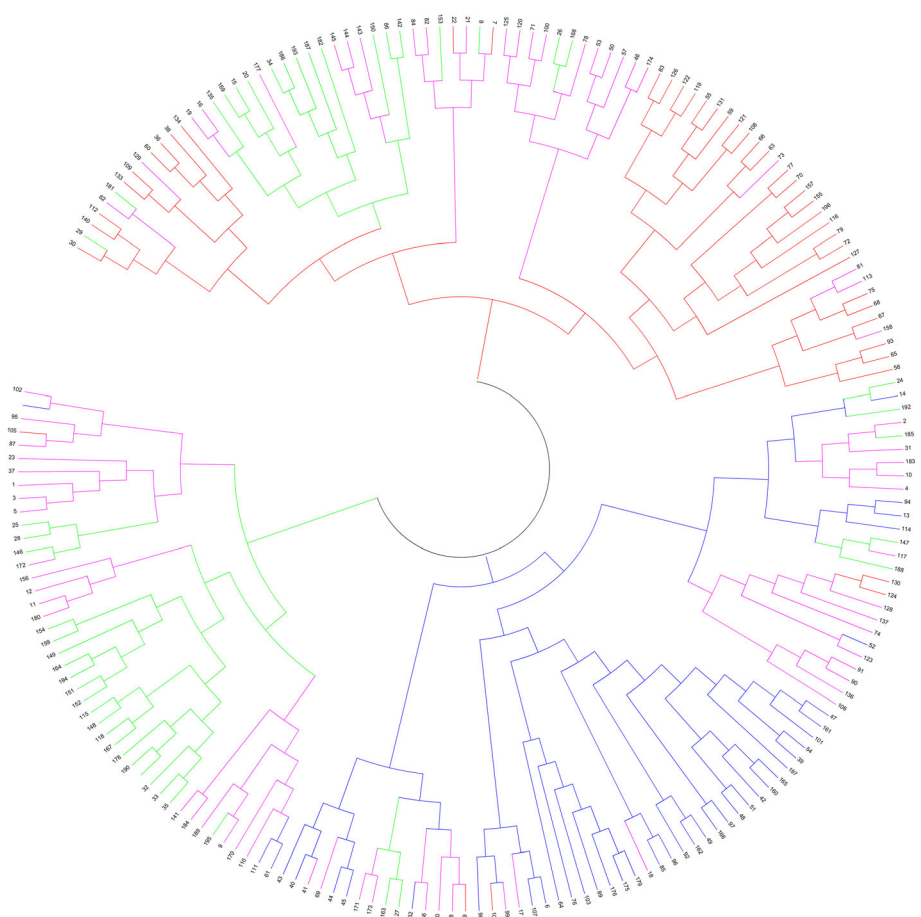
**Table 2** Genetic distance estimates calculated using Nei et al. distance matrix within and between *Gossypium arboreum* groups identified by STRUCTURE analysis

	Group 1	Group 2	Group 3
Group 1	0.253		
Group 2	0.203	0.274	
Group 3	0.196	0.205	0.308

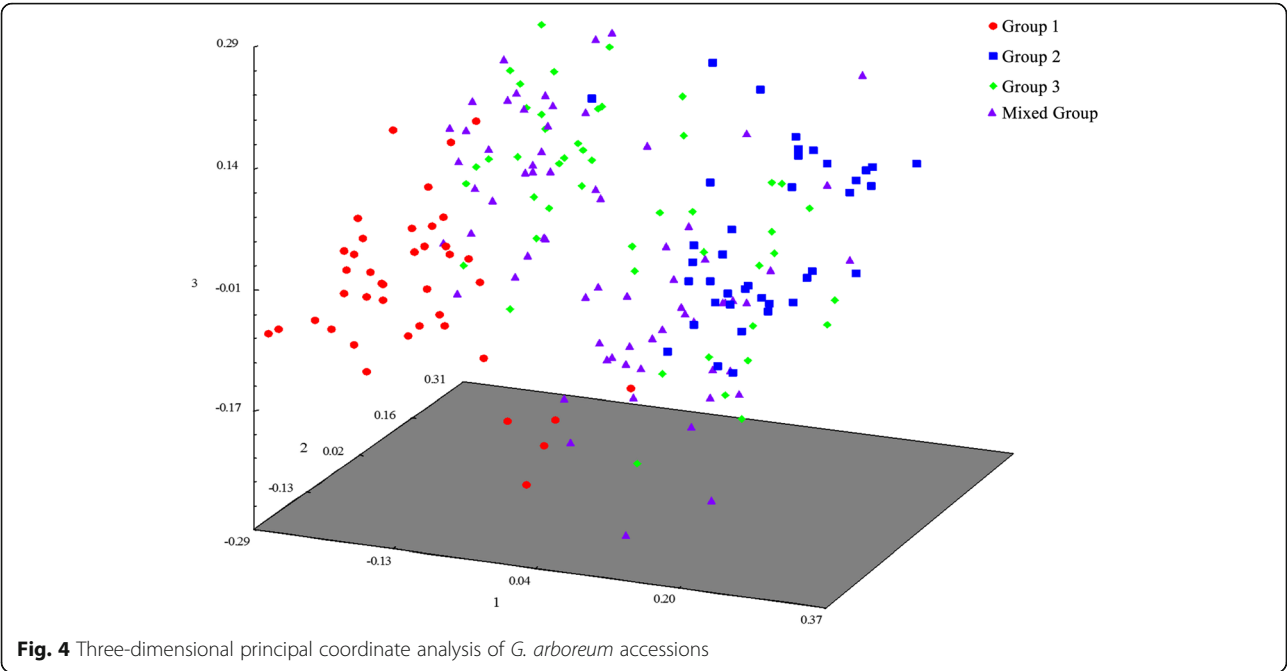
marker. PIC values were found to range from 0.17 to 0.79 with an average of 0.47. Kantartzi et al. (2009) used 115 SSR primers to characterize 96 *G. arboreum* accessions and found 2.40 alleles per locus. The average PIC value in Kantartzi et al. (2009) was 0.42, which is in agreement with the present study. Guo et al. (2006) observed higher number of alleles per locus and a higher value of polymorphism information content with 60 SSR markers and 108 accessions. Liu et al. (2006) reported genetic similarity coefficients ranged from 0.58 to 0.87 with 39 *G. arboreum* accessions. The diversity of each population and alleles observed per marker deeply

depends on markers, germplasm and the platform used for the resolution of amplified products (Lacape et al. 2007). Lower diversity was found in tetraploid type of *G. hirsutum* though it owned more alleles per locus with the average alleles 3.9–4 and average PIC 0.13–0.17 in different *G. hirsutum* populations (Abdurakhmonov et al. 2008; Tyagi et al. 2014; Cai et al. 2014).

Three differentiated subgroups in *G. arboreum* accessions that were identified with the analysis of population structure were congruent with the major *G. arboreum* growing regions in China: the Yangtze River Valley, North China and Southwest China. Mixed group embraced 69 accessions that were spread across the three regions. A few accessions that were classified in one subgroup by structure analysis were found to be from outside the geographical region of accession origin. For example, Group 1 was considered to represent the Yangtze River Valley area, although it included seven accessions from Southwest China and one accession from North China. Group 2 composed of nine accessions from the Yangtze River Valley and one from Southwest



**Fig. 3** Neighbor-joining tree of 197 *Gossypium arboreum* accessions. Colors in the tree correspond to subgroups identified in Structure analysis (Red for Group 1, Blue for Group 2, Green for Group 3, Purple for mixed group)



China. Group 3 had three accessions from the Yangtze River Valley and one accession from North China region. It might be a result of germplasm migrating or gene flow by local people among different regions.

The phylogenetic tree based on the estimates of genetic distance revealed the relationships among the *Gossypium arboreum* accessions. Three subgroups identified by the Structure software were observed in clusters which were presented in three colored lines. Accessions in the mixed group were found within the main clusters in the phylogenetic tree. This result was mostly in accordance with their pedigree, although a few discrepancies were observed between pedigree information and genetic relationships for some accessions. The highest genetic distance between the Group 2 and Group 3 supported the fact of the farthest geographic distance between North China and Southwest China. The lowest genetic distance between Group 1 and Group 3 corroborated the historical fact that *G. arboreum* was imported to the Yangtze River Valley from South China, which then spread over the cotton growing regions of the country. Group 3 was far from Group 1 and Group 2 in both genetics and geography suggesting that *G. arboreum* grew in Southwest China for a

long time before it was introduced to the Yangtze River Valley across the Yunnan-Guizhou Plateau and the Five Ridges Mountains. This result corroborates the studies of previous researchers who reported that *G. arboreum* spread from Southern to Northern China (Guo et al. 2006). However, Silow (1939) found that the lack lintless modifiers type of *G. arboreum* were common in both the Yellow River Valley (belong to North China) and the Yangtze River Valley, and pointed to two primary routes of importation that were overland from Bengal-Assam to the Yellow River basin, and by sea from Indo-China to the Yangtze River Valley. Watt (1907) also thought that cotton was introduced into China several times from various sources and was domesticated into diverse types, including the early-fruited type each time. In the present study, accessions in the mixed group were scattered in the three regions. Accessions from Group 3 overlapped with accessions from Group 1 and Group 2 in the phylogenetic tree and by principal component analysis (PCA), but there was little overlap between Group 1 and Group 2, which probably means that accessions from Group 3 (from Southwest China) are the common ancestors of accessions from Group 1 (from the Yangtze River Valley) and Group 2

**Table 3** Analysis of molecular variance for *Gossypium arboreum* accessions among and within three groups corresponding to three major regions of *Gossypium arboreum* growing in China as identified by STRUCTURE

Source of variation	df	Sum of squares	Variance components	Percentage of variation
Among groups	2	1 059.815	3.185 43 *Va	13.61
Within groups	194	8 745.708	22.657 27*Vb	86.39
Total	196	9 805.523	25.842 70	

\* Significant at  $P < 0.0001$

**Table 4** Pairwise  $F_{ST}$  estimates for the three groups corresponding to three major regions of cotton production as identified by STRUCTURE

	Group 1	Group 2	Group 3
Group 2	0.235 55***		
Group 3	0.191 96***	0.191 05***	

\*\*\* Significant at  $P < 0.0001$ 

(from North China). Moreover, the highest genetic differentiation was observed between Group 1 and Group 2, which corroborated the genetic relationships and confirmed the various routes of introduction. The lowest genetic differentiation observed between Group 2 and Group 3, further validating the introducing route of *G. arboreum* from Bengal-Assam to the Yellow River Valley. The difference between these two kinds of analyses maybe caused by recombination and selection of *G. arboreum* under the environment or by the choice of *Gossypium arboreum* samples.

The differentiation among groups obtained from Structure analysis was also validated by analysis of molecular variance. The marker variation among groups was observed to be 13.61%, whereas variation within groups was 86.39%. This was in agreement with 0.137 of  $F_{ST}$  caused by differences among accessions in genetic structure analyses.  $F_{ST}$  values observed in this study ranged from 0.18 to 0.23 which are different with the results in upland cotton (ranged from 0.29 to 0.42) (Tyagi et al. 2014). Deep population differentiation could reduce the efficiency of the successful Genome-wide association studies (GWAS) (Tyagi et al. 2014). Moreover, large variations would decrease the power of structure-based association studies to detect the effects of single genes (Flint-Garcia et al. 2005). Because of low differentiation, *Gossypium arboreum* population may be more suitable for GWAS analysis than *Gossypium hirsutum* for screening important genes associated with traits. Moreover, *G. arboreum* has a relatively smaller genome than *G. hirsutum*, which is beneficial for characterization at the molecular level and facilitates the use of this resource in developing superior cotton cultivars with favorable agronomic traits. Therefore, evaluation of genetic diversity and population differentiation is desirable for the efficient utilization of valuable genes of *G. arboreum*.

## Conclusions

Genetic structure was studied within the *Gossypium arboreum* accessions collected in China. Three subgroups identified from the analysis of Structure agreed with the main regions of *G. arboreum* growing regions in China. Genetic diversity of the panel corroborated the result of genetic structure analysis, however a few discrepancies were observed between pedigree information and genetic relationships. From the results of AMOVA and genetic structure, the genetic differentiation among

and within the groups was a reality. Recombination and selection caused by the environment and farmers may have led to the occurrence of valuable traits that can be useful for breeding programs in cotton. These valuable traits will be beneficial to the breeding programs of cotton through the generation of inter-specific hybrids.

## Additional files

**Additional file 1:** Table S1. Description of accessions used in this research. (XLS 36 kb)

**Additional file 2:** Table S2. A summary of marker statistics. (XLSX 18 kb)

## Abbreviations

AMOVA: Analysis of molecular variance; CMD: Cotton marker database; GD: Genetic distance; GWAS: Genome-wide association study; PAGE: Polyacrylamide gel electrophoresis; PCA: Principal coordinate analysis; PCR: Polymerase chain reaction; PIC: Polymorphic information content; SSR: Simple sequence repeat

## Funding

This research was supported by the National Natural Science Foundation of China Agriculture (Grant No. 2015NWB039).

## Availability of data and materials

Please contact author for data requests.

## Authors' contributions

Jia YH carried out the molecular genetic studies, and drafted the manuscript. Pan ZE participated in the sequence alignment. He SP participated in the sequence alignment. Gong WF participated in the design of the study and performed the statistical analysis. Geng XL conceived of the study, and helped to draft the manuscript. Pang BY participated in its design and coordination. Wang LR helped to draft the manuscript. Du XM conceived the study, and helped to draft the manuscript. All the authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

Received: 9 July 2018 Accepted: 19 September 2018

Published online: 28 September 2018

## References

- Abdurakhmonov IY, Kohel RJ, Yu JZ, et al. Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics*. 2008;92:478–87. <https://doi.org/10.1016/j.ygeno.2008.07.013>.
- Brubaker CL, Bourland FM, Wendel JF. The origin and domestication of cotton. In: Smith CW, Cothren JT, editors. *Cotton: origin, history, technology and production*. New York: Wiley; 1999. p. 3–31.
- Cai CP, Ye WX, Zhang TZ, et al. Association analysis of fiber quality traits and exploration of elite alleles in upland cotton cultivars/accessions (*Gossypium hirsutum* L.). *J Integr Plant Biol*. 2014;56:51–62. <https://doi.org/10.1111/jipb.12124>.
- Chen ZJ, Scheffler BE, Dennis E, et al. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol*. 2007;145:1303–10. <https://doi.org/10.1104/pp.107.107672>.
- Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol*. 2005;14:2611–20. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>.

- Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Eco Res*. 2010;10:564–7. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>.
- Flint-Garcia SA, Thuitel AC, Yu J, et al. Maize association population: a high resolution platform for quantitative trait locus dissection. *Plant J*. 2005;44: 1054–64.
- Fryxell PA. The natural history of the cotton tribes. College Station, USA: Texas A & M University Press; 1979. p. 245.
- Fryxell PA. A revised taxonomic interpretation of *Gossypium* L., (Malvaceae). *Rheeda*. 1992;2:108–65.
- Guo WZ, Zhou BL, Yang LM, et al. Genetic diversity of landraces in *Gossypium arboreum* L. race sinense assessed with simple sequence repeat markers. *J Integr Plant Biol*. 2006;48:1008–17. <https://doi.org/10.1111/j.1744-7909.2006.00316.x>.
- Hardy OJ, Vekemans X. SpaGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Mol Ecol Notes*. 2002; 2:618–20. <https://doi.org/10.1046/j.1471-8286.2002.00305.x>.
- Huang ZK. Cotton varieties and their genealogy in China. Beijing: China Agricultural Press; 1996 (in Chinese).
- Hutchinson JB. New evidence on the origin of the old world cotton. *Heridity*. 1954;8:225–41. <https://doi.org/10.1038/hdy.1954.20>.
- Kantartzi S, Ulloa M, Sacks E. Assessing genetic diversity in *Gossypium arboreum* L. cultivars using genomic and EST-derived microsatellites. *Genetica*. 2009;136: 141–7. <https://doi.org/10.1007/s10709-008-9327-x>.
- Lacape JM, Dessauw D, Rajab M, et al. Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. *Mol Breed*. 2007;19:45–58. <https://doi.org/10.1007/s11032-006-9042-1>.
- Liu DQ, Guo XP, Lin ZX, et al. Genetic diversity of Asian cotton (*Gossypium arboreum* L.) in China evaluated by microsatellite analysis. *Genet Resour Crop Evol*. 2006;53:1145–52. <https://doi.org/10.1007/s10722-005-1304-y>.
- Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*. 2005;21:2128–9.
- Liu Q, Brubaker CL, Green AG, et al. Evolution of the FAD2-1 fatty acid desaturase 5'UTR intron and the molecular systematics of *Gossypium* (Malvaceae). *Am J Bot*. 2001;88:92–102.
- Mehetre SS, Aher AR, Gawande VL, et al. Induced polyploidy in *Gossypium*: a tool to overcome interspecific incompatibility of cultivated tetraploid and diploid cottons. *Curr Sci*. 2003;84:1510–2.
- Moulherat C, Tengberg M, Haquet JF, et al. First evidence of cotton at neolithic Mehrgarh, Pakistan: analysis of mineralized fibres from a copper bead. *J Archaeol Sci*. 2002;29:1393–401.
- Park YH, Alabady MS, Ulloa M, et al. Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred (RIL) cotton population. *Mol Gen Genomics*. 2005;274:428–41. <https://doi.org/10.1007/s00438-005-0037-0>.
- Paterson AH, Smith RH. Future horizons: biotechnology for cotton improvement. In: Smith WC, Cothren JT, editors. *Cotton: origin, history, technology, and production*. New York: Wiley; 1999. p. 415–32.
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
- Rahman M, Hussain D, Zafar Y. Estimation of genetic divergence among elite cotton (*Gossypium hirsutum* L.) cultivars/genotypes by DNA fingerprinting technology. *Crop Sci*. 2002;42:2137–44. <https://doi.org/10.2135/cropsci2002.2137>.
- Renny-Byfield S, Page JT, Udall JA, et al. Independent domestication of two old world cotton species. *Genome Biol Evol*. 2016;8(6):1940–7. <https://doi.org/10.1093/gbe/eww129>.
- Rohlf F. NTSYS-pc: numerical taxonomy and multivariate analysis system, version 2.2. Exeter software. New York: Setauket; 2000.
- Sambrook J, Fritsch EF, Maniatis T. *Molecular cloning: a laboratory manual*. Cold Spring Harbour, New York: Cold Spring Harbor Laboratory Press; 1989.
- Seelanan T, Schnabel A, Wendel JF. Congruence and consensus in the cotton tribe (Malvaceae). *Syst Bot*. 1997;22:259–90. <https://doi.org/10.2307/2419457>.
- Silow RA. The genetics and taxonomic distribution of some species lint quantity genes in Asiatic cottons. *J Genet*. 1939;38:277–98. <https://doi.org/10.1007/BF02982174>.
- Silow RA. The genetics of species development in Old World cottons. *J Genet*. 1944;46:62–77. <https://doi.org/10.1007/BF02986694>.
- Stephens SG. Phenogenetic evidence for the amphidiploids origin of New World cottons. *Nature*. 1944;153:53–4. <https://doi.org/10.1038/153053a0>.
- Sunilkumar G, Campbell LAM, Puckhaber L, et al. Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. *Proc Natl Acad Sci U S A*. 2006;103:18054–9. <https://doi.org/10.1073/pnas.0605389103>.
- Tyagi P, Gore MA, Bowman DT, et al. Genetic diversity and population structure in the US upland cotton (*Gossypium hirsutum* L.). *Theor Appl Genet*. 2014; 127:283–95. <https://doi.org/10.1007/s00122-013-2217-3>.
- Watt G. The wild and cultivated cotton plants of the world. London: Longmans; 1907.
- Wendel JF, Albert VA. Phylogenetics of the cotton genus (*Gossypium* L.): character-state weighted parsimony analysis of chloroplast DNA restriction site data and its systematic and biogeographic implications. *Syst Bot*. 1992; 17:115–43. <https://doi.org/10.2307/2419069>.
- Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. *Adv Agron*. 2003;78:139–86. [https://doi.org/10.1016/S0065-2113\(02\)78004-8](https://doi.org/10.1016/S0065-2113(02)78004-8).
- Wendel JF, Olson PD, Stewart JM. Genetic diversity, introgression, and independent domestication of old world cultivated cottons. *Am J Bot*. 1989; 76(12):1795–806. <https://doi.org/10.1002/j.1537-2197.1989.tb15169.x>.
- Xiang XL. Study and utilization on Asiatic cotton (*G. arboreum*) in China. *Sci Agr Sin*. 1988;21:94.
- Zhang J, Wu YT, Guo WZ, et al. Fast screening of microsatellite markers in cotton with PAGE/silver staining. *Cotton Sci*. 2000;12:267–9.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

