Genetica (2016) 144:535–545
DOI 10.1007/s10709-016-9921-2

Genetic diversity analysis of *Gossypium arboreum* germplasm accessions using genotyping-by-sequencing

Ruijuan Li[1] · **John E. Erpelding**[1]

Received: 3 November 2015 / Accepted: 31 August 2016 / Published online: 7 September 2016

Abstract The diploid cotton species *Gossypium arboreum* possesses many favorable agronomic traits such as drought tolerance and disease resistance, which can be utilized in the development of improved upland cotton cultivars. The USDA National Plant Germplasm System maintains more than 1600 *G. arboreum* accessions. Little information is available on the genetic diversity of the collection thereby limiting the utilization of this cotton species. The genetic diversity and population structure of the *G. arboreum* germplasm collection were assessed by genotyping-by-sequencing of 375 accessions. Using genome-wide single nucleotide polymorphism sequence data, two major clusters were inferred with 302 accessions in Cluster 1, 64 accessions in Cluster 2, and nine accessions unassigned due to their nearly equal membership to each cluster. These two clusters were further evaluated independently resulting in the identification of two sub-clusters for the 302 Cluster 1 accessions and three sub-clusters for the 64 Cluster 2 accessions. Low to moderate genetic diversity between clusters and sub-clusters were observed indicating a narrow genetic base. Cluster 2 accessions were more genetically diverse and the majority of the accessions in this cluster were landraces. In contrast, Cluster 1 is composed of varieties or breeding lines more recently added to the collection. The majority of the accessions had kinship values ranging from 0.6 to 0.8. Eight pairs of accessions were identified as potential redundancies due to their high kinship relatedness. The genetic diversity and genotype data from this study are essential to enhance germplasm utilization to identify genetically diverse accessions for the detection of quantitative trait loci associated with important traits that would benefit upland cotton improvement.

Keywords Genotyping-by-sequencing · *Gossypium arboreum* · Single nucleotide polymorphism · Genetic diversity

Introduction

Genetic diversity is significant for agricultural sustainability and serves as the basis for plant improvement. The goal of successful adaptation to changing environmental conditions and success in plant breeding cannot be achieved without diverse crop genetic resources (Govindaraj et al. 2015; Ulukan 2011). Determining the genetic diversity maintained in germplasm collections will assist in the utilization of germplasm in breeding programs.

Various methods have been utilized to evaluate genetic diversity of crop plants, including morphological, biochemical, and molecular marker analyses. Morphological and biochemical markers are easy to detect, robust, and highly reproducible; however, their application is limited by the small number of available markers. In contrast, molecular markers such as RFLPs (restriction fragment length polymorphism), SSRs (single sequence repeat), and SNPs (single nucleotide polymorphism) are abundantly distributed across plant genomes. Among the different molecular markers, SNPs are the most abundant type, which have become widely used especially with the rapid

Electronic supplementary material The online version of this article (doi:10.1007/s10709-016-9921-2) contains supplementary material, which is available to authorized users.

✉ John E. Erpelding
John.Erpelding@ARS.USDA.GOV

[1] Crop Genetics Research Unit, USDA-ARS, 141 Experiment Station Road, PO Box 345, Stoneville, MS 38776, USA

advancements in sequencing technology (Govindaraj et al. 2015; He et al. 2014).

Genotyping-by-sequencing (GBS) was developed as a rapid and robust approach for genotyping and SNP discovery (Poland and Rife 2012). The protocol employs a reduced representation sequencing strategy using one or more restriction enzymes to digest genomic DNA with a subset of restriction fragments sequenced to produce partial but genome-wide sequencing reads. The SNPs identified from the GBS generated sequencing data can be subsequently implemented in genomic diversity and genome-wide association (GWAS) studies. With the advantages of increased efficiency and cost effectiveness, GBS shows great promise in accelerating the progress of plant breeding and the rate of developing new crop cultivars (He et al. 2014; Poland and Rife 2012). As such, GBS has been widely applied in genetic studies of various species, including chickpea (Kujur et al. 2015), wheat (Lin et al. 2015), cotton (Logan-Young et al. 2015), lentil (Wong et al. 2015), maize (Elshire et al. 2011; Pace et al. 2015; Romay et al. 2013), barnyard millet (Wallace et al. 2015), guinea yams (Girma et al. 2014), and blue catfish (Li et al. 2014a). However, one factor to consider in conducting a GBS assay is missing data, which can lead to biased results (Glaubitz et al. 2014). Genotype imputation, a process to predict the genotypes that were not assayed in a sample of individuals, is frequently used to boost the number of SNPs for GWAS analysis (Marchini and Howie 2010).

Cotton (*Gossypium spp.*), widely grown in more than 70 countries worldwide, is the leading fiber crop and serves as important source of oil and seed meal. China, India, United States, and Pakistan are the leading cotton producing nations accounting for more than 70 % of world production (http://apps.fas.usda.gov/psdonline/psdDataPublications.aspx). Among the 50 species of cotton, two tetraploid species (*G. hirsutum* and *G. barbadense*) and two diploid species (*G. arboreum* and *G. herbaceum*) are cultivated with *G. hirsutum* (upland cotton) accounting for more than 90 % of cotton production worldwide. As a tetraploid species, *G. hirsutum* arose from the interspecific hybridization between putative *G. arboreum* and *G. raimondii* ancestors followed by chromosome doubling. Although *G. hirsutum* cultivars produce superior quality fiber and have greater lint yield, they are more susceptible to abiotic and biotic stresses (Maqbool et al. 2010; Robinson 2007). Therefore, plant breeders have been evaluating diploid cotton species to enhance genetic diversity for specific traits.

Gossypium arboreum was widely cultivated in Asia, the Middle East, and north-eastern Africa prior to the introduction of *G. hirsutum* cultivars (Guo et al. 2006). Because of its ability to adapt to extreme environmental conditions such as drought and saline soils (Guo et al. 2006; Maqbool et al. 2010), *G. arboreum* cultivars are more frequently grown in regions not suitable for *G. hirsutum* cultivation or under low input management practices (Iqbal et al. 2015). In addition, *G. arboreum* cultivars have a number of other traits that could benefit the breeding of new *G. hirsutum* cultivars including enhanced resistance to leaf curl virus (Ahmad et al. 2011), tobacco budworm (Hedin et al. 1992), thrips (Stanton et al. 1992), and reniform nematode (Erpelding and Stetina 2013; Sacks and Robinson 2009; Yik and Birchfield 1984). Some cultivars of *G. arboreum* also produce greater fiber strength, higher oil content, and increased seed indexes (Mehetre et al. 2003). While transferring genes from *G. arboreum* to *G. hirsutum* is achievable, specialized breeding approaches (Gill and Bajaj 1987; Sacks and Robinson 2009) are required to overcome hybridization barriers. Therefore, information on the genetic diversity of the *G. arboreum* collection is essential to aid in the selection of accessions for introgression and the utilization of this germplasm in breeding programs.

Several diversity assessments have been reported for *G. arboreum*, (Guo et al. 2006; Kantartzi et al. 2009; Kebede et al. 2007; Liu et al. 2006; Rahman et al. 2008), but these studies employed a small number of markers on a limited number of accessions or cultivars. In this study, GBS was performed using 375 *G. arboreum* accessions from the United States Department of Agriculture (USDA), National Plant Germplasm System (NPGS) cotton collection for large-scale SNP discovery to evaluate genetic diversity and population structure for the collection. Marker data produced from the GBS platform and information from this study can be further applied in GWAS and genomic selection evaluations for cotton improvement.

Materials and methods

Plant materials

More than 1600 *G. arboreum* accessions are maintained in NPGS cotton collection (Wallace et al. 2009) and the majority of these accessions were planted at the USDA in Stoneville, Mississippi during the 2011 and 2012 cropping seasons for seed production. Each accession was planted in a single row plot and plants in each row were self-pollinated to ensure seed purity. For this study, 380 accessions were selected for genotyping, which included landraces added to the collection prior to 1970 and cultivars or breeding lines added to the collection prior to 2001. Additionally, accessions PI529740 and PI630019 were included in the evaluation to determine whether these were duplicate accessions. Except for these two accessions, phenotypic data was not used as a criteria in selecting accessions. One plant from each accession was used and self-pollinated seed was planted in the greenhouse during the summer of 2014. Leaf samples were collected from a single seedling for each accession at approximately the six leaf stage. For DNA extraction, approximately 0.1 g of fresh leaf tissue was placed into 2 mL microfuge tubes and lyophilized. The dried leaf tissue was ground to a fine power, DNA extracted using the DNeasy Plant Mini kit (Qiagen, Valencia, CA) following the manufacturer's protocol, and DNA quantification was conducted using Quant-iT PicoGreen dsDNA Assay kit (Molecular Probes, Inc., Eugene, OR). The quality of DNA samples was assayed by electrophoresis of a 1 µL sample on a 1 % agarose gel using 1× TBE running buffer.

Genotyping-by-sequencing library construction and sequencing

Library construction and sequencing were conducted at the Institute for Genomic Diversity, Cornell University, Ithaca, New York using the protocol described by Elshire et al. (2011). DNA libraries for the accessions were prepared using the restriction enzyme *Ape*KI. This enzyme has been optimized for cotton and was selected to generate a greater number of SNPs. Each plate of 95 samples included a randomly placed blank well. A 96-plex library per flowcell channel was used for sequencing with four channels used to sequence the libraries for the 380 accessions.

Single-nucleotide polymorphism (SNP) calling

Single-nucleotide polymorphisms were identified using the TASSEL-GBS pipeline (Glaubitz et al. 2014) in TASSEL: 3.0.166. Briefly, raw fastQ sequences were trimmed of barcodes and reads from each of the four fastQ files were collapsed into one master TagCounts file containing unique tags along with their associated read count information. The resulting unique tags were then mapped to the *G. arboreum* (cultivar Shixiyal, SXY1) reference genome (Li et al. 2014b) by Burrows-Wheeler Aligner (Li and Durbin 2009) using default parameters on eight threads.

Tags aligned to unique positions on the genome were retained for SNP calling. The cultivar used for the reference genome was not included as one of the accessions in the present study. SNP discovery was performed for each set of tags that aligned to the exact same starting genomic position and strand. The genotype of the SNP was then determined by the default binomial likelihood ratio method of quantitative SNP calling in TASSEL: 3.0.166 (Glaubitz et al. 2014). Specifically, SeqToTBTHDF5Plugin, TagsToSNPByAlignmentPlugin, and MergeDuplicateSNPsPlugin inside TASSEL were used sequentially with all default parameters except that callHets option was set to true.

After genotypes were obtained for a potential SNP, initial filtering of SNPs was performed with minimum minor allele frequency of 0.01, minimum locus coverage of 0.1, and minimum taxa coverage of 0.8. Additional filtering steps were conducted in TASSEL to generate discriminating SNPs for subsequent population structure analysis using minimum minor allele frequency of 0.05 and maximum heterozygosity ratio of 0.2.

Population structure analysis

To infer the population structure for the sequenced accessions, STRUCTURE version 2.3.4 (Pritchard et al. 2000) was used to identify the number of populations (K). Briefly, a 10^5 length of burn-in period, followed by 10^5 Markov Chain Monte Carlo (MCMC) replications was set as the parameter using the admixture model with correlated allele frequency. K-values from 2 to 6 were tested for 5 independent runs. The STRUCTURE generated results were imported into Structure Harvester (Earl and vonHoldt 2012) to determine the best K-value using the Delta K method (Evanno et al. 2005). After determining the number of populations (K), individual accessions were assigned to each cluster if they had at least 60 % membership in that cluster. Then, the Q-matrix under the best K-value from Structure Harvester was exported to CLUMPP version 1.1.2 (Jakobsson and Rosenberg 2007) and subsequently the bar plot representing clusters with the best K-value was generated using DISTRUCT version 1.1 (Rosenberg 2004).

To compare the results from the STRUCTURE analysis, multidimensional scaling (MDS) analysis was conducted in TASSEL (Glaubitz et al. 2014), where the discriminating SNP dataset was converted into a distance matrix. The TASSEL built-in MDS analysis produced covariant data were used to construct the different MDS plots. PHYLIP 3.695 (Felsenstein 1989) was used for the phylogenetic analysis using 100 bootstrap iterations and maximum parsimony algorithm. The resulting trees from PHYLIP 3.695 were imported into SplitsTree4 (Huson and Bryant 2006) for the construction of phylogenetic networks. To determine the differentiation between clusters, pairwise *F*st values were assessed using GENEPOP version 4.3 (Rousset 2008). Locus level *F*st values and Nei's genetic distance calculations were determined using the R package adegenet (Jombart 2008). Kinship analysis and figure construction used the EMMA algorithm in the R package GAPIT (Lipka et al. 2012).

PGDSpider 2.0.8.2 (Lischer and Excoffier 2012) was used for the conversion of data format used in STRUCTURE version 2.3.4 (Pritchard et al. 2000), PHYLIP 3.965 (Felsenstein 1989), and GENEPOP 4.3 (Rousset 2008). Missing data rate, minor allele frequency (MAF), and heterozygosity ratio were generated in TASSEL and their distributions were calculated in R.

Results

Genotyping of *G. arboreum* germplasm accessions and genome-wide SNP discovery

The GBS assay was performed for 380 *G. arboreum* accessions (Online Resource Table 1) that generated 777,680,809 raw sequence reads, which produced 730,123,644 (93.9 % of raw reads) good barcoded reads. Good barcoded reads were reads with an exact match to one of the barcodes provided and with no unknown reads (Ns) in the sequence following the matched barcodes up to 64 bp. Five accessions had less than 50,000 reads and were eliminated from genetic diversity analysis. The number of reads for the 375 accessions included in the diversity analysis ranged from 115,901 to 3,323,123 with an average of 1,946,864 reads. The 730,123,644 good barcoded reads were collapsed into 39,801,247 unique sequences, which are denoted as tags. Tags that were present in three out of the four sequencing flowcell channels were retained resulting in 4,832,643 filtered tags (12.1 % of all tags). Alignment of filtered tags to the *G. arboreum* reference genome (Li et al. 2014b) resulted in 3,815,473 tags (79.0 % of filtered tags) that aligned to unique positions on the reference genome, 446,566 tags (9.2 % of filtered tags) that aligned to multiple positions, and 570,604 tags (11.8 % of filtered tags) that were not aligned to the reference genome. SNP calling was performed on uniquely mapped tags resulting in the identification of 61,822 SNPs. Filtering of these SNPs with minor allele frequency of 0.01, minimum locus coverage of 0.1, and minimum taxa coverage of 0.8 generated 18,571 SNPs. Most of the SNPs removed by this step had a high level of missing data (Online Resource Fig. 1).

In order to infer the population structure of the sequenced *G. arboreum* accessions using genome-wide SNP data, the 18,571 SNPs were further filtered to generate a set of discriminating SNPs. Using a MAF less than 0.05 over 10,000 rare alleles were removed. Given the fact that the *G. arboreum* accessions are inbred cultivars and that *G. arboreum* is predominantly self-fertilized, highly heterozygous sites with a heterozygosity ratio greater than 0.2 were also eliminated. This step removed over 1000 SNPs which can be associated with sequencing errors or sequence polymorphisms generated from paralogous sites. After filtering, 6224 discriminating SNPs were retained for the 375 accessions (Online Resource Table 2). The locus level distributions of missing data rate, MAF, and heterozygosity for the 61,822 SNPs, 18,571 SNPs, and 6224 SNPs are shown in Online Resource Fig. 1. For the 6224 discriminating SNPs, missing data rate ranged from 0 to 0.152 with most markers having ~0 missing rate. At the taxa level, missing data rate ranged from 0 to 0.675 with most accessions missing between 2 and 60 SNPs. This low level of missing data resulted from the two-step filtration of the original data based on several parameters including MAF, sequencing coverage, and site heterozygosity ratio, which removed many SNPs with high levels of missing data from the initial dataset of 61,822 SNPs. MAF for the 6224 SNPs ranged from 0.05 to 0.5 with a peak at ~0.05 and heterozygosity ratio ranged from 0 to 0.2 with a peak at ~0.

The genome-wide distribution of discriminating SNPs on the 13 *G. arboreum* chromosomes is presented in Fig. 1. Chromosomes 1, 6, and 12 had a greater number of SNPs; whereas, chromosomes 2, 5, and 8 had fewer SNPs. There were 160 SNPs identified from scaffolds, of which the sequences were not anchored to the 13 *G. arboreum* reference chromosomes. The average marker density per Mb for the 13 reference chromosomes was 3.88. The density of SNPs was maximum on chromosome 5 (~5.36/Mb) and minimum on chromosome 2 (~2.41/Mb). The number of SNPs also showed variation within each chromosome. Chromosomes 9 and 12 had the highest peaks where SNPs aggregate within particular regions.

Population structure for the *G. arboreum* accessions

Population structure analysis using the 6224 discriminating SNPs inferred two major sub-populations for the 375 accessions (Fig. 2a). While K values of 2–6 were tested, the delta K method (Online Resource Fig. 2) and visual inspection of the results (Online Resource Fig. 3) showed that K = 2 best explained the population structure.

The majority of the accessions were grouped into a single cluster referred to as Cluster 1, which included 302 accessions, and the second cluster referred to as Cluster 2 consisted of 64 accessions (Online Resource Table 1). The remaining 9 accessions were unassigned because they had almost even membership in each cluster. Different distribution patterns of missing data rate, MAF, and heterozygosity were observed for Cluster 1 and Cluster 2 accessions; however, Cluster 1 showed a similar distribution pattern as observed for the 375 accessions due to the large representation of accessions (302 accessions) in this cluster (Online Resource Fig. 4).

Phylogenetic analysis (Fig. 2b) and multidimensional scaling (Fig. 2c) showed a similar pattern of population structure for the sequenced *G. arboreum* accessions. Kinship analysis also showed a similar clustering pattern of accessions for population structure, where accessions within each cluster exhibited a higher relative kinship

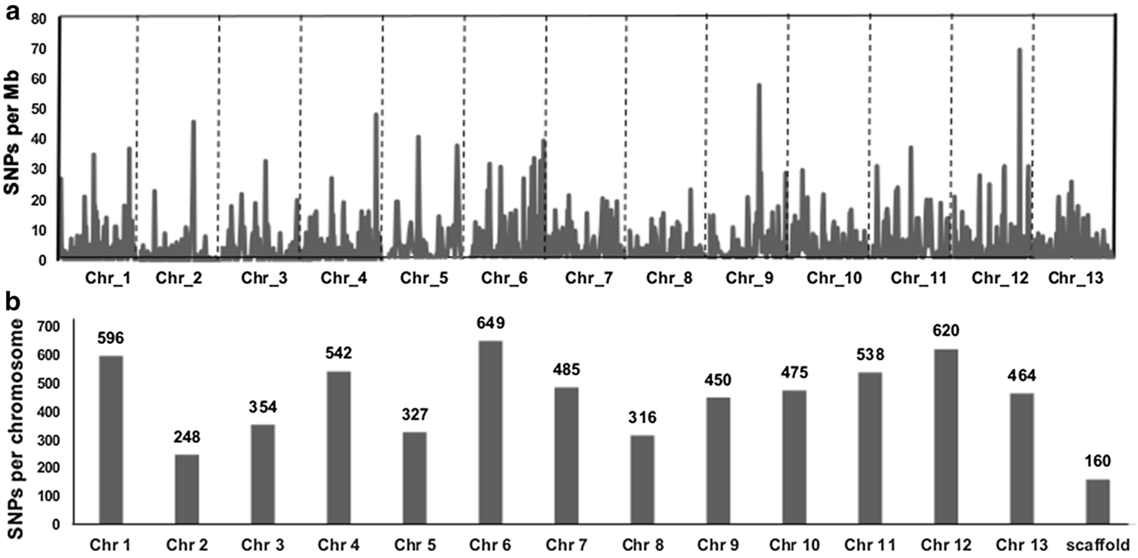

Fig. 1 The chromosomal distribution of 6224 discriminating SNPs used to evaluate genetic diversity for the 375 *Gossypium arboreum* accessions. **a** Discriminating SNPs per Mb; **b** discriminating SNPs per chromosome. The 13 reference *G. arboreum* chromosomes are presented in each panel along the x-axis

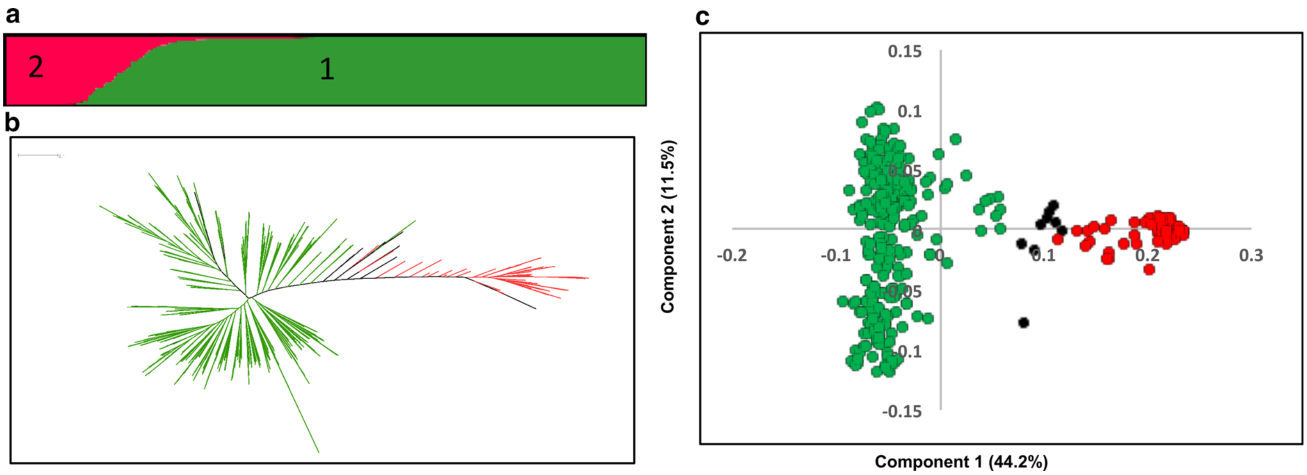

Fig. 2 Population structure classification for 375 sequenced *Gossypium arboreum* germplasm accessions based on STRUCTURE results. **a** *Bar plot* representation of STRUCTURE results indicates two clusters (K = 2) for the sequenced *G. arboreum* accessions; **b** phylogenetic analysis and **c** multidimensional scaling analysis showed similar pattern as the STRUCTURE analysis results. Accessions in Cluster 1 are represented in *green*, Cluster 2 accessions in *red*, and unassigned accessions are in *black*

compared to a lower kinship observed between the two clusters (Fig. 3). Most of the accessions had a relative kinship value between 0.6 and 0.8. Eight pairs of accessions showed a very high relative pairwise kinship with values above 0.95 and included accessions PI417892 with PI615758, PI529643 with PI529738, PI529762 with PI615902, PI529762 with PI616282, PI529780 with PI615743, PI615786 with PI616226, PI615902 with PI616282, and PI616160 with PI616005. Five pairs of these accessions were added to the collection at different times and may represent redundancies. The other pairs were added to the collection at the same time and may represent accessions with a common pedigree. Two accessions (PI529740 and PI630019) included in the study as redundant accessions based on phenotypic data showed a pairwise kinship value 0.92.

To gain more knowledge on the structure stratification of the accessions, each of the clusters was further evaluated by STRUCTURE analysis to dissect the substructure within the clusters. For the 302 Cluster 1 accessions, two sub-clusters were identified with 203 accessions as one sub-cluster, 77 accessions as another sub-cluster, and 22

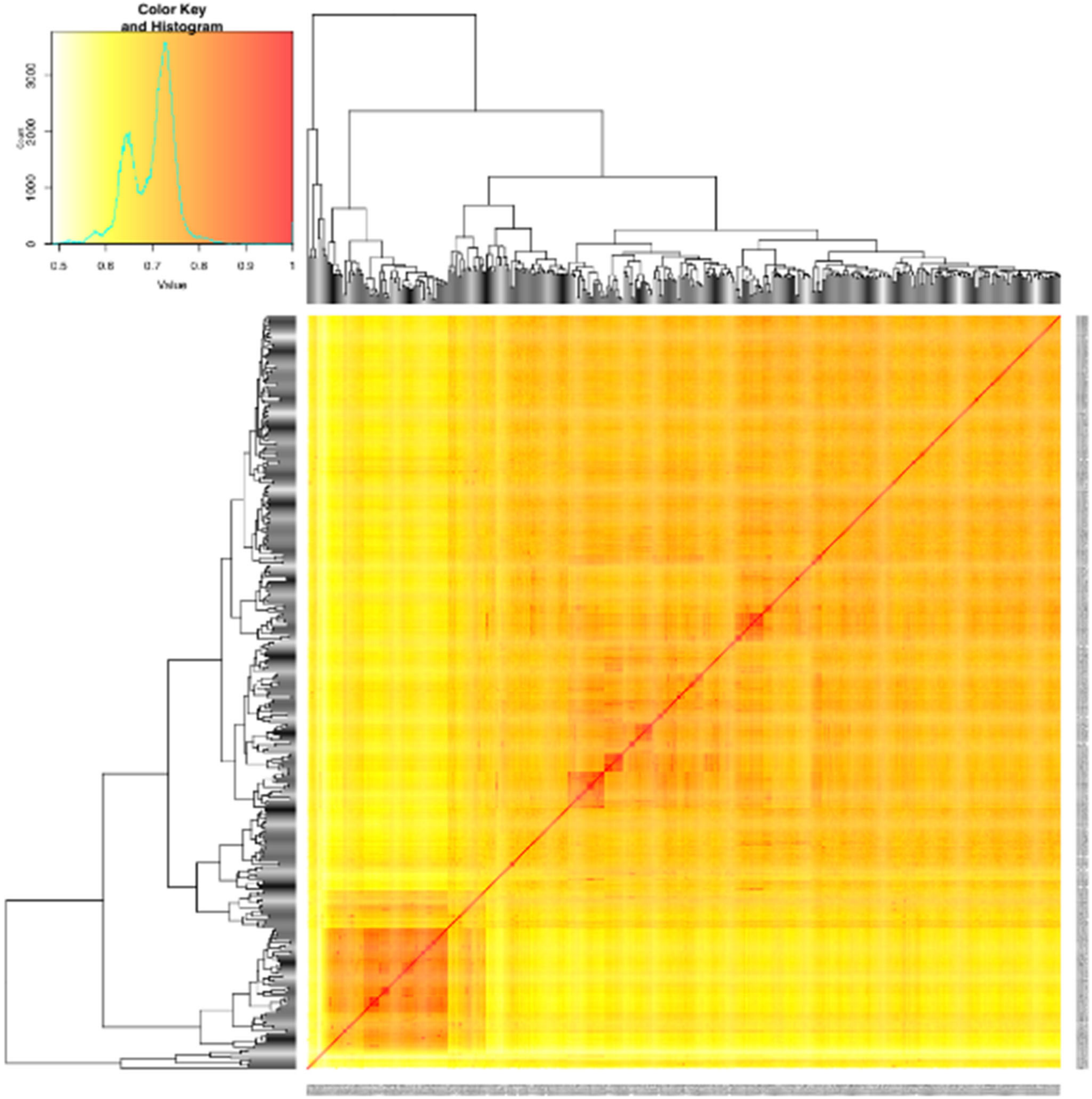

Fig. 3 Kinship matrix among the 375 *Gossypium arboreum* accessions. The majority of the accessions showed high levels of kinship (*yellow* to *red* colored portion of matrix)

accessions unassigned (Fig. 4; Online Resource Fig. 2b; Online Resource Table 1). The 64 Cluster 2 accessions were divided into three sub-clusters containing 10, 40, and 6 accessions with 8 accessions unassigned (Fig. 5; Online Resource Fig. 2c; Online Resource Table 1). Results generated from STRUCTURE, phylogenetic analysis, and multidimensional scaling showed a similar pattern for the sub-clustering (Figs. 4, 5).

To evaluate the degree of differentiation and genetic divergence, pairwise *F*st and Nei's diversity index were calculated for each cluster and sub-cluster (Table 1). A moderate degree of differentiation and divergence were observed between the two major clusters. For the two sub-clusters of Cluster 1 accessions, low levels of genetic differentiation and divergence were found. Low to moderate levels of genetic differentiation and divergence were also

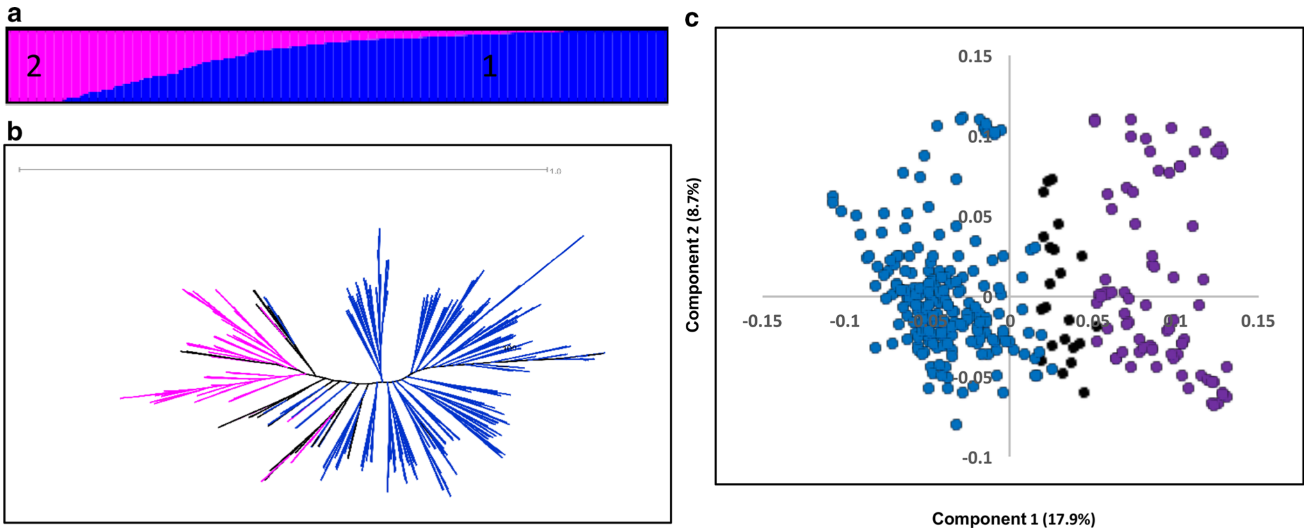

Fig. 4 Population substructure classification for 302 Cluster 1 *Gossypium arboreum* accessions based on STRUCTURE results. **a** *Bar plot* representation of STRUCTURE results indicates two sub-clusters (K = 2) for Cluster 1 accessions; **b** phylogenetic analysis and **c** multidimensional scaling analysis showed similar pattern as the STRUCTURE analysis results. Sub-cluster 1 accessions are represented in *blue*, sub-cluster 2 accessions in *pink*, and unassigned accessions in *black*

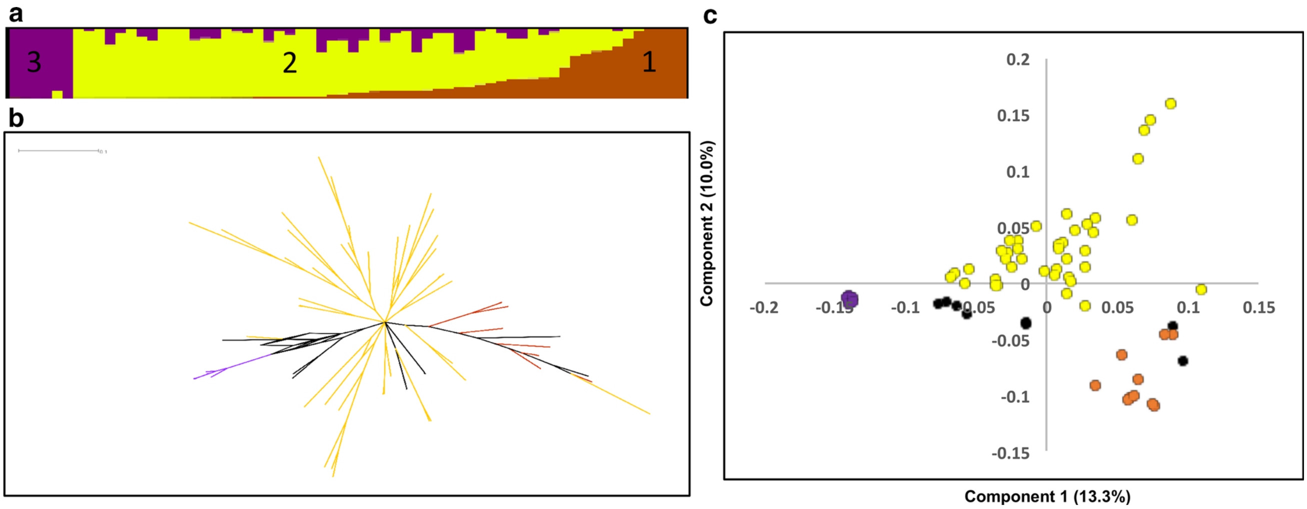

Fig. 5 Population substructure classification for 64 Cluster 2 *Gossypium arboreum* accessions based on STRUCTURE results. **a** *Bar plot* representation of STRUCTURE results indicates three sub-clusters (K = 3) for Cluster 2 accessions; **b** phylogenetic analysis and **c** multidimensional scaling analysis showed similar pattern as the STRUCTURE analysis results. Sub-cluster 1 accessions are represented in *brown*, sub-cluster 2 accessions in *yellow*, sub-cluster 3 accessions in *purple*, and unassigned accessions in *black*

Table 1 Pairwise *F*st values (below diagonal) and Nei's genetic distance (above diagonal) matrix among clusters and sub-clusters for the 375 *Gossypium arboreum* accessions

	C1	C2	C1_sub1	C1_sub2	C2_sub1	C2_sub2	C2_sub3
C1	–	0.12	–	–	–	–	–
C2	0.16	–	–	–	–	–	–
C1_sub1	–	–	–	0.05	–	–	–
C1_sub2	–	–	0.05	–	–	–	–
C2_sub1	–	–	–	–	–	0.07	0.37
C2_sub2	–	–	–	–	0.07	–	0.09
C2_sub3	–	–	–	–	0.19	0.13	–

C1 Cluster 1, *C2* Cluster 2, *C1_sub1* sub-cluster 1 and *C1_sub2* sub-cluster 2 for Cluster 1 accessions, *C2_sub1* sub-cluster 1, *C2_sub2* sub-cluster 2, and *C2_sub3* sub-cluster 3 for Cluster 2 accessions observed for sub-clusters of Cluster 2 accessions; however, relatively higher values of *F*st and Nei's genetic distance suggested sub-cluster 3 as the most differentiated subpopulation. Locus level *F*st distribution was illustrated in Online Resource Fig. 5 and showed low levels of genetic differentiation with 44.4 % of the *F*st values less than 0.05 and 12.0 % greater than or equal to 0.5. Fourteen loci had *F*st values greater than 0.9 suggesting different fixation at these loci between the two populations. A blast search of a 1 kb region at these loci identified seven mRNA sequences involved in plant growth and signaling transduction, which may play a role in the differentiation between the two clusters.

Discussion

There is increasing interest in using *G. arboreum* germplasm in breeding programs to transfer desirable traits to *G. hirsutum*, due to the lack of tolerance and/or resistance of *G. hirsutum* cultivars to abiotic and biotic stresses (Maqbool et al. 2010; Robinson 2007). In the United States, reniform nematode (*Rotylenchulus reniformis*) has emerged as a major disease challenge (Robinson 2007) and *G. arboreum* accessions have been identified as a source of resistance (Erpelding and Stetina 2013; Sacks and Robinson 2009). Resistance to leaf curl virus is absent in *G. hirsutum* and transferring genes from *G. arboreum* demonstrated higher resistance levels (Ahmad et al. 2011). Introgression of traits between the two species is difficult; therefore, more information is needed on the genetic diversity of the *G. arboreum* germplasm collection to aid in the selection of accessions for breeding. In this study, a subset 380 *G. arboreum* accessions were sampled from the collection to evaluate genetic diversity and population structure. These accessions included landraces added to the collection prior to 1970 and breeding lines or varieties more recently added to the collection. In order to generate a large number of molecular markers, GBS was conducted on the accessions. More than 777 million sequencing reads were generated for the accessions. Five accessions with low read counts were eliminated from the diversity analysis; whereas, the remaining 375 accessions averaged nearly 2 million reads. To identify discriminating SNPs, all highly heterozygous sites were excluded to eliminate error-prone SNPs and spurious SNPs detected from paralogues sites (Glaubitz et al. 2014) resulting in 6224 SNPs selected for genetic diversity and population structure analyses. The 6224 discriminating SNPs were distributed across the 13 chromosomes of the *G. arboreum* reference genome with inter- and intra-chromosomal variation observed, which has also been reported for GBS assays in *G. hirsutum* (Thyssen et al. 2015).

Population structure analysis for the 375 accessions inferred two major clusters. The differentiation between the two major clusters was demonstrated by a pairwise *F*st value of 0.12 and Nei's genetic distance of 0.16. The 302 accessions in Cluster 1 were further divided into two subclusters with 203 accessions in a single sub-cluster. These results would suggest a narrow genetic base for Cluster 1 accessions, which was supported by small *F*st and Nei's genetic distance values. The majority of the accessions in this cluster were more recently added to the collection representing varieties or breeding lines with a narrow genetic base resulting in lower genetic diversity. Similarly, Kebede et al. (2007) reported low to moderate genetic diversity with a mean value of 0.89 for *G. arboreum* germplasm accessions.

Greater genetic diversity was observed for the 64 Cluster 2 accessions and this cluster was further divided into three sub-clusters with low to moderate differentiation between sub-clusters. In accordance, many of the accessions in Cluster 2 were some of the first accessions added to the collection and higher levels of genetic diversity are expected. Nine accessions had nearly equal membership in the two major clusters suggesting that they may be hybrids between the two major genetic groups.

Data generated from this study can aid in the management of the germplasm collection, such as the identification of redundancies or misclassified accessions and the selection of accessions for the development of a core collection. In the present study, PI630019 and PI529740 were included as potential redundant accessions, because the accessions were phenotypically similar and exhibited the ovule fiberless phenotype. Accession PI630019 is listed as *G. herbaceum*; however, PI630019 and *G. arboreum* accession PI529740 clustered together suggesting genetic similarity. These two accessions have a pairwise kinship value of 0.92, indicating their high genetic similarity. Any genetic variation between the two accessions could have resulted from genetic drift. The data suggested that PI630019 should be classified as *G. arboreum* and considered a duplicate of PI529740. High pairwise kinship values were also observed for an additional eight pairs of accessions and further evaluation of these accessions will be required to determine if they are redundancies. Only one accession in the present study may be misclassified. Accession PI408751 is phenotypically similar to *G. herbaceum* accessions; however, this accession clustered with other *G. arboreum* accessions in Cluster 1. Because of the genetic similarity between *G. arboreum* and *G. herbaceum* (Kebede et al. 2007) and no barriers to hybridization between the species, accessions for these two species may cluster together. Genetic characterization of the *G. herbaceum* collection would be desirable.

Extensive phenotypic characterization of the 375 accessions in the present study has not been conducted to determine if a similar clustering pattern is observed. However, Stanton et al. (1994) reported few recognizable groupings for 169 accession evaluated using 53 phenotypic characteristics. In the present study, there are several groups of accessions that are phenotypically similar, but GBS data indicated they are genetically different. Seven phenotypically similar accessions (PI529708, PI529712, PI452096, PI529774, PI529776, PI615742, and PI615747) clustered in the more genetically diverse sub-cluster 3 of Cluster 2 with one accession unassigned in Cluster 2 suggesting phenotypically similar accessions may not be genetically similar.

Compared to genetic diversity studies conducted with SSR or RFLP markers, the highly abundant distribution of SNPs enables the capture of sequence polymorphism across the whole genome. While SNP chips are an alternative technology, their application is limited to a few species for which SNP chips are available. Additionally, the development of SNP chips for a specific organism is time consuming and expensive. In contrast, GBS does not reply on previous knowledge of SNPs and has the ability to detect more uncommon or rare alleles when increasing the number of investigated samples at high confidence level (Romay et al. 2013). Also, there is a strong argument that rare alleles play a major role in quantitative traits, which explains a fraction of the missing heritability. For GBS assays, missing data and sequencing errors commonly occur and can lead to biased results for association studies (Glaubitz et al. 2014); genotype imputation is frequently used to boost the number of SNPs for GWAS analysis (Marchini and Howie 2010). Nevertheless, for the kinds of analyses conducted in this study, imputation of missing data is usually not necessary (Glaubitz et al. 2014), especially considering that no significant differences in population structure were reported for imputed and unimputed data (Samad et al. 2015) and no reference samples are available for *G. arboreum*. Thus, special attention should be paid to the few accessions with high levels of missing data, as their cluster assignment can be biased due to their low level of call rate.

The present study provides a more comprehensive evaluation of the NPGS *G. arboreum* germplasm collection, which included more genotypes and a greater number of molecular markers as compared to previous genetic diversity analyses (Kantartzi et al. 2009; Kebede et al. 2007). SSR markers have been frequently used for genetic diversity studies and genetic diversity assessments can be conducted with a small number (30–50) of SSR markers; however, similar results have been reported from comparisons between SSR and SNP markers (Filippi et al. 2015) and the cost of genotyping would be similar to GBS. GBS has the advantage in that large data sets of sequencing data are generated with genome-wide coverage that can be used for other applications. The SNP data will be useful for the development of molecular markers for QTL mapping of important traits and marker-assisted selection for the introgression of traits into *G. hirsutum* cultivars. Specifically, the *G. arboreum* collection is being evaluated for reniform nematode resistance (Erpelding and Stetina 2013) with the introgression of resistance into *G. hirsutum*. The genetic diversity and population structure data will serve as an important foundation for the selection of accessions for phenotypic evaluation to enhance the utilization of the collection for cotton improvement. Further, data generated from this study can be used in GWAS and genomic selection studies when various phenotypic data are incorporated.

Acknowledgments This research was funded by the USDA-ARS project 6066-22000-074-00D. This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DOE) and the US Department of Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, ARS, DOE, or ORAU/ORISE. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

Compliance with ethical standards

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

References

Ahmad S, Mahmood K, Hanif M, Nazeer W, Malik W, Qayyum A, Hanif K, Mahmood A, Islam N (2011) Introgression of cotton leaf curl virus-resistant genes from Asiatic cotton (*Gossypium arboreum*) into upland cotton (*G. hirsutum*). Genet Mol Res 10:2404–2414

Earl DA, von Holdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Erpelding JE, Stetina SR (2013) Genetics of resistance to *Rotylenchulus reniformis* in *Gossypium arboreum* accession PI529728. World J Agric Res 1:48–53

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

Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5:164–166

Filippi CV, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, Moreno MV, Fusari CM, Alvarez D, Heinz RA, Hopp HE, Paniego NB, Lia VV (2015) Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC Plant Biol 15:52

Gill MS, Bajaj YPS (1987) Hybridization between diploid (*Gossypium arboreum*) and tetraploid (*Gossypium hirsutum*) cotton through ovule culture. Euphytica 36:625–630

Girma G, Hyma KE, Asiedu R, Mitchell SE, Gedil M, Spillane C (2014) Next-generation sequencing based genotyping, cytometry and phenotyping for understanding diversity and evolution of guinea yams. Theor Appl Genet 127:1783–1794

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346

Govindaraj M, Vetriventhan M, Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. Genet Res Int 2015:431487

Guo WZ, Zhou BL, Yang LM, Wang W, Zhang TZ (2006) Genetic diversity of landraces in *Gossypium arboreum* L. race *sinense* assessed with simple sequence repeat markers. J Integr Plant Biol 48:1008–1017

He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484. doi:10.3389/fpls.2014.00484

Hedin PA, Jenkins JN, Parrott WL (1992) Evaluation of flavonoids in *Gossypium aboreum* (L) cottons as potential source of resistance to tobacco budworm. J Chem Ecol 18:105–114

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Iqbal MA, Abbas A, Zafar Y, Rahman MU (2015) Characterization of indigenous *Gossypium arboreum* L. genotypes for various fiber quality traits. Pak J Bot 47:2347–2354

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Jombart T (2008) *adegenet*: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405

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

Kebede H, Burow G, Dani RG, Allen RD (2007) A-genome cotton as a source of genetic variability for upland cotton (*Gossypium hirsutum*). Genet Resour Crop Evol 54:885–895

Kujur A, Bajaj D, Upadhyaya HD, Das S, Ranjan R, Shree T, Saxena MS, Badoni S, Kumar V, Tripathi S, Gowda CLL, Sharma S, Singh S, Tyagi AK, Parida SK (2015) A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. Sci Rep 5:11166

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Li C, Waldbieser G, Bosworth B, Beck BH, Thongda W, Peatman E (2014a) SNP discovery in wild and domesticated populations of blue catfish, *Ictalurus furcatus*, using genotyping-by-sequencing and subsequent SNP validation. Mol Ecol Resour 14:1261–1270

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S (2014b) Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet 46:567–572

Lin M, Cai S, Wang S, Liu S, Zhang G, Bai G (2015) Genotyping-by-sequencing (GBS) identified SNP tightly linked to QTL for pre-harvest sprouting resistance. Theor Appl Genet 128:1385–1395

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Lischer HE, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28:298–299

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 Resour Crop Evol 53:1145–1152

Logan-Young CJ, Yu JZ, Verma SK, Percy RG, Pepper AE (2015) SNP discovery in complex allotetraploid genomes (*Gossypium* spp., Malvaceae) using genotyping by sequencing. Appl. Plant Sci 3:1400077. doi:10.3732/apps.1400077

Maqbool A, Abbas W, Rao AQ, Irfan M, Zahur M, Bakhsh A, Riazuddin S, Husnain T (2010) *Gossypium arboreum GHSP26* enhances drought tolerance in *Gossypium hirsutum*. Biotechnol Prog 26:21–25

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511

Mehetre SS, Aher AR, Gawande VL, Patil VR, Mokate AS (2003) Induced polyploidy in *Gossypium*: a tool to overcome interspecific incompatibility of cultivated tetraploid and diploid cottons. Cur Sci 84:1510–1512

Pace J, Gardner C, Romay C, Ganapathysubramanian B, Lübberstedt T (2015) Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genom 16:47

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Gen 5:92–102

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

Robinson AF (2007) Reniform in U.S. cotton: when, where, why, and some remedies. Annu Rev Phytopathol 45:263–288

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55. doi:10.1186/gb-2013-14-6-r55

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8:103–106

Sacks EJ, Robinson AF (2009) Introgression of resistance to reniform nematode (*Rotylenchulus reniformis*) into upland cotton (*Gossypium hirsutum*) from *Gossypium arboreum* and a *G. hirsutum/Gossypium aridum* bridging line. Field Crop Res 112:1–6

Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG (2015) Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. PLoS Genet 11:e1005131

Stanton MA, Stewart JM, Tugwell NP (1992) Evaluation of *Gossypium arboreum* L. germplasm for resistance to thrips. Genet Resour Crop Evol 39:89–95

Stanton MA, Stewart JM, Percival AE, Wendel JF (1994) Morphological diversity and relationships in the A-genome cottons, *Gossypium arboreum* and *G. herbaceum*. Crop Sci 34:519–527

Thyssen GN, Fang DD, Turley RB, Florane C, Li P, Naoumkina M (2015) Mapping-by-sequencing of Ligon-lintless-1 (*Li_1*) reveals a cluster of neighboring genes with correlated expression in developing fibers of upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet 128:1703–1712

Ulukan H (2011) Plant genetic resources and breeding: current scenario and future prospects. Int J Agric Biol 13:447–454

Wallace TP, Bowman D, Campbell BT, Chee P, Gutierrez OA, Kohel RJ, McCarty J, Myers G, Percy R, Robinson F, Smith W, Stelly DM, Stewart JM, Thaxton P, Ulloa M, Weaver DB (2009) Status of the USA cotton germplasm collection and crop vulnerability. Genet Resour Crop Evol 56:507–532

Wallace JG, Upadhyaya HD, Vetriventhan M, Buckler ES, Hash CT, Ramu P (2015) The genetic makeup of a global barnyard millet germplasm collection. Plant Gen. doi:10.3835/plantgenome2014.10.0067

Wong MML, Gujaria-Verma N, Ramsay L, Yuan HY, Caron C, Diapari M, Vandenberg A, Bett KE (2015) Classification and characterization of species within the genus *Lens* using genotyping-by-sequencing (GBS). PLoS One 10:e0122025

Yik CP, Birchfield W (1984) Resistant germplasm in *Gossypium* species and related plants to *Rotylenchulus reniformis*. J Nematol 16:146–153