The *Journal of Cotton Science 23:38–47 (2019)*
http://journal.cotton.org, © The Cotton Foundation 2019

BREEDING AND GENETICS

Genetic Diversity and Population Structure in Elite U.S. and Race Stock Accessions of Upland Cotton (*Gossypium hirsutum*)

Linglong Zhu, Priyanka Tyagi, Baljinder Kaur, and Vasu Kuraparthy*

ABSTRACT

Assessing genetic diversity and population structure is prerequisite to the systematic utilization and conservation of germplasm resources of crop plants. The genetic diversity and population structure in a combined panel of 557 *Gossypium hirsutum* L. accessions including 375 cultivars and 182 race stocks using 114 pairs of simple sequence repeat primers were evaluated in the current study. Six *G. barbadense* L. accessions were included as an out-group. Genotyping the diversity panel of 563 accessions with the markers identified a total of 819 alleles and 662 alleles within *G. hirsutum*. Population structure analysis identified one *G. barbadense* group and five *G. hirsutum* groups corresponding to southwestern cultivars, Mexican collections, western cultivars, southeastern and mid-south cultivars, and Guatemalan collections. Average genetic distance of 0.253 indicated a moderate level of genetic diversity in this panel. Analysis of molecular variance revealed a low level of differentiation among cultivated cotton groups compared to landrace accessions. Genetic diversity and population structure analyses suggest landraces of Guatemala could be a potential source of novel genetic variability for U.S. cotton. Further, multiple core sets with different levels of allele richness were identified. The diversity panel and the core sets identified could be a good resource for broadening the genetic base of U.S. cotton and for genetic analysis of agronomic traits.

Cotton (*Gossypium* spp.) is the most important natural fiber source in the world (Carmichael, 2015). The genus *Gossypium* includes more than 50 species that are widely distributed throughout three major geographic regions: Africa-Asia, Australia, and central and southern Mexico (Campbell et al., 2009; Wendel et al., 1992). Recently, two new species of *Gossypium* have been discovered with the application of molecular biological approaches (Gallagher et al., 2017; Grover et al., 2015). Modern cultivated cotton consists of four species: *G. arboreum* L., *G. herbaceum* L., *G. hirsutum* L., and *G. barbadense* L. Among these, *G. hisutum* and *G. barbadense* dominate cotton production, contributing 98% (96% and 2%, respectively) of world cotton production. *G. arboreum* and *G. herbaceum*, which are mainly grown in Asia, contribute the remaining 2% of production. *G. arboreum* and *G. herbaceum* are diploid species with two copies of A genome (2n = 2x = 26), whereas *G. hirsutum* and *G. barbadense* are allotetraploid species with two copies of both A and D genome (2n = 4x = 52). Allotetroploid cotton is believed to have formed by a hybridization event between A genome species *G. herbaceum* and D genome species *G. raimondii* Ulbrich, which occurred one to two million years ago (Brubaker et al., 1999; Wendel, 1989; Wendel and Cronn, 2003).

The genetic diversity of *G. hirsutum* is believed to be greater than the other three cultivated cotton species (Wendel et al., 1992). However, the modem *G. hirsutum* gene pool is relatively narrow according to multiple studies, including the results from allozyme analyses (Wendel et al., 1992), cluster analysis of coefficient of parentage (May et al., 1995), DNA fingerprinting studies using restriction fragment length polymorphism (RFLP) (Iqbal et al., 2001; Wendel and Brubaker, 1993), simple sequence repeat (SSR) markers (Bertini et al., 2006; Tyagi et al., 2014), and single nucleotide polymorphisms (SNP) markers (Hinze et al., 2017). Crossing and re-selection within a limited set of lines that are adapted to the local environment is a common strategy in the development of cotton cultivars (May et al., 1995; Van Esbroeck et al., 1998; Wendel et al., 1992). In addition, the initial bottleneck caused by domestication limits the genetic diversity of the modern cotton gene pool. This narrow genetic base potentially makes elite cotton in the U.S. genetically vulnerable to diseases, pests, and environmental stresses. Limited genetic diversity raises a serious

L. Zhu, P. Tyagi, B. Kaur, and V. Kuraparthy*, Crop & Soil Sciences Department, North Carolina State University, Raleigh, NC 27695.
*Corresponding author: vasu_kuraparthy@ncsu.edu

concern for continued genetic gain in cultivar development in cotton. Introduction of a subtropical gene pool has been important in the development of Upland cotton cultivars (Wendel et al., 1992). Therefore, assessing, maintaining, and enhancing the genetic diversity of U.S. cotton germplasm is an important task for U.S. breeding programs.

Molecular marker-based methods have been used to estimate genetic diversity in Upland cotton (Fang et al., 2013; Hinze et al., 2017; Iqbal et al., 2001; Tyagi et al., 2014). However, most studies included a limited number of accessions and markers (Iqbal et al., 1997, 2001; Van Becelaere et al., 2005); and some studies focused on specific germplasm (Campbell et al., 2009; Zhang et al., 2005). Only a few studies estimated genetic diversity using large collections (Fang et al., 2013; Hinze et al., 2016, 2017; Tyagi et al., 2014). Previously, Tyagi et al. (2014) evaluated 378 elite Upland cotton cultivars and breeding lines using 120 SSR markers, whereas Kaur et al. (2017) reported the genetic diversity in a set of 185 race-stock accessions. However, a combined analysis of the two sets of Upland accessions is lacking. The current study combined the datasets from Tyagi et al. (2014) and Kaur et al. (2017). This study was designed to estimate the genetic diversity and population structure in a large and diverse collection of major U.S. cotton cultivars and landrace collections from tropical and subtropical regions using SSR marker alleles resolved through capillary-based gel electrophoresis. The main objectives of this study were to (1) estimate genetic diversity among *G. hirsutum* cultivars and landraces, (2) investigate population structure, and (3) identify core sets of lines that maximize the genetic diversity in the *G. hirsutum* collection.

MATERIALS AND METHODS

Plant Material. In this study, a combined panel of 569 accessions comprised a diversity panel of 384 *G. hirsutum* cultivars and a collection of 185 landrace accessions was used to study genetic diversity and population structure in a broad collection of Upland cottons of the Americas. The 384 *G. hirsutum* cultivars represent the cultivars that were released between the early 1900s and 2005, whereas the 185 landrace accessions were randomly selected from the U.S. National Cotton Germplasm Collection, USDA-ARS, College Station, TX (Kaur et al., 2017; Tyagi et al., 2014). The majority of landrace accessions used in this study were collected from Mexico and Guatemala (Kaur et al., 2017). Six accessions from the complete panel, Puerto Rican Regular, SA 2466, SA 2512, TX-0612, TX-0347, and TX-0604 were classified as *G. barbadense* lines (Kaur et al., 2017; Tyagi et al., 2014). They were included in the complete panel as an out-group. Detailed information for the 569 accessions used in this study is provided in Supplemental Table S1. Seeds for most of the accessions were obtained from the U.S. National Cotton Germplasm Collection, USDA-ARS. All accessions were further self-pollinated for two generations with single seed descent method at the Central Crops Research Station, Clayton, NC and at the winter nursery in Mexico.

Genotyping. The diversity panel of 384 *G. hirsutum* cultivars was genotyped using SSR markers by Tyagi et al. (2014). The race stock collection of 185 accessions was genotyped using the same set of SSR markers by Kaur et al. (2017). Description of genotyping of the Upland cotton accessions used in the current study is detailed in Tyagi et al. (2014) and Kaur et al. (2017). List of SSR markers used for genotyping is listed in Supplemental Table S2.

Analysis of Genotypic Data. Initial statistical summary of the genotypic data was calculated using Powermarker software version 3.25 (Liu and Muse, 2005). Major allele frequency, number of genotypes, number of alleles, heterozygosity, and polymorphism information content (PIC) values were calculated for each marker. Heterozygosity is estimated as the proportion of heterozygous individuals in the population, ranging from 0 to 1. A heterozygosity value of 0 indicates that all accessions are homozygous; 1 means all accessions are heterozygous. The PIC of SSR marker was calculated based on the method described by Botstein et al. (1980). A locus is of high diversity when PIC value is larger than 0.5, low when PIC value is less than 0.25, and intermediate diversity when PIC is between 0.25 and 0.5.

Analysis of Population Structure. STRUCTURE software version 2.3.4 (Pritchard et al., 2000) was used to analyze population structure among the 563 *Gossypium* accessions. STRUCTURE software assigns individuals into clusters based on genotypic data using model-based methods. In this study, admixture model with the option of correlated allele frequencies between clusters was used to identify the number of clusters (Falush et al., 2003). Admixture model assumes that each individual has inherited some fraction of its genome from different clusters. Key parameter settings to run STRUCTURE are

number of clusters (K), length of burn-in period, and number of replications. In this study, 10 runs were conducted for each number of K ranging from 2 to 12 with 10,000 burn-in period and 10,000 replications.

The number of clusters was estimated by plotting the distribution of ΔK on the Structure Harvester website (Earl and vonHoldt, 2012). The ΔK is an ad hoc statistic based on the rate of change in the log probability of data between K values (Evanno et al., 2005). The modal value of the distribution of ΔK most accurately detects the best value of K, which is the uppermost hierarchical level of structure (Evanno et al., 2005). Accessions with membership probability higher than 60% were assigned to one of the K clusters. Accessions with membership probability less than 60% were referred to the mixed cluster.

Analysis of Genetic Diversity. Frequency-based genetic distance between the accessions was calculated using Nei et al.'s (1983) D_A distance in Powermarker software. The distance matrix obtained from the previous step was used to construct a phylogenetic tree using the neighbor joining (NJ) tree option in Powermarker. Visualization and editing of the phylogenetic tree were done using Dendroscope version 3.5.9 (Huson and Scornavacca, 2012). Analysis of molecular variance (AMOVA) was performed using GenAlEx 6.5 (Peakall and Smouse, 2006, 2012) to study the genetic variance among and within groups.

Core Set Assembly. The core sets of accessions are a limited number of accessions that contain maximum allele richness from the collection used in the diversity analysis. In the current study, the core sets were assembled by finding the minimum number of accessions with maximal genetic diversity using a simulated annealing algorithm in Powermarker based on the genotypic data. The probability of finding the core set is determined by three parameters: number of evaluations per each annealing schedule (R), cooling coefficient (ρ), and initial temperature (T_0). In the current study, R was set to 2500, ρ was set to 0.95, and T_0 was 1.00. Core sets of different sizes, ranging from k = 10 to k = 80 were assembled, with increments of five accessions per core set in ascending order starting with a core set of 10.

RESULTS

Marker Statistics. Out of 137 SSR primer pairs used for genotyping, 23 were either found to be monomorphic or difficult to score, leaving 114 SSRs as assayable markers. SSR primer pairs frequently generate multiple loci in cotton due to its allopolyploid genome. In the current experiment, 26 markers generated two loci, making the total number of loci 140. Based on the marker statistics from Powermarker, markers BNL1604_Fa, CIR372_Fb, and BNL3029_Fa, had a missing data proportion of > 5% and were removed from the list. Another monomorphic marker, BNL 1153_Fa, was also excluded from the data set. Six accessions, ACALA 1064, Cleveland 54, Coker's Clevewilt 3, Ewings Long Staple, PD 3246, and PD 4381, with missing data > 10%, were also removed from the data set. The final data set included data on 563 accessions including six *G. barbadense* accessions for 136 marker loci.

All 563 accessions were highly homozygous (average heterozygosity rate = 2.1%). Among the 563 accessions with both cultivated tetraploid species, a total of 819 alleles were detected across 136 loci with an average of 6.02 alleles per SSR locus. PIC values for SSRs ranged from 0.0036 to 0.7308 with an average of 0.2346. For the 557 *G. hirsutum* accessions, four of the 136 SSR markers (BNL3474_Fb, CIR181_F, CIR119_Fa, and CIR187_Fb) were monomorphic and therefore removed from the list. This resulted in the detection of 662 alleles across 132 loci with an average of 5.02 alleles per SSR locus within *G. hirsutum*. PIC values ranged from 0.0036 to 0.7290 with an average of 0.2283, which approximated the number observed in the complete panel with six *G. barbadense*. The most informative markers in this study were BNL3800 and BNL3594_b with PIC values of 0.7244 and 0.7308, respectively. Detailed summary of marker statistics for both the complete panel and *G. hirsutum* accessions is presented in Supplemental Table S3.

Unique Alleles. In the current study, unique alleles are defined as alleles that are found in only one accession. Out of 819 alleles detected in the complete panel, 169 (20.6%) were unique alleles (Table 1). Fifty-two accessions had one unique allele; 24 accessions contained two or more unique alleles. Out of these 76 accessions containing putative unique alleles, 21 were cultivars, 49 were landraces, and six were *G. barbadense* accessions. The six *G. barbadense* accessions contained 37.9% (64 out of 169) unique alleles of the complete panel. In *G. hirsutum* accessions, 136 (20.5%) out of 662 alleles were unique alleles (Table 1). Fifty-eight accessions had one unique allele and 27 accessions contained two or more unique alleles. Among those 85 accessions, 27 were cultivars and 58 were landraces. Based on geographical distribution, eight accessions were collected from Southeastern

states, five from the Mid-South, five from the Southwest, two from the West, 29 from Mexico, 15 from Guatemala, seven from Caribbean countries, and 14 from other or unspecified regions. Landraces from Mexico and Guatemala showed more unique alleles (68.2%) compared to U.S. elite accessions. Within the U.S., accessions from the Southeast and Mid-South (48.1%) contained more unique alleles. Higher unique allele percentage in the landraces suggested that the genetic base of subtropical cotton is broader than the elite U.S. cotton germplasm and significant allelic diversity exists in landraces from Mexico. Detailed information of accessions containing unique alleles is presented in Supplemental Tables S4 and S5.

Population Structure. Analysis of population structure was performed with the complete panel of 563 accessions using STRUCTURE. Six clusters were identified based on ΔK value (Fig. 1). Out of 563 accessions, 429 could be successfully assigned to clusters based on the 60% membership threshold; 134 accessions could not be assigned to any subgroup (Fig. 2, Supplemental Table S6). The first group (red color in Fig. 2) only contained six *G. barbadense* accessions, which were assigned to Group 1 with higher than 99% membership probabilities. Those accessions were previously assigned to *G. barbadense* (Kaur et al., 2017; Tyagi et al., 2014). By phenotyping these in the field, we confirmed them as *G. barbadense* accessions. Those six accessions also carried the most unique alleles (Supplemental Table S4). The remaining five *G. hirsutum* groups corresponded with their growing regions. Group 2 (green color in Fig. 2) included 103 accessions that were mostly developed in the southwestern cotton growing region. Group 4 (yellow color in Fig. 2) was composed of 46 accessions, one landrace accession, and 45 cultivars. The majority of accessions were from the West region. Group 5 (pink color in Fig. 2) was the largest group with 11 landrace accessions and 149 cultivars. This group was a combination of two major cotton growing regions, Mid-South and Southeast. Group 3 (dark blue color in Fig. 2) included 59 landraces with most (46 accessions) originally collected from Mexico. Finally, Group 6 (light blue color in Fig. 2) was defined as Guatemalan cotton, as most of the accessions in this group were originally collected from Guatemala. The complete list of proportional membership of individual accessions is presented in Supplemental Table S6.

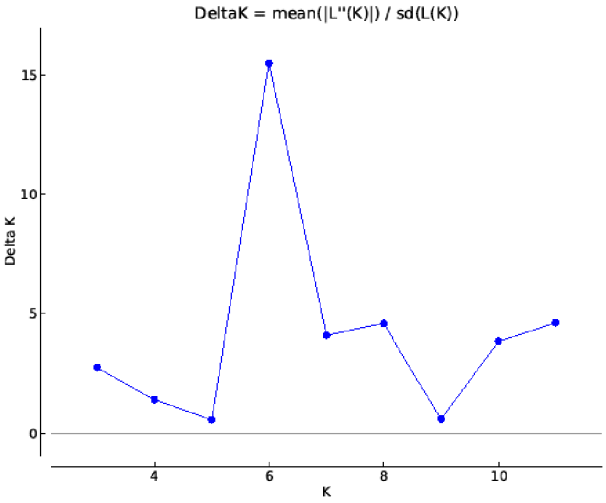

Figure 1. Plot showing the distribution of ΔK values for K ranging from 2 to 12 using method developed by Evanno et al. (2005). ΔK was calculated as ΔK = mean (|L″ (K)|) / sd [L (K)]. The modal value of this distribution is the true K value that identified six clusters.

Table 1. Summary of unique alleles (present in only one accession) observed in the complete panel and *G. hirsutum* panel[z]

Dataset	Total allele number	Unique allele number	Total number of accessions	Number of accessions with unique allele(s)
Complete panel	**819**	**169 (20.6%)**	**563[¶]**	**76**
***G. hirsutum* panel**	**662**	**136 (20.5%)**	**557**	**85**

[z] **Complete panel contained 557 *G. hirsutum* and 6 *G. barbadense* accessions.**

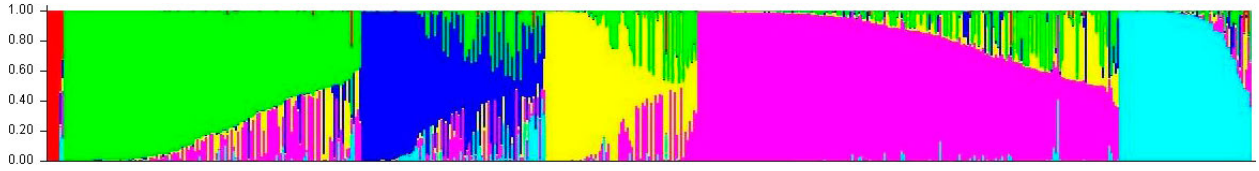

Figure 2. Q-plot showing clustering of 563 cotton accessions based on genotypic data of SSR markers using Structure v2.3.4. Each accession is represented by a vertical bar. The colored subsections within each vertical bar indicate membership coefficient of the accession to different clusters. Identified subgroups are: Group 1 (red color; *G. barbadense* accessions), Group 2 (green color; Southwestern cotton), Group 3 (dark blue color; Mexican cotton), Group 4 (yellow color; Western cotton), Group 5 (pink color; Mid-South and Southeastern cotton), and Group 6 (light blue color; Guatemalan cotton).

Genetic Diversity and Phylogenetic Relationships. Frequency-based genetic distance analysis showed that the average genetic distance was 0.253 for 557 *G. hirsutum* accessions. Accession TX-0078 had the highest average genetic distance from other accessions, averaging 0.637 with a range from 0.452 to 0.691. TX-0078 also had the most unique alleles within the *G. hirsutum* accessions. A phylogenetic tree was constructed using the distance matrix (Figs. 3 and 4). The accessions were also manually colored based on grouping results from STRUCTURE analysis to study how phylogenetic relationships corresponded to population structure results. The phylogenetic tree corresponded to the grouping results although some level of disagreement existed (Fig. 3). In Fig. 3, all six groups identified in STRUCTURE closely clustered in groups, although the group of Mid-South and Southeast cotton (in yellow) split into two major groups. The circular phylogram view of the phylogenetic tree gave a clearer view of the genetic distance between accessions or groups (Fig. 4). TX-0078, which had the highest average genetic distance (0.637) from other accessions, was also apart from the major *G. hirsutum* cluster (Fig. 4). Southwest, West, Mid-South, and Southeast groups had relatively low genetic distance between each other but had a higher genetic distance value against the Mexican group and Guatemalan group (Table 2). This was also true in the phylogenetic tree, where accessions from the Southwest, West, Mid-South, and Southeast groups were closely clustered, whereas accessions from the Mexican group and Guatemalan group spread wider and *G. barbadense* accessions were farther away from the mix (Fig. 4).

AMOVA showed that 32.1% of the total variation was due to among-group difference, indicating that there was significant variation between groups (Supplemental Table S7). However, 67.9% of the total variation was due to the diversity between the accessions within the groups (Supplemental Table S7). Pairwise F_{ST} values indicated that accessions from Southwest, Mid-South, and Southeast groups were genetically close to each other and accessions from Groups 3 and 6 shared a lower level of variance from each other but had significantly higher genetic distance from the other three groups (Table 3).

Core Set Assembly of *G. hirsutum* Accessions. Core sets of accessions with minimum number of accessions were assembled by maximizing allelic richness using a simulated annealing algorithm in Powermarker. The percentage of alleles captured was increased as the size of the core set increasing (Fig. 5). The smallest core set with 10 accessions covered 66% of the total 662 alleles detected from *G. hirsutum* accessions; whereas the largest core set with 80 accessions captured 96% of the total alleles (Fig. 5). Complete list of accessions for each *G. hirsutum* core set identified by Powermarker are listed in Supplemental Table S8.

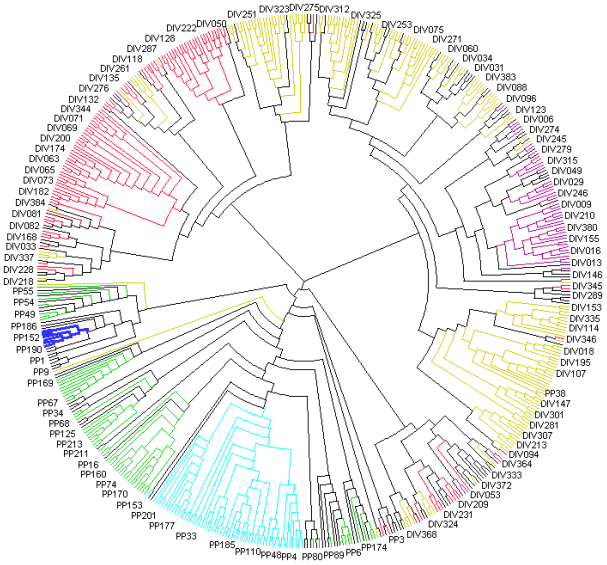

Figure 3. Circular cladogram view of the phylogenetic tree obtained from NJ analysis on complete panel of 563 accessions. Spikes are colored based on groups identified from STRUCTURE using 60% membership probabilities. Groups identified are: Group 1 (blue color; *G. barbadense* accessions), Group 2 (red color; Southwestern cotton), Group 3 (green color; Mexican cotton), Group 4 (purple color; Western cotton), and Group 5 (yellow color; Mid-South and Southeastern cotton), and Group 6 (light blue; Guatemalan cotton). Accessions assigned to "mixed" are indicated in black.

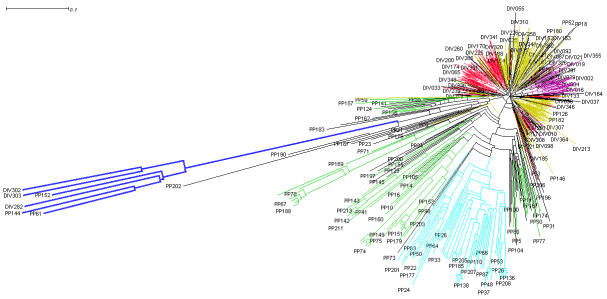

Figure 4. Circular phylogram view of the phylogenetic tree obtained from NJ analysis on complete panel of 563 accessions. Spikes are colored based on groups identified from STRUCTURE using 60% membership probabilities. Identified groups are: Group 1 (blue color; *G. barbadense* accessions), Group 2 (red color; Southwestern cotton), Group 3 (green color; Mexican cotton), Group 4 (purple color; Western cotton), Group 5 (yellow color; Mid-South and Southeastern cotton), and Group 6 (light blue; Guatemalan cotton). Accessions assigned to "mixed" are indicated in black.

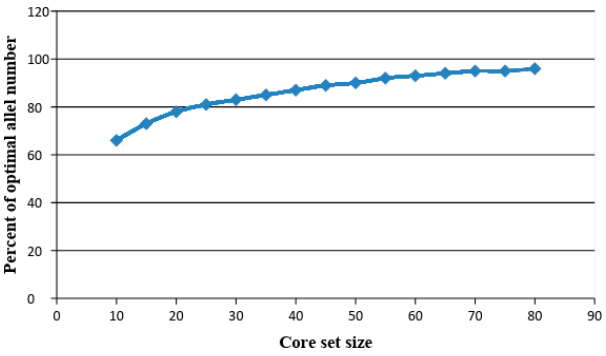

Figure 5. Plot showing optimum percent of allele number obtained against different core set sizes. For a given core set size, 2500 evaluations of simulated annealing were performed on *G. hirsutum* accessions. Core set sizes are in increments of five starting from 10 through 80 accessions.

DISCUSSION

The current study estimated the population structure and genetic diversity using genotypic data generated by SSR markers. An average of 6.02 and 5.02 alleles per SSR locus were detected in the complete panel with 563 accessions and *G. hirsutum* panel with 557 accessions, respectively. Compared to our study, lower numbers of alleles per locus were reported in many published reports. Zhang et al. (2005) discovered 189 alleles on 86 SSR loci with an average of 2.20 alleles per locus; Bertini et al. (2006) detected 2.13 alleles per SSR locus; and Campbell et al. (2009) detected 4.2 alleles per locus. These studies used specific germplasm or cultivars from a specific region in the genetic diversity analyses. A few studies reported higher number of alleles amplified per marker. For example, Lacape et al. (2007) identified 1128 alleles out of 201 loci with an average of 5.61 per locus and Kaur et al. (2017) observed 819 alleles across 143 loci with an average of 5.73 per locus. McCarty et al. (2018) reported an average of 7.9 alleles per locus in a day-neutral converted landrace population, whereas only three alleles per locus were found in a group of U.S. cultivars. This was expected because these later studies used landraces and wild accessions in the diversity analyses. This difference in allelic spectrum between the two types of germplasms suggest and confirm the higher genetic diversity in the landraces compared to the elite germplasm of the U.S. This further suggests that landraces could be a good source of novel alleles for broadening the genetic diversity of elite U.S. germplasm.

Generally, the number of alleles detected per locus could be affected by the selections of markers and plant materials. In the current study, the high number of 6.02 alleles per locus detected from the complete panel was caused by the addition of six *G. barbadense* accessions. The PIC values calculated to estimate the informative content of primers in this study varied from 0.0036 to 0.7308 with an average of 0.2346. This value is comparable with previous reports, which showed a range of PIC values, from low (0.122) (Abdurakhmonov et al., 2008), medium (0.55) (Lacape et al., 2007), to high (0.80) (Zhang et al., 2011). Higher PIC values in Lacape et al. (2007) and Zhang et al. (2011) studies were due to the low number of accessions and more diverse accessions used in their studies. Similar PIC values were reported by Fang et al. (2013), where relatively large *G. hirsutum* collections were used as plant materials. The lower level of PIC values estimated in the current study suggest a narrow genetic base in Upland cotton.

In this study, the model-based Bayesian method identified one group as an out-group with six *G. barbadense* accessions (due to limited sample size) and five groups consisting of *G. hirsutum* accessions. Among the five *G. hirsutum* groups, three groups consisted mainly of released cultivars and the other two consisted of only landrace collections. Group 2, corresponding to Southwest cotton growing region, consist mainly of Texas-derived accessions. Group 4 corresponded to the West region and included most of the Acala lines, which were developed in New Mexico, Arizona, and California. Group 5 was likely the combination of Mid-South and Southeast regions in which 55 members were cultivars developed from the Mid-South and 43 were from the Southeast. This indicated that Southeast accessions were genetically close to Mid-South accessions. This was also observed by Hinze et al. (2016), where accessions from Mid-South and Southeast regions had the highest genetic similarity among different breeding regions within the U.S. Pairwise F_{ST} data from Tyagi et al. (2014) made a similar conclusion. Groups 3 and 6 could be defined as Mexican cotton and Guatemalan cotton, as most of the accessions from those groups were Mexican and Guatemalan landrace collections, respectively. This result corresponded with geographical locations of the accessions, indicating that Mexican cotton and Guatemalan cotton retained relatively large genetic differentiation between each other. It

was notable that some landrace collections were also included in Group 2 (TX-0062, TX-0039, TX-2367), Group 4 (TX-0180), and Group 5 (TX-0043, TX-0101, TX0164, TX-0167, TX-0238, TX-0401, TX1131, TX-1464, TX-2368). Those accessions might be genetically close to the subtropical introductions during early cotton cultivar development as the earliest cultivars were introductions or selections of tropical landraces (Meredith, 1991; Moore, 1956; Ware, 1951; Wendel et al., 1992). Apart from those five groups, 134 accessions, which had mixed parentage, could not be assigned to any group based on 60% membership threshold. Among those, 81 accessions were advanced cultivars, indicated that there was a decent level of germplasm sharing between different U.S. breeding programs.

The phylogenetic tree was congruent with the grouping results (Figs. 3 and 4). The average genetic distance for 557 *G. hirsutum* accessions was 0.253, which was higher than many earlier reports on cultivated cotton germplasm (Abdurakhmonov et al., 2008, Tyagi et al., 2014; Ulloa et al., 1999). However, 0.253 was less than the average genetic distance among landrace collections (0.36) reported by Kaur et al. (2017). Average genetic distances between individuals in the same group were relatively low in all three cultivated cotton groups, but high in landrace collection groups (Table 2). Those results suggested that the genetic base of modern cultivated cotton germplasm is narrow, and a much higher genetic diversity exists in the subtropical cotton gene pool. This observation also is supported by the data on unique alleles, where 52 (61.2%) out of 85 unique alleles identified in the current study were found in the landrace accessions (Supplemental Table S5).

AMOVA was performed to study the genetic variance among and within groups. Within the three cultivated cotton groups, Southwest, Mid-South, and Southeast groups had lower genetic variance compared with West group (Table 3). This suggested that germplasm exchanges happened more frequently between East, Mid-South, and Southwest cotton growing regions in the U.S. cotton belt. This result was consistent with the statement made by May et al. (1995) that certain germplasm lines had been freely exchanged in developing early Texas stripper-type cotton and southeastern cotton. The average F_{ST} value was 0.48 between Guatemalan landrace collections and U.S. cultivars, which is higher than the F_{ST} value of 0.32 between Mexican landrace collections and U.S. cultivars (Table 3). This result was not surprising as Mexican stocks were the main introduction sources in early development of U.S. cotton cultivars (Wendel et al., 1992). This further suggested that Guatemalan race stocks could be an excellent source of variability for broadening the genetic base of U.S. cotton.

Table 2. Allele-frequency based genetic distance among and within six groups (net nucleotide distance) estimated using Nei et al. (1983) distance

	G. barbadense group	Southwest group	Mexican group	West group	Mid-South and Southeast group	Guatemalan group
***G. barbadense* group**	**0.5682**					
Southwest group	**0.4734**	**0.1136**				
Mexican group	**0.3482**	**0.1451**	**0.3610**			
West group	**0.4768**	**0.0689**	**0.1545**	**0.1023**		
Mid-South and Southeast group	**0.4684**	**0.0351**	**0.1326**	**0.0705**	**0.1283**	
Guatemalan group	**0.4113**	**0.1954**	**0.0873**	**0.2045**	**0.1915**	**0.2398**

Table 3. Pairwise fixation index F_{ST} among five groups of *G. hirsutum* identified by STUCTURE

	Southwest group	Mexican group	West group	Mid-South and Southeast group	Guatemalan group
Southwest group	-				
Mexican group	**0.336**	-			
West group	**0.264**	**0.314**	-		
Mid-South and Southeast group	**0.147**	**0.315**	**0.230**	-	
Guatemalan group	**0.494**	**0.184**	**0.481**	**0.476**	-

Finally, core sets from size 10 to 80 were selected from the *G. hirsutum* panel of 557 accessions (Supplemental Table S8). A core set of 25 accessions represented 81% of the total 662 alleles in *G. hirsutum* panel (Fig. 5). Among those 25 accessions, only three cultivars were presented. This result confirmed the high level of diversity that exists in the subtropical gene pool. Group-wise, two accessions were from the Mid-South and Southeast groups, seven accessions were from the Mexican group, six were from the Guatemalan group, and 10 did not belong to any group. None of the accessions from the West and Southwest groups were represented in the core set of 25. The accessions in the mixed group had mixed parentage; therefore, a single accession from the mixed group carries more diverse alleles than a member from grouped accessions. This could explain why 10 out of 25 accessions in the core set of 25 were from the mixed group. Tyagi et al. (2014) identified a core set of 23 cultivated accessions capturing 74% of total alleles in the diversity panel of 324 accessions. This is an excellent resource for breeding and genetic analysis on U.S. cotton cultivars. However, the core set of 23 accessions from Tyagi et al. (2014) could capture only 41% of total alleles in the current study indicating the untapped potential of tropical race-stock accessions as donors of novel alleles for U.S. cotton improvement. The narrow genetic base of modern cultivated cotton poses a challenge for developing cultivars with new and desired traits. The core sets of 25 identified in this study could be a readily available genetic source for broadening the genetic base in U.S. cotton.

ACKNOWLEDGMENTS

The authors acknowledge the generous funding support provided by North Carolina Cotton Producers Association (NCCPA) and Cotton Incorporated (CI) towards the research and assistantship to Linglong Zhu, Priyanka Tyagi, and Baljinder Kaur. We thank Gina Brown-Guedira for giving access to the genotyping platforms. We acknowledge Candace Haigler for giving access to NanoDrop 2000 spectrophotometer.

REFERENCES

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

Bertini, C.H., I. Schuster, T. Sediyama, E.G.D. Barros, and M.A. Moreira. 2006. Characterization and genetic diversity analysis of cotton cultivars using microsatellites. Genet. Mol. Biol. 29:321–329.

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Amer. J. Human Genet. 32:314–331.

Brubaker, C.L., F.M. Bourland, and J.F. Wendel. 1999. The origin and domestication of cotton. p. 3–31 *In* C.W. Smith and J.T. Cothren (ed.) Cotton: Origin, History, Technology and Production. Wiley, New York, NY.

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

Carmichael, A. 2015. Man-made fibers continue to grow. Textile World. 165:20–22. Cotton Incorporated. Properties of the Growing Regions [Online]. Available at https://www.cottoninc.com/fiber/quality/US-Fiber-Chart/Properties-of-the-Growing-Regions/ (verified 24 Dec. 2018).

Earl, D.A., and B.M. vonHoldt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4:359–361.

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

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

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

Gallagher, J.P., C.E. Grover, K. Rex, M. Moran, and J.F. Wendel. 2017. A new species of cotton from Wake Atoll, *Gossypium stephensii* (Malvaceae). Syst. Bot. 42:115–123.

Grover, C.E., X. Zhu, K.K. Grupp, J.J. Jareczek, J.P. Gallagher, E. Szadkowski, J.G. Seijo, and J.F. Wendel. 2015. Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. Genet. Resour. Crop Evol. 62:103–114.

Hinze, L.L., E. Gazave, M.A. Gore, D.D. Fang, B.E. Scheffler, J.Z. Yu, D.C. Jones, J. Frelichowski, and R.G. Percy. 2016. Genetic diversity of the two commercial tetraploid cotton species in the *Gossypium* diversity reference set. J. Hered. 107:274–286.

Hinze, L.L., A.M. Hulse-Kemp, I.W. Wilson, Q.H. Zhu, D.J. Llewellyn, J.M. Taylor, A. Spriggs, D.D. Fang, M. Ulloa, J.J. Burke, and M. Giband. 2017. Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. BMC Plant Biol. 17:37.

Huson, D.H., and C. Scornavacca. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61:1061–1067.

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

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

Kaur, B., P. Tyagi, and V. Kuraparthy. 2017. Genetic diversity and population structure in the landrace accessions of *Gossypium hirsutum*. Crop Sci. 57:2457–2470.

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

Liu, K., and S.V. Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129.

May, O.L., D.T. Bowman, and D.S. Calhoun. 1995. Genetic diversity of US upland cotton cultivars released between 1980 and 1990. Crop Sci. 35:1570–1574.

McCarty, J.C., D.D. Deng, J.N. Jenkins, and L. Geng. 2018. Genetic diversity of day-neutral converted landrace *Gossypium hirsutum* L. accessions. Euphytica 214:173.

Meredith, W.R. 1991. Contributions of introductions to cotton improvement. p. 127–146 *In* H.L. Shands and L.E. Wiesner (ed.), Use of Plant Introductions in Cultivar Development Part 1. Crop Science Society of America, Madison, WI.

Moore, J.H. 1956. Cotton breeding in the old south. Agric. Hist. 30:95–104.

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

Peakall, R., and P.E. Smouse. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes. 6:288–295.

Peakall, R., and P.E. Smouse. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537–2539.

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

Tyagi, P., M.A. Gore, D.T. Bowman, B.T. Campbell, J.A. Udall, and V. Kuraparthy. 2014. Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). Theor. Appl. Genet. 127(2):283–295.

Ulloa, M., W.R. Meredith Jr., R. Percy, and H. Moser. 1999. Genetic variability within improved germplasm of *Gossypium hirsutum* and *G. barbadense* cottons. Agronomic Abstract ASA. Madison, WI.

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

Van Esbroeck, G.A., D.T. Bowman, D.S. Calhoun, and O.L. May. 1998. Changes in the genetic diversity of cotton in the USA from 1970 to 1995. Crop Sci. 38:33–37.

Ware, J.O. 1951. Origin, Rise and Development of American Upland cotton Varieties and Their Status at Present. Univ. Arkansas Coll. Agric., Agricultural Experiment Station.

Wendel, J.F. 1989. New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. 86:4132–4136.

Wendel, J.F., and C.L. Brubaker. 1993. RFLP diversity in *Gossypium hirsutum* L. and new insights into the domestication of cotton. Am. J. Bot. 80:71.

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

Wendel, J.F., and R.C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Adv. Agron. 78: 139–186.

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

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

SUPPLEMENTAL TABLES

Supplemental Table S1. List of *G. hirsutum* accessions with identification number and their geographical location. http://www.cotton.org/journal/2019-23/1/upload/TableS1.htm

Supplemental Table S2: List of SSR markers used to genotype the panel of 569 cotton accessions. http://www.cotton.org/journal/2019-23/1/upload/TableS2.htm

Supplemental Table S3. Summary statistics for SSR markers used to genotype *G. hirsutum* accessions. http://www.cotton.org/journal/2019-23/1/upload/TableS3.htm

Supplemental Table S4. List of accessions with unique alleles (present in only one accession) in complete panel. Group numbers are corresponding with the result obtained from Structure software. Identified subgroups are group 1 (*G. barbadense* accessions), group 2 (Southwest cotton), group 3 (Mexican cotton), group 4 (Western cotton), group 5 (Mid-south and Southeast cotton) and group 6 (Guatemalan cotton). http://www.cotton.org/journal/2019-23/1/upload/TableS4.htm

Supplemental Table 5. List of accessions with unique alleles (present in only one accession) in *G. hirsutum* accessions. Group numbers are corresponding with the result obtained from Structure software. Identified subgroups are group 1 (*G. barbadense* accessions), group 2 (Southwest cotton), group 3 (Mexican cotton), group 4 (Western cotton), group 5 (Mid-south and Southeast cotton) and group 6 (Guatemalan cotton). http://www.cotton.org/journal/2019-23/1/upload/TableS5.htm

Supplemental Table S6. Membership of accessions to groups is determined by model based analysis using STRUCTURE. Accessions were assigned to a group based on membership probability greater than 60%. The identified groups roughly correspond to follwing geographical regions : group 1 - G. barbadense, group 2 - South west, group 3 - Mexican, group 4 - Western, group 5 - Mid-south and South east and group 6 - Guatemalan. http://www.cotton.org/journal/2019-23/1/upload/TableS6.htm

Supplemental Table S7. Analysis of molecular variance between and within groups for *G. hirsutum* accessions corresponding to 5 groups based on STRUCTURE analysis. http://www.cotton.org/journal/2019-23/1/upload/TableS7.htm

Supplemental Table S8. List of accessions for each *G. hirsutum* core set identified by maximizing allelic richness using a simulated annealing algorithm in Powermarker. http://www.cotton.org/journal/2019-23/1/upload/TableS8.htm