Theor Appl Genet (2014) 127:283–295
DOI 10.1007/s00122-013-2217-3

ORIGINAL PAPER

Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.)

Priyanka Tyagi · Michael A. Gore · Daryl T. Bowman · B. Todd Campbell · Joshua A. Udall · Vasu Kuraparthy

Received: 11 May 2013 / Accepted: 14 October 2013 / Published online: 30 October 2013

Abstract

Key message **Genetic diversity and population structure in the US Upland cotton was established and core sets of allelic richness were identified for developing association mapping populations in cotton.**

Abstract Elite plant breeding programs could likely benefit from the unexploited standing genetic variation of obsolete cultivars without the yield drag typically associated with wild accessions. A set of 381 accessions comprising 378 Upland (*Gossypium hirsutum* L.) and 3 *G. barbadense* L. accessions of the United States cotton belt were genotyped using 120 genome-wide SSR markers to establish the genetic diversity and population structure in tetraploid cotton. These accessions represent more than 100 years of Upland cotton breeding in the United States. Genetic diversity analysis identified a total of 546 alleles across 141 marker loci. Twenty-two percent of the alleles in Upland accessions were unique, specific to a single accession. Population structure analysis revealed extensive admixture and identified five subgroups corresponding to Southeastern, Midsouth, Southwest, and Western zones of cotton growing areas in the United States, with the three accessions of *G. barbadense* forming a separate cluster. Phylogenetic analysis supported the subgroups identified by STRUCTURE. Average genetic distance between *G. hirsutum* accessions was 0.195 indicating low levels of genetic diversity in Upland cotton germplasm pool. The results from both population structure and phylogenetic analysis were in agreement with pedigree information, although there were a few exceptions. Further, core sets of different sizes representing different levels of allelic richness in Upland cotton were identified. Establishment of genetic diversity, population structure, and identification of core sets from this study could be useful for genetic and genomic analysis and systematic utilization of the standing genetic variation in Upland cotton.

Introduction

Cotton is the leading natural fiber crop. Cotton belongs to the genus *Gossypium,* which has extensive phenotypic diversity among the approximately 50 species representing this genus (Campbell et al. 2009). Worldwide, four species are cultivated: two of these cultivated species are diploids ($2n = 2x = 26$) and two are allotetraploids ($2n = 4x = 52$). Most of the global cotton production comes from the two

Communicated by A. J. Bervillé.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2217-3) contains supplementary material, which is available to authorized users.

P. Tyagi · V. Kuraparthy (✉)
Crop Science Department, North Carolina State University,
Raleigh, NC 27695, USA
e-mail: vasu_kuraparthy@ncsu.edu

M. A. Gore
Department of Plant Breeding and Genetics, Cornell University,
Ithaca, NY 14853, USA

D. T. Bowman
North Carolina Foundation Seed Producers Inc., 8220 Riley Hill
Road, Zebulon, NC 27597, USA

B. T. Campbell
Coastal Plains Soil, Water and Plant Research Center,
USDA-ARS, 2611 West Lucas St., Florence, SC 29501, USA

J. A. Udall
Department of Plant and Wildlife Sciences, Brigham Young
University, Provo, UT 84602, USA

allotetraploid species, *G. hirsutum* and *G. barbadense* (Wendel et al. 1992).

Allotetraploid cotton evolved approximately 1.5 million years ago from a hybridization event between Old world cotton *G. herbaceum* (A1 genome) and New world cotton *G. raimondii* (D5 genome), followed by subsequent diploidization creating five tetraploid species (Wendel et al. 1992; Brubaker et al. 1999). *Gossypium hirsutum,* also called Upland cotton, represents 95 % of global cotton fiber production. *Gossypium barbadense* (also known as Pima cotton) is valued for its higher fiber quality and contributes around 3 % of global cotton production. The other three tetraploid species (*G. mustelinum*, *G. darwinii,* and *G. tomentosum*) are wild and are not grown commercially (Wendel et al. 1994; Wendel and Percy 1990; DeJoode and Wendel 1992).

Domesticated in the Yucatan peninsula about 5,000 years ago, (Wendel et al. 1992) primitive *G. hirsutum* strains were photoperiod sensitive. During the course of domestication day-neutral stocks were selected which allowed cotton to be eventually grown in the U.S. Cotton cultivation in the U.S. dates back to the early seventeenth century (Smith et al. 1999). One of the most important events in US cotton breeding history was the introduction of Mexican highland stocks in the early 1,800s, which contributed to the foundation of current Upland germplasm (Wendel et al. 1992). Several introductions from outside the US were also incorporated into different breeding programs. Further development of cultivars was directed by the need for locally adapted cultivars and events like the Boll weevil (*Anthonomous grandis* Boh.) outbreak that necessitated the demand for early maturing varieties (Niles and Feaster 1984). The cotton growing area in the US can be broadly divided into four regions: Western, Southwestern plains, Mid-south or the delta, and Southeast.

Within species, *G. hirsutum* shows great phenotypic diversity (Wendel et al.1992). Level of genetic diversity within *G. hirsutum* has been found to be higher than the other three cultivated cotton species (Wendel et al. 1992; Abdurakhmonov et al. 2012). Yet, studies have indicated that this diversity is not represented in the present cultivated germplasm of Upland cotton (Van Esbroeck et al. 1999). Apart from the initial bottleneck encountered during domestication process, cotton breeding has frequently involved crossing and re-selections within small sets of breeding materials which has led to the loss in genetic diversity (May et al. 1995; Bowman et al. 1996; Wendel et al. 1992; Brubaker et al. 1999). The narrow genetic base of Upland cotton has become a serious concern since limited genetic diversity translates to limited allelic availability for continued genetic gain (Brown 1983). With a heightened risk of genetic vulnerability to disease epidemics and climate change, elite breeding programs could benefit from the unexploited standing genetic variation of obsolete cultivars without the yield drag 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 (Van Esbroeck and Bowman 1998). 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 genetic analysis studies. Thus, assessment of genetic diversity and population structure is important in the US Upland cotton.

Genetic diversity estimates have been made using pedigree and morphological data (May et al. 1995; Van Esbroeck et al. 1999), biochemical markers (Wendel et al. 1992), 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. They are usually assumed to be equally unrelated even if otherwise, this often leads to overestimation of diversity (Bowman et al. 1996; Van Esbroeck et al. 1999). Molecular markers, on the other hand, are more reliable and informative since they can directly measure allelic diversity and give robust estimates of genetic distances. A multitude of DNA-based marker systems including restriction fragment length polymorphism (RFLP) (Van Becelaere et al. 2005), random amplified polymorphic DNA (RAPD) (Multani and Lyon 1995; Iqbal et al. 1997; Rahman et al. 2008), 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. SSRs have proven to be a very good marker system due to their codominant nature, reproducibility and convenience of use (Powell et al. 1996).

Although several studies have employed marker-based estimation of genetic diversity in cotton, most of these studies are limited in the number of accessions included or the number of markers used to characterize genetic diversity. Some of these studies have been conducted using germplasm specific for a breeding program, for example, the Pee Dee program (Campbell et al. 2009), New Mexico Acala program (Zhang et al. 2005) or cultivars from a specific geographical area, such as cultivars grown in Greece (Kalivas et al. 2011), Brazil (Bertini et al. 2006) and China (Liu et al. 2006), whereas others are more focused towards exotic material (Abdurakhmonov et al. 2008) and interspecific relationships (Abdalla et al. 2001). Besides, most of these studies used gel-based platforms for resolving marker allele diversity. However, capillary-based resolution of amplified product is more effective in separating different alleles than gel-based systems, thus increasing the efficiency and utility of genetic diversity and population structure studies. 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 genomics studies in cotton. The objectives of this study were to (1) estimate US Upland cotton genetic diversity; (2) analyze population structure; and (3) identify core sets of lines that maximize allelic diversity in the Upland cotton germplasm.

Materials and methods

Plant material

We sampled 378 *G. hirsutum* accessions, covering genotypes from all 14 states that constitute the US cotton belt. The selected accessions represent cotton cultivars from the early 1900s to 2005. These accessions include most of the important lines that have been used as parents in different breeding programs. Also included were 11 accessions used as parents to generate a random mated population in Upland cotton (Jenkins et al. 2008). In addition to public breeding lines and cultivars, the set also includes obsolete accessions from Delta and Pine Land Company, Stoneville, and Coker's Pedigreed Seed Company. Most of the seed material was obtained from the US National Cotton Germplasm Collection, USDA-ARS, College Station, TX, USA. Three accessions of *G. barbadense* were also included as out-group in the complete panel. Detailed information about the 381 lines used in this study is provided in Supplemental Table S1. To reduce residual heterozygosity within the accessions, all entries were selfed for three generations with single plant selections at the Central Crops Research Station, Clayton, NC, USA, during the summers of 2010 and 2011.

SSR genotyping

Leaf tissue was collected from a single plant of each accession and DNA was extracted using the procedure as described by Li et al. (2001). As a preliminary study we used a panel of 12 genotypes to identify SSR markers that gave reproducible amplification and could be confidently scored (data not shown). Out of 160 SSR primer pairs initially tested, 135 were selected to genotype the whole panel. These selected markers are uniformly distributed across the genome, with a minimum of four markers per chromosome. Primer sequences for all SSR markers are publically available and were obtained from Cotton Marker Database now housed in CottonGen (http://www.cottongen.org). Supplemental Table S2 includes the list of 135 SSR primers with their repeat motif and chromosomal locations as reported in literature. All forward primers were modified by adding a M13 sequence of 19 bases to their 5′end. A fluorescent 6-FAM or HEX labeled M13 primer was separately added to the PCR mix to generate fluorescent-labeled amplified product. PCR reaction volume was 6 μl and reaction mix included 20 ng DNA, 0.2 mM dNTP mix, 0.08 μM modified forward primer, 0.6 μM reverse primer and fluorescent-labeled M13 primer each, and 0.6 unit of Taq polymerase with 1 × reaction buffer containing 15 mM $MgCl_2$. All primers were amplified using a Touchdown protocol with amplification conditions as follows: 95 °C for 5 min, 15 cycles of 94 °C for 45 s, 65 °C for 45 s with a reduction of 1 °C per cycle, 72 °C for 1 min, followed by 25 cycles of 95 °C for 5 min, 50 °C for 45 s, 72 °C for 1 min, with a final step of 72 °C for 10 min. Amplified products were separated on an ABI 3730 capillary electrophoresis system (Applied Biosystems Inc., Foster City, CA, USA) with GeneScan™ 500 LIZ® used as internal size standard. GeneMarker software version 1.91 (Softgenetics, LLC, State College, PA, USA) was used to analyze ABI output. Amplicons with different fluorescent labels were multiplexed during ABI runs to increase throughput.

Preliminary analysis of genotypic data and genetic diversity

Since upland cotton is an allotetraploid crop with two different genomes it is possible that some markers could produce amplicons from both genomes giving rise to multilocus data. We employed two criterion for separating such multilocus data into different loci. First, if one of the alleles was monomorphic across all entries it was considered to be an individual locus. Second, because all entries were selfed for three generations we did not expect high residual heterozygosity so alleles were separated into two different loci to reduce overall heterozygosity for the marker. Basic summary statistics for biallelic data were calculated using POWERMARKER software package version 3.25 (Liu and Muse 2005). The polymorphism information content (PIC) of SSR marker was determined according to the method described by Botstein et al. (1980). A PIC value of 1 indicates that the marker can differentiate each line, and 0 indicates a monomorphic marker. Informative potential of a marker is high if its PIC value is more than 0.5, moderate if PIC is between 0.5 and 0.25, and only slightly informative if PIC value is below 0.25. Other statistics calculated were number of alleles, allele diversity, and heterozygosity for each marker.

POWERMARKER software package version 3.25 was used to calculate pairwise genetic distance between the accessions using Nei et al.'s (1983) D_A distance. The distance matrix was used to construct a dendrogram using neighbor joining method in POWERMARKER software. Dendroscope version 3.2.2 was used to visualize and edit the dendrogram. Further analysis of genetic structure was done by means of Principal co-ordinate analysis (PCA). Dominant data (0, 1 binary data) was used for PCA analysis in NTSYS-pc software version 2.2 using DCENTER and EIGEN functions (Rohlf 2000). Partitioning of genetic variance among and within groups was performed using Arlequin ver 3.5 software (Excoffier and Lischer 2010).

Analysis of genetic structure

STRUCTURE software version 2.3.4 (Pritchard et al. 2000) which is a model-based Bayesian method was used to delineate 381 accessions into clusters of individuals based on co-dominant genotypic data. Admixture model was used with the option of correlated allele frequencies between populations as recommended by Falush et al. (2003). 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. The degree of admixture (α) was determined from data. A value of α near 0 indicates no admixture is present, whereas α more than 1 suggests that most of the individuals are admixed (Falush et al. 2003). 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.

The number of sub-populations was estimated using the method proposed by Evanno et al. (2005) by plotting a distribution of ΔK, an ad hoc quantity based on second-order rate of change of the likelihood function with respect to K. The value of ΔK was calculated as mean of absolute values of difference between successive likelihood values of K divided by the standard deviation of L (K). The modal value of this distribution of ΔK best represents the underlying value of K, which is the uppermost hierarchical level of population structure. This method is also successful in identifying the true value of population number when there is little genetic differentiation between populations (Evanno et al. 2005). Structure harvester software was used to estimate ΔK (Dent and Bridgett 2012). Accessions were assigned to a 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.

Construction of core sets

The genotypic data were used to identify core sets of accessions that maximize genetic diversity in a limited number of accessions. Phenotypic data were collected for days to flowering, node of first fruiting branch, and fiber quality at three locations (data not presented). In order to identify a core set, any accessions with unfavorable traits like mutant phenotype, colored lint or late maturity were removed from analysis. Also, accessions with more than 1 % introgression from *G. barbadense* based on membership probability estimates from population structure analysis were eliminated from analysis. This reduced the number of accessions in final set to 324. Core sets of lines were assembled by maximizing allelic richness using a simulated annealing algorithm in POWERMARKER software package using the following parameters: $R = 2500$, $p = 0.95$, and $T_0 = 1$. Core sets of different sizes, ranging from $k = 8$ to $k = 53$ were assembled, in increments of five accessions.

Results

SSR marker analysis

Out of 135 SSR primer pairs used for genotyping, 12 were found to be monomorphic, whereas three could not be scored with confidence. These 15 SSRs were dropped from analysis, leaving data for 120 SSR primer pairs. Three accessions with more than 5 % missing data were also removed from analysis. The final data set included data for 378 accessions and 120 SSR primer pairs. Due to an allopolyploid genome, SSR primer pairs generating multi-locus markers are frequently observed in cotton (Fang et al. 2013). In the panel of accessions used in the present study, 17.5 % of the polymorphic primer pairs generated multi-locus markers, thus identifying 141 polymorphic and 21 monomorphic loci across the panel. Data for monomorphic loci were excluded from analysis.

Preliminary results show that these 378 lines are highly homozygous (average $H = 2.3$ %). Among 378 accessions a total of 546 SSR alleles were detected across 141 loci with an average of 3.9 alleles per SSR locus. Only 120 of the 141 SSR loci were found to be polymorphic exclusively among *G. hirsutum* accessions. For *G. hirsutum* accessions, observed heterozygosity was 2.6 %. Total 367 alleles were generated at 120 loci with an average of 3.1 alleles per SSR locus. Eighty-five alleles were common between *G. hirsutum* and *G. barbadense* accessions. Average PIC value for SSRs was 0.17 for Upland cotton accessions and 0.16 for the complete panel. A summary of marker statistics for *G. hirsutum* accessions is presented in Supplementary Table S3.

Unique alleles

Out of 546 alleles detected in the complete panel, 134 were unique, i.e., alleles found in only one accession (Table 1). Most of the alleles had very low allele frequency (Fig. 1a, b). Seventy-eight percent of these unique alleles were present in *G. barbadense* accessions (Table 1). Among *G. hirsutum* accessions, 80 unique alleles were observed in 54 accessions. A list of accessions containing unique alleles is presented in Supplement Table S4. Maximum number of unique alleles present in a single cultivar was seven, found in Rugose Indore. Eight of the 54 accessions that had unique alleles were introductions. Of the other 46 accessions, 14 were from mid-south or delta region, 14 from Southeast, 15 from Southwest or Texas plains, and only three were from Western region representing Acala breeding program. Release dates of these accessions with unique alleles ranged from early 1900s to 1998. Interestingly, among accessions from delta and eastern region more accessions were released prior to 1980. However, out of 15 accessions from plains region, 10 were released after 1980. These results suggest an improvement in the genetic diversity of accessions from plains region between 1980 and 1998.

Table 1 Summary of unique (present in one accessions) and rare alleles (present in <5 % accessions) observed in a combined panel of *G. hirsutum* and *G. barbadense* accessions versus only the Upland cotton accessions

Panel	Total alleles	Unique alleles	Rare alleles (freq <5 %)
Combined panel	546	134 (24 %)	199 (36 %)
G. hirsutum panel	367	80 (22 %)	94 (25 %)

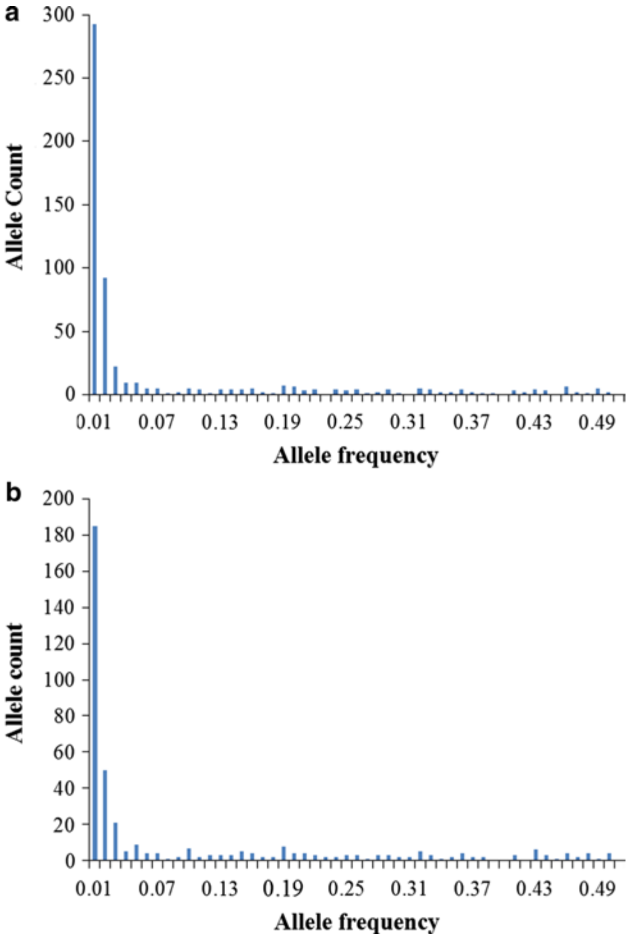

Fig. 1 **a** Histogram of allele frequencies for complete panel of *G. hirsutum* and *G. barbadense* accessions. **b** Histogram of allele frequencies for alleles amplified in *G. hirsutum* accessions

Population structure

Analysis of population structure was performed in the complete set of 378 accessions using software STRUCTURE. The number of subpopulations could not be identified from the plot of Ln (probability of data) for K (Fig. 2a). However, the number of clusters was successfully identified to be five based on ΔK value (Fig. 2b). Substantial admixture was found to occur between clusters (Fig. 3). Out of 378 accessions, only 184 could be assigned to subgroups based on 70 % membership threshold, meaning that more than half of cultivars were considered to have admixed parentage. Detailed description of membership probabilities of individual accessions is presented in Supplementary Table S5.

Accessions from *G. hirsutum* were separated into four groups (Fig. 3). Identified groups roughly correspond to the four zones of cotton breeding belt in United States. Cluster 1 (indicated red in Fig. 3) included 36 accessions. These accessions have Acala germplasm in their pedigrees and most belong to the western zone of the Cotton Belt. Tashkent 1, an introduction from Uzbekistan, was also included as a member of this group. Inclusion of Lone Star in group 1 was surprising and could not be explained. Lone Star, which was developed from Jackson Round Boll, is a historically important cultivar being the founder line for Stoneville cultivars. Group 2 (indicated green in Fig. 3) included 38 accessions mostly from eastern cotton belt. Most of the accessions included in group 2 were adapted to southeast zone of cotton growing area in the United States and included Sealand accessions and lines from Coker breeding program. Group 2 had highest number of rare alleles (alleles with frequency less than 5 %) detected within a group (Table 2). Group 3 (indicated blue in Fig. 3) mostly included accessions from the southwest region of cotton belt representing cotton breeding program for plains cotton. The accessions from Multi-Adversity Resistance program from Texas were included in group 3. Group 4 (indicated yellow in Fig. 3) primarily included the accessions from the midsouth or delta region. These include accessions from Delta and Pine Land Company and Stoneville Pedigreed Seed Company which are the major breeding programs for this area. Group 5 (shown pink in Fig. 3) had the three accessions from *G. barbadense*; all grouped together in a distinct cluster with more than 99 % membership probabilities.

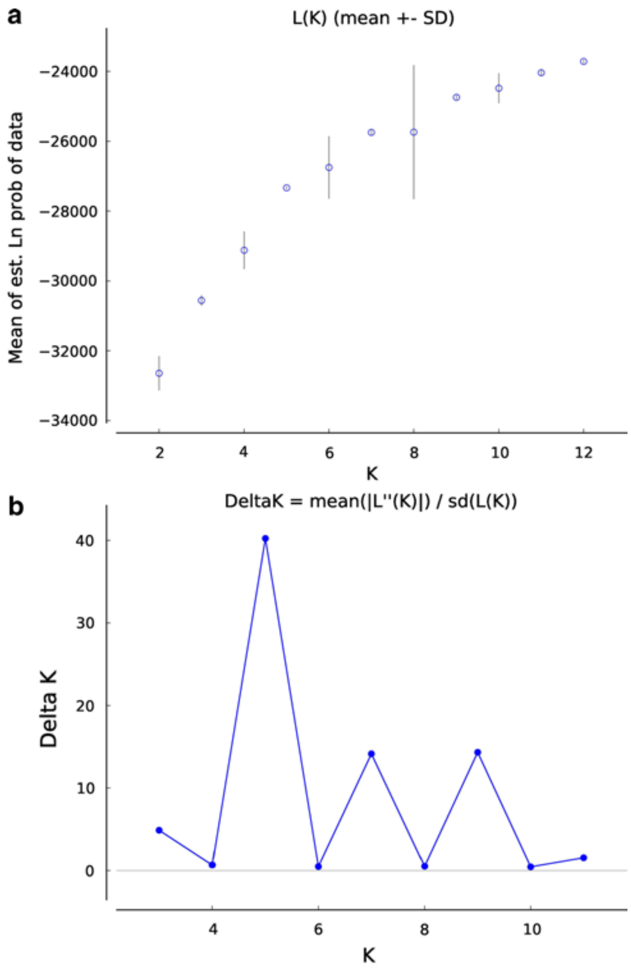

Fig. 2 **a** Ln (probability of data) for *K* ranging from 2 to 12. **b** Estimating number of subpopulations using delta *K* values for *K* ranging from 2 to 12 using method proposed by Evanno et al. (2005)

Genetic diversity and phylogenetic analysis

Neighbor joining analysis of genotypic data for *G. hirsutum* accessions using Powermarker software package indicated that the average genetic distance was 0.195 and ranged from 0.00 to 0.37. The highest genetic distance of 0.37 was between M4 and Tashkent 1. Among the four groups of accessions identified based on membership probabilities, group 2, which included accessions from southeastern United States had the highest average genetic distance of 0.186 (Table 3). Using the distance matrix, a phylogenetic tree was constructed. Six major clusters were identified in the NJ tree. In order to see how STRUCTURE results correspond to the phylogenetic analysis the dendrogram was manually edited to show STRUCTURE grouping (Fig. 4). Four groups of *G. hirsutum* accessions identified in STRUCTURE formed distinct clusters in the tree. Overall, there was good agreement between the two estimates. Clustering pattern was found to be in general agreement with relationships based on pedigree studies (Bowman et al. 2006), although in some cases lines that were selections or direct descendants from another lines were not grouped close to their parents, which is contradictory to the pedigree information. Still, in most cases they were in the same major cluster. For example, Lankart 611 grouped with Stoneville 5A instead of grouping with Lankart 57 of which it is a direct descendent (Bowman et al. 2006). Deltapine 14 and TM1 which is a selection from Deltapine 14 were also in the same major cluster but in different subgroups. Phylogenetic tree with accession identifiers is provided as Supplementary Figure S1.

Genetic relationships between *G. hirsutum* accessions were further studied using Principal coordinate analysis (Fig. 5). The first two axes of PCO accounted for 59.2 % of the variation. This indicates low level of genetic diversity in *G. hirsutum* germplasm with continuous variation between the subgroups. Analysis of molecular variance (AMOVA) revealed highly significant variation between the four groups identified by structure analysis with 31.4 % of the total variation contributing to between-group differences (Table 4). However, a larger amount of variation

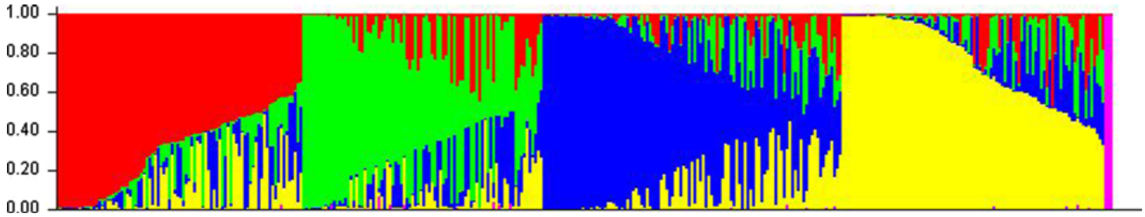

Fig. 3 *Q*-plot showing clustering of 381 *Gossypium* accessions based on analysis of genotypic data using STRUCTURE. Each accession is represented by a *vertical bar*. The colored subsections within each *vertical bar* indicate membership coefficient (*Q*) of the accession to different clusters. Identified subgroups are group 1 (*red color*), group 2 (*green color*), group 3 (*blue color*), group 4 (*yellow color*) and group 5 (*pink color*) (color figure online)

Table 2 Summary of rare alleles found in Upland cotton accessions grouped in clusters based on STRUCTURE analysis

Region	Number of lines	Total alleles	Rare allele (present in <5 % lines)
Western (red)	36	233	29 (12 %)
Eastern (green)	38	288	60 (21 %)
Southwest (blue)	57	237	32 (14 %)
Midsouth (yellow)	50	228	24 (11 %)

The color name in bracket shows the color assigned to the group in STRUCTURE *Q*-plot

Table 3 Genetic distance estimates calculated using Nei et al. (1983) distance within and between *G. hirsutum* groups identified by STRUCTURE analysis

	Western	Eastern	Southwest	Midsouth
Western	0.124			
Eastern	0.245	0.186		
Southwest	0.207	0.232	0.131	
Midsouth	0.209	0.212	0.193	0.109

(65.84 %) was due to diversity within the groups (Table 4). Overall population differentiation estimate (F_{ST}) among the groups was 0.342 highly significant at $P < 0.0001$. Pairwise F_{ST} revealed that accessions from group 2 (eastern region) are closer to accessions from midsouth (group 4) and southwest region (group 3) as compared with western Upland accessions represented mostly by Acala germplasm (Table 5). Highest genetic differentiation was observed between accessions from western (group 1) and midsouth (group 4) regions of Upland cotton growing area with a pairwise F_{ST} of 0.4196 ($P < 0.0001$) (Table 5).

Core sets of Upland cotton diversity panel

Genotypic data from 120 SSR loci were analyzed using Powermarker software to identify core sets of accessions based on allele number. Core group selection was constrained to positively include Acala Maxxa (as group I representative), Wannamaker Cleveland (as group II representative), Dixie King (as group III representative), Deltapine 14 (as group IV representative), DES56 (selected as female parent for population development), and Paymaster HS200. A plot of percentage of alleles in 324 *G. hirsutum* accessions captured in different core set sizes is shown in Fig. 6. The smallest set with 8 accessions captured 64 %, whereas the largest set of 53 accessions captured 96 % of all alleles detected using 120 SSR loci on 324 accessions (Fig. 6). Complete list of accessions included in different sets is presented in Table 6. A core set of 23 accessions represents 84 % of the alleles in 324 accessions and 74 % of alleles identified in complete panel of 381 accessions.

Discussion

In the current study, 120 SSR primer pairs produced 141 polymorphic loci in the complete panel of 381 accessions and 120 polymorphic loci in a panel of 378 Upland cotton accessions. An average of 3.1 alleles was amplified per locus in *G. hirsutum* accessions and 3.9 alleles per loci for the complete panel. Similar results were observed by Hinze et al. (2012) on allele number using a smaller sub set of SSR markers on improved US cotton germplasm. However, other studies showed variable allele number per locus. For example, Bertini et al. (2006) used 31 SSR primers to characterize 53 cultivars and reported 2.13 alleles per SSR locus. While 80 SSR primer pairs amplified 4.2 alleles per SSR across 72 Pee Dee lines (Campbell et al. (2009). Few studies have reported more alleles amplified per marker, for example, 5.8 alleles per primer were detected in a panel of 59 cotton cultivars of China (Zhang et al. 2011), and 5.6 alleles per marker were detected in a study using landraces and wild accessions of *G. hirsutum* (Lacape et al. 2007). The higher allele number observed in these studies is more likely a result of diverse germplasm used. Overall, the number of alleles observed per marker depends upon the selection of markers, collection of germplasm to be genotyped as well as the platform used for resolution of amplified products (Lacape et al. 2007). Fewer alleles per locus in Upland cotton are similar to the trend observed in other self-pollinated crops like rice (*Oryza sativa*) with 3.14 to 5.1 alleles per locus (Garris et al. 2005) and wheat (*Triticum aestivum* L.) with four alleles per locus (Oliveira et al. 2012). While the average alleles per locus for cross-pollinated crops are higher, for example, 14.1 alleles per locus reported in alfalfa (*Medicago sativa*) (Flajoulot et al. 2005) and 21.7 alleles per locus in maize (*Zea mays* L.) (Liu et al. 2003).

In the current study average PIC was 0.17, whereas in literature average PIC value for cotton SSRs can range from 0.122 (Abdurakhmonov et al. 2008) to 0.80 (Zhang et al. 2011). Lower number of alleles per locus and low PIC values in Upland cotton as shown in the current study further substantiate previous reports on narrow genetic base in cotton (Zhang et al. 2005; Campbell et al. 2009). We identified 21.5 % unique alleles in the elite germplasm (Table 1) which is much higher than 3 % reported in an earlier study (Abdurakhmonov et al. 2008). Among the cultivars released after 1998 none had any unique alleles (Supplementary Table S4). Fang et al. (2013) also reported that the average number of unique alleles in the US cultivars has declined from 0.53 in 1899–1950 to 0.24 in 1981–2011.

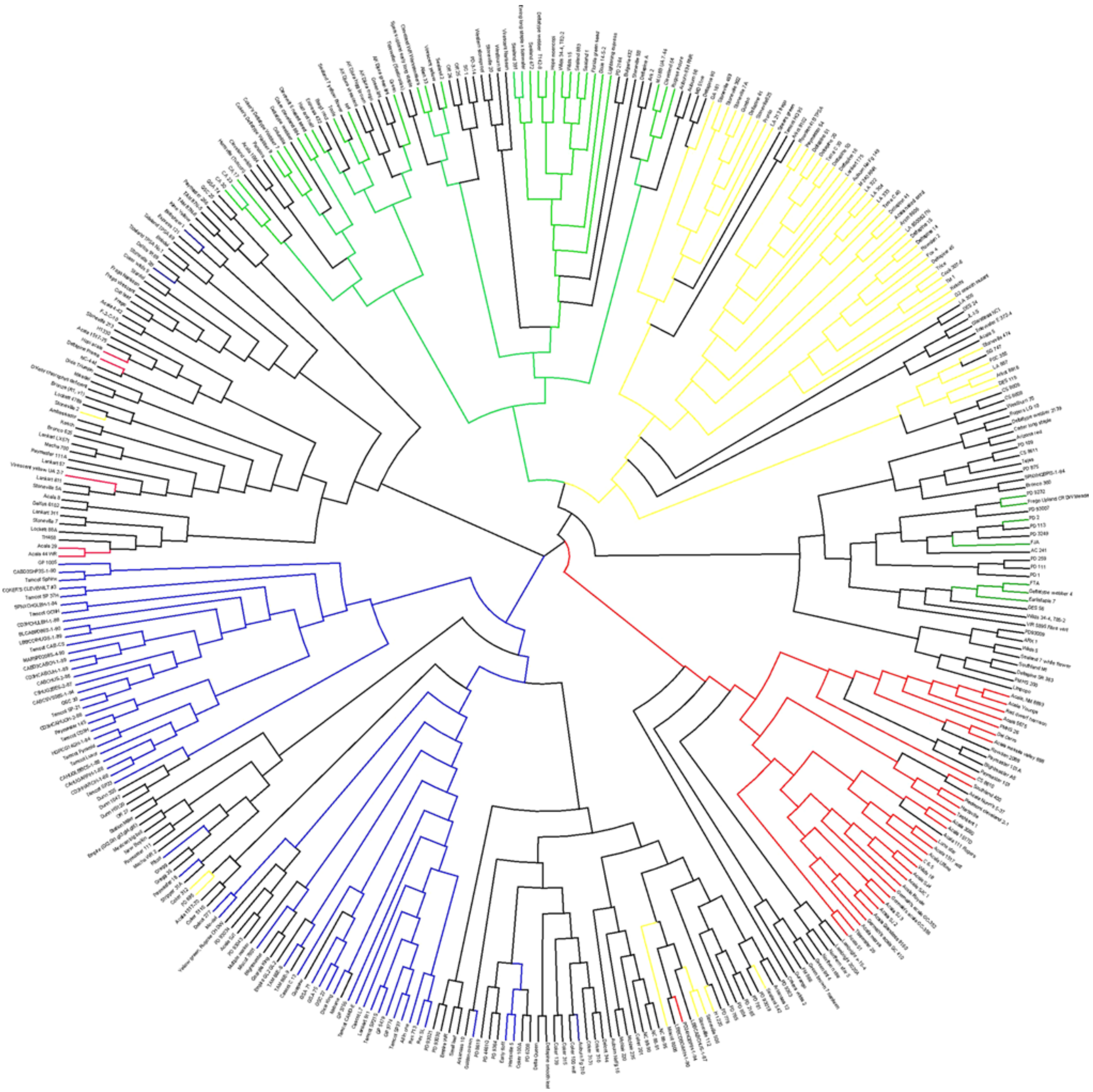

Fig. 4 Dendrogram of 381 *G. hirsutum* accessions by NJ analysis. *Colors* in the dendrogram correspond to population structure as identified in structure analysis. Membership threshold of 70 % was used to assign accessions to different cluster. Accessions with <70 % membership to any cluster were considered as mixed and are indicated in *black* in this dendrogram

These findings suggest a trend of declining genetic diversity in Upland cotton.

Model-based population structure analysis identified four differentiated sub-populations in Upland cotton accessions congruent with major cotton growing regions: Western, Southwestern, Midsouth, and Eastern (Fig. 3, Supplementary Table S5). One hundred and ninety-one accessions were assigned to mixed group indicating significant admixture (Supplementary Table S5). This admixture is possibly a result of germplasm sharing among different breeding programs. Another reason could be frequent appearance of a few lines with favorable agronomic traits in multiple breeding programs (Van Esbroeck and Bowman 1998). For example, between 1970 and 1990, Stoneville, Coker, and New Mexico lines were found in pedigrees of lines of other breeding programs more frequently than

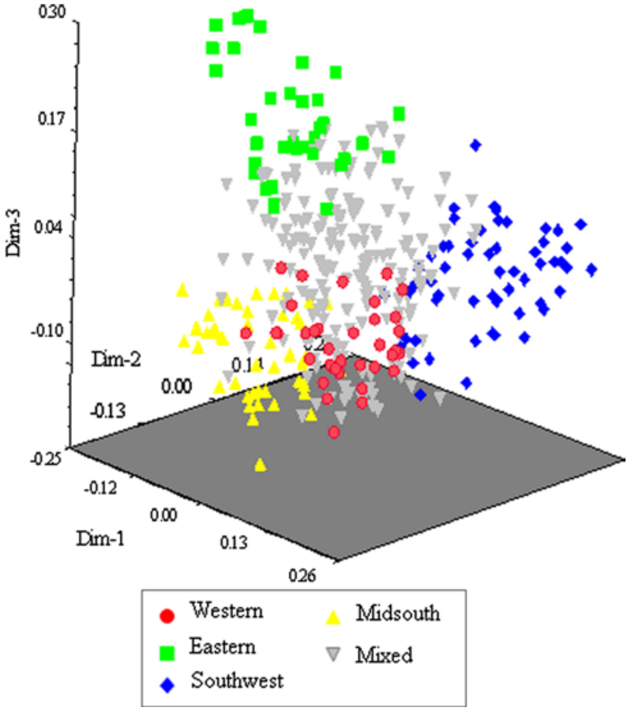

Fig. 5 Three-dimensional principal coordinate analysis (PCA) of Upland cotton diversity panel genotyped with SSR markers

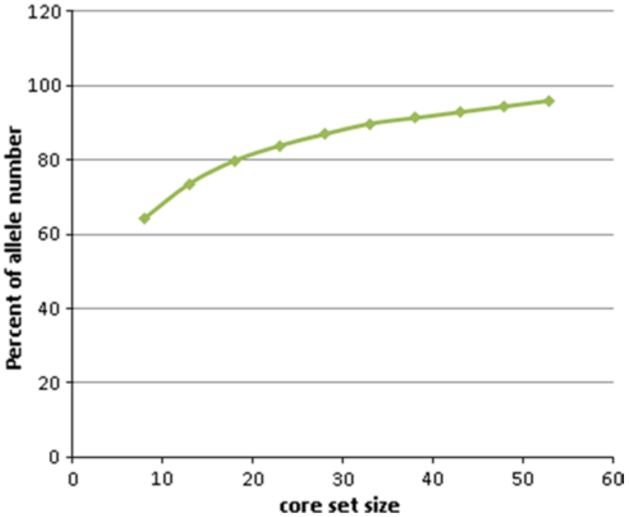

Fig. 6 Plot showing percentage of alleles included in core sets of different sizes ranging from 8 to 53 accessions of *G. hirsutum*

Table 4 Analysis of molecular variance for Upland cotton accessions between and within four groups corresponding to four major regions of cotton production in United States as identified by STRUCTURE

Source of variation	*df*	Sum of squares	Variance components	Percentage of variation
Among groups	3	1221.62	4.45879***Va	34.16
Within groups	358	3076.46	8.59347***Vb	65.84
Total	361	4298.09	13.05226	

*** Significant at $P < 0.0001$

Table 5 Pairwise F_{ST} estimates for the four groups corresponding to four major regions of cotton production in United States as identified by STRUCTURE

	Western	Eastern	Southwest	Midsouth
Eastern	0.3456			
Southwest	0.3581	0.2995		
Midsouth	0.4196***	0.2936	0.3494	

*** Significant at $P < 0.0001$

other lines (Bowman et al. 1996; Kuraparthy and Bowman 2013). Admixture was also observed between the two *Gossypium* species. Such admixture between *G. hirsutum* and *G. barbadense* is expected since introgressions from *G. barbadense* have been used for cultivar development, specifically in the development of Acala and Pee Dee germplasms. Out of 85 alleles common between the two species, Eastern group had the highest number of alleles (78) common with *G. barbadense* accessions.

The Western group of accessions included lines that had Acala germplasm in their pedigree (Supplementary Table S5). Original Acala accession was introduced in 1907 and breeding for better fiber quality traits helped shape the family of Acala cottons (Staten 1970). One of the Paymaster lines, PaymasterHS26, was also included in this group; its pedigree Acala SJ-4/5B9-184 shows that it has Acala germplasm and thus may explain this grouping (Bowman et al. 2006). The Southeastern group of accessions had representation of two different breeding programs, germplasm from Coker Seed Company and Pee Dee program. A large number of Pee Dee lines were included in this study. But most of the Pee Dee lines developed post boll weevil era could not be assigned to a cluster and were considered mixed. This finding seems acceptable given the ancestry of these lines, since Pee Dee germplasm was derived from a complex series of crosses using Triple hybrid, Sealand, Earlistaple, C 6-5, AHA, Dixie King, Auburn 56 and Coker 421 (Bowman et al. 2006). Early Sealand cultivars that were long staple Upland cotton cultivars adapted to southeastern US were also included in group 2 based on membership probability (Supplementary Table S5). The third group included lines from region west to the Mississippi delta and Texas plains. Only two Paymaster cultivars which were bred for Texas high plains were included in this group. Paymaster HS26 clustered with group 1 whereas Paymaster 54 clustered with group 4. Other six cultivars from Paymaster Seed Company were assigned to the mixed group. These results were not surprising given that most of the Paymaster

Table 6 Core sets of Upland cotton accessions identified by simulated annealing algorithm using Powermarker software

Set size	Accessions in core sets of reduced panel	Allele number	% of alleles in reduced panel	% of alleles in complete *G. hirsutum* panel
8	Acala 5, M.U.8B UA 7-44	207	64	56
13	Acala 5, M.U.8B UA 7-44, NC 88-95, PD 0113, PD 785, Auburn 634RNR, Toole	237	74	65
18	Acala 5, M.U.8B UA 7-44, NC 88-95, PD 0113, PD 785, Auburn 634RNR, Toole, Allen 33, Arkansas 10, PD 2164, Southland M1, Tashkent 1	257	80	70
23	Acala 5, Allen 33, CD3HCABCUH-1-89, DES 24, FJA, M.U.8B UA 7-44, NC 88-95, Paymaster HS26, PD 2164, PD 785, Auburn 634RNR, Sealand #2, Southland M1, Station Miller, Tashkent 1, Tidewater 29, Toole	270	84	74
28	Acala 5, Allen 33, BJAGL NECT, CD3HCABCUH-1-89, CS-8610, DES 24, Earlistaple 7, FJA, LBBCDBOAKH-1-90, M.U.8B UA 7-44, NC 88-95, Paymaster 101, Paymaster 266, PD 2164, PD 785, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Southland M1, Station Miller, Tashkent 1, Toole	280	87	76
33	Acala 5, Allen 33, BJAGL NECT, CA17, CABD3CABCH-1-89, CS-8610, Earlistaple 7, FJA, GSA 74, LBBCDBOAKH-1-90, M.U.8B UA 7-44, NC 88-95, Paymaster 101, Paymaster HS26, PD 0113, PD 2164, PD 785, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Southland M1, SPNXCHGLBH-1-94, Station Miller, Tashkent 1, Tidewater 29, Toole, Wilds 18	289	90	79
38	Acala 5, Allen 33, BJAGL NECT, CA23, CABD3CABCH-1-89, Coker's Wilds #2, CS-8610, Earlistaple 7, FJA, GSA 74, La.850082FN, LBBCDBOAKH-1-90, M.U.8B UA 7-44, NC 88-95, Paymaster 266, PD 0113, PD 2164, PD 785, PD 93009, PD 93021, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Southland M1, SPNXCHGLBH-1-94, Station miller, Tashkent 1, Tidewater 29, Toole, Western Stormproof, Wilds 18, Wilds 34-4(411) T85-2	294	91	80
43	Acala 5, Allen 33, BJAGL NECT, CA23, CD3HCABCUH-1-89, Coker 201, CS-8610, Earlistaple 7, Empire WR, FJA, GSA 74, La.850082FN, LBBCDBOAKH-1-90, Lightning Express, Lockett 88, M.U.8B UA 7-44, NC 88-95, New Boykin, Paymaster 266, Paymaster HS26, PD 0113, PD 2, PD 2164, PD 785, Piedmont Cleveland, Rowden 41B TPSA, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Southland M1, SPNXCHGLBH-1-94, SPNXHQBPIS-1-94, Station Miller, Stoneville 20, Tidewater 29, Toole, Wilds 18	299	93	81
48	Acala 5, Allen 33, Arkansas 10, BJAGL NECT, CA17, CABD3CABCH-1-89, Coker 201, Coker's Wilds #2, CS-8610, Earlistaple 7, Express 121, FJA, FOX 4, GSA 74, H1220, La.850082FN, LBBCDBOAKH-1-90, M.U.8B UA 7-44, NC 88-95, NC-4-M(3), Paymaster 266, PD 0113, PD 2, PD 2164, PD 785, PD 93009, PD 93021, PSC 355, Rogers LG-10, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Sealand 883, Southland M1, SPNXCHGLBH-1-94, Station Miller, Tamcot SP-23, Tashkent 1, Tidewater 29, Toole, Wilds 18, Wilds 34-4(411) T85-2	304	94	83
53	Acala 111 Rogers, Acala 5, Allen 33, Arkansas 10, Arkot 8102, BJAGL NECT, CA23, CABD3CABCH-1-89, CAHUGLBBCS-1-88, Coker 201, CS-8610, Earlistaple 7, Empire, Express 121, FJA, Gregg 35, GSA 74, H1220, Half and Half, Hopi Moencopi, La.850082FN, LBBCDBOAKH-1-90, LOCKETT 88, M.U.8B UA 7-44, NC 88-95, New Boykin, Paymaster 101, Paymaster HS26, PD 0113, PD 2164, PD 2165, PD 781, PD 785, PD 93009, PD 93030, Auburn 634RNR, Sealand #2, Sealand #7 Yellow Flower, Southland M1, SPNXCHGLBH-1-94, Station Miller, Tamcot luxor, Tamcot SP-23, Tashkent 1, Tidewater 29, Toole, Wilds 18	309	96	84

Complete panel has 375 accessions, whereas reduced panel contains 324 accessions after excluding agronomically unfavorable accessions. Acala Maxxa, Wannamaker Cleveland, Dixie King, Deltapine 14, DES 56 and Paymaster HS200 are used as positive constrains in core sets of all sizes

cultivars were mixed in ancestry (Smith et al. 1999). Historically, these cultivars were developed from a series of complex crosses between accessions from delta region (Stoneville), Acala cottons, Macha as well as Kekchi which was introduced from Guatemala (Smith et al. 1999).

A phylogenetic tree made using genotypic data broadly corroborated clustering of accessions detected by STRUCTURE (Fig. 4). The estimates of genetic distance (GD = 0.195) revealed overall level of genetic diversity to be low, a finding similar to earlier reports (Zhang et al.

2005; Campbell et al. 2009; Fang et al. 2013). However, this estimate may be inflated since data from monomorphic SSR loci were excluded in the current study. Most of the accessions in mixed group were located between major clusters in the Neighbor-joining tree (Fig. 4, Supplementary Figure S1). There was good agreement between this study and pedigree information. However, for some accessions there were discrepancies between pedigree information and marker-based relationships. Similar observations have been made in previous studies where discrepancies were observed between pedigree information and genetic relationships based on SSR markers (Zhang et al. 2005; Fang et al. 2013). Genetic diversity within the group was lowest for Midsouth group and highest for eastern group (Table 3). Eastern accessions in group 2 were closer to accessions from Midsouth or Southwest than Western accessions. The above observations suggest that, although different breeding programs developed cultivars suitable to specific geographic locations in the US cotton belt, germplasm exchanges between different breeding programs were not uncommon. This also could explain that in spite of narrow genetic base in cotton, breeders were able to develop improved cotton cultivars. Thus, current research results could help breeders to determine the selection of appropriate parental combinations in germplasm enhancement programs and conservation of genetic diversity.

The differentiation between groups was further validated by high F_{ST} value, with 34 % of the marker variation being explained by population structure of Upland cotton germplasm (Table 4). F_{ST} values for cotton observed in this study (0.29–0.42) are closer to another self-pollinating crop like rice (0.20–0.46) than to an out-crossing crop like corn (0.06–0.31) (Courtois et al. 2012; Garris et al. 2005; Liu et al. 2003). The presence of profound population differentiation could pose a challenge to successful Genome-Wide Association Mapping (GWAS) studies in Upland cotton germplasm for traits that are associated with population structure. The power of structure-based association studies to detect the effects of single genes would be reduced if a large fraction of variation was explained by population structure (Flint-Garcia et al. 2005). In such cases, alternative association mapping populations would be more useful (Flint-Garcia et al. 2005). Joint linkage-association mapping, especially the Nested Association Mapping (NAM) populations developed from core sets of allelic richness, could be used to detect and map agronomically desirable variation in crop plants (Wu and Zeng 2001; Meuwissen et al. 2002; Wu et al. 2002; Blott et al. 2003, Yu et al. 2008).

Core sets are a small subset of accessions that retain most of the genetic diversity present in an original collection of germplasm (Frankel 1984). They facilitate efficient utilization of overall genetic diversity while dealing with fewer accessions. Core sets are also excellent germplasm sets for developing association mapping populations. Molecular marker-based core sets have been identified in other crops, including maize (Liu et al. 2003), rice (Courtois et al. 2012), soybean (Kuroda et al. 2009), and Chinese wheat (Hao et al. 2008). In maize, the core sets identified from a panel of 260 lines led to the development of NAM populations, which have been used extensively in dissecting the genetic architecture of quantitative traits in corn (Liu et al. 2003; McMullen et al. 2009; Buckler et al. 2009). Genotypic values for agronomic traits have been used to identify core sets of *G. barbadense* accessions in China (Xu et al. 2006). However, no systematic efforts, utilizing molecular marker-based genotyping methods, were made to identify core sets for the US Upland cotton. In the current study, using 324 accessions that represent 322 alleles in the US upland cotton, core sets were assembled from the cotton diversity panel with sizes ranging from 8 to 53 lines in increments of 5 lines by maximizing allelic richness. A core set of 23 accessions that captured 74 % of the 322 alleles was selected for developing NAM population in upland cotton for establishing the genetic architecture of quantitative traits in cotton. Genetic diversity and population structure established in the present study would be informative to select parental accessions for breeding and genetic analysis as well as for efficient management and conservation of Upland cotton genetic diversity. Further, the current diversity panel of Upland cotton will be invaluable as a community resource for measuring linkage disequilibrium (LD) and for fine scale mapping of traits through LD mapping or Genome-Wide Association Study (GWAS) that can be streamlined for genomics-assisted plant breeding programs.

Acknowledgments We thank Dr. Gina Brown-Guedira for providing access to the genotyping facility and Jared Smith, Kim Howell and Blake Bowen for their technical assistance. We are grateful to Cotton Incorporated, NC Agricultural Research Service and NC Cotton Producers Association for funding support. The authors would like to thank NC State University Plant Breeding Center and Monsanto Company for providing PhD assistantship to Priyanka Tyagi.

Conflict of interest The authors declare that there are no conflicts of interest in the reported research.

Ethical standards The authors note that this research is performed and reported in accordance with ethical standards of the scientific conduct.

References

Abdalla AM, Reddy OUK, El-Zik KM, Pepper AE (2001) Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theor Appl Genet 102:222–229

Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov LB, Buriev ZT, Saha S, Scheffler

BE, Jenkins JN, Abdukarimov A (2008) Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. Genomics 92:478–487

Abdurakhmonov IY, Buriev ZT, Shermatov SE, Abdullaev AA, Urmonov K, Kushanov F, Egamberdiev SS, Shapulatov U, Abdukarimov A, Saha S, Jenkins JN, Kohel RJ, Yu JZ, Pepper AE, Kumpatla SP, Ulloa M (2012) Genetic Diversity in G*ossypium* genus. In: Caliskan M (ed) Genetic Diversity in Plants, ISBN: 978-953-51-0185-7, InTech, pp 313–338. doi:10.5772/2640

Bertini CHCD, Schuster I, Sediyama T, Barros EG, Moreira MA (2006) Characterization and genetic diversity analysis of cotton cultivars using microsatellites. Genet Mol Biol 29:321–329

Blott S et al (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 16:253–266

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Bowman DT, May OL, Calhoun DS (1996) Genetic base of upland cotton cultivars released between 1970 and 1990. Crop Sci 36:577–581

Bowman DT, Gutierrez OA, Percy RG, Calhoun DS, May OL (2006) Pedigrees of upland and pima cotton cultivars released between 1970 and 2005. Miss Agric For Exp Stn Bull 1155

Brown WL (1983) Genetic diversity and genetic vulnerability: an appraisal. Econ Bot 37:4–12

Brubaker CL, Bourland FM, Wendel JF (1999) The origin and domestication of cotton. In: Smith CW, Cothren JT (eds) Cotton: origin, history, technology, and production. Wiley, New York, pp 3–32

Buckler ES et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Campbell BT, Williams VE, Park W (2009) Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. Euphytica 169:285

Courtois B, Frouin J, Greco R, Bruschi G et al (2012) Genetic diversity and population structure in a European collection of rice. Crop Sci 52:1663–1675

Dejoode D, Wendel J (1992) Genetic diversity and origin of the hawaiian-islands cotton, *Gossypium tomentosum*. Am J Bot 79:1311–1319

Dent AE, Bridgett MV (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Eco Res 10:564–567

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fang DD, Hinze LL, Percy RG, Li P, Deng D, Thyssen G (2013) A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in Upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. Euphytica 1–11

Flajoulot S, Ronfort J, Baudouin P, Barre P, Huguet T, Huyghe C, Julier B (2005) Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. Theor Appl Genet 111:1420–1429

Flint-Garcia SA et al (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber W, Llimensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Hao C, Dong Y, Wang L, You G, Zhang H, Ge H, Jia J, Zhang X (2008) Genetic diversity and construction of core collection in Chinese wheat genetic resources. Chin Sci Bull 53:1518–1526

Hinze LL, Dever JK, Percy RG (2012) Molecular variation among and within improved cultivars in the U.S. cotton germplasm collection. Crop Sci 52:222–230

Iqbal MJ, Aziz N, Saeed NA, Zafar Y, Malik KA (1997) Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. Theor Appl Genet 94:139–144

Jenkins JN, McCarty JC Jr, Gutierrez OA, Hayes RW, Bowman DT, Watson CE, Jones DC (2008) Registration of RMUP-C5, a random mated population of upland cotton germplasm. J Plant Reg 2:239–242

Kalivas A, Xanthopoulos F, Kehagia O, Tsaftaris AS (2011) Agronomic characterization, genetic diversity and association analysis of cotton cultivars using simple sequence repeat molecular markers. Genet Mol Res 10:208–217

Kuraparthy V, Bowman DT (2013) Gains in breeding Upland cotton for fiber quality. J Cotton Sci (in press)

Kuroda Y, Tomooka N, Kaga A, Wanigadeva SMSW, Vaughan DA (2009) Genetic diversity of wild soybean (*Glycine soja* Sieb. et Zucc.) and Japanese cultivated soybeans [*G. max* (L.) Merr.] based on microsatellite (SSR) analysis and the selection of a core collection. Genet Res Crop Evol 56:1045–1055

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. Mol Breeding 19:45–58

Li H, Luo J, Hemphill JK, Wang JT (2001) A rapid and high yielding DNA miniprep for cotton (*Gossypium* spp.). Plant Mol Biol Rep 19:183a

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Liu B, Wendel JF (2001) Intersimple sequence repeat (ISSR) polymorphisms as a genetic marker system in cotton. Mol Ecol Notes 1:205–208

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

Liu D, Guo X, Lin Z, Nie Y, Zhang X (2006) Genetic diversity of Asian cotton (*Gossypium arboreum* L.) in china evaluated by microsatellite analysis. Genet Res Crop Evol 53:1145–1152

May OL, Bowman DT, Calhoun DS (1995) Genetic diversity of U.S. upland cotton cultivars released between 1980 and 1990. Crop Sci 35:1570–1574

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 16:373–379

Multani DS, Lyon BR (1995) Genetic fingerprinting of australian cotton cultivars with RAPD markers. Genome 38:1005–1008

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170

Niles GA, Feaster CV (1984) Breeding. In: Kohel RJ, Lewis CF (eds) Cotton, agronomy monograph no. 24. CSSA, Madison, pp 201–231

Oliveira HR, Campana MG, Jones H, Hunt HV, Leigh F, Redhouse DI, Lister DL, Jones MK (2012) Tetraploid wheat landraces in the Mediterranean basin: taxonomy, evolution and genetic diversity. PLoS One 7:e37063. doi:10.1371/journal.pone.0037063

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends in Plant Sci 1:215–222

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

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

Rohlf FJ (2000) Numerical taxonomy and multivariate analysis system, ver. 2.11. Applied Biostatistics, New York

Smith CW, Cantrell RG, Moser HS, Oakley SR (1999) History of cultivar development in the United States. In: Smith CW, Cothren JT (eds) Cotton: origin, history, technology, and production. Wiley, New York, pp 99–171

Staten G (1970) Breeding Acala 1517 cottons, memoir series no. 4. New Mexico State University, Las Cruces, pp 1926–1970

Van Becelaere G, Lubbers EL, Paterson AH, Chee PW (2005) Pedigree- vs. DNA marker-based genetic similarity estimates in cotton. Crop Sci 45:2281–2287

Van Esbroeck GA, Bowman DT (1998) Cotton improvement. Cotton germplasm diversity and its importance to cultivar development. J Cotton Sci 2:121–129

Van Esbroeck GA, Bowman DT, May OL, Calhoun DS (1999) Genetic similarity indices for ancestral cotton cultivars and their impact on genetic diversity estimates of modern cultivars. Crop Sci 39:323–328

Wendel J, Percy R (1990) Allozyme diversity and introgression in the galapagos-islands endemic *Gossypium darwinii* and its relationship to continental *Gossypium barbadense*. Biochem Syst Ecol 18:517–528

Wendel J, Brubaker C, Percival A (1992) Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. Am J Bot 79:1291–1310

Wendel J, Rowley R, Stewart J (1994) Genetic diversity in and phylogenetic-relationships of the brazilian endemic cotton, *Gossypium mustelinum* (malvaceae). Plant Syst Evol 192:49–59

Wu R, Zeng ZB (2001) Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909

Wu R, Ma CX, Casella G (2002) Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. Genetics 160:779–792

Xu H, Mei Y, Hu J, Zhu J, Gong P (2006) Sampling a core collection of Island cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. Genet Res Crop Evo 53:515–521

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Yu JZ, Fang DD, Kohel RJ, Ulloa M, Hinze LL, Percy RG, Zhang J, Chee P, Scheffler BE, Jones DC (2012) Development of a core set of SSR markers for the characterization of gossypium germplasm. Euphytica 187:203–213

Zhang J, Lu Y, Cantrell R, Hughs E (2005) Molecular marker diversity and field performance in commercial cotton cultivars evaluated in the southwestern USA. Crop Sci 45:1483–1490

Zhang Y, Wang XF, Li ZK, Zhang GY, Ma ZY (2011) Assessing genetic diversity of cotton cultivars using genomic and newly developed expressed sequence tag-derived microsatellite markers. Genet Mol Res 10:1462–1470