Theor Appl Genet (2015) 128:313–327
DOI 10.1007/s00122-014-2431-7

ORIGINAL PAPER

Molecular characterization of the *Gossypium* Diversity Reference Set of the US National Cotton Germplasm Collection

Lori L. Hinze · David D. Fang · Michael A. Gore · Brian E. Scheffler · John Z. Yu · James Frelichowski · Richard G. Percy

Received: 5 December 2013 / Accepted: 14 November 2014 / Published online: 28 November 2014

Abstract

Key message **A core marker set containing markers developed to be informative within a single commercial cotton species can elucidate diversity structure within a multi-species subset of the** ***Gossypium*** **germplasm collection.**

Abstract An understanding of the genetic diversity of cotton (*Gossypium* spp.) as represented in the US National Cotton Germplasm Collection is essential to develop strategies for collecting, conserving, and utilizing these germplasm resources. The US collection is one of the largest world collections and includes not only accessions with improved yield and fiber quality within cultivated species, but also accessions possessing sources of abiotic and biotic stress resistance often found in wild species. We evaluated the genetic diversity of a subset of 272 diploid and 1,984 tetraploid accessions in the collection (designated the *Gossypium* Diversity Reference Set) using a core set of 105 microsatellite markers. Utility of the core set of markers in differentiating intra-genome variation was much greater in commercial tetraploid genomes (99.7 % polymorphic bands) than in wild diploid genomes (72.7 % polymorphic bands), and may have been influenced by pre-selection of markers for effectiveness in the commercial species. Principal coordinate analyses revealed that the marker set differentiated interspecific variation among tetraploid species, but was only capable of partially differentiating among species and genomes of the wild diploids. Putative species-specific marker bands in *G. hirsutum* (73) and *G. barbadense* (81) were identified that could be used for qualitative identification of misclassifications, redundancies, and introgression within commercial tetraploid species. The results of this broad-scale molecular characterization are essential to the management and conservation of the collection and provide insight and guidance in the use of the collection by the cotton research community in their cotton improvement efforts.

Communicated by Alan H. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2431-7) contains supplementary material, which is available to authorized users.

L. L. Hinze · J. Z. Yu · J. Frelichowski · R. G. Percy (✉)
Crop Germplasm Research Unit, Southern Plains Agricultural
Research Center, USDA-ARS, College Station, TX, USA
e-mail: richard.percy@ars.usda.gov

D. D. Fang
Cotton Fiber Bioscience Research Unit, Southern Regional
Research Center, USDA-ARS, New Orleans, LA, USA

M. A. Gore
Plant Breeding and Genetics Section, School of Integrative Plant
Science, Cornell University, Ithaca, NY, USA

B. E. Scheffler
Genomics and Bioinformatics Research Unit, Jamie Whitten
Delta States Research Center, USDA-ARS, Stoneville, MS, USA

Introduction

The *Gossypium* genus is composed of more than 50 species of differing ploidy levels and contains a wealth of genetic variability ranging from wild diploid species to highly improved allotetraploid species (Fryxell 1992; Wendel and Cronn 2003). The US National Cotton Germplasm Collection (NCGC) contains much of the diversity of the genus—with 10,276 accessions originating from five continents and even some endemic to several tropical islands (Campbell et al. 2010). This diversity has been documented first with passport and pedigree information (germplasm information is available online at Germplasm Resources Information Network, http://www.ars-grin.gov/) and further through routine characterization of accessions using a descriptor list of morphological traits (IBPGR 1980; Percival 1987; Toll 1995). Although replicated agronomic evaluations in different US environments are possible for improved, non-photoperiodic accessions, the genotyping of accessions provides enhanced power to detect and document diversity among highly similar accessions without environmental influence (Tanksley and McCouch 1997). Using a set of standardized molecular markers, groups of accessions can be characterized at different times depending on funding and relevancy, and at a later date molecular data from these groups can be combined, analyzed, and compared as a whole. Characterization of the NCGC at the DNA level would also help its management by identifying redundancy (i.e. duplication), as well as monitoring the integrity of accessions in the collection.

Molecular characterization of cotton germplasm has lagged behind other major crops because of the slow progress in developing cost-efficient and highly polymorphic DNA markers (Kohel et al. 2001; Lacape et al. 2007; Rungis et al. 2005). Other obstacles have been the varying ploidy levels, genome sizes, and demography of *Gossypium* species that required developing a large set of widely applicable easy-to-use markers to enable evaluations across all species (Han et al. 2006; Lacape et al. 2003; Zhang et al. 2013). In recent years, significant strides have been made in the development and validation of numerous highly polymorphic simple sequence repeat (SSR) markers in *G. barbadense* and *G. hirsutum* (Blenda et al. 2012; Yu et al. 2012b). Cotton SSR markers have been developed from microsatellite-enriched genomic libraries (Hoffman et al. 2007; Nguyen et al. 2004; Reddy et al. 2001; Xiao et al. 2009), expressed sequence tag (EST) libraries (Guo et al. 2007; Jena et al. 2012; Park et al. 2005; Xiao et al. 2009; Yu et al. 2011), and bacterial artificial chromosome (BAC) libraries (Frelichowski Jr. et al. 2006; Guo et al. 2008) and directly mining the genome sequences (Wang et al. 2012). At present, CottonGen (http://www.cottongen.org), the cotton community genomics, genetics and breeding database—a consolidation and expansion of data from the Cotton Genome Database (http://www.cottondb.org) and the Cotton Marker Database (http://www.cottonmarker.org)—contains about 20,000 SSR markers.

With the development and validation of more SSR markers, genetic diversity studies have been pursued with varying objectives, including understanding the evolutionary process of interspecific gene flow within *G. aridum* (Alvarez and Wendel 2006), the genetic relationships of geographically diverse *G. arboreum* cultivars (Kantartzi et al. 2009; Liu et al. 2006), and the utility of microsatellite primers developed in a tetraploid species (i.e. G. *hirsutum*) and applied to a diploid species (i.e. G. *davidsonii*) (Han et al. 2004; Kuester and Nason 2012). Since cultivars possess the most readily accessible genetic variation for use in breeding programs, a number of genetic diversity investigations have occurred using this type of Upland cotton germplasm (Fang et al. 2013; Hinze et al. 2012; Rungis et al. 2005; Zhang et al. 2013). Many of these diversity studies have characteristically been limited in the scope of accessions investigated or the number of markers used. There has been minimal standardization in these studies, as different sets of less characterized or randomly selected SSR markers were used for each set of accessions.

Attempts to characterize larger genomic groups using a standardized marker set have been made in the cotton germplasm collections of Uzbekistan (Abdurakhmonov et al. 2008) and France (Lacape et al. 2007). Recognizing the need to describe the molecular diversity within the NCGC (one of the largest cotton collections in the world) and to determine the utility of markers in managing its diversity, this project was initiated to characterize a major portion of the NCGC utilizing SSR markers. In initiating this effort, it was recognized that the development and use of a standardized core marker set would allow for comparative studies to be performed in different germplasm collections and, using this information, to build more diverse, safer, backed up collections through exchange. A core set of 105 highly polymorphic SSR markers that uniformly cover the tetraploid genome was developed to accomplish this task (Yu et al. 2012a). These markers were characterized by being amenable to high-throughput assay via multiplex polymerase chain reaction (PCR) bins on high-resolution automated genetic analyzers, making them an excellent genetic marker system for large-scale germplasm characterization. A potential limitation of any core marker set developed for cotton is its possible limited informative capabilities across a wide range of diploid and tetraploid species, due to the majority of markers being initially developed to be polymorphic and informative among individuals of a single commercial tetraploid species, *G. hirsutum*. A complementary goal of the present investigation has been to establish the capabilities of a core marker set in elucidating genetic structure across the *Gossypium* genus.

The ultimate objective of this research is to increase the utilization efficiency of the NCGC by providing a molecular description of its accessions. Molecular marker data would provide additional bases for managing and conserving the diversity of the NCGC. Also of particular interest is the understanding of the genetic structure of the NCGC and its potential exploitation for cotton cultivar improvement. To that end, the core marker set was used to genotype and analyze a set of 2,256 cotton accessions known as the

Gossypium Diversity Reference Set (GDRS). Comparisons were made by genome groups to determine the discriminating power of the core marker set and to analyze the structure of genetic diversity in the NCGC. Ancillary objectives included determining the efficacy of the SSR markers in identifying possible introgression and redundancies in the commercial tetraploid species of the NCGC. Our intention is that this broad-scale molecular analysis will demonstrate the importance of characterization for the management and conservation of the NCGC and for the cotton research community in their cotton improvement efforts.

Materials and methods

Gossypium accessions

Two thousand two hundred fifty-six (2,256) *Gossypium* accessions were selected from the NCGC to make up the GDRS for standardized molecular description. These accessions represent approximately 22 % of the collection and encompass all nine cytogenetic genome groups and 33 species (Table 1). The tetraploid cytogenetic group (AD) is composed of five species and is represented in the GDRS by 1,984 accessions. The eight cytogenetic groups (A, B, C, D, E, F, G, and K) of the diploid genomes are collectively represented by 272 accessions from 28 species. The data set was heavily biased towards tetraploid accessions, specifically *G. hirsutum* and *G. barbadense*; however, this reflects the distribution of accessions available within the NCGC (Table 1) and the economic importance of Upland cotton (*G. hirsutum*) and Pima cotton (*G. barbadense*). Accessions within taxons were selected to capture maximum diversity of phenotype and geographic distribution within the taxon, as known from passport data. Information about these accessions (plant introduction number, accession name, genome group, species group, and geographic origin) is given in Online Resource 1.

Genomic DNA extraction

For each accession, a sample of seed was germinated on sterile water agar (9 g agar in 1 L distilled water) media in an incubator at 36 °C until root tips emerged approximately 1 cm. Root tips from ten seeds were bulked to capture the diversity of each accession. DNA was extracted and column-purified following a modified protocol using the E-Z 96® Plant DNA Kit from Omega (Omega Bio-tek, Norcross, Georgia USA) (Fang et al. 2010, 2013). After purification, DNA was quantified using a NanoDrop2000 (Thermo Scientific, Wilmington, DE, USA) and normalized at 50 ng/μL.

Table 1 Distribution of accessions in the US National Cotton Germplasm Collection (NCGC) and in the subset of accessions known as the *Gossypium* Diversity Reference Set (GDRS)

Genome	Species	No. accessions	
		NCGC[a]	GDRS[b] (%)
Diploid			
A	*G. arboreum*	1,729	145 (8.4)
	G. herbaceum	194	49 (25.3)
B	*G. anomalum*	7	5 (71.4)
C	*G. nandewarense*	6	1 (16.7)
	G. sturtianum	7	3 (42.9)
D	*G. aridum*	14	4 (28.6)
	G. armourianum	10	2 (20.0)
	G. davidsonii	31	9 (29.0)
	G. gossypioides	7	7 (100.0)
	G. klotzschianum	59	1 (1.7)
	G. laxum	2	1 (50.0)
	G. lobatum	4	1 (25.0)
	G. raimondii	56	4 (7.1)
	G. thurberi	37	9 (24.3)
	G. trilobum	11	6 (54.5)
E	*G. areysianum*	2	1 (50.0)
	G. somalense	3	2 (66.7)
	G. stocksii	4	2 (50.0)
F	*G. longicalyx*	4	4 (100.0)
G	*G. australe*	11	3 (27.3)
	G. bickii	5	4 (80.0)
	G. nelsonii	4	3 (75.0)
K	*G. costulatum*	2	1 (50.0)
	G. exiguum	1	1 (100.0)
	G. marchantii	1	1 (100.0)
	G. nobile	1	1 (100.0)
	G. populifolium	4	1 (25.0)
	G. pulchellum	1	1 (100.0)
Tetraploid			
AD	*G. barbadense*	1,584	430 (27.1)
	G. darwinii	138	4 (2.9)
	G. hirsutum	6,302	1,541 (24.5)
	G. mustelinum	19	7 (36.8)
	G. tomentosum	16	2 (12.5)
Overall			
9	33	10,276	2,256 (22.0)

[a] Number of accessions in the NCGC based on Campbell et al. (2010)

[b] Percent of accessions sampled from the NCGC is indicated in parentheses

SSR genotyping

A core set of 105 well characterized and genetically mapped SSR markers was identified to analyze the cotton genome at a frequency of two markers per chromosome arm (Yu et al. 2012a). These markers were primarily identified on the basis of a uniform distribution across the 26 tetraploid cotton chromosomes (Wang et al. 2013; Blenda et al. 2012; Yu et al. 2012b; Zhao et al. 2012) with additional criteria including purported lack of locus duplication, PCR reproducibility, polymorphism information content (PIC), and source representation. These markers, many of which are single copy, were multiplexed prior to PCR to form 35 sets of reaction bins (Yu et al. 2012a). Multiplex PCR was performed according to Fang et al. (2010). DNA fragments were separated using an ABI 3730XL Genetic Analyzer (Life Technologies, Carlsbad, CA, USA), and SSR data acquisition was accomplished using the software GeneMapper 4.0 (Life Technologies, Carlsbad, CA, USA) with further verification by manual checking of all SSR products. Due to the tendency of some of these SSRs to amplify duplicate loci from both A and D subgenomes of the tetraploid species, and the presence of multiple alleles (heterogeneous and heterozygous nature of ten individuals in a bulked sample), these primer pairs often yielded multiple PCR products in an individual accession sample. Such circumstances generated a complexity that a marker fragment could not be unambiguously assigned to a locus. Consideration of these factors led us to score and analyze our SSR data like a dominant marker system where the fragment data were recorded as ‘1’ (present, *AA* or *Aa* genotype), ‘0’ (absent, *aa* genotype), or ‘−1’ (missing).

Descriptive diversity statistics

To estimate genetic diversity and differentiation of genomes and species, the summary statistics were computed using GenAlEx, version 6.5b3 (Peakall and Smouse 2012) software and manual calculations in an Excel 2010 spreadsheet. The ‘Frequency…’ option and ‘binary (diploid)’ data format in GenAlEx were used to calculate descriptive statistics based on phenotypic (i.e. band presence/absence) data for individual accessions, genome groups, and species groups.

Genome- and species-specific bands were identified using a methodology developed in an unpublished pilot study that included five tetraploid species and the diploid species *G. arboreum*, *G. herbaceum*, and *G. raimondii.* Markers with genome- or species-specific bands in the pilot study were also found to have specific bands in the current study. To avoid the false identification of genomic and species-specific bands due to sampling error resulting from small sample sizes, band specificity calculations were only implemented for A, D, and AD genomes and the *G. barbadense* and *G. hirsutum* species. Markers that did not generate PCR product in 50 % of the accessions within a genome or species group were eliminated from consideration. Bands were identified that occurred at a frequency of 50 % or greater in one genome or species, but whose occurrence was lower than 10 % in another genome or species in pairwise comparisons. These bands were considered to be species- or genome- specific for the particular pairwise comparison. Shared bands among these genomes were also assessed. Shared bands were defined as those bands present at a frequency of at least 10 % within all groups being compared.

In a dominant marker system, the PIC value is equivalent to calculations of genetic diversity (Weir 1996). This value was calculated as:

$$\mathrm{PIC}_b = D_b = 1 - (p^2 + q^2)$$

where p is the frequency of the band presence and q the frequency of band absence of the bth band, and PIC_b is the polymorphism information content (equivalent to D_b, genetic diversity) of band b. PIC for dominant marker bands is a maximum of 0.5 when $p = q = 0.5$. Estimates of genetic diversity (or PIC) are calculated for each band, and the mean over all bands in a given group is the overall estimate of diversity for the genome of interest.

Comparative diversity statistics

When calculating the following diversity statistics, we did not eliminate markers with a rare band frequency of ≤ 5 % as is frequently practiced in diversity studies, because a primary objective of this study was to characterize the extent of allelic diversity in the GDRS. To have filtered data based on an allele frequency threshold would have eliminated taxonomically informative bands that were unique (or private) to a *Gossypium* genome or species represented in the collection by 112 or fewer accessions (5 % of 2,254 total accessions).

Principal coordinate analysis (PCoA) was applied to discover and plot patterns of data structure using distance-type measures. A pairwise matrix of genetic similarity (GS) values was calculated in NTSYS (Rohlf 2000) using Jaccard’s coefficient (Jaccard 1908) which is commonly applied with dominant-type data where allele frequencies cannot be calculated (Reif et al. 2005). The Jaccard coefficient (J) between two genotypes was calculated by:

$$J = a/(a + b + c),$$

where ‘a’ is the number of bands common to both accessions, ‘b’ is the number of bands only present in accession 1, and ‘c’ is the number of bands only present in accession 2. This matrix was double-centered (DCENTER module) by subtracting the row and column means of the matrix from its elements and adding the grand mean. This standardized matrix was used in the EIGEN module to calculate eigenvectors which were plotted for PCoA. PCoA was conducted separately on four sets of data: the nine *Gossypium* genome groups, the five tetraploid species groups, the eight diploid genome groups, and the diploid genomes minus the A genome accessions. For the tetraploids, the data set was reduced to 1,982 accessions with all bands remaining for analysis. For the eight diploid genomes, the data set was reduced to 272 accessions, and 21 bands were removed that were monomorphic within this set. When the A genome was removed from the diploid data set, 78 accessions and 950 bands remained for the PCoA.

Results

Descriptive diversity statistics

Due to multiple failures of amplification of one SSR marker and missing data for two accessions, data were collected on 2,254 accessions using 104 SSR markers (Online Resource 1 and 2). All 104 SSR markers were polymorphic, and 1,702 unique bands were observed across the data set, with 1,362 bands in the AD genome, the most, and 86 in the F genome, the least (Table 2). For each SSR marker, the overall number of bands ranged from four (CIR169) to 34 bands (TMB2295; Online Resource 2).

The percentage of polymorphic bands in the nine genomic groups varied from a high of 98.2 % (D genome) to a low of 17.4 % (F genome) among the diploids, and equaled 99.7 % for the tetraploid AD genome (Table 2). This statistic compared the number of polymorphic bands to the total number of bands that amplified in accessions of a given genome, and indicated that the SSR markers worked well in identifying polymorphism among the tetraploid accessions and among their A and D genome progenitors. However, there was a dramatic reduction in their ability to detect polymorphisms among other diploid accessions, particularly in the African B and F genomes.

The overall average PIC value or genetic diversity was 0.08 and, when averaged within genomic groups, ranged from a high of 0.07 in the AD genome to a low of 0.003 in the F genome (Table 2). PIC values were very low in this study relative to values reported in other studies of cotton germplasm collection diversity (Abdurakhmonov et al. 2008; Lacape et al. 2007). However, the values of the present investigation cannot be directly compared to previous investigations due to different marker analysis methods. In other collections, SSRs were analyzed as co-dominant markers (Lacape et al. 2007) or as dominant markers (Abdurakhmonov et al. 2008; Campbell et al. 2009). The PeeDee collection (Florence, SC, USA) of *G. hirsutum* genotypes was genotyped with SSR markers which were evaluated in a dominant manner similar to the current study; however, PIC values were not reported (Campbell et al. 2009).

Revealed intra-accession genetic variation

As many as eight bands per SSR marker were identified in the bulk DNA of an individual accession, indicating probable presence of intra-accession heterozygosity and/or heterogeneity (Table 3). Among the diploid genomes, the D, K, and A genomes had band range maximums of four, five, and six bands per SSR marker, respectively. On average in the diploids, amplification of up to two bands per SSR was most common while four or fewer bands were most frequently amplified per SSR marker in tetraploid accessions. All but one SSR marker detected three or more bands across the five tetraploid species, which was 1.3 times more common than their putative diploid progenitors (with an average of 81 SSR markers with three or more bands) and 4.3 times more common than the remaining diploid genomes (with an average of 24 SSR markers with three or more bands) (Online Resource 3).

Genetic structure of germplasm collection

A multivariate approach of PCoA was used to visualize the capacity of the core marker set to resolve genetic relationships among the accessions of the collection (Abdalla et al. 2001; Rohlf 2000). The first two coordinates explain 27 % of the variance in the data set (Fig. 1). In this analysis, the clusters exposed by PCoA fail to discretely differentiate all the known taxonomic and genomic organization within the genus, but the visualized clusters do generally conform to known genome groups. Two distinct tetraploid clusters can be discerned, as well as a distinct cluster comprised A genome accessions. Upon separate analysis of the five tetraploid species, the PCoA readily differentiated *G. hirsutum* and *G. barbadense* (Fig. 2a) as well as accessions with putative introgression or misclassification (Fig. 2b, c). When viewing the diploids only, the A genome accessions generally formed two clusters largely explained by *G. arboreum* and *G. herbaceum* species dissimilarities (Fig. 3). The D genome and all other diploid accessions combined to form a very tight cluster in which individual genomes were indistinguishable from one another. Upon further analysis of the non-A diploid genomes only, minor separation and subsequent clustering of D genome species became more apparent (Fig. 4). Accessions from three species (*G. trilobum, G. thurberi,* and *G. gossypioides*) can be assigned to discrete clusters while the remaining seven D genome species group into three clusters (*G. davidsonii*and *G. klotzschianum* in one distinct cluster; *G. aridum*, *G. laxum*, and *G. lobatum* in a second cluster; and *G. raimondii* and *G. armourianum* in a cluster immediately adjacent to the second cluster). The failure to fully differentiate genome and species clusters among the non-A diploid accessions indicates the limitations of using the current

Table 2 Summary of SSR marker polymorphism statistics by genome group and for the overall data set

Summary statistic	Diploid								Tetraploid	Overall
	A	B	C	D	E	F	G	K	AD	
Sample size	194	5	4	44	5	4	10	6	1,982	2,254
Average no. of observations	157.0 (72.3)	3.4 (2.3)	2.9 (1.7)	30.2 (17.1)	3.1 (2.1)	2.6 (1.8)	6.1 (4.2)	4.0 (2.4)	1,951.9 (42.3)	2,161.1 (97.5)
No. of amplified SSRs	104	77	80	92	78	74	78	91	104	104
No. of polymorphic SSRs	101	28	41	82	61	10	62	70	104	104
Percentage of polymorphic SSRs (%)	97.1	36.4	51.3	89.1	78.2	13.5	79.5	76.9	100	100.0
Average band frequency (%)	6.7	5.0	5.0	6.1	5.0	4.6	4.9	7.4	10.0	9.7
No. of bands	488	112	132	616	180	86	208	317	1,362	1,702
No. of polymorphic bands	474	43	86	605	156	15	188	279	1,358	1,702
Percentage of polymorphic bands (%)	97.1	38.4	65.2	98.2	86.7	17.4	90.4	88.0	99.7	100.0
Average no. of bands/SSR	4.7 (2.5)	1.1 (0.9)	1.3 (0.9)	5.9 (4.3)	1.7 (1.3)	0.8 (0.6)	2.0 (1.6)	3.0 (2.5)	13.1 (6.7)	16.4 (7.1)
Average no. of bands/polymorphic SSR	4.8 (2.4)	2.3 (0.6)	2.3 (0.4)	7.4 (3.6)	2.7 (0.8)	2.2 (0.4)	3.1 (1.1)	4.2 (2.2)	13.1 (6.7)	16.4 (7.1)
Maximum no. of bands/SSR	12	5	3	21	6	3	6	10	30	34
Maximum no. of bands/SSR/accession	6	5	3	4	2	3	3	5	8	8
Average PIC value (genetic diversity)/band	0.04 (0.11)	0.01 (0.05)	0.02 (0.08)	0.05 (0.10)	0.03 (0.10)	0.00 (0.04)	0.03 (0.10)	0.04 (0.11)	0.07 (0.13)	0.08 (0.13)

Standard deviations are indicated in parentheses

Table 3 Intra-accession variability summarized by genome for the number of accessions that have an SSR locus with band numbers ranging from zero to eight

Number of bands/ SSR marker	Diploid								Tetraploid	Overall
	A	B	C	D	E	F	G	K	AD	
0	194	5	4	44	5	4	10	6	1,103	1,375
1	194	5	4	44	5	4	10	6	1,982	2,254
2	193	2	4	43	5	4	10	6	1,982	2,249
3	34	0	1	28	0	1	3	6	1,981	2,054
4	173	0	0	8	0	0	0	3	1,645	1,829
5	8	1	0	0	0	0	0	1	297	307
6	2	0	0	0	0	0	0	0	42	44
7	0	0	0	0	0	0	0	0	7	7
8	0	0	0	0	0	0	0	0	2	2

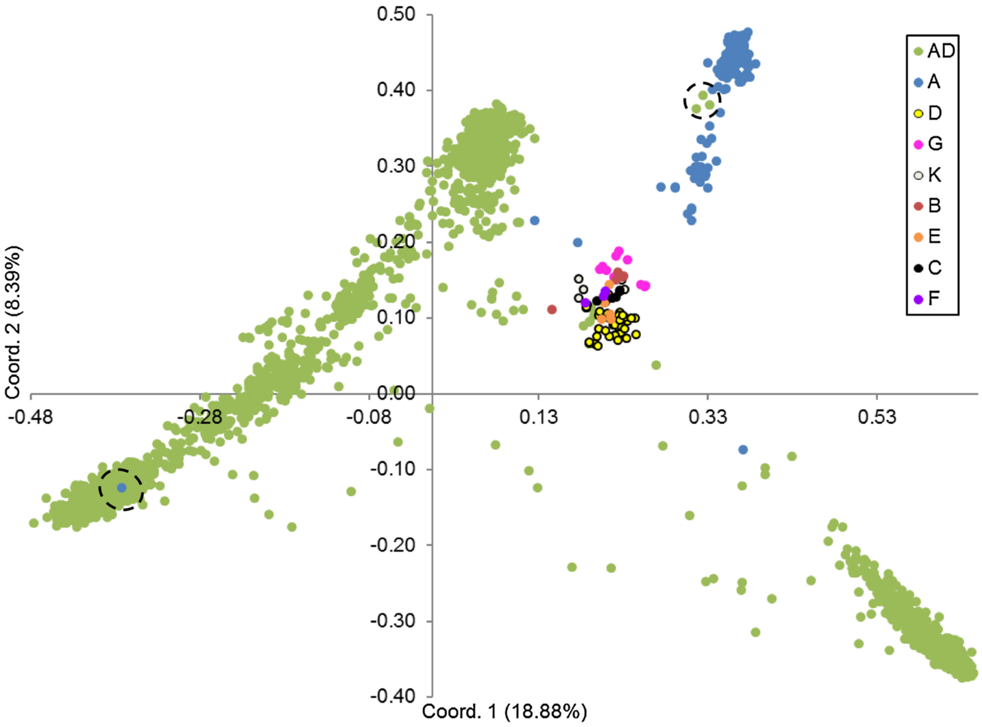

Fig. 1 Principal coordinate plot of the first two coordinate axes showing patterns of separation that reflect 27.3 % of the variation among nine *Gossypium* genome groups. Jaccard's coefficient was used to measure genetic distance among the accessions. The four accessions within *black circles* are putative genome misclassifications

core marker set pre-selected to exhibit polymorphism in the AD genome of cultivated cotton.

Of some interest for the maintenance and management of the collection were the potential outlier accessions (misclassifications, hybrids, or admixtures) that did not appear to associate with any clusters or appeared to coalesce with a genome group that contradicted the passport data. In addition to the *G. hirsutum* and *G. barbadense* accessions that appeared to cluster within the "wrong" species, there also were three accessions (SA-1415, SA-1416, and SA-1417) designated as tetraploid AD genotypes that clustered with the A diploid genome (Fig. 1; Online Resource 1). Similarly, there is one accession (A_2-0253) labeled as A genome that appeared directly within in the AD genome cluster. These four accessions, along with the putative misclassified or hybrid tetraploid accessions, are candidates for examination of original records to determine intended classification, phenotypic re-evaluation of taxonomic traits, and cytogenetic screening to verify ploidy levels. To that end, the three AD accessions that lie within the A genome cluster (Fig. 1) were found to be misclassified in passport documents but have since been verified and reclassified as true *G. arboreum* accessions. PCoA also reveals a number of accessions that did not appear associated with any cluster (Figs. 1, 2, 3, 4). These accessions may be the product of varying levels of introgression or they may possess a unique genetic background worthy of further investigation.

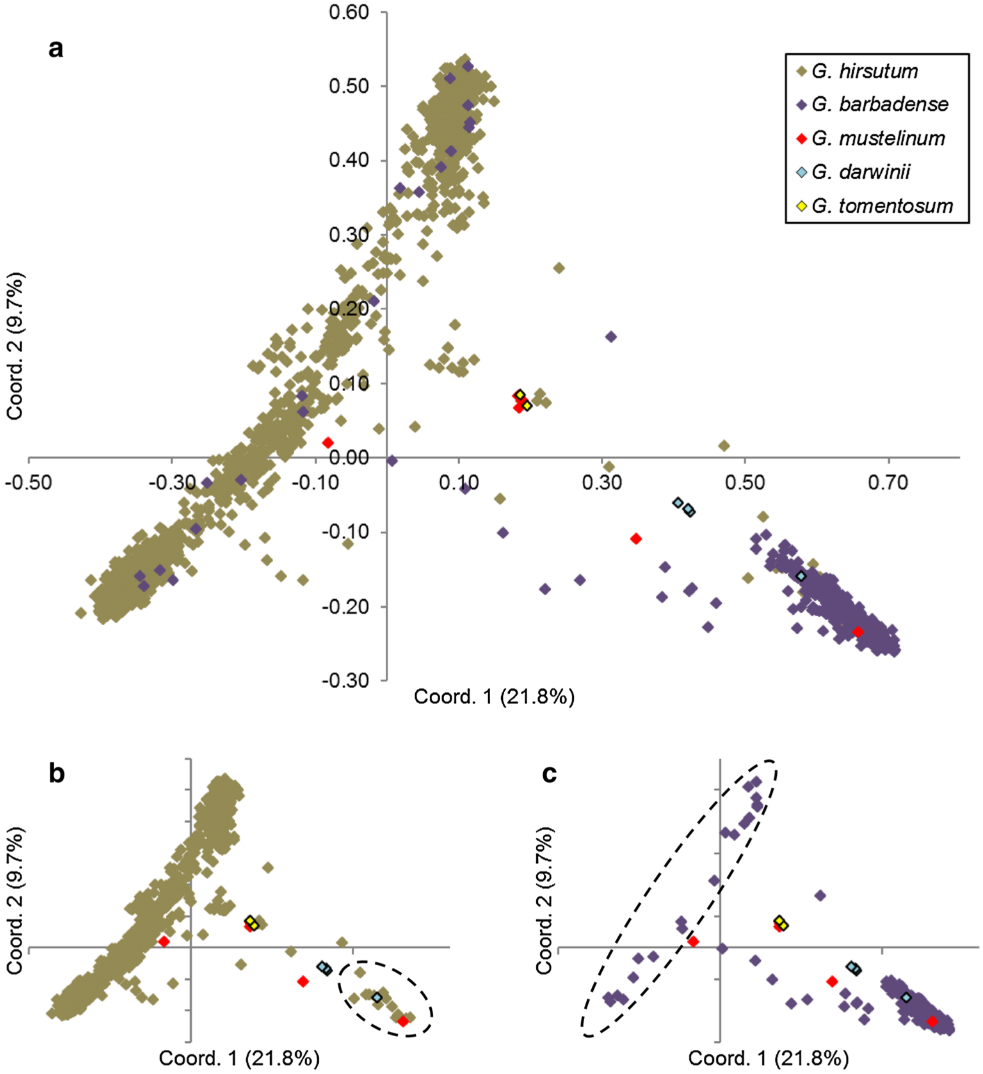

Fig. 2 Principal coordinate plots showing the relationship between the first two coordinate axes which reflect 31.5 % of variation among AD genome accessions. **a** Five *Gossypium* tetraploid species. **b** *G. barbadense* accessions removed (*dashed line* encircles 14 putative *G. hirsutum* clustering as *G. barbadense*). **c** *G. hirsutum* accessions removed (*dashed line* encircles 19 putative *G. barbadense* accessions clustering as *G. hirsutum* introgressions). Jaccard's coefficient was used to measure genetic distance among the accessions

Genetic diversity within genome groups

Prior to this investigation, it was assumed that the genetic diversity observed within genome groups would be significantly affected by the varying levels of representation of the genomes within the collection. As expected, the tetraploid (AD genome) group, with two subgenomes and the largest accession representation, was the most diverse with 13.1 bands per SSR and a genetic diversity of 0.07 (Table 2). Among the diploids, the D genome was the most diverse with 5.9 bands per SSR marker and a genetic diversity of 0.05. In contrast, the F genome was the least diverse with 0.8 bands per SSR marker and a genetic diversity of near zero (0.003). The F genome was represented in this study by one species, *G. longicalyx*, with only four accessions, and thus the small number of alleles and low diversity were expected.

Genetic similarities revealed comparable patterns of diversity across genomes as discussed above (Table 4). Among diploids, the highest similarity was observed in the F genome (four accessions, 0.911) and the B genome (seven accessions, 0.805) likely because each genome was represented by a single species and only a few accessions. The two genomes representing the most species, D genome (ten species, 0.197) and K genome (six species, 0.268), showed the greatest differentiation among their representative accessions. The D genome likely originated in western Mexico with species dispersal primarily in Mexico plus long-range dispersal events that established *G. raimondii* in Peru and *G. klotzschianum* in the Galapagos Islands (Wendel and Cronn 2003). The K genome is the most disjunct genomic group of species from Australia in terms of geographical distribution and morphology (Wendel and Cronn 2003). In comparison, the GS value within the AD genome was 0.364.

The average GS value for pairwise comparisons of *Gossypium* genomes was 0.082, with values ranging from 0.043 (G/D) to 0.142 (K/C; Table 4). The pairwise comparison of

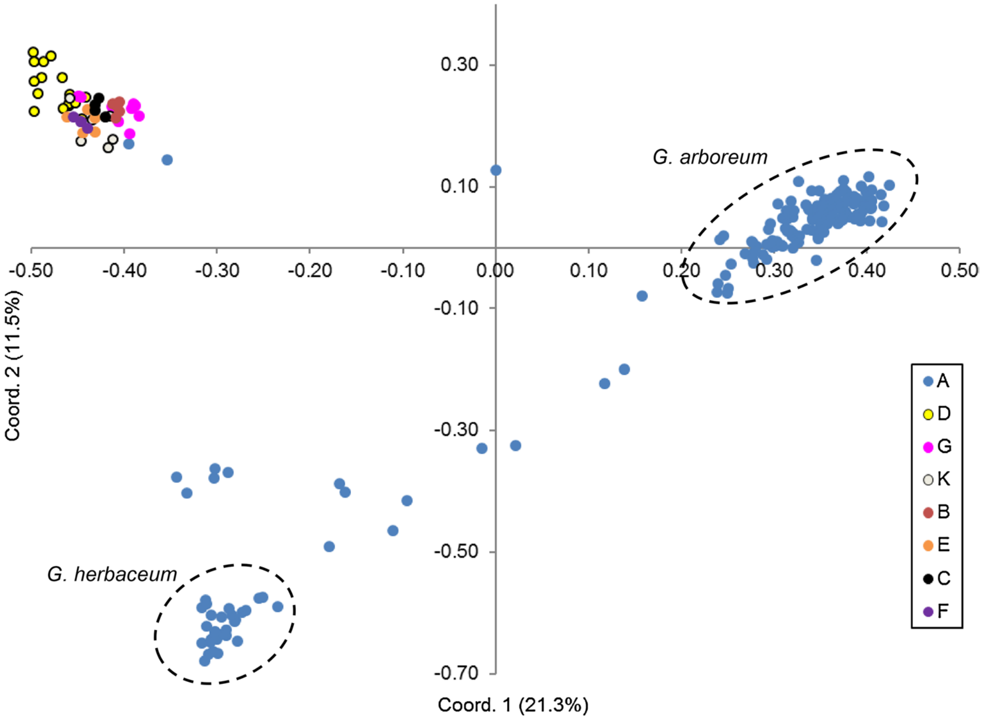

Fig. 3 Principal coordinate plot of the first two coordinate axes showing patterns of separation that reflect 32.8 % of the dissimilarity among eight *Gossypium* diploid genome groups. The *dashed lines* encircle two A genome clusters representing the majority of *G. arboreum* and *G. herbaceum* accessions. Jaccard's coefficient was used to measure genetic distance among the accessions

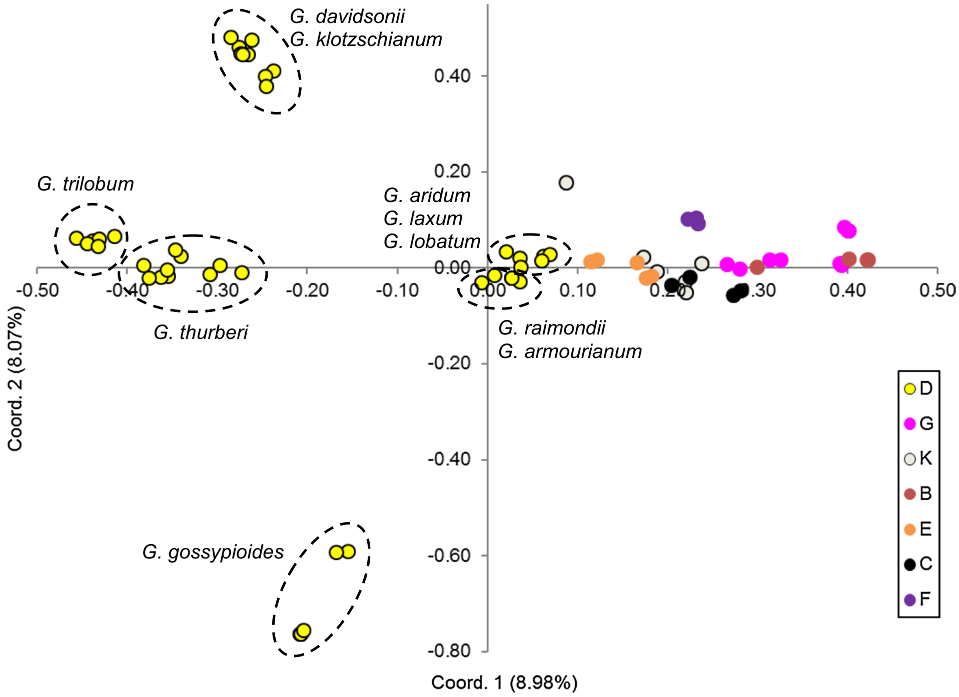

Fig. 4 Principal coordinate plot showing 17.1 % of the variation among the seven non-A diploid genome groups. Jaccard's coefficient was used to measure genetic distance among the accessions. The *six black circles* distinguish clusters of the ten D genome species. The *G. thurberi* cluster contained all nine *G. thurberi* accessions plus one *G. davidsonii* accession. The other eight *G. davidsonii* accessions clustered with *G. klotzschianum*. Remaining clusters contain all accessions of the indicated species

D (originating in the Americas) and G (originating in Australia) genomes showed the highest level of differentiation (0.043). On average, the relationships among the Australian genomes (C, G, and K) showed the highest degree of similarity (0.132) with C and K genomes most similar (0.142). This is in contrast to earlier studies of these diploids showing C and G to be the most similar and most recently diverged of the Australian genomes (Liu et al. 2001; Wendel et al. 1991).

Duplication of accessions

A feature of the NCGC that is shared with other collections is that over time it has accumulated putative duplicates of various genotypes. This is often the result of germplasm exchanges between collections, but can result from slight variations in genotype names or designations. In this investigation, 15 sets of *G. barbadense* and 37 sets of

Table 4 Genetic similarities based on Jaccard's coefficient are averaged across accessions within (along diagonal) and between (off-diagonal) *Gossypium* genomes

Genome	Diploid								Tetraploid
	A	B	C	D	E	F	G	K	AD
A	0.572 (0.200)								
B	0.096 (0.021)	0.805 (0.146)							
C	0.077 (0.013)	0.078 (0.005)	0.537 (0.202)						
D	0.062 (0.018)	0.052 (0.023)	0.057 (0.022)	0.197 (0.210)					
E	0.068 (0.013)	0.064 (0.019)	0.075 (0.018)	0.066 (0.022)	0.310 (0.279)				
F	0.088 (0.012)	0.084 (0.015)	0.075 (0.011)	0.071 (0.018)	0.083 (0.023)	0.911 (0.088)			
G	0.087 (0.025)	0.098 (0.013)	0.129 (0.017)	0.043 (0.021)	0.087 (0.027)	0.094 (0.012)	0.322 (0.301)		
K	0.093 (0.016)	0.084 (0.026)	0.142 (0.024)	0.076 (0.042)	0.083 (0.018)	0.075 (0.011)	0.125 (0.021)	0.268 (0.032)	
AD	0.123 (0.040)	0.074 (0.045)	0.062 (0.011)	0.090 (0.018)	0.075 (0.015)	0.078 (0.015)	0.051 (0.016)	0.091 (0.030)	0.364 (0.189)

Standard deviations are shown in parentheses

Table 5 Six sets of accessions that are identical based on Jaccard's genetic similarity value

Genome	Species	Accession designation
A	*G. herbaceum*	A_1-135, A_1-139
D	*G. raimondii*	D_5-3, D_5-1
	G. gossypioides	D_6-1, D_6-3, D_6-4
F	*G. longicalyx*	F_1-3, F_1-4
AD	*G. hirsutum*	TX-0367, TX-2231
	G. hirsutum	TX-1810, SA-1334

G. hirsutum accessions with synonymous names or designations were evaluated to determine their true relatedness. Comparisons were made using Jaccard's GS values and the absolute number of differences (non-matching bands) across SSR loci (van Treuren et al. 2010). The GS value of reputed duplicates ranged from a high of 0.988 (Rowden 41B) to a low of 0.186 (Pima S-2; Online Resource 4). High GS values call for a re-evaluation of the necessity of maintaining both individuals of a duplicate set, while low GS values call for phenotypic and genotypic re-evaluation of the duplicates and their original source germplasm, if still existing, to determine "true" type. Putative duplicates with GS values of less than 0.80 (Ashmouni, B233, B-237, Pima S-2, Acala SJ-4, etc.) would be prime candidates for the latter type of re-evaluation. When viewing relatedness as the absolute number of non-matching bands, we can see the actual allelic differences between any subset of accessions. The two Rowden 41B accessions with a high GS value had only one band unique to each accession (DPL0196_135 in SA-0493 and DPL0196_137 in SA-0522) while the pair of Pima S-2 accessions with the lowest GS value had 111 and 121 bands that only amplified in one of the accessions.

While the potential duplicates in the above discussion were identified on the basis of accession names, six sets of accessions with Jaccard GS values of unity (GS = 1.000) were actually identified in the GDRS (Table 5). Four of the sets are from diploid genomes and likely are identical due to the core marker set being less effective in discriminating accessions within the diploid genomes. The currently available data provide no notable rationale for these accessions to have very comparable fingerprints. However, these results serve to apprise users of the germplasm collection of the fact that samples with similar names may not be genetically identical while the opposite may also be true (i.e. samples with different names may, on rare occasion, be very similar).

Potential to identify genome- and species-specific bands and measure introgression between species

Prior to defining genome-specific bands, 616 bands (41 SSR markers) with <50 % amplification in the A, D, or AD

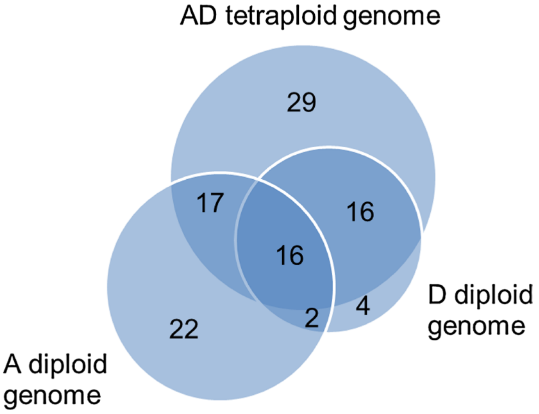

Fig. 5 Venn diagram displaying amplification patterns of the 106 bands that were genome-specific within or shared among a combination of the tetraploid AD genome and its progenitor diploid genomes, A and D. The *numbers in shared regions of each circle* represent the number of bands that were present in at least 10 % of the accessions of corresponding genomes. The graphical sizes of the displayed regions approximate the number of bands within each region

Table 6 Levels of introgression based on percentage of *G. hirsutum* species-specific bands within *G. barbadense* accessions and percentage of *G. barbadense* species-specific bands within *G. hirsutum* accessions

Introgression (%)	Accessions (no.)	
	G. barbadense	*G. hirsutum*
≤10	397	1,495
11–20	2	23
21–30	1	2
31–40	3	3
41–50	8	0
51–60	8	1
61–70	2	3
71–80	3	8
81–90	4	4
91–100	1	1

Eighty-one (81) bands were *G. barbadense* specific and 73 bands were specific to *G. hirsutum*

genome were removed. From this reduced data set, a total of 55 genome-specific bands were detected in the A, D, and AD genomes. The tetraploids had 29 genome-specific bands, while the A and D genome diploid progenitors of the tetraploids had 22 and 4 genome-specific bands, respectively. The tetraploids and A genome diploids had a similar number of shared bands in common (17 bands) when compared to the tetraploids and D genome diploids (16 bands; Fig. 5). These results are in contrast to the pairwise GS values obtained for these genomes (Table 4). When considering all bands in the GS analysis, the AD and A genomes (GS = 0.123) were more similar than the AD and D genomes (GS = 0.090). This overall similarity agrees with current results from cotton genome sequencing efforts (Page et al. 2013; Paterson et al. 2012). This also suggests that there are more bands of low frequency (rather than genome-specific or shared as defined here) in common between the AD and A genomes than are mutual between the AD and D genomes. All bands amplified well in both commercial tetraploid species, and we were able to identify 81 bands specific to *G. barbadense* and 73 specific to *G. hirsutum*.

With the identification of species-specific bands, we are able to get an indication of the amount of introgression between species, specifically *G. hirsutum* and *G. barbadense*, in the NCGC. TM-1 and 3-79, which are highly inbred examples of their respective species, show 2.06 and 0.00 % introgression, respectively. This would indicate that the rate of falsely identifying introgressed bands is about 2 % in our study. When averaged across accessions, *G. barbadense* shows about 4.8 % introgression from *G. hirsutum*. In contrast, *G. hirsutum* has about half as much (2.8 %) introgression from *G. barbadense*. The great majority of *G. barbadense* and *G. hirsutum* accessions have from 0 to 10 % introgression (Table 6). However, ten *G. barbadense* accessions and 16 *G. hirsutum* accessions have upwards of 60 % introgression. Nineteen *G. barbadense* accessions with >40 % *G. hirsutum* introgression were observed clustering tightly with *G. hirsutum* accessions (Fig. 2c). Similarly, 14 *G. hirsutum* accessions (>65 % *G. barbadense* introgression) lie within the tight cluster of *G. barbadense* accessions (Fig. 2b).

Discussion

Utility of marker core set

The analyses and interpretation of results from this investigation have proven to be challenging due to purposeful capture of intra-accession variability through sampling, varying ploidy levels, and a large number of diverse species. Therefore, we have adopted methodologies seldom used in cotton studies (Campbell et al. 2009) but more frequently used in other polyploid taxa, primarily sugarcane (*Saccharum* spp.) (Cordeiro et al. 2003; Oliveira et al. 2009; Silva et al. 2012). In a polyploid, it is often a challenge to distinguish alleles from homoeologous chromosomes; thus, this method prevented incorrect assignments of allelic relationships (Oliveira et al. 2009). For these reasons we have treated these markers as dominant molecular markers when applying statistical methods.

Intentional capture of intra-accession variability was performed to reveal true variability within the collection and determine where it resides. When collected during plant explorations, *Gossypium* accessions may represent a single plant or, more commonly, a localized cluster of plants, often assumed to be a sibling colony. During collection seed increases, accessions are self-pollinated and seeds are obtained from a bulk of ~14 plants. Bulk harvests have been performed in the collection not only to recover adequate seed amounts, but also to preserve variation present in the accession. Until the present investigation, it was assumed that varying levels of allelic diversity existed within accessions, but this variation was ignored in diversity studies where homogeneity within accessions was assumed and single individuals were used to represent accessions. The DNA obtained from our bulk of ten seeds represents the true nature of an accession but resulted in the complexity that rendered co-dominant marker analysis unfeasible or very difficult, especially in a relatively recently formed tetraploid (Cronn and Wendel 2003). The tendency of SSR markers to amplify multiple loci further complicates this analysis.

This core set of SSR markers detects a moderate rate of diversity among the *G. hirsutum* and *G. barbadense* species, but this was not observed in the diploid species. Views of the first two principal coordinates show that the A and D genome accessions can be differentiated to some extent while other diploid accessions remain closer to each other than to other clusters. Biological interpretation of the reduced diversity in diploids is two-fold. First, the lower number of accessions evaluated for the diploids would suggest that more representatives of these genomes need to be analyzed. For that to happen, more plant exploration should occur to sample the naturally occurring diversity. Second, the markers selected were originally developed from three species: primarily *G. hirsutum* and its putative diploid progenitor species, *G. arboreum* and *G. raimondii* (Blenda et al. 2012; Fang et al. 2013). In addition, the markers were selected based on the chromosomal locations in the *G. hirsutum* × *G. barbadense* interspecific map (Yu et al. 2012a). Markers residing on the A subgenome of the tetraploid might not exist in the D genome species, and likewise, those on the D subgenome likely failed to amplify products in the A genome species. For the more distant genomes (B, C, E, F, G, and K genomes), these markers appeared even less specific. This phenomenon is not unusual. Kuester and Nason (2012) screened 50 SSR markers previously developed for *G. hirsutum*, and only ten were identified that could be dependably amplified and scored in *G. davidsonii* (a D genome species). In the present study, 57 of the 104 core SSR markers amplified alleles in *G. davidsonii* giving us a higher success rate than that of Kuester and Nason (2012) yet providing only half of the genetic discrimination information as available in *G. hirsutum*. Many of the diploid accessions showed incomplete DNA profiles due to lack of PCR amplification or non-informative profiles due to the amplification of monomorphic DNA fragments. From these results, we determined that further characterization of diploids would require additional markers that were genome- or species-specific and additional representation within the diploid species (i.e. more accessions to characterize, if available).

Diversity trends

The cotton germplasm collection of Uzbekistan used a random set of 95 SSR markers for molecular characterization of ~300 *G. hirsutum* accessions (Abdurakhmonov et al. 2008). A similar study of the cotton germplasm collection of CIRAD proposed an informative set of 201 SSR markers to characterize tetraploid *Gossypium* accessions (Lacape et al. 2007). Both of these analyses have recognized useful variation within the tetraploids of their respective collections, specifically *G. hirsutum*, but are not comparable to the range of genomes and species surveyed here. Other independent studies have, to a lesser degree, reported on variability across diploid and tetraploid species. Abdalla et al. (2001) used 16 dominant markers (AFLPs) to estimate genetic and evolutionary relationships among 29 accessions from five *Gossypium* species including the AD, A, and D genomes. The AFLP markers revealed 368 polymorphic bands, with 143 bands (38.9 %) shared between the tetraploids and A genome accessions, and 84 bands (22.8 %) shared between the tetraploids and D genome accessions. This is in contrast to our study which revealed 1,086 polymorphic bands across these three genomes, with 17 (1.6 %) and 16(1.5 %) shared bands, respectively. While AFLP technology usually generates many non-specific bands, another possible explanation for this discrepancy may be the different sample sizes between the two studies and possible imprecise estimates that could result from a very small sample size. In the investigation of Abdalla et al. (2001), only one *G. raimondii* accession represented the D genome, and two *G. herbaceum* accessions and one *G. arboreum* accession represented the A genome. The shared bands identified here are supported by the cryptic introgression observed in duplicated loci in the tetraploids originating from *G. raimondii* and the A genome species, *G. arboreum* and *G. herbaceum* (Cronn and Wendel 2003).

The genetic diversity (PIC values) reported here are very low relative to other cotton analyses using SSR markers. However, this is the first report of PIC estimates in a cotton study scoring SSRs as dominant markers and comparing intra- to inter-genetic diversity across species with a marker set optimized for mainly two (*G. hirsutum* and *G. barbadense*) of the 33 species. In the present investigation, the tetraploid species of the NCGC display greater diversity than their A and D progenitor genomes (Table 2). However, these results may not reflect the true relationship between the tetraploid and diploid species' diversity. The diploids are much underrepresented in the NCGC (Table 1). Therefore, the tetraploids will show higher diversity by simply having a larger number of individuals that capture many more alleles. The extant diversity levels of the diploids can be better determined through unbiased genome sequencing of multiple individuals from each species that capture the geographic range of the species.

Germplasm collection management

An ancillary objective of this investigation has been to determine the utility of the core marker set in detecting errors in assigning accession designations and in maintaining the integrity of accessions. One aspect of maintaining the integrity of accessions in the NCGC is the related problems of purported redundancy and undetected redundancy. If true duplications exist, all accessions would be kept in the NCGC but only one copy (accession) would be routinely increased for distribution. This approach reduces the regeneration costs to the NCGC. Through this investigation, probable false duplicates could be easily identified by extremely low GS values (less than 0.50). The cutoff GS value for accepting a true duplication would be somewhat arbitrary and set by a trained curator in combination with phenotypic evaluations. In the present study, a GS value of 0.80 or above appeared to be an acceptable threshold for declaring duplication, but under these circumstances, users of the germplasm should be aware that reputed duplicate accessions may not be genetically identical and care should be taken to maintain accurate acquisition records. In addition to reputed duplicates, GS values of 1.0 were found between accessions having no morphological or passport evidence available to ascertain why these accessions would be similar. As mentioned previously, a set of markers optimized for maximal polymorphism within diploid species is needed to confirm potential duplicates among the diploids. For the *G. hirsutum* and *G. barbadense* accessions that were identical (GS = 1.0), morphological data on the twin accessions will be compared and additional analyses with more markers will be subsequently conducted to confirm or refute the potential duplications. A minimum of two independent data sources including visual assessment of accessions should be available to identify and validate potential duplicate accessions (van Treuren et al. 2010).

From PCoA, we have identified several putative misclassifications of accessions to species (Online Resource 1; Fig. 1). Within *G. hirsutum* and *G. barbadense*, accessions showing significant numbers of putative introgressed bands were identified and found to cluster with the opposite species (Fig. 2). Specifically, 14 *G. hirsutum* and 19 *G. barbadense* accessions possessing significant introgression were found to cluster with accessions of the other species. Although we cannot demonstrate that our method of determining introgressant bands accurately presents a quantitative estimate of introgression, use of species-specific bands and PCoA to identify introgressed accessions are mutually supportive.

The misclassified, duplicated, or introgressant accessions identified in this study have been documented for further investigation. Although marker analyses may be a powerful and relatively cheap tool for the initial identification of classification errors or purity issues, confirmation and remediation of these situations will entail phenotypic characterization. Proper pedigree and species identification and the presence of introgression are of extreme importance to breeders when selecting parental materials for improvement efforts.

Intra-accession variability was also noted in the high number of bands per SSR marker. Some of these bands were likely due to the aforementioned detection of multiple loci by SSR markers. The bulk of ten root tips could also have been comprised individuals heterozygous for a given locus, individuals homozygous for different alleles at a given locus, or a combination of both. The source of this variability would need to be confirmed by comparing the DNA profiles of individual root tips rather than a bulk of root tips.

In conclusion, molecular markers are a versatile tool in characterizing the diversity of the NCGC and in its maintenance. Marker information readily identified accession assignment in tetraploid species of the NCGC, and in the A and D diploid species. Marker information will be used to prioritize regeneration efforts, identify potential redundancy and uniqueness in the NCGC, and monitor integrity of accessions through regeneration cycles. Identified weaknesses of the core marker set have led to efforts to expand it to improve its discriminatory precision in wild diploid species. Although limitations have been identified, the availability of this core set of SSR markers facilitates further efforts to establish a common protocol for germplasm molecular analysis, particularly as coordination of efforts and exchange of information among all researchers in the field become more commonplace.

Author contribution statement R.P. conceived of the project, organized the research team, and developed the germplasm set, L.H. collected all DNA samples, J.F. provided all germplasm, J.Y. created the core marker set, D.F. developed all multiplex marker sets, B.S. oversaw all genotyping reactions, and D.F. and J.Y. performed primary scoring of marker data. Statistical analyses were performed by L.H. and M.G., evaluation and interpretation of results were performed by L.H. and R.P., the first draft of the manuscript was written by L.H. with assistance from R.P., and D.F., M.G., B.S., J.Y., and J.F. participated in the revision of the manuscript.

Acknowledgments The authors wish to acknowledge the contributions of S. Simpson of the USDA-ARS Genomics and Bioinformatics Research Unit who managed the whole process of the PCRs from setting up and proofing the plates to monitoring the data output. Without Ms. Simpson's efforts and diligence this project could not have been completed. The authors also recognize the contributions of members of the USDA-ARS cotton breeding, genetics, and genomics project teams for their valuable technical assistance. This research project was supported by funding from Cotton Incorporated and USDA CRIS Nos. 6202-21000-031-00D (L.H., J.Y., J.F., R.P.), 6435-21000-016-00D (D.F.), 5347-21000-011-00D (M.G.), and 6402-21310-003-00D (B.S.). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

Conflict of interest The authors declare that they have no conflict of interest.

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

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 IB, 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

Alvarez I, Wendel JF (2006) Cryptic interspecific introgression and genetic differentiation within *Gossypium aridum* (*Malvaceae*) and its relatives. Evolution 60:505–517

Blenda A, Fang DD, Rami JF, Garsmeur O, Luo F, Lacape JM (2012) A high density consensus genetic map of tetraploid cotton that integrates multiple component maps through molecular marker redundancy check. PLoS One 7:e45739

Campbell BT, Williams VE, Park W (2009) Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. Euphytica 169:285–301

Campbell BT, Saha S, Percy R, Frelichowski J, Jenkins JN, Parka W, Mayee CD, Gotmare V, Dessauw D, Giband M, Du X, Jia Y, Constable G, Dillon S, Abdurakhmonov IY, Abdukarimov A, Rizaeva SM, Abdullaev A, Barroso PAV, Pádua JG, Hoffmann LV, Podolnaya L (2010) Status of the global cotton germplasm resources. Crop Sci 50:1161–1179

Cordeiro GM, Pan YB, Henry RJ (2003) Sugarcane microsatellites for the assessment of genetic diversity in sugarcane germplasm. Plant Sci 165:181–189

Cronn R, Wendel JF (2003) Cryptic trysts, genomic mergers, and plant speciation. New Phytol 161:133–142

Fang DD, Xiao J, Canci PC, Cantrell RG (2010) A new SNP haplotype associated with blue disease resistance gene in cotton (*Gossypium hirsutum* L.). Theor Appl Genet 120:943–953

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 191:391–401

Frelichowski JE Jr, Palmer MB, Main D, Tomkins JP, Cantrell RG, Stelly DM, Yu J, Kohel RJ, Ulloa M (2006) Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. Mol Genet Gen 275:479–491

Fryxell PA (1992) A revised taxonomic interpretation of *Gossypium* L. (*Malvaceae*). Rheedea 2:108–165

Guo WZ, Cai CP, Wang CB, Han ZG, Song XL, Wang K, Niu XW, Wang C, Lu KY, Shi B, Zhang TZ (2007) A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*. Genetics 176:527–541

Guo W, Cai C, Wang C, Zhao L, Wang L, Zhang T (2008) A preliminary analysis of genome structure and composition in *Gossypium hirsutum*. BMC Genom 9:314

Han ZG, Guo WZ, Song XL, Zhang TZ (2004) Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol Genet Gen 272:308–327

Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theor Appl Genet 112:430–439

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

Hoffman SM, Yu JZ, Grum DS, Xiao JH, Kohel RJ, Pepper AE (2007) Identification of 700 new microsatellite loci from cotton (*Gossypium hirsutum*). J Cotton Sci 11:208–241

IBPGR (1980) Descriptors for cotton species. In: International Board for Plant Genetic Resources Working Group, 1979, Rome

Jaccard P (1908) Nouvelles rescherches sur la distribution florale. Bull Soc Vaud Sci Nat 44:223–270

Jena SN, Srivastava A, Rai KM, Ranjan A, Singh SK, Nisar T, Srivastava M, Bag SK, Mantri S, Asif MH, Yadav HK, Tuli R, Sawant SV (2012) Development and characterization of genomic and expressed SSRs for levant cotton (*Gossypium herbaceum* L.). Theor Appl Genet 124:565–576

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

Kohel R, Yu J, Park Y-H, Lazo G (2001) Molecular mapping and characterization of traits controlling fiber quality in cotton. Euphytica 121:163–172

Kuester AP, Nason JD (2012) Microsatellite loci for *Gossypium davidsonii* (*Malvaceae*) and other D-genome, Sonoran Desert endemic cotton species. Am J Bot 99:e91–e93

Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B, Hau B (2003) A combined RFLP-SSR-AFLP map of tetraploid cotton based on a *Gossypium hirsutum* × *Gossypium barbadense* backcross population. Genome 46:612–626

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 Breed 19:45–58

Liu Q, Brubaker CL, Green AG, Marshall DR, Sharp PJ, Singh SP (2001) Evolution of the FAD2-1 fatty acid desaturase 5′ UTR intron and the molecular systematics of *Gossypium* (*Malvaceae*). Am J Bot 88:92–102

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

Nguyen TB, Giband M, Brottier P, Risterucci AM, Lacape JM (2004) Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. Theor Appl Genet 109:167–175

Oliveira KM, Pinto LR, Marconi TG, Mollinari M, Ulian EC, Chabregas SM, Falco MC, Burnquist W, Garcia AA, Souza AP (2009) Characterization of new polymorphic functional markers for sugarcane. Genome 52:191–209

Page JT, Gingle AR, Udall JA (2013) PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 Genes Genomes Genet 3:517–525

Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, Stelly DM, Kohel RJ, El-Shihy OM, Cantrell RG (2005) Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol Genet Gen 274:428–441

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J et al (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28:2537–2539

Percival AE (1987) The National Collection of *Gossypium* Germplasm. In: Southern cooperative series bulletin no. 321

Reddy OUK, Pepper AE, Abdurakmonov I, Saha S, Jenkins JN, Brooks T, El-Zik KM (2001) New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. J Cotton Sci 5:103–113

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Rohlf FJ (2000) NTSYSpc: numerical taxonomy and multivariate analysis system, version 2.1. Exeter Software, Setauket

Rungis D, Llewellyn D, Dennis ES, Lyon BR (2005) Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton (*Gossypium hirsutum* L.) cultivars. Aust J Agric Res 56:301–307

Silva DC, Costa Duarte Filho LS, Dos Santos JM, de Souza Barbosa GV, Almeida EC (2012) DNA fingerprinting based on simple sequence repeat (SSR) markers in sugarcane clones from the breeding program RIDESA. Afr J Biotech 11:4722–4728

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066

Toll JA (1995) Processing of germplasm, associated material and data. In: Guarino L, Ramanatha RV, Reid R (eds) Collecting plant genetic diversity: technical guidelines. CABI, Wallingford, pp 577–595

van Treuren R, de Groot EC, Boukema IW, van de Wiel CCM, van Hintum TJL (2010) Marker-assisted reduction of redundancy in a genebank collection of cultivated lettuce. Plant Genet Resour Charact Util 8:95–105

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q, Yuan Y, Lu C, Wei H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu Y-X, Wang J, Yu S (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1104

Wang Z, Zhang D, Wang X, Tan X, Guo H, Paterson AH (2013) A whole-genome DNA marker map for cotton based on the D-genome sequence of *Gossypium raimondii* L. G3 Gene Genome Genet 3:1759–1767

Weir B (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates Inc, Sunderland

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Wendel JF, Stewart JM, Rettig JH (1991) Molecular evidence for homoploid reticulate evolution among Australian species of *Gossypium*. Evolution 45:694–711

Xiao J, Wu K, Fang DD, Stelly DM, Yu J, Cantrell RG (2009) New SSR markers for use in cotton (*Gossypium* spp.) improvement. J Cotton Sci 13:75–157

Yu Y, Yuan DJ, Liang SG, Li XM, Wang XQ, Lin ZX, Zhang XL (2011) Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC1 population between *Gossypium hirsutum* and *G. barbadense*. BMC Genom 12:15

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

Yu JZ, Kohel RJ, Fang DD, Cho J, Van Deynze A, Ulloa M, Hoffman SM, Pepper AE, Stelly DM, Jenkins JN, Saha S, Kumpatla SP, Shah MR, Hugie WV, Percy RG (2012b) A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. G3 Gene Genome Genet 2:43–58

Zhang T, Qian N, Zhu X, Chen H, Wang S, Mei H, Zhang Y (2013) Variations and transmission of QTL alleles for yield and fiber qualities in upland cotton cultivars developed in China. PLoS One 8:e57220

Zhao L, Yuanda L, Caiping C, Xiangchao T, Xiangdong C, Wei Z, Hao D, Xiuhua G, Wangzhen G (2012) Toward allotetraploid cotton genome assembly: integration of a high-density molecular genetic linkage map with DNA sequence information. BMC Genom 13:539