G3, 2021, 11(7), jkab145
DOI: 10.1093/g3journal/jkab145
Advance Access Publication Date: 29 April 2021
Investigation

Population structure and genetic diversity of the Pee Dee cotton breeding program

Grant T. Billings [1,2] Michael A. Jones,[1] Sachin Rustgi [1] Amanda M. Hulse-Kemp,[2,3] and B. Todd Campbell [4,*]

[1]Clemson University, Pee Dee Research and Education Center, Florence, SC 29501, USA
[2]North Carolina State University, Crop Science Department, Raleigh, NC 27695, USA
[3]USDA-ARS, Genomics and Bioinformatics Research Unit, Raleigh, NC 27695, USA
[4]USDA-ARS, Coastal Plains, Soil, Water, and Plant Research Center, Florence, SC 29501, USA

*Corresponding author: Email: todd.campbell@usda.gov

Abstract

Accelerated marker-assisted selection and genomic selection breeding systems require genotyping data to select the best parents for combining beneficial traits. Since 1935, the Pee Dee (PD) cotton germplasm enhancement program has developed an important genetic resource for upland cotton (*Gossypium hirsutum* L.), contributing alleles for improved fiber quality, agronomic performance, and genetic diversity. To date, a detailed genetic survey of the program's eight historical breeding cycles has yet to be undertaken. The objectives of this study were to evaluate genetic diversity across and within-breeding groups, examine population structure, and contextualize these findings relative to the global upland cotton gene pool. The CottonSNP63K array was used to identify 17,441 polymorphic markers in a panel of 114 diverse PD genotypes. A subset of 4597 markers was selected to decrease marker density bias. Identity-by-state pairwise distance varied substantially, ranging from 0.55 to 0.97. Pedigree-based estimates of relatedness were not very predictive of observed genetic similarities. Few rare alleles were present, with 99.1% of SNP alleles appearing within the first four breeding cycles. Population structure analysis with principal component analysis, discriminant analysis of principal components, fastSTRUCTURE, and a phylogenetic approach revealed an admixed population with moderate substructure. A small core collection ($n < 20$) captured 99% of the program's allelic diversity. Allele frequency analysis indicated potential selection signatures associated with stress resistance and fiber cell growth. The results of this study will steer future utilization of the program's germplasm resources and aid in combining program-specific beneficial alleles and maintaining genetic diversity.

Introduction

The Pee Dee (PD) cotton germplasm enhancement program in Florence, South Carolina, was formalized in 1935 as part of the USDA Agricultural Research Service's goal to revitalize Sea Island cotton (*Gossypium barbadense* L.) cultivation (Harrell 1974). Over time, the PD program transitioned into a long-term Upland cotton (*Gossypium hirsutum* L.) breeding effort focused mainly on the improvement of fiber strength and other quality traits, insect resistance, and other key agronomic traits (Campbell *et al.* 2011). Complex intercrossing, mating schemes, and germplasm recycling have led to the development of unique breeding materials and cultivars throughout the history of the program. Sources of genetic diversity for the PD program include obsolete cultivars of *G. barbadense*, *G. hirsutum*, and the triple hybrid series composed from *G. hirsutum*, *Gossypium arboreum* L., and *Gossypium thurberi* Tod (Beasley 1940). Germplasm releases from the PD program have been distributed and utilized across many public and private cotton breeding programs, especially as a source for combined fiber length and strength (Calhoun *et al.* 1997; Bowman and Gutierrez 2003).

From 1935 to 2000, the PD program completed eight breeding cycles, generating dozens of elite lines released as cultivars and/or germplasm lines in each cycle (Campbell *et al.* 2011). Group one started with the crossing of founding parents to generate new intercrossed, recombinant lines with interspecific sources of fiber length and strength alleles. Groups two, three, and four were developed through the progressive intercrossing and subselection of materials generated in the first three cycles. Groups five and six represented a change in breeding objectives as efforts were made to develop host plant resistance to the boll weevil (*Anthonomus grandis* Boh.). Group seven began another change in the PD program, where materials from outside of the breeding program were incorporated as breeding parents in an effort to bring new sources of genetic variation for increased yield potential. Group eight resulted from a combination of intercrossing of materials developed in prior breeding cycles, along with the introduction of more breeding parents from outside the PD program. The program's history is summarized graphically in Figure 1.

A retrospective accounting of the breeding resources produced from the program following 65 years of breeding was undertaken to better understand the breeding history of the PD program and to aid us in efforts to accelerate present breeding efforts. In 2009, data from a multi-site-year field experiment were combined with

Received: February 19, 2021. Accepted: April 19, 2021

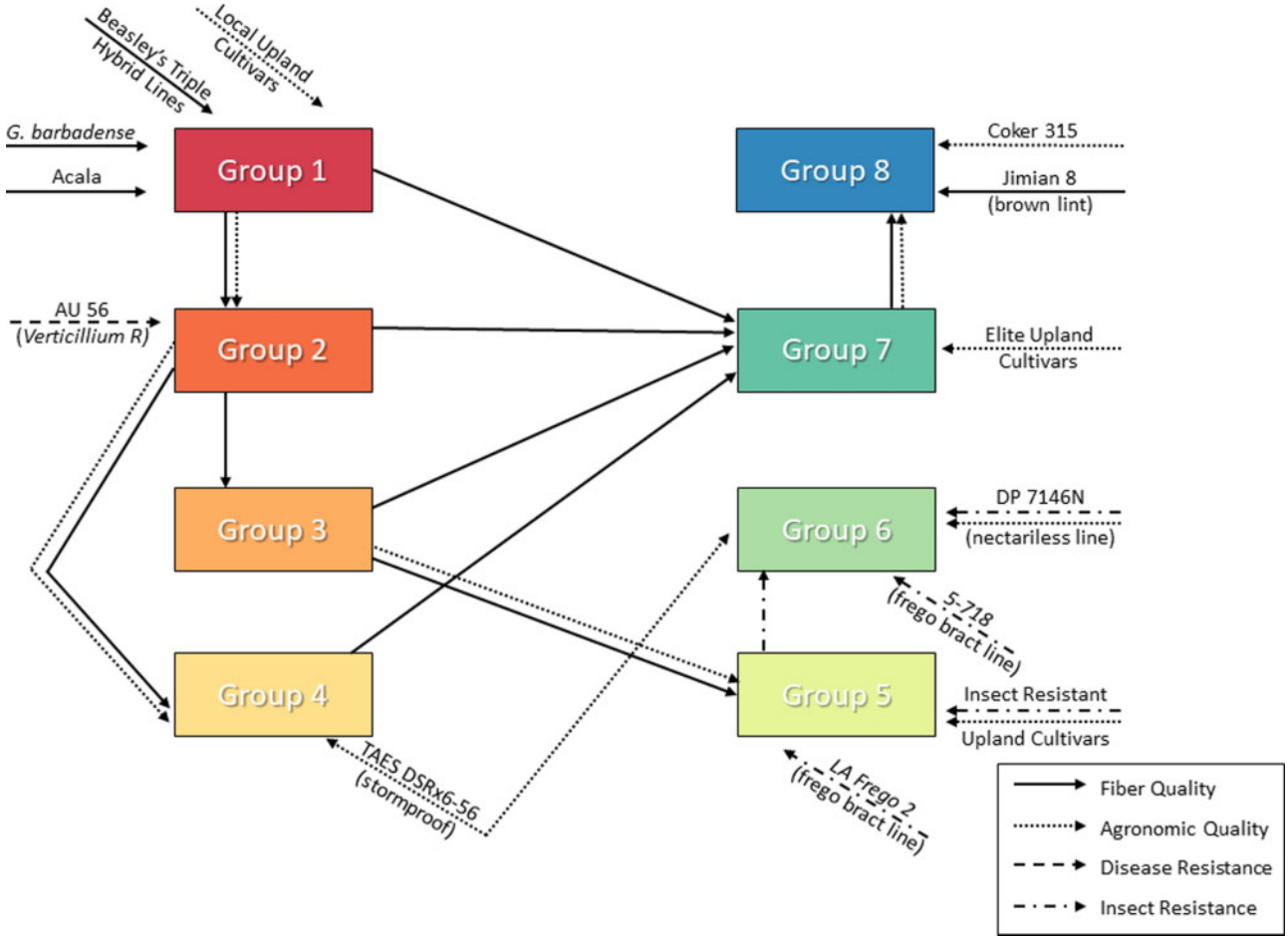

Figure 1 The historical relationships between PD breeding groups. The first four groups share a common gene pool primarily established in the first two breeding groups and focused on the improvement of fiber and agronomic characteristics. Groups five and six focused on the development of host plant insect resistant breeding material and saw the introduction of new genetic diversity and background incorporated from group three. Groups seven and eight were formed from the combination of older, high-quality material from the first four groups and new elite upland cultivars released from other breeding programs.

80 polymorphic simple sequence repeat (SSR) markers to characterize the phenotypic and genetic variability across these eight breeding groups (Campbell *et al.* 2009). They found variation for multiple fiber quality and yield components, including fiber length, fiber strength, fiber fineness, and lint percent, among others. However, the study was limited by the low density of molecular markers and genotyping techniques available at the time. Modern genotyping technologies, like the CottonSNP63K array (Hulse-Kemp *et al.* 2015), have enabled a host of new experiments and discoveries in cotton.

Population structure and diversity, assessed by the scoring of genome-wide genetic markers such as single nucleotide polymorphisms (SNPs), is crucial to generating an unbiased picture of the genomic landscape before undertaking genome-wide association studies or genomic selection (Hamblin *et al.* 2011). A wide range of methods are available for evaluating population structure, ranging from the classic phylogenetic model, which uses nucleotide substitution or genetic similarity to group similar individuals (Odong *et al.* 2011), to other more complex models of population differentiation (Bourgeois *et al.* 2017). Principal component analysis (PCA) has long been used to correct for population structure in further genomic analyses (Price *et al.* 2006). Other methods, such as discriminant analysis of principal components (DAPCs) and fastSTRUCTURE, enable the visualization and evaluation of complex stratification in such panels as nested association mapping or breeding populations (Jombart *et al.* 2010; Raj *et al.* 2014; Huang *et al.* 2015; Maurer *et al.* 2015; Deperi *et al.* 2018).

Marker-trait association experiments have resulted in the discovery of dozens of quantitative trait loci (QTL) underlying diverse traits including salt tolerance, fiber quality, and wilt resistance (Gapare *et al.* 2017; Sun *et al.* 2018; Abdelraheem *et al.* 2020). Efforts to characterize the genetic diversity and population structure in the US upland cotton gene pool have also been undertaken. Tyagi *et al.* (2014) used a set of 122 polymorphic SSR marker, which were sufficient to distinguish 378 cultivars and breeding lines originating from the western, southwestern, mid-south, and eastern US cotton growing regions. They observed similar correspondence between PCA, STRUCTURE, and allele frequency methods, noting an overall low level of genetic diversity relative to other crop species. Hinze *et al.* (2017) evaluated germplasm from the upland cotton core collection, with a focus on comparing a catalogue of phenotypic traits to SNP genotypes from the CottonSNP63K array. Multidimensional scaling analysis revealed overlap between germplasm originating from the United States and other places in the world, and a moderate ability to distinguish germplasm by US cotton growing region. However, they did not observe meaningful clustering within improved upland cotton germplasm with the fastSTRUCTURE method.

The goal of this study was to evaluate genetic diversity across and within PD breeding groups and relate these findings to the worldwide improved upland cotton germplasm. We hypothesized that this closed (largely inbreeding) breeding program, with long breeding cycles, complex intermating, and repeated shuffling of potentially unique alleles would provide an interesting population genetics model for studying the effects of genetic drift and artificial selection. Hence, the objectives of this study were to evaluate genetic diversity across and within PD program breeding groups by utilizing genome-wide SNP markers from the CottonSNP63K array, examine population structure, and contextualize these findings relative to the global upland cotton gene pool.

Materials and methods

Description of plant genotypes and genotyping

Representative plant genotypes from each of eight PD breeding groups were selected for examination, covering 96 released breeding lines and cultivars. Seeds were requested from the US National Cotton Germplasm Collection in College Station, TX (https://npgsweb.ars-grin.gov/gringlobal/site?id=1), and grown in a greenhouse in Florence, SC, during Winter 2018. Newly emerged leaves were collected in 1.5 ml centrifuge tubes and immediately placed on ice. Leaf tissue was stored at −80°C until processing for DNA extraction. Frozen leaves were lysed in a tissue homogenizer with two added glass beads. Genomic DNA extraction was performed using the DNeasy Plant Mini Kit (Qiagen Inc, Germantown, MD, USA) according to manufacturer instructions. Sample DNA concentration was measured using a NanoDrop Spectrophotometer (Thermo Fisher Scientific Inc, Waltham, MA, USA). A vacuum centrifuge was used to concentrate samples with concentration below 100 ng/μl. Samples of 25 μl were loaded onto a 96-well plate and shipped on dry ice overnight to the Texas A&M Institute for Genomic Sciences and Society (College Station, TX, USA). Upon receipt, samples were quality checked using the PicoGreen assay (Ahn *et al.* 1996), and adjusted to a DNA concentration of 50 ng μl^{-1}. The standardized DNA samples were hybridized with the CottonSNP63K array, a custom Infinium iSelect HD Genotyping Assay (Illumina Inc., San Diego, CA), developed by Hulse-Kemp *et al.* (2015). Marker probe sequences were mapped (further detail in Supplementary Methods) to the UTX_v2.1 reference genome (Chen *et al.* 2020), and any SNPs that could not be mapped were excluded.

The experimental dataset also included 18 PD genotypes from Hinze *et al.* (2017) available on CottonGen (Yu *et al.* 2014). The entire experimental dataset therefore included a total of 114 PD genotypes (Supplementary Table S1: Pee Dee Genotypes). The final report file from Illumina GenomeStudio was filtered using plink 1.9 (Chang *et al.* 2015) retaining only (1) SNPs listed as functional polymorphic (Hulse-Kemp *et al.* 2015), (2) minor allele frequency (MAF > 2.5%), and (3) call rate (CR > 90%) to generate Dataset One. Putative linkage disequilibrium (LD) blocks were discovered with the "–indep-pairwise" command in plink 1.9 and used to generate Dataset Two, with SNPs culled until neighboring SNPs were only moderately correlated ($r^2 < 0.8$).

The SNP data of 249 improved upland cotton samples genotyped on the CottonSNP63K array were downloaded from the array project page on CottonGen. A total of 249 improved upland cotton lines (non-PD lines) were included in the analysis, as well as 114 PD lines (96 from the present study and 18 from CottonGen). Markers were filtered to include those with MAF > 2.5% and CR > 90%.

Population structure analysis

Breeding group designations were selected based on parentage and the breeding history of the PD program (Campbell *et al.* 2011). These group designations were used *a priori* for population structure analyses. Two PCA variants, classic PCA from plink 1.9 (Chang *et al.* 2015) and double-centered PCA (DC-PCA) from the R script in Gauch *et al.* (2019), were applied to identify a consensus between individual clustering. Biplots of individuals for the SNP × Individual interaction were generated for Datasets One and Two with individuals color coded by the prior breeding group number. To test for differences between-breeding groups, DAPCs was performed on Dataset Two with the R package adegenet (Jombart *et al.* 2010).

Population structure was also evaluated with the maximum likelihood tree in MEGA X (Kumar *et al.* 2018). Phylogenetic analysis was carried out and branches with <50% bootstrap support were collapsed into polytomies. The tree was plotted as a phylogram with the "plot.phylo" function in the R package ape (Paradis and Schliep 2019).

To test for the number of groups and group membership of each genotype, the "chooseK.py" function in fastSTRUCTURE was used for $k = 1$–10 (Raj *et al.* 2014). To identify DAPC clusters, the "find.clusters.genlight" command was used, with 40 PCs retained. These identified clusters from DAPC were retained and plotted in a Sankey diagram to examine the relationship between the three classification methods.

Core collection analysis

Core collection analysis was performed using GenoCore at 80%, 95%, and 99% SNP allele coverage levels (Jeong *et al.* 2017). The core collection analysis was compared with a random SNP allele sampling method and one that includes breeding group number designations. The probability of discovering each allele was calculated as (1 − MAF) ^(*c*). The 50th percentile of the Poisson binomial distribution was used to determine the expected value for the number of alleles observed after seeing the *c*th individual. A breeding-group informed algorithm was also employed, which cycled between an individual within each breeding group rather than selecting completely at random. The number of alleles observed was plotted as a function of the number of individuals sampled to determine if the growth in diversity was approaching an asymptote.

The minimum number of SNPs needed to separate out individuals in the PD program was found using a custom R script. The algorithm we used identified the two SNPs with the highest pairwise distance, and then successively added the most different SNP until each individual could be distinctly identified. This set of SNPs was then tested for discrimination capacity among the improved upland cotton germplasm from CottonGen.

Signatures of selection in the PD program

To test for putative signatures of selection in the PD program versus other improved Upland cotton genotypes, a marker-specific Bayes factor (BF), analogous to Wright's F_{ST}, was estimated for each marker with the function in BayEnv2 (Coop *et al.* 2010; Gunther and Coop 2013). Samples were classified as PD or World (non-PD). The $\log_{10}$ of resulting BFs were plotted in a Manhattan plot with a threshold of $\log_{10}(BF) > 2$, and allele frequency plots for the each of the significant markers were generated. Putative regions under selection were determined as chromosomal segments containing at least one significant marker. A list of genes and their gene ontology (GO) terms in these regions was

identified using the GFF3 annotation file for the annotation of the UTX_v2.1 reference genome assembly (Chen *et al.* 2020). The list of genes was subjected to gene enrichment analysis with the weight-count method ($P < 0.05$) and ranked by Fisher's exact test with the R package topGO (Alexa and Rahenfuhrer 2020).

Data availability

The SNP data from all 96 newly genotyped individuals is provided in Supplementary Tables S3 and S4 will be made publicly available at CottonGen database (cottongen.org). All Supplemental materials are accessible at https://github.com/USDA-ARS-GBRU/PeeDeeCottonBreedingProgram_Diversity.

Results

Dataset One, the filtered set of markers without market density correction, contained 17,441 markers (Supplementary Table S3: Dataset 1 SNPs) anchored to a position on the UTX_v2.1 reference genome (Chen *et al.* 2020). During filtering, an initial set of 38,869 known polymorphic markers across any *Gossypium* spp. had 19,952 markers excluded with MAF < 2.5%, 280 markers excluded with CR < 90%, and 1196 markers excluded due to no determined reference genome position. After thinning to account for marker redundancy due to high LD, Dataset Two reduced this number to 4597 markers (Supplementary Table S4: Dataset 2 SNPs). The marker density across 15 of the 26 chromosomes differed significantly between Datasets One and Two (Supplementary Figure S1A). In Dataset One, the number of markers ranged from 1629 on chr A08 to 247 on chr A02. After reducing SNPs to account for variable marker density, the number of markers per chromosome was more uniform, ranging from a maximum of 268 SNPs on chr D05 to 116 on chr A02 (Supplementary Figure S1B).

Of the 9194 alleles (2 alleles for each of 4597 SNPs) present in at least 2 of the 114 individuals in Dataset Two, 95% were introduced (first detected in at least one individual) in group one. The remaining 5% were introduced as follows: 2.9% in group two, 1.1% in group three, 0.5% in group four, and <0.4% in each of groups five through eight. This indicated that most of the genetic diversity present in the PD germplasm pool was introduced in the first few cycles of breeding development. Almost all SNP alleles were present in at least two groups. However, group eight contained five unique SNP alleles, two of which flanked a haploblock present in the denser set of variants in Dataset One, corresponding to a 408 kb region of chr A05 (109.48–109.89 Mb) containing 17 group unique alleles. Heterozygosity varied substantially between genotypes (Supplementary Table S5: Heterozygosity of 114 PD Genotypes), meaning few SNPs were completely fixed in any breeding group. Of the 9194 alleles in Dataset Two, 457 alleles were fixed (present in at least one copy in every genotype) in breeding group one, 764 in group two, 854 in group three, 702 in group four, 816 in group five, 561 in group six, 569 in group seven, and 273 in group eight.

Both datasets exhibited similar distributions of identity-by-state (IBS) scores. The mean pairwise genetic distance was highly similar, 0.661 in Dataset One and 0.665 in Dataset Two. Pairwise IBS genetic distances ranged in Dataset Two from 0.553 for Sealand-3 (AHK) and Sealand-542 (AHK), the two most dissimilar individuals, to 0.967 for PD 762 and PD 948, the two most similar individuals. Comparison of the additive genetic relationship matrix derived from these two datasets, which is analogous to IBS distance except it ranges from around zero to a maximum of two, also indicated high concordance (Supplementary Figure S2 and Supplementary Table S6: IBS and IBD Estimates). When compared with the generalized numerator relationship matrix from NumericwareN (see calculation in Supplementary Methods and input in Supplementary Table S7: NumericwareN Input), which is the comparable estimate from pedigree data, the values calculated from Dataset Two were in higher agreement ($R^2 = 0.20$) than those of Dataset One ($R^2 = 0.13$) with the pedigree-based scores (Supplementary Figure S3). Average within group genetic similarities were generally higher (*i.e.*, pair of genotypes were more similar) than between group comparisons (Table 1).

Both classic PCA and DC-PCA as well as DAPC all showed similar results across the two datasets with the exception of classic PCA on Dataset One (Figure 2). Classic PCA and DC-PCA supported the same relationship between-breeding groups in Dataset Two. To mitigate the effect of variable marker density across the chromosomes, further analyses on the PD genotypes was performed with only Dataset Two.

Both fastSTRUCTURE and phylogenetic analysis were consistent across both datasets, so the output from Dataset Two is discussed here. The results from fastSTRUCTURE supported the existence of multiple groups ($k = 6$), and 55 of 114 individuals were classified at the ≥80% level of probability (Figure 3). *De novo* group assignments, either through DAPC or fastSTRUCTURE, supported the original eight groups with the novel groups representing a superset, or overlap, of the historical breeding groups (Figure 4). The consensus phylogenetic tree also identified the same subgroups as fastSTRUCTURE and DAPC (Figure 5).

Allele richness was plotted for the three core collections and two sampling methods (Figure 6). The core collections from GenoCore grew in allele richness much more quickly than the random sampling or sampling by breeding group methods did, and the breeding group method was only moderately better than sampling individuals randomly. These results show that 80%, 95%, and 99% of alleles can be captured by selecting three, nine, and nineteen individuals, respectively, from the PD breeding program (Supplementary Table S8. Core Collections for 80%, 95%, and 99% Alleles Covered). Similarly, a set of 19 SNPs provided enough information to uniquely identify each PD genotype (Supplementary Table S9. List of Minimum Number of SNPs to Discriminate PD Genotypes). This set of SNPs was able to discriminate between 229 improved upland cotton entries from CottonGen, while another 20 individuals could not be discriminated between.

We noted that generally, PD genotypes clustered together relative to other improved upland cotton genotypes (Supplementary Figure S4). To explore the genetic differentiation of the PD

Table 1 Identity-by-state genetic distance for between- and within-breeding group comparisons, corrected for variable marker density

		Breeding group							
Breeding group		1	2	3	4	5	6	7	8
	1	0.673	0.680	0.669	0.667	0.636	0.633	0.650	0.646
	2	—	0.687	0.672	0.671	0.637	0.636	0.647	0.642
	3	—	—	0.686	0.689	0.663	0.659	0.668	0.662
	4	—	—	—	0.702	0.665	0.663	0.679	0.672
	5	—	—	—	—	0.682	0.698	0.662	0.658
	6	—	—	—	—	—	0.713	0.659	0.657
	7	—	—	—	—	—	—	0.685	0.672
	8	—	—	—	—	—	—	—	0.676

A higher number indicates that the individuals compared are more similar to each other, and lower numbers indicate individuals between groups are more different.

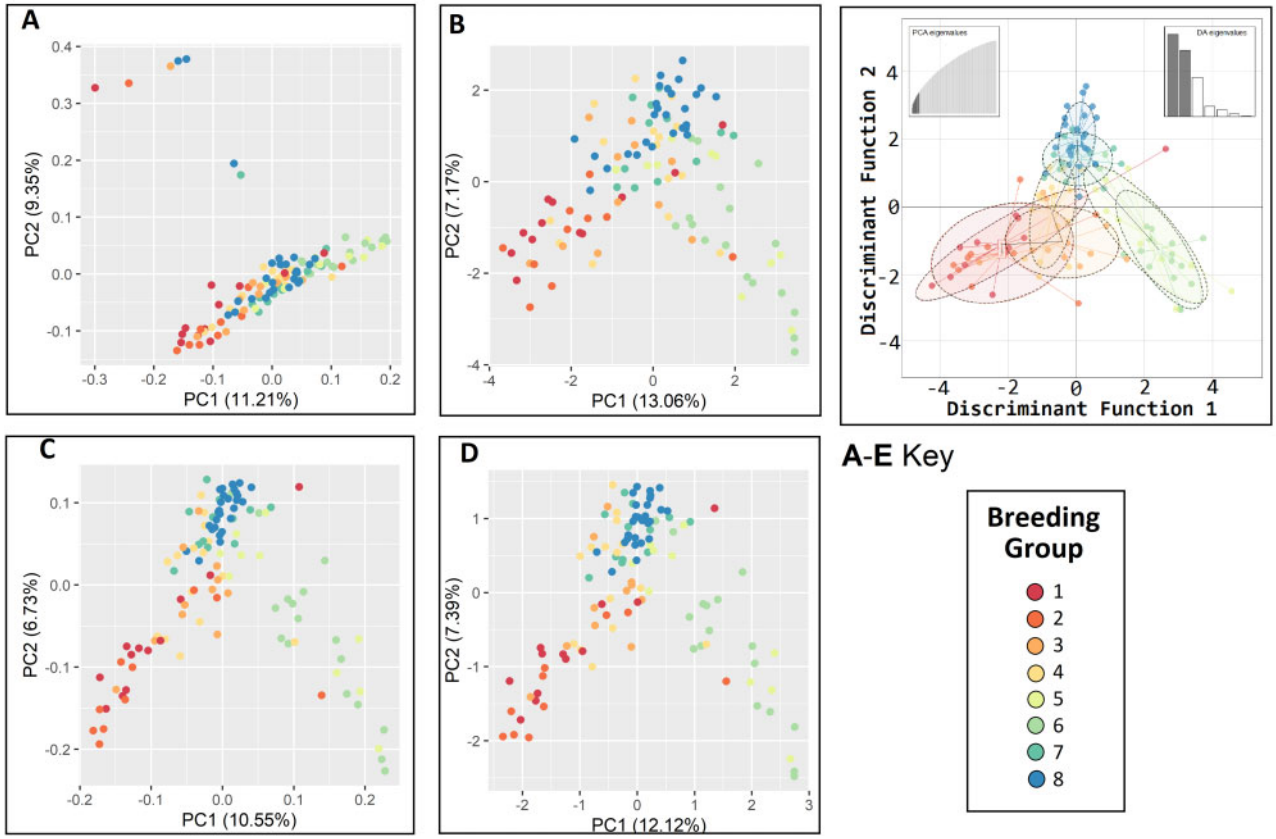

Figure 2 Comparison between two PC estimation methods before and after correcting for variable marker density. (A–D) The SNP × individual biplots of the PC coordinates for individuals, colored by breeding group, in PC1 (horizontal axis) and PC2 (vertical axis); (E) the discriminant analysis of PCs results for Dataset Two. (A) Plink PCA with 17,441 SNPs, (B) DC-PCA with 17,441 SNPs, (C) plink PCA with 4597 SNPs out of strong LD (R2 < 0.8), (D) DC-PCA with 4,597 SNPs out of strong LD (R2 < 0.8), and (E) discriminant analysis of PCs of 4597 SNPs out of strong LD.

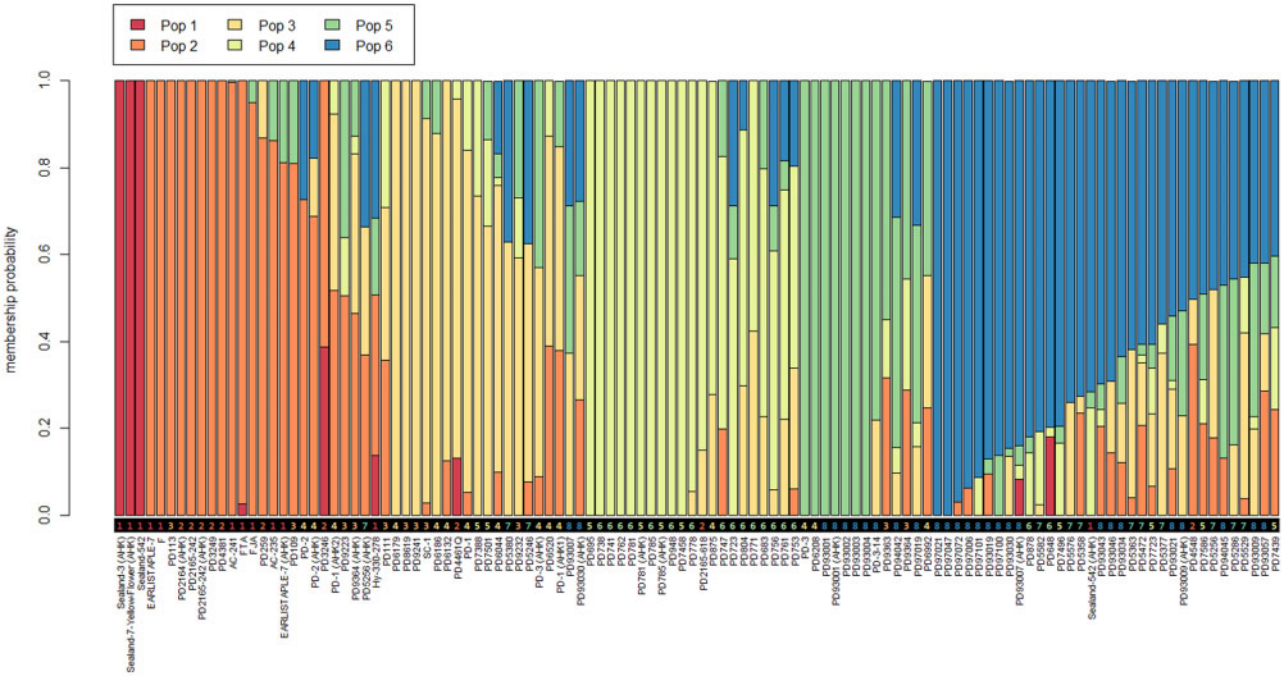

Figure 3 The Q plot for six fastSTRUCTURE subpopulations. Membership probability plot for probability of group assignment, sorted by the likeliest group assignment for each individual. The most likely number of populations (*k*), as determined by the model complexity that maximizes marginal likelihood, is 6. The individual names are given along the bottom of the horizontal axis, with the breeding group number given above it in the same color scheme as other figures.

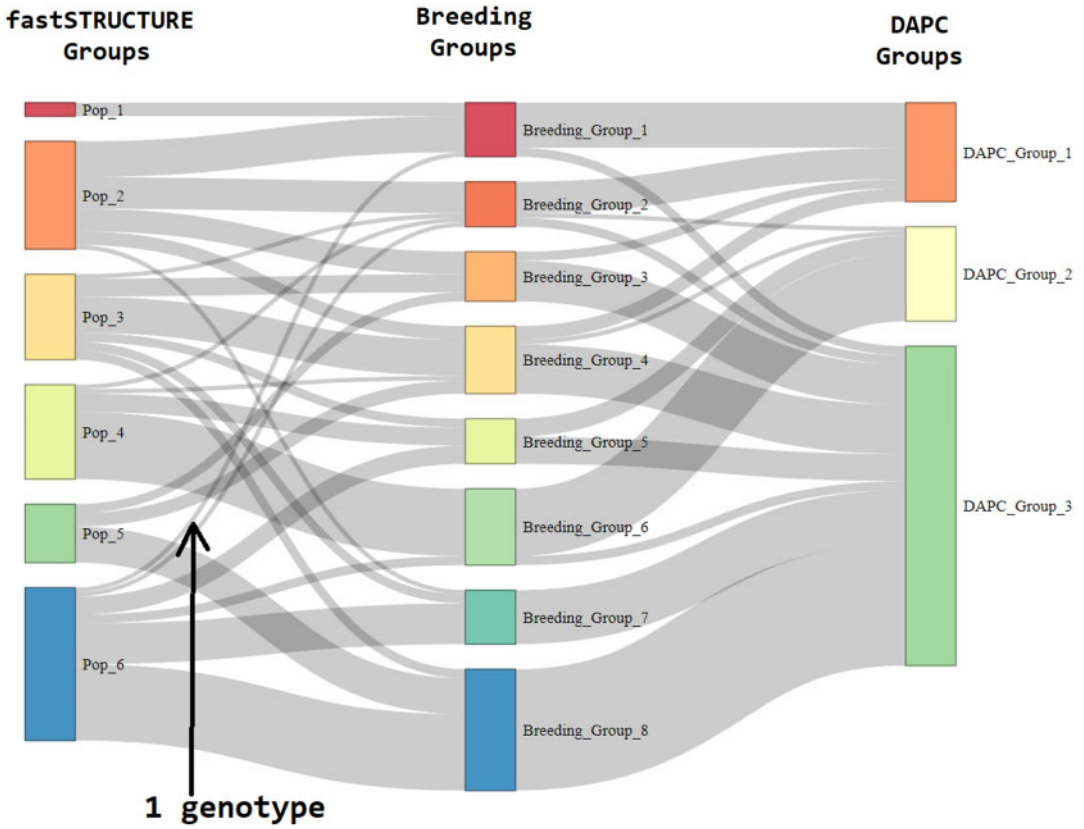

Figure 4 Overlap between three group designation methods. Sankey diagram showing how individuals in each of the prior breeding groups (center) are classified in fastSTRUCTURE (left) and in DAPC (right). In both DAPC and fastSTRUCTURE, the number of populations or clusters ($k = 6$ for fastSTRUCTURE, $k = 3$ for DAPC) is less than the number of breeding groups ($k = 8$).

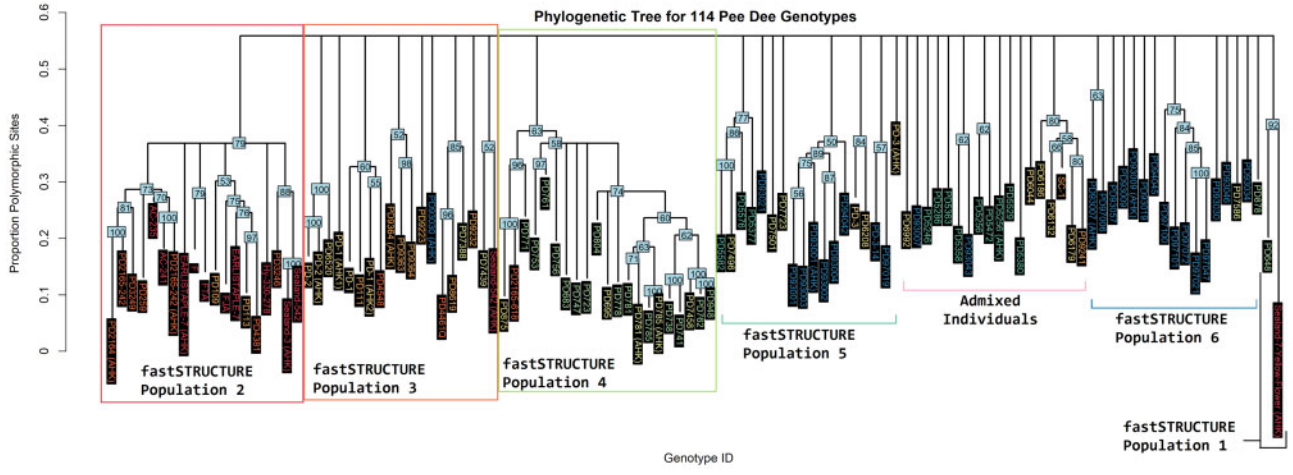

Figure 5 Unrooted consensus phylogenetic tree for 114 PD genotypes. Bootstrap values are given for branches with >50% support based on 1000 replicates and all other branches are collapsed into polytomies. Branch length is proportional to the evolutionary distance between sub-branches. Highlighted clades correspond to populations discovered with fastSTRUCTURE.

germplasm (PD Group) from other improved *G. hirsutum* cultivars (World Group), a BF was calculated to compare genetic differentiation relative to the background level of genetic differentiation between the groups at each of 20,566 polymorphic SNPs (Supplementary Table S10: SNPs for PD *vs* World). The BF was log10-transformed and plotted for each SNP, with allele frequencies at six example SNPs for the eight breeding groups and world group plotted (Figure 7). Thirty-six of the SNP markers were significant (BF > 10) for the test for selection. These SNPs were located at 32 genetic locations distributed across 13 chromosomes (Supplementary Table S11: Markers Under Selection). The regions near the significant SNPs contained 118 genes (Supplementary Table S12: Genes in Selection Windows) enriched for GO terms related to response to stimuli, translation, actin, and glutathione metabolic process (Table 2 and Supplementary Table S13: Matchup Between Genes and GO Terms).

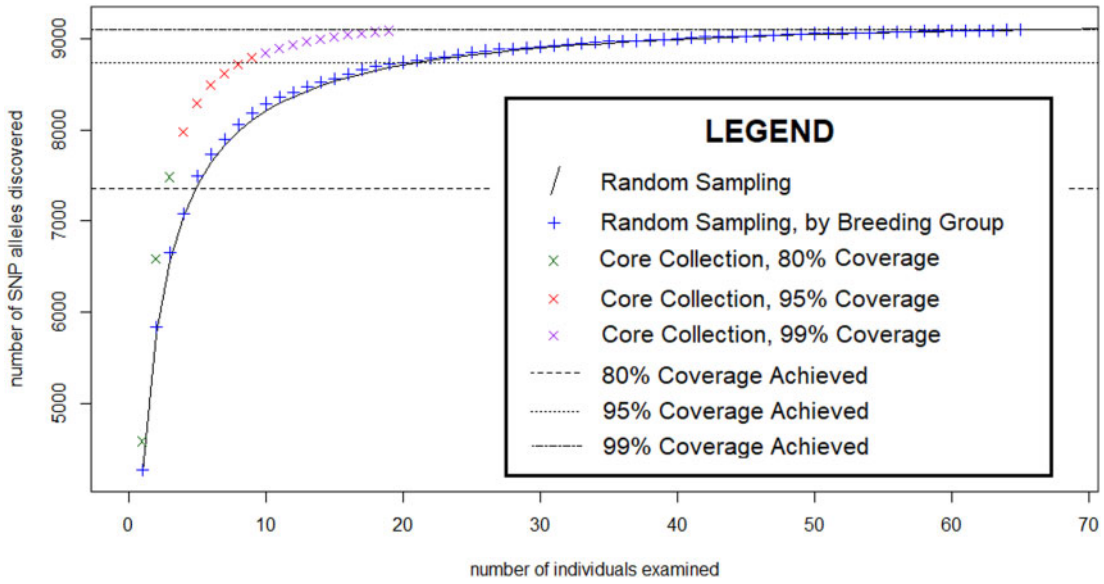

Figure 6 Collector's curve for three different-sized core collections from and random sampling of SNP alleles. The horizontal lines indicate represent 80%, 95%, and 99% allele coverage, respectively.

Discussion

Between-breeding group genetic variation

PD breeding groups one through four have common parentage composed of ~12 diverse founders (Culp *et al.* 1979). Most of the allelic diversity was introduced in these first four breeding groups, accounting for 99.5% of the total SNP alleles in Dataset Two. Most later introductions into the program also originated from the United States, meaning they probably came from the same original gene pool as the PD founders. Hence, recent diversity was mostly associated with new combinations of the same alleles.

We hypothesized that within-breeding group genetic variation would be lower than between-breeding group variation, since members of a breeding group tended to have similar parents and selection regimes (Table 1). Given the IBS distance scores calculated from Dataset Two, individuals within each group were on average more similar to one another than to members of any other group. Interestingly, individuals in each of three groups (groups one, three, and five) were more similar to individuals in the respective subsequent group (groups two, four, and six) than they were to each other, perhaps indicating additional subselection and/or drift among genotypes in these groups across generations. These pairs of breeding groups (one and two, three and four, and five and six) also clustered together in DAPC (Figure 4). Based on average genetic distance, breeding groups one through four separated out together, groups five and six together, and groups seven and eight represented out groups (Supplementary Figure S5). The breeding groups were also conserved in a phylogenetic model (Figure 5) and admixture-based model from fastSTRUCTURE (Figure 3).

In terms of genetic diversity, our results are consistent with other prior studies. The average IBS genetic distance of genotype pairs in this study (~0.66) was similar to other recent studies [~0.67 for Hinze *et al.* (2017) and ~0.80 for Tyagi *et al.* (2014)], neither of whom performed a corrective procedure to account for variable marker density across the genome. Because of ascertainment bias in the construction of genetic arrays or marker sets, additional processing is necessary to reduce bias that may not be present in whole genome datasets (Albrechtsen *et al.* 2010; Moragues *et al.* 2010; Lachance and Tishkoff 2013; Malomane *et al.* 2018). Another study which assayed 100 SSR loci found an average IBS genetic distance of 0.80 within the New Mexico Acala breeding program (Zhang *et al.* 2005). Differences in IBS genetic distance estimates reflect changes in the number and types of genetic markers used, population sizes, distribution of markers, type of genotypes used in the study (*i.e.*, obsolete *vs* elite), and differences in how rare alleles change genetic distance. Overall, the change corresponded to a reduction in SNP overrepresentation in low recombination pericentromeric regions; the difference in diversity relative to the worldwide germplasm also reflected the removal of monomorphic SNPs when calculating pairwise IBS values.

Pairwise genetic distance alone was inadequate to fully capture the genetic diversity present within- and between-breeding groups. Both methods of PCA (classic PCA and DC-PCA) for Datasets One and Two captured underlying genetic structure by summarizing the differences between individuals at the SNP × individual interaction level (Price *et al.* 2006; Gauch *et al.* 2019). In all four cases, once flipped for sign changes in PC1, the primary dimension of PC showed a gradient of separation between the earlier groups, one through four, in one extreme (Figure 2). The host-plant insect resistant breeding groups, five and six, were in the other extreme; and the most recent groups, seven and eight, were in the middle. The primary dimension, PC1, explained between 10.6% and 13.1% of the variance included in the first 40 PCs. The second dimension, PC2, was the same for all plots except for the classic PCA of Dataset One. In all other plots, the newer groups, seven and eight, clustered together on one pole and the other six groups in the other pole.

The outliers for the plink PCA plots, without marker density correction, in PC2 included PD 3246 (AC 239/FJA 348), PD 9232 ("Coker 421"/PD 2164), PD 93034 (PD 5285/PD 5485), PD 93004 (Brown Accession/PD-3) and Sealand 3 (resel. "Sealand": "Coker Wilds"/"Bleak Hall") at the furthest extreme, and PD 93001 (Brown Accession/PD-3) and PD 5576 ("Deltapine 41"/PD 3246) near the center of the two large clusters. PD 93001 and PD 93004 are brown lint cottons. PD 3246 is the pollen donor for the original cross for PD 5576 and is also a full sib of PD 2164, one of the

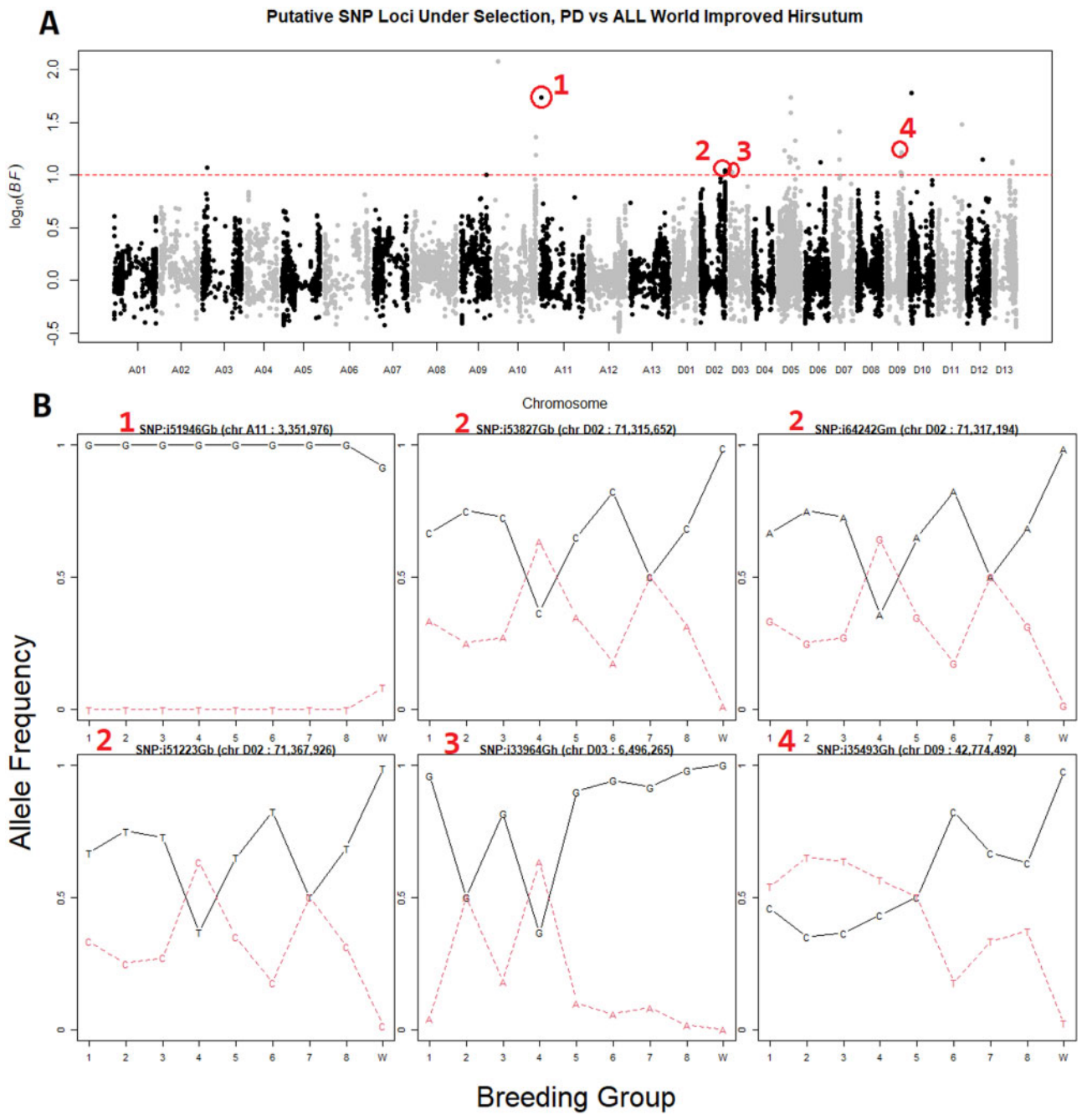

Figure 7 Identifying loci under selection in the PD Breeding Program. (A) The $\log_{10}$ BF from BayEnv2 for genetic differentiation between the114 PD from the 249 other improved upland cotton genotypes, estimated for 20,566 SNPs. (B) Allele frequency for six significant SNPs in PD breeding groups one through eight (1–8) or other genotypes (W) are given on the vertical axis. The red numbers in (A and B) indicate significant SNPs that are near genes annotated with significant GO terms.

parents of PD 9232. There were other individuals in the study with highly similar parentage and selection strategies, suggesting that common pedigrees and brown lint do not alone explain these outliers. Therefore, we considered the possibility that there is a genomic feature shared between these individuals that is obscuring genome-wide variability in the plink PCA plot of Dataset One.

The loadings for variant weights in PC2 of plink PCA for Dataset One (Supplementary Figure S6) revealed significant contribution (27.8% of total variant loadings) from a run of 911 markers in high LD on chromosome A08 (16.46–79.48 Mb). After removing SNPs in Dataset One based on putative haploblocks, this segment was reduced in Dataset Two to include only 21 markers. Pedigree analysis indicated a possible common breeding program origin for this chromosomal segment from "Hopi Moencopi" via C-6-5, a California breeding line used early in the development of the PD program (Supplementary Figure S7). Another potential origin was Coker Wilds or Bleak Hall (*G. barbadense*) via Sealand. Interestingly, the pericentromeric region of chr A08 has been noted as exhibiting low recombination frequency (Shen *et al.* 2017; Chen *et al.* 2020), which may be due to gametic incompatibility associated with multiple large scale inversions in this region of chr A08 (Yang *et al.* 2019). Others have recently evaluated the extent of *G. barbadense* introgression in

Table 2 Significant GO—biological process terms in regions under selection

GO number (biological process)	GO term (1,341 terms > 5 genes)	Number of genes with this GO term			Fisher's exact test rank	P-value	
		Count in whole genome (n = 24,647 genes with GO annotation)	Count in Selection Windows (n = 52 genes with GO annotation)	Expected (of 52 randomly chosen genes)		Weight method	Fisher's exact test
GO:0006749	*Glutathione metabolic process*	9	2	0.02	3	0.00016	0.00016
GO:0006412	*Translation*	1,494	11	3.15	4	0.00024	0.00075
GO:0009733	*Response to auxin*	263	4	0.55	11	0.00230	0.00230
GO:0046907	*Intracellular transport*	489	5	1.03	14	0.00364	0.00364
GO:0045010	*Actin nucleation*	44	2	0.09	16	0.00390	0.00390
GO:0044743	*Protein transmembrane import into intracellular organelle*	18	1	0.04	67	0.03732	0.03732
GO:0006452	*Translational frameshifting*	20	1	0.04	69	0.04138	0.04138
GO:0009416	*Response to light stimulus*	20	1	0.04	70	0.04138	0.04138
GO:0045901	*Positive regulation of translational elongation*	20	1	0.04	71	0.04138	0.04138
GO:0045905	*Positive regulation of translational termination*	20	1	0.04	72	0.04138	0.04138

GO terms for Biological Process enriched in the set of genes in genomic regions (detected with BayEnv) that differentiate PD genotypes (n = 114) from other improved worldwide *G. hirsutum* material (n = 249) filtered to include only those terms significant by the graph weight method for Fisher's exact test ($P < 0.05$).

upland cotton (Brown *et al.* 2019; He *et al.* 2020). However, our analysis did not allow for an investigation of *G. barbadense* introgression due to our filtering of SNPs based on MAF in upland cotton, and our choice to exclude SNPs that could not be mapped to the v2.0 *G. hirsutum* reference genome.

The two individuals near the center of the two major clusters in plink PC2 for Dataset One, PD 93001 and PD 5576, were heterozygous for >90% of these 911 markers, indicating a potential region of fixed heterozygosity. These regions accounted for a 70% and 27% increase in observed heterozygosity for PD 93001 and PD 5576, respectively, between Datasets One and Two (Supplementary Table S5: Heterozygosity of 114 PD Genotypes). Five other individuals from the improved germplasm set ("Coker 315," "Reba P279," "Acala 5," "Lockett BXL," and "Deltapine 16") shared this region of heterozygosity. All other individuals were >95% homozygous in this region, except for "Sicala-3-2" and "Namcala" which had a high number of no-calls in this region. In total, 51 of the 249 improved upland cotton samples from CottonGen are homozygous for the minor haplotype (Supplementary Table S14: Individuals by A08 Haplotype).

The other three PCA biplots showed a much clearer picture of the interrelatedness of individuals in terms of the entire genome (Figure 2). Examination of variant weights did not indicate highly weighted genomic regions, a potential indicator of bias as the case had been with plink PCA of Dataset One, suggesting that polymorphism across the genome was responsible for separation between individuals (Supplementary Figure S6). Plots of additional dimensions of PCA did not reveal any obvious population structure relative to the original breeding group classifications (data not shown).

One possible biological interpretation of these results is that PC1 and PC2 captured two allele frequency gradients (Novembre and Stephens 2008). The primary axis, PC1, may have captured alleles associated with high frequency in breeding groups five and six, perhaps associated with the genetic background of their parents. In this model, the earlier breeding groups may have had low levels of this genetic background, the newest groups seven and eight had moderate levels, and groups five and six had the highest amount. This genetic background may be associated with the insect resistance in groups five and six, or with highly improved fiber quality characteristics in breeding groups one through four, moderate fiber quality characteristics in groups seven and eight, and poor fiber quality in groups five and six. Campbell *et al.* (2011) showed that groups five and six had a drag in fiber quality, perhaps at the expense of host plant resistance features. Similarly, the secondary axis, PC2, may have involved the SNP alleles associated with elite, modern cultivars, with individuals from groups seven and eight having the highest frequency of these alleles. From a historic perspective this finding makes sense, since the program's breeders focused on plant productivity, fiber quality, and host plant resistance during different time periods.

Another possibility is that the plink PCA plots of Dataset One reveals the "true" population structure and the other three plots are examples of PCA "arch distortion." Arch distortion results from the projection of a single gradient onto the first two, dominant dimensions of PCA (Gauch *et al.* 2019). For example, perhaps PC1 and PC2 in the other three PCA plots are simply capturing the same information as PC1 in the other two plots. However, these three plots do not have the characteristic closed arch at the bottom of the plot, and both dimensions have plausible biological interpretations.

Three other methods reflected the same basic relationships between breeding groups, including DAPC and fastSTRUCTURE which resulted in the identification of de novo genotype clusters (Figure 4). It is worth considering, however, that a significant number of outliers existed in each analysis. In fastSTRUCTURE, 59 of the 114 genotypes could not be classified into a single population at a probability ≥80%, providing evidence for the existence of significant admixture among groups (Figure 3). In DAPC, there were multiple individuals that plotted far away from other members of the breeding group (Figure 2E). In the phylogenetic approach the ability to resolve branches was fairly low, and most branches collapsed into polytomies due to low (<50%) bootstrap support, except for in cases with simple, unidirectional breeding schemes with noncyclic pedigrees (Figure 5). For example, unique clades containing the majority of fastSTRUCTURE populations one, three, and four were obvious. Within-clade genetic variation was still relatively high, with branch lengths (proportional to genetic distance) > 0.1 usually present between sister lines, indicating that gene flow across generations has contributed to the construction of multiple (10 clades with >3 member individuals), small (each clade < 20), diverse populations within the entire breeding program. For this reason, we explored outliers in our analysis to examine how these individuals could inform our understanding of the program's breeding history.

Between individual genetic variation

Datasets One and Two exhibited strong agreement ($R^2 = 0.77$; Supplementary Figure S2) in the additive genomic relationship matrix (GRM), showing that between-individual comparisons were not significantly affected. However, when fit to the pedigree-based relationship estimate (Supplementary Figure S3), pairwise comparisons calculated from Dataset Two ($R^2 = 0.19$) fit the expected value better than those for Dataset One ($R^2 = 0.09$). Because the procedure we performed for Dataset Two reduced the high weight from redundant alleles, the dispersion of the GRM was higher in Dataset One (SD = 0.044) than Dataset Two (SD = 0.036), likely contributing to better fit to the pedigree-based scores by reducing the distance of each data point from the line of best fit. The between-individual relationships were maintained even after reducing the number of SNPs by a factor of four, and they more closely reflected the pedigree expectations after applying this correction.

For some genotype pairs we hypothesized a high level of genetic similarity; however, some reselection pairs of lines, published as separate germplasm releases purportedly from the same gene pool, were more genetically distinct than other completely unrelated pairs. For example, "PD-3" and PD-3-14, released as a reselection of PD-3, had a pedigree-based kinship ~1.00 but a genetic distance of 0.76, indicating they were only somewhat more different from each other than the average pair of genotypes (Supplementary Table S6: IBS and IBD Estimates).

Relationships between individuals could usually be interpreted as the consequence of shared ancestry. In the fastSTRUCTURE membership probability plot (Figure 3), population one included three of the Sealand germplasm lines, resulting from the interspecific cross between Coker Wilds and Bleak Hall, a *G. barbadense* cultivar. Population two is composed entirely of founding lines and intercrosses between them. Population three includes mostly early crosses between founding lines and elite introductions "Coker 421," "MO-DEL," and "AU-56." Population four includes PD 695, PD 875, and 18 selections from their progeny, all sharing a common grandparent LA Frego 2, an insect-resistant frego-bract line. Population five includes a subtree of the entire PD pedigree centered around the cultivar PD-3, all six of its descendants included in this study and two of its ancestors, and PD 6992, an outlier for this group with a low probability of true membership (43.9%). Population six was the most diverse group, including germplasm releases resulting from crosses with elite materials from the Delta Experiment Station, McNair, Deltapine, and Stoneville breeding programs, as well as a line developed in China, "Jimian-8" (May 1999).

Clearly, there is genetic redundancy in the PD breeding program, as there is in all breeding programs. To reduce the redundancy in terms of individuals, we generated three core collections with GenoCore (Jeong *et al.* 2017) at differing SNP coverage (80%, 95%, and 99%) and compared them with the collector's curves generated by randomly sampling the population and sampling by breeding group (Figure 6). Collector's curves, or species-accumulation curves, are used in ecology to evaluate the rate at which diversity increases as a function of the number of individuals sampled (Ugland *et al.* 2003; McGill *et al.* 2006). The random allele sampling method took examining 68 individuals to reach 99% allele coverage at least half the time, compared with 64 for the method for sampling by breeding group. Comparatively, the GenoCore algorithm was able to guarantee 99% allele coverage after only 19 individuals. That is to say, a core collection of <20 individuals is large enough to capture nearly all of the SNP diversity present in the PD program. Core collections of size three and nine, respectively were large enough to capture 80% and 95% of the SNP diversity, demonstrating that there is a quickly decreasing rate of gain for adding additional individuals to the core collection. Breeding groups were not equally represented in any of the core collections, with earlier releases from breeding groups one through four representing a significant majority in all three collection sizes, representing 100%, 78%, and 63% of the 80%, 95%, and 99% SNP coverage collections. Groups seven and eight were only included in the 99% coverage core collection, probably due to the presence of more recently introduced allelic diversity from elite germplasm sources. Therefore, our findings show that older germplasm from the PD program is a better resource for allelic diversity than newer germplasm due to a higher number of minor alleles present together in the genome of a few obsolete breeding lines and cultivars.

The results of our core collection analysis differ from another core set reported by Tyagi *et al.* (2014). Their analysis, which used SSR markers on a set of 375 accessions and PowerMarker software, was dominated by rare SSR alleles, which made generating a core collection much more difficult. They found that a core set of 18 individuals was necessary to capture 80% of allelic diversity, reaching 95% only after including 53 individuals. Our results were significantly different, in that the PD core collection grew in allelic richness much more quickly, benefitting from a few individuals that had a significant combination of multiple rare alleles.

PD versus world germplasm

Following our analysis of the genetic variation within the PD germplasm, we identified genomic segments that distinguished PD genotypes from other improved *G. hirsutum* cultivars and breeding lines. Generally, PD genotypes tended to cluster together based on pairwise genetic distance (Supplementary Figure S4). For SNP loci passing filtering (CR > 90%, MAF > 2.5%), 3.5% of alleles were absent entirely from surveyed PD genotypes despite being present in the other improved *G. hirsutum* cultivars and breeding lines, whereas only 0.05% were private to the PD program, indicating that most of the SNP diversity present in the

improved Upland cotton gene pool can be found in the PD program as well. However, this result is worth considering carefully since the number of individuals between the two groups (n_{PD} = 114, n_{WORLD} = 272) was significantly different.

Thirty-five putative selection windows were identified across 14 chromosomes, ranging from a single SNP with nonsignificant SNPs 25 bp away to a larger region spanning 291 kb in length, and these concentrated in the telomeric regions of each respective chromosome (Figure 7). Most of the SNPs under selection were common in the PD genotypes (~50% frequency) and at low frequency (<5%) in the other improved *G. hirsutum* germplasm. Minor alleles for each of the 35 significant SNPs ($P < 0.05$) were present in every PD breeding group with low preference towards one breeding group over the others. Therefore, these chromosomal segments may be associated with the genetic background of the PD genotypes, regional adaptation, or the cumulative results of efforts to improve fiber quality traits, especially fiber strength (Harrell 1974; Campbell *et al.* 2011).

We further explored these regions by subjecting the genes in the putative selection window to gene enrichment analysis using GO biological process annotations. We identified 10 significant GO terms (Fisher's exact test $P < 0.05$) in five chromosomal regions associated with four categories of biological function: (1) response to auxin, (2) glutathione metabolic process, (3) actin nucleation, and (4) cellular localization and translation.

There were four genes in the enrichment set annotated with the GO term "response to stimulus" localized to a single 50 kb segment of chromosome D02 (near 71.394 Mb). Although the role of auxin is ubiquitous across an array of morphological and immunological traits in plants, other genes in this enrichment set may provide evidence of how the PD programs breeding history has changed allele frequency in these particular regions. Gene expression studies in multiple plant species have exposed the potential for crosstalk between auxin biochemical pathways and other biotic and abiotic stress pathways (Lekshmy *et al.* 2017). These four genes were annotated as auxin-responsive protein small auxin up RNA (*SAUR*)-like, coding for small polypeptides (~140 amino acids) with an auxin-inducible motif. Other members of the *SAUR* gene family colocalized with fiber length and strength QTL (Li *et al.* 2017), and an association with fiber strength has been found nearby on D02 (qFS-Chr14-1.E1.XZV-RIL; Shang *et al.* 2016). The minor alleles for these SNPs are found at about 40% frequency across PD breeding groups and is at <5% frequency in other improved cotton germplasm.

Two adjacent genes on chromosome D03 (6.39–6.40 Mb) under selection were annotated with glutathione metabolic process. These two genes (*D03G045000* and *D03G045100*) have not been previously identified as having a specific role in any gene pathways in cotton. The minor alleles at the nearby significant SNP was more prevalent in the earlier breeding groups than later breeding groups, suggesting a role in early germplasm development. Genes in the glutathione metabolic pathway in cotton have been found to associate with resistance to wilt caused by *Verticillium dahliae* and mediate salt stress (Meloni *et al.* 2003; Li *et al.* 2019).

A pair of tandem-repeat "formin-like protein 20" genes, annotated with the GO term "actin nucleation," were located near a significant SNP on chr A11 (at 3.35 Mb). Genes that affect the actin network that forms the cellular skeleton have been characterized as expressing in cotton fiber development and elongation (Li *et al.* 2005), and another gene that influences the actin network in cotton has been located in a selective sweep during domestication (Fang *et al.* 2017). Further work is needed to identify genes that influence cotton fiber formation and to determine if this locus is important for fiber production.

Five genes with the GO term "intracellular transport" and eleven with "translation" were also identified on chromosomes A11, D02, D03, and D09. Most of these genes have not been well characterized in cotton, although a few seem to be involved with host plant resistance. Seven of the eleven "translation" genes were annotated as involved in the "ribosome" pathway. One of the genes, *A11G030881*, a homolog of the *Arabidopsis ERF1* gene has been found to play a role in resistance to Verticillium wilt (Xu *et al.* 2011). One of the "intracellular transport" genes, *A11G032100*, is annotated as "vesicle transport v-SNARE 11-like," a member of family of genes that controls the transport of precursor molecules during gossypol production (Lang and Jahn 2008; Ting 2014). Gossypol levels are under genetic control and are thought to play a role in cotton host plant insect resistance (Liu *et al.* 2015).

Final remarks

Overall, we found evidence for sustained genetic diversity throughout eight breeding cycles of the PD program. Genetic signatures demarcating shifting breeding goals were evident after controlling for variable marker density across the genome associated with genotyping array ascertainment bias. We also found SNP alleles with increased frequency in the PD program relative to in other improved upland cotton germplasm, with nearby genes enriched for biological functions including response to auxin, glutathione biosynthesis, translation, and cellular localization, implicating genetic drift for QTLs underlying host plant resistance. An additional locus under selection was found for actin nucleation, which may be a site that participated in fiber improvement in the PD program. The results of this study contribute to the growing body of knowledge regarding the breeding history of upland cotton in the southeastern United States and the world. In addition, our findings in this study inform future breeding efforts based on PD program materials by establishing the basis for ongoing development of marker-assisted selection and genomic selection. The PD cotton germplasm enhancement program, an 85+-year-old cotton improvement experiment, serves as a model system to study population genetics in the context of continued cotton improvement over the course of multiple breeders, breeding goals, and sources of genetic material.

Acknowledgments

We acknowledge the Palmetto Cluster at Clemson University for providing the computational resources to run BayEnv2. We also acknowledge Dr. Jason Holliday for his suggestion to use Discriminant Analysis of Principal Components, and Dr. Chris Saski for his suggesting of R packages to use in this project. We thank Ms. Reid Stephens for her assistance with DNA extraction.

Funding

Technical Contribution No. 6946 of the Clemson University Experiment Station. This material is based upon work supported by NIFA/USDA, under project number SC-1700561. In addition, we acknowledge financial support from CRIS No. 6082-21000-008-00D of the U.S. Department of Agriculture and additional support from Cotton Incorporated Project No. 19-872. 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 U.S. Department of Agriculture.

Conflicts of interest

None declared.

Literature cited

Abdelraheem A, Elassbli H, Zhu Y, Kuraparthy V, Hinze L, *et al.* 2020. A genome-wide association study uncovers consistent quantitative trait loci for resistance to Verticillium wilt and Fusarium wilt race 4 in the US Upland cotton. Theor Appl Genet. 133:563–577.

Ahn SJ, Costa J, Emanuel JR. 1996. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic Acids Res. 24:2623–2625.

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 27:2534–2547.

Alexa A, Rahenfuhrer J. 2020. R package 'topGO': Enrichment Analysis for Gene Ontology (version 2.40.0).

Beasley JO. 1940. The origin of American tetraploid *Gossypium* species. Am Nat. 74:285–286.

Bourgeois Y, Hazzouri KM, Warren B. 2017. Going down the rabbit hole: a review on methods characterizing selection and demography in natural populations. bioRxiv 052761. https://www.biorxiv.org/content/10.1101/052761v3.

Bowman DT, Gutierrez OA. 2003. Sources of Fiber Strenth in the U.S. Upland Cotton Crop from 1980 to 2000. J Cotton Sci. 7:164–169.

Brown N, Kumar P, Singh R, Lubbers E, Campbell BT, *et al.* 2019. Evaluation of a chromosome segment from *Gossypium barbadense* harboring the fiber length QTLqFL-Chr.25 in four diverse upland cotton genetic backgrounds. Crop Sci. 59:2621–2633.

Calhoun DS, Bowman DT and May OL. 1997. Pedigrees of Upland and Pima Cotton Cultivars Released Between 1970 and 1995.Starkville, MS: Mississippi Agricultural & Forestry Experiment Station.

Campbell BT, Chee PW, Lubbers E, Bowman DT, Meredith WR, *et al.* 2011. Genetic improvement of the pee dee cotton germplasm collection following seventy years of plant breeding. Crop Sci. 51:955–968.,

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.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, *et al.* 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4:7.

Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, *et al.* 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet. 52:525–533.

Coop G, Witonsky D, Rienzo AD, Pritchard JK. 2010. Using environmental correlations to identify loci underlying local adaptation. Genetics. 185:1411–1423.

Culp TW, Harrell DC, Kerr T. 1979. Some genetic implications in the transfer of high fiber strength genes to upland cotton. Crop Sci. 19:481–484.

Deperi SI, Tagliotti ME, Bedogni MC, Manrique-Carpintero NC, Coombs J, *et al.* 2018. Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. PLoS One. 13:e0194398.

Fang L, Wang Q, Hu Y, Jia Y, Chen J, *et al.* 2017. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet. 49:1089–1098.

Gapare W, Conaty W, Zhu Q-H, Liu S, Stiller W, *et al.* 2017. Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. Euphytica. 213: 66.

Gauch HGJ, Qian S, Piepho HP, Zhou L, Chen R. 2019. Consequences of PCA graphs, SNP codings, and PCA variants for elucidating population structure. PLoS One. 14:e0218306.

Gunther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. Genetics. 195:205–220.

Hamblin MT, Buckler ES, Jannink JL. 2011. Population genetics of genomics-based crop improvement methods. Trends Genet. 27: 98–106.

Harrell DC. 1974. ARS-S-30: Breeding Quality Cotton and the Pee Dee Experiment Station Florence S.C. Florence, SC: USDA.

He S, Wang P, Zhang YM, Dai P, Nazir MF, *et al.* 2020. Introgression leads to genomic divergence and responsible for important traits in upland cotton. Front Plant Sci. 11:929.

Hinze LL, Hulse-Kemp AM, Wilson IW, Zhu QH, Llewellyn DJ, *et al.* 2017. Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. BMC Plant Biol. 17:37.

Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, *et al.* 2015. MAGIC populations in crops: current status and future prospects. Theor Appl Genet. 128:999–1017.

Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, *et al.* 2015. Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. G3 (Bethesda). 5:1187–1209.

Jeong S, Kim JY, Jeong SC, Kang ST, Moon JK, *et al.* 2017. GenoCore: a simple and fast algorithm for core subset selection from large genotype datasets. PLoS One. 12:e0181420.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35:1547–1549.

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays. 35:780–786.

Lang T, Jahn R. 2008. Core proteins of the secretory machinery. In: TC Sudhof, K Starke, editors. Pharmacology of Neorotransmitter Release. Berlin Heidelberg: Springer-Verlag, p. 107–127.

Lekshmy S, Krishna GK, Jha SK, Sairam RK. 2017. Mechanism of auxin mediated stress signaling in plants. In: G, Pandey editor. Mechanism of Plant Hormone Signaling under Stress. Hoboken, NJ: John Wiley & Sons, Inc.

Li ZK, Chen B, Li XX, Wang JP, Zhang Y, *et al.* 2019. A newly identified cluster of glutathione S-transferase genes provides Verticillium wilt resistance in cotton. Plant J. 98:213–227.

Li XB, Fan XP, Wang XL, Cai L, Yang WC. 2005. The cotton *ACTIN1* gene is functionally expressed in fibers and participates in fiber elongation. Plant Cell. 17:859–875.

Li X, Liu G, Geng Y, Wu M, Pei W, *et al.* 2017. A genome-wide analysis of the small auxin-up RNA (SAUR) gene family in cotton. BMC Genomics. 18:815.

Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, *et al.* 2015. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci. Rep. 5:14139.

Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, *et al.* 2018. Efficiency of different strategies to mitigate ascertainment

bias when using SNP panels in diversity studies. BMC Genomics. 19:22.

Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, *et al.* 2015. Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics. 16: 290.

May OL. 1999. Registration of PD 94042 germplasm line of upland cotton with high yield and fiber maturity. Crop Sci. 39:597–598.

McGill BJ, Maurer BA, Weiser MD. 2006. Empirical evaluation of neutral theory. Ecology. 87:1411–1423.

Meloni DA, Oliva MA, Martinez CA, Cambraia J. 2003. Photosynthesis and activity of superoxide dismutase, peroxidase and glutathione reductase in cotton under salt stress. Environ Exp Bot. 49:69–76.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, *et al.* 2010. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor Appl Genet. 120:1525–1534.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40:646–649.

Odong TL, van Heerwaarden J, Jansen J, van Hintum TJ, van Eeuwijk FA. 2011. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet. 123:195–205.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35: 526–528.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, *et al.* 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38:904–909.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 197:573–589.

Shang L, Wang Y, Wang X, Liu F, Abduweli A, *et al.* 2016. Genetic analysis and QTL detection on fiber traits using two recombinant inbred lines and their backcross populations in upland cotton. G3 (Bethesda). 6:2717–2724.

Shen C, Li X, Zhang R, Lin Z. 2017. Genome-wide recombination rate variation in a recombination map of cotton. PLoS One. 12: e0188682.

Sun Z, Li H, Zhang Y, Li Z, Ke H, *et al.* 2018. Identification of SNPs and candidate genes associated with salt tolerance at the seedling stage in cotton (*Gossypium hirsutum* L.). Front Plant Sci. 9:1011.

Ting HM. 2014. Biosynthesis and transport of terpenes [Doctoral dissertation]. Wageningen, The Netherlands: Graduate School of Experimental Plant Sciences, Wageningen University.

Tyagi P, Gore MA, Bowman DT, Campbell BT, Udall JA, *et al.* 2014. Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet. 127:283–295.

Ugland KI, Gray JS, Ellingsen KE. 2003. The species-accumulation curve and estimation of species richness. J Anim Ecol. 72:888–897.

Xu L, Zhu L, Tu L, Guo X, Long L, *et al.* 2011. Differential gene expression in cotton defence response to *Verticillium dahliae* by SSH. J Phytopathol. 159:606–615.

Yang Z, Ge X, Yang Z, Qin W, Sun G, *et al.* 2019. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nat Commun. 10:2989.

Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, *et al.* 2014. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 42:D1229–D1236.

Zhang JF, Lu Y, Adragna H, Hughs E. 2005. Genetic improvement of New Mexico Acala cotton germplasm and their genetic diversity. Crop Sci. 45:2363–2373.

Communicating editor: J. Wendel