Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GRAS-Di SNP-based molecular characterization and fingerprinting of Upland cotton germplasm: insights into genetic diversity and structure

Asena Akköse Baytar
Izmir Institute of Technology
Ertuğrul Gazi Yanar
Izmir Institute of Technology
Volkan Sezener
Nazilli Cotton Research Institute
Anne Frary
Izmir Institute of Technology
Sami Doğanlar
samidoganlar@iyte.edu.tr
Izmir Institute of Technology

Research Article

Keywords: Genetic diversity, Gossypium hirsutum, GRAS-Di genotyping, SNP fingerprinting, Population structure

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11124828/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GRAS-Di SNP-based molecular characterization and fingerprinting of Upland cotton germplasm: insights into genetic diversity and structure

Asena Akköse Baytar[1], Ertuğrul Gazi Yanar[1], Volkan Sezener [3], Anne Frary[1], Sami Doğanlar[1,2]

[1]Department of Molecular Biology and Genetics, Faculty of Science, Izmir Institute of Technology, Izmir, TÜRKİYE

[2]Plant Science and Technology Application and Research Center, Izmir Institute of Technology, Izmir, TÜRKİYE

[3]Nazilli Cotton Research Institute, Nazilli, Aydın, TÜRKİYE

Corresponding author: Sami Doğanlar

E-mail: samidoganlar@iyte.edu.tr

ORCID

Asena Akköse Baytar https://orcid.org/0000-0002-1068-4544

Ertuğrul Gazi Yanar https://orcid.org/0000-0002-1325-0346

Volkan Sezener https://orcid.org/0000-0001-7046-2065

Anne Frary https://orcid.org/0000-0002-8973-0100

Sami Doğanlar https://orcid.org/0000-0002-5712-0826

Key Message

Genome-wide analysis of 200 Upland cotton genotypes revealed population structure and linkage disequilibrium patterns, established a representative core collection and identified a minimal 20 SNP fingerprinting panel that uniquely discriminated all genotypes, providing practical genomic resources for germplasm management and future cotton breeding.

Abstract

Upland cotton (*Gossypium hirsutum* L.) breeding has been shaped by domestication bottlenecks and intensive selection resulting in a narrowed genetic base that challenges the effective utilization of germplasm resources. In this study, we investigated genetic diversity, population structure, linkage disequilibrium patterns and fingerprinting in a panel of 200 Upland cotton genotypes using a high-quality SNP dataset comprising 8,013 markers. Genetic diversity analysis revealed moderate variation ranging from 0.06 to 0.37 with a mean of 0.27. Population structure analysis based on Bayesian clustering supported the

presence of five genetic clusters while discriminant analysis of principal components resolved the panel into four genetic groups comprising 41, 48, 74, and 37 genotypes, respectively. The population structure clusters showed strong concordance with DAPC group assignments whereas individuals classified as admixed by Bayesian inference were further resolved by multivariate analysis. Genome-wide LD analysis showed extended LD decay with r^2 decreasing gradually with physical distance and remaining above background levels over megabase scale intervals (~2.0 Mb, r = 0.35). Identification of the most genetically divergent and closely related genotype pairs in this study may provide practical insights into parent selection and germplasm management. Furthermore, we applied a minimal marker strategy and identified a panel of 20 SNPs that uniquely discriminated all genotypes within the panel. Overall, although genetic diversity in cultivated Upland cotton remains constrained, sufficient variation is available to support further genomic analyses and breeding. The genomic framework established here may contribute to the characterization and utilization of Upland cotton germplasm in future genetic improvement.

Key words Genetic diversity, *Gossypium hirsutum*, GRAS-Di genotyping, SNP fingerprinting, Population structure

Abbreviations

BIC Bayesian Information Criterion
DAPC discriminant analysis of principal components
GBS genotyping by sequencing
GWAS genome wide association analysis studies
LD linkage disequilibrium
SNP single nucleotide polymorphism

Introduction

Cotton is a globally important natural fiber crop derived from genus *Gossypium* L. and is valued for its economic, industrial, and agricultural significance. *Gossypium* belongs to the Malvaceae (mallow) family which comprises 250 accepted genera and approximately 4,230 species. Within this family, *Gossypium* includes 55 species (Oliveira Da Silva et al. 2010; Wendel and Grover 2015) that are distributed across tropical and subtropical regions worldwide. However, only four species have been domesticated and cultivated for economic purposes. These species are the diploids *Gossypium herbaceum* L. ($n = x = 13$, A_1) and *Gossypium arboreum* L. ($n = x = 13$, A_2) and the tetraploids *Gossypium hirsutum* L. [$n = 2x = 26$, $(AD)_1$] and *Gossypium barbadense* L. [$n = 2x = 26$, $(AD)_2$]. Of these, *G. hirsutum*, commonly known as Upland cotton, dominates global cotton production accounting for more than 90% of total fiber production

due to its superior fiber yield and adaptability (Wendel and Cronn 2003; Chen et al. 2020). Currently, the top three cotton producing countries are China accounting for 26.0% of global production (18.7 Mt), India with a 20.9% share (15.0 Mt) and the United States of America (9.5 Mt) contributing 13.3%. Türkiye ranks eighth in global cotton production with a 3.1% share of global production (2.2 Mt) and holds third place in yield achieving 4,803.7 kg/ha (FAO 2026).

G. hirsutum is native to the tropical and subtropical regions of Central America, Mexico, and the Caribbean (Wendel et al. 2009). Wild populations are typically found in coastal and inland areas where warm temperatures and seasonal rainfall patterns create favorable conditions for growth (Fryxell 1986; Coppens d'Eeckenbrugge and Lacape 2014). These wild forms are important reservoirs of genetic diversity having adapted to a variety of environmental conditions over thousands of years (Vega et al. 2023; Ning et al. 2024). Natural introgression patterns suggest that *G. barbadense* (Pima cotton) alleles were likely transferred to *G. hirsutum* in regions where their growing areas overlapped such as the Caribbean (Wang et al. 1995). In modern elite cultivars, however, gene flow usually happens in the opposite direction from *G. hirsutum* to *G. barbadense* (Wang et al. 1995). According to molecular studies, 8.9% of alleles in current *G. barbadense* cultivars may have originated from *G. hirsutum*. Compared to other commercial cotton species, *G. hirsutum* has a notably diverse genetic background as shown by many molecular studies (Hinze et al. 2017; Wang et al. 2022b; Safiullina et al. 2024). However, cultivation and domestication processes have significantly reduced genetic diversity with only about half of the original genetic variation being represented in today's *G. hirsutum* cotton cultivars (Brubaker and Wendel 1994; Lacape et al. 2006). During domestication, some traits were selected based on the different needs of wild versus cultivated environments. For example, wild cotton favors seeds with traits like easy dispersal and hard seed coats which are less desirable for cultivation since reduced seed dispersal and softer seed coats facilitate processing and germination (Lee 1984; Fryxell 1986; Westengen et al. 2005).

In the early 1900s, a severe boll weevil (*Anthonomus grandis*) infestation spread from Mexico into the United States devastating major cotton fields and causing the loss of many cotton types (Tyler 1910; Cook 1912; Lubbers et al. 2006; Ware 2023). In response, breeding programs were initiated to develop cultivars capable of withstanding boll weevil infestations resulting in the release of lines with enhanced tolerance and agronomic performance (such as Lone Star, Stoneville, Coker-100, and Deltapine) derived from the Western Big Boll type (Cook 1912; Lubbers et al. 2005; Ware 2023; Beasley 2024). Since then, the economic importance of Upland cotton has increased and new cultivars suitable to regional environments have been developed leading to the differentiation of various cotton types based on morphological traits (Duggar 1907; Tyler 1910; Lubbers et al. 2006).

Germplasm is an invaluable source of alleles for creating new lines with specific traits (Tanksley and McCouch 1997; Hinze et al. 2025). Broad genetic diversity within germplasm is necessary for

developing new cotton cultivars with traits of interest (Van Esbroeck and Bowman 1998) and molecular genetic studies help characterize this diversity. When pedigree information is missing, this type of analysis can be used to select parent lines by deciphering the genetic relationships among accessions. Therefore, selecting genetically diverse cotton genotypes or those with unique characteristics such as resistance to biotic and abiotic stresses or superior fiber quality, can help foster greater variability and adaptability in breeding programs.

Despite extensive characterization efforts within cotton germplasm collections, much work is still required. For example, only 5% of accessions in the U. S. National Plant Germplasm System have been genotyped using single nucleotide polymorphism (SNP) arrays, while 7% have been characterized through resequencing-based SNP discovery and just 0.2% have undergone de novo whole-genome sequencing (Hinze et al. 2025). Thus, only a limited fraction of the available germplasm has been surveyed using high-density genomic tools to assess genetic diversity, population structure and core collection development. For example, Hinze et al. (2017) used the CottonSNP63K array to evaluate genetic diversity in 395 *Gossypium* accessions collected worldwide and housed at the USDA National Cotton Germplasm Collection. They reported that the SNP markers clearly distinguished *G. hirsutum* from other species separating wild and cultivated types. Loci associated with cottonseed protein content were identified and aligned well with previous simple sequence repeat (SSR)-based studies. Ai et al. (2017) analyzed 288 elite *G. hirsutum* cultivars that were primarily of Chinese origin using genome wide SNP markers to assess genetic diversity and population structure. They determined a low average PIC value of 0.25 and identified three subgroups with extensive admixture pattern. Most genetic variation existed within groups supporting future breeding applications. J. Wang et al. (2022) analyzed 273 elite, mainly Chinese accessions using 1,313,331 SNP markers to evaluate population structure, genetic diversity and relationships. The material had an average genetic distance of 0.25 and low differentiation (Fst: 0.024–0.107) among six regional groups with admixture dividing the accessions into 16 subgroups revealing regionally influenced clustering patterns. Han et al. (2022) used 503 Upland cotton genotypes (477 from China and 26 introduced lines from America and the former Soviet Union) to construct a representative core germplasm collection for breeding. Three preliminary germplasm banks were developed based on phenotypic traits (ten yield and fiber traits), genome-wide SNP data (11,975 SNPs from the CottonSNP63K chip), and favorable alleles (89 significant trait-associated SNPs). In this way, a core set of 85 accessions (an 83% reduction of the original population) was constructed which retained most of the genetic diversity (PIC and He > 80% of the original). This set was proposed as a compact and diverse genetic resource to increase cotton breeding efficiency (Han et al. 2022). Gowda et al. (2023) analyzed 657 U.S. accessions using 26,952 SNPs from the CottonSNP63K array to evaluate genetic diversity, structure, and to perform genome-wide association analysis (GWAS) of photoperiod sensitivity. They reported a decline in genetic diversity from landraces (0.23) to elite lines

(0.12) identifying seven subpopulations in the germplasm. They also detected 14 marker-trait associations across eight chromosomes including candidate genes such as VOZ1 and PIE1.

In the present study, we investigated genetic diversity, population structure, ancestral background, linkage disequilibrium and fingerprinting in a panel of 200 *G. hirsutum* genotypes which were officially registered in Türkiye as well as those developed in other regions. The panel was analyzed using a high-quality SNP dataset comprising 8,013 markers. This work aimed to provide a robust genomic framework for characterizing cotton germplasm by integrating multivariate and model-based clustering approaches with genome-wide linkage disequilibrium (LD) analysis. The resulting insights into genetic differentiation, admixture patterns, and LD decay are directly relevant for optimizing germplasm management, guiding parent selection and defining appropriate marker densities for GWAS studies and molecular breeding applications in Upland cotton.

Materials and methods

Plant material

The germplasm panel consisted of 200 Upland cotton (*G. hirsutum*) genotypes (Supplementary Table S1) supplied by the Nazilli Cotton Research Station in Aydın, Türkiye. This collection included both cultivars bred and officially registered within Türkiye as well as those developed in other regions. Of them, 107 originated from Türkiye, 60 from the USA, 13 from Australia, 5 from Spain, 4 from Greece, 3 from Uzbekistan, 2 each from Pakistan and Syria, and 1 each from Albania, Iran, Israel, and South Africa.

DNA isolation of plant material and sequencing by GRAS-Di

DNA was isolated from leaf samples using the MN NucleoSpin Plant II kit. The quality and quantity of the isolated samples were measured using a Thermo Scientific™ Multiskan™ GO Microplate Spectrophotometer. The concentration of the samples was adjusted to a minimum of 30 ng/µl and a total quantity of at least 600 ng in a volume of no less than 25 µl. The DNA samples were sent to Eurofins Genomics (Ebersberg, Germany) for library preparation and subsequent sequencing. The genotyping-by-random-amplicon sequencing method (Hosoya et al. 2019) was used for GBS of the DNA samples. In brief, an indexed sequencing library was constructed through two successive polymerase chain reactions (PCR). First, genomic DNA was amplified in a primary PCR using a mixture of 63 randomly designed primers (Supplementary Table S2). Each of these random primers included a 10-base Illumina Nextera adapter sequence and a 3-base random oligomer. Next, the products of the first PCR were further amplified using a second PCR reaction with barcoded adapter primers that enabled multiplex sequencing. The resulting PCR products were pooled to construct a multiplex sequencing library. Sequencing of this library was performed using the latest chemistry (v4) on an Illumina HiSeq4000 platform (100 bp x 2). The raw sequencing reads were demultiplexed using the unique 8-bp dual-index codes in each sample's P5–P7

adapter combination allowing for a single-base mismatch. Eurofins Genomics then provided the demultiplexed reads as compressed FASTQ archives (fastq.gz).

Genotyping

Demultiplexed reads were processed with dDocent v2.9.8 using default settings for SNP genotyping (Puritz et al. 2014). First, low-quality regions and adaptor contamination were removed with fastp (Chen et al. 2018). The trimmed reads were then aligned to the cotton reference genome (NCBI RefSeq assembly *G. hirsutum* V2, GCF_007990345.1) with BWA-MEM and SAMtools (Li et al. 2009; Li 2013). Variant calling with FreeBayes (Garrison and Marth 2012) produced a raw VCF containing both SNPs and INDELs. Filtering followed the workflow of O'Leary et al. (2018) and was automated with a modified R script from TinHan et al. (2020) using VCFtools plus vcfallelicprimitives and vcffilter in vcflib (Danecek et al. 2011; Garrison 2016). Complex alleles were decomposed into primitive variants using vcfallelicprimitives, and INDELs were removed. Loci with quality scores < 20 were excluded and genotypes with genotype call quality (GQ) < 30 or genotype depth (DP) < 5 were recorded as missing. Loci showing > 50% missing data or a mean depth < 15 were discarded. Next, loci that had high variation (99th-quantile: 87) in mean coverage between individuals were removed. Finally, loci with > 25% missing data were removed.

The resulting loci were then filtered based on the INFO-field parameters. Loci were first filtered based on allele balance (AB), keeping loci with AB < 0.01, > 0.99 or between 0.20 and 0.80. Next, the loci with quality to depth ratio (QUAL/DP) < 0.20 were removed. Then, loci with mapping-quality ratios outside 0.25–1.75 were discarded, and filters based on properly-paired status were applied to retain balanced PAIRED and PAIREDR metrics, as described by O'Leary et al. (2018). Additional depth-versus-quality filtering was applied based on the recommendation of H. Li (2014). Briefly, loci with $depth > mean\ depth + 3 \times \sqrt{mean\ depth}$ were flagged and removed if their loci quality scores were less than two times of their loci depth scores. An upper mean-depth threshold was set to remove loci with excessive mean depth (95th quantile: 486), indicative of paralogous loci. Finally, multiallelic loci and loci with > 10% missing genotypes were removed, and a minor-allele count filter (MAC < 10) was applied to yield a high-quality dataset for downstream analyses.

Linkage disequilibrium analysis

LD statistics were calculated using PLINK v1.9 (Chang et al. 2015; Purcell and Chang 2019) for pairwise SNPs within each chromosome using a 1 Mb sliding window across the cotton genome. LD was quantified as the squared correlation coefficient ($0 < r^2 < 1$) between SNP pairs. LD decay was assessed by plotting pairwise r^2 values against the physical distance between SNPs in R. LD decay curves were generated using a custom R script that calculated mean r^2 values across distance bins and fit a nonlinear regression model to describe LD decay with increasing physical distance. SNP thinning was performed after LD decay estimation using the half-decay distance derived from the LD decay curve to reduce marker redundancy.

Genetic diversity analysis

Genetic relationships of the cotton genotypes were examined by constructing a dendrogram. First, TASSEL v5.2.59 was used for the final SNP dataset to compute an identity-by-state (IBS) genetic-distance matrix. This matrix was then converted to DIS format by TASSEL and imported into DARwin v6.0.21 where an unweighted neighbor-joining tree was generated (Perrier and Jacquemoud-Collet 2006; Glaubitz et al. 2014). Discriminant analysis of principal components (DAPC) was conducted using the adegenet R package (Jombart 2008; Jombart et al. 2010) with SNP data converted from VCF format to a genlight object using dartR (Gruber et al. 2018).

Population structure analysis

Population structure was inferred using the Bayesian clustering algorithm implemented in STRUCTURE v2.3.4. StrAuto pipeline was performed to automate multiple independent runs under identical parameter settings. SNP data in VCF format were converted to STRUCTURE input format using PGDSpider v2.1.1.5 (Lischer and Excoffier 2012). STRUCTURE was run under the admixture model with correlated allele frequencies. The number of genetic clusters (K) was evaluated for values ranging from K = 1 to K = 10 with ten independent replicates for each K value. Each replicate consisted of a burn-in period of 100,000 iterations followed by 100,000 Markov Chain Monte Carlo (MCMC) iterations. The optimal number of clusters was determined using the Evanno method based on the rate of change in the log probability of the data between K values (ΔK). ΔK values were calculated using STRUCTURE HARVESTER and the K value with the high ΔK peak was considered to be the optimal cluster number. Among the ten replicated runs at the selected K, the run with the highest mean log-likelihood value was selected for downstream analyses. Individual membership coefficients (Q values) from this run were used to assign genotypes to genetic clusters. Genotypes with a membership coefficient of ≥ 0.60 were assigned to a single cluster whereas genotypes with lower values were classified as admixed. Population structure results were visualized using bar plots generated from the selected Q-matrix in R employing the ggplot2 package (Wickham 2016) for graphical representation.

Core collection construction

A core subset was selected from the SNP dataset using the R package Core Hunter 3 (De Beukelaer et al. 2018). The dataset was converted into a Core Hunter genotype object using the biparental format. Core subset selection was performed using a combined objective function integrating average entry-to-nearest-entry distance (EN) and Modified Rogers' distance (MR) to maximize allelic diversity and genetic distance simultaneously. The subset size was predefined as 50 genotypes from the original germplasm panel. Selection was performed under the default search mode with normalization enabled to balance objective contributions. A fixed random seed was used to ensure reproducibility of the optimization procedure. Genetic diversity of the core collection was evaluated using the genetic distance matrix in DARwin v6.0.21.

Fingerprinting

A minimal SNP subset capable of uniquely discriminating all genotypes in the panel was identified using a custom Python implementation of the Minimal Marker algorithm (Fujii et al. 2013; Yanar et al. 2023). The script evaluated all pairwise genotype combinations and constructed a binary discrimination matrix in which homozygous genotype differences between pairs are coded as 1 whereas identical calls, heterozygous calls, or missing data are coded as 0.

SNPs were ranked according to their discriminatory power which was defined as the number of genotype pairs they resolved. A heuristic procedure was then applied, iteratively selecting the SNP that maximized the number of unresolved pairwise differences until all distinguishable genotype combinations were resolved. Two output datasets were generated: a masked dataset in which heterozygous calls were converted to missing (N) to ensure conservative fingerprinting, and a full dataset retaining original heterozygous codes. Pairwise IBS distances among genotypes were subsequently calculated using the reduced SNP set, and the resulting distance matrix was used to construct a neighbor-joining dendrogram in DARwin v6.0.21 for visualization of genetic relationships.

Results

Genotyping

A stringent variant filtering approach yielded 8,013 SNPs in 200 cotton genotypes. No multi-nucleotide polymorphisms, indels, or other variant types remained after filtering. The transition/transversion ratio (ts/tv) was 2.78 (5,891 transitions vs. 2,122 transversions) reflecting a strong excess of A↔G and C↔T changes as expected and indicating high data quality. Examination of substitution classes showed that C→T (20.5%), G→A (19.7%), A→G (16.5%), and T→C (16.9%) transitions dominated (Supplementary Table S3), while the most frequent transversion was A→C (3.9%).

High quality SNP markers were distributed across the 26 chromosomes of the *G. hirsutum* reference genome spanning both the A subgenome with 5,504 SNPs (68.7%) (A01–A13) and D subgenome with 2,425 SNPs (30%) (D01–D13) (Fig. 1a and 1b). A total of 88 SNPs were located on unplaced genomic scaffolds. The mean distance between SNPs was consistently lower in the A subgenome (312.1 kb) than in the D subgenome (382.5 kb) (Fig. 1c). SNP density varied substantially among chromosomes with SNP counts ranging from 103 on D03 to 857 on A08 with a mean of 305.0 ± 185.9 SNPs (Supplementary Fig. S1, Table 1). Chromosome A08 exhibited the highest marker density with an average distance between SNPs of 148.77 kb whereas the lowest density was observed on A02 (632.40 kb) and D11 (587.29 kb). These results indicated that some chromosomes accumulated nearly eight-fold more SNPs than others suggesting uneven patterns of genetic variation across the genome. Overall coverage was high across many

chromosomes exceeding 95% in most cases and reaching up to 99.9% on A08. Reduced coverage was observed on a limited number of chromosomes, notably D01 (92.3%).

The genomic distribution of SNPs was then examined based on their localization within major genomic regions. The majority of SNPs were in intergenic regions (90.2%) and a smaller proportion of SNPs mapped to exonic (7.2%) or intronic regions (2.5%). Very few SNPs were located in splice junctions (0.1%) (Table 2). Further functional annotation indicated that most SNPs were in intergenic (66.8%) or upstream regulatory regions (18.8%) suggesting predominantly indirect effects on gene function (Supplementary Table S4 and Supplementary Table S5). Consistent with this pattern, most variants were predicted to have minimal functional consequences (95.6% with modifier impact) whereas a smaller proportion was associated with altered amino acid sequences (2.2% with moderate impact, 2.1% with low impact) and only a few variants were inferred to have potential high impact consequences on gene structure or function such as loss of function events (0.1% with high impact) (Supplementary Table S6). Transcript type classification further showed that most genic SNPs were associated with protein coding genes (19.4%) with fewer variants detected in pseudogenes and non-coding RNA classes (Supplementary Table S7).

Linkage disequilibrium (LD) decay

LD decay was evaluated using pairwise r^2 values calculated across all chromosomes. As expected, LD declined rapidly with increasing physical distance (Supplementary Fig. S2). Mean r^2 values were high at short distances approaching approximately 0.70 at very close marker intervals and decreased sharply within the first few kilobases. The nonlinear decay curve indicated that r^2 dropped to approximately half of its initial value ($r^2 = 0.35$) within approximately 2.0 to 2.5 Mb after which the rate of decay slowed markedly. Beyond 7 to 8 Mb, LD reached a plateau with r^2 values stabilizing at around 0.15. The estimated half-decay distance (2.0–2.5 Mb) was used to guide SNP thinning for genetic diversity and population structure analyses to reduce the influence of tightly linked markers and to avoid overrepresentation of high LD genomic regions (Kaeuffer et al. 2007; Zheng et al. 2012; Chang et al. 2015). In this way, a panel of 805 nearly independent SNPs was retained for multivariate and clustering analyses.

Genetic diversity and population structure

Genetic diversity and population structure were assessed to characterize the extent of variation and the underlying relationships among the cotton genotypes. The unweighted neighbor-joining dendrogram based on pairwise IBS genetic distances revealed four major groups comprising 47, 15, 112, and 26 genotypes assigned to D1, D2, D3, and D4, respectively (Fig. 2). The panel had genetic diversity values ranging from 0.06 to 0.37 with an average of 0.27 (Supplementary Fig. S3). The greatest genetic distance was observed between genotypes TKY9309 and SAHRA09 whereas the lowest genetic distance was identified between DP493 and DPL882 (Supplementary Table S8).

Within-origin analyses revealed similar variation in different countries' materials. The highest mean genetic distance was observed in the Greek subset (0.30; n = 4), followed by Uzbekistan (0.26; n = 3) and USA (0.26; n = 60). Turkish genotypes (n = 107) exhibited moderate variation (0.25), while Spain (0.24; n = 5) and Australia (0.24; n = 13) had slightly lower genetic divergence (Supplementary Fig. S4).

The Evanno method was applied to the STRUCTURE results to determine the most likely number of genetic clusters. A peak in ΔK was observed at K = 5 (Supplementary Fig. S5) indicating that five clusters best explained the population structure of the analyzed cotton panel. Accordingly, the results at K = 5 were examined and the fourth replicate run was selected for further analysis due to the highest mean log-likelihood. Using a 60% membership coefficient threshold, 106 (53%) genotypes were assigned to a cluster whereas 94 (47%) genotypes were admixed. Among the non-admixed genotypes, clusters C1, C2, C3, C4, and C5 contained 12 (11.3%), 11 (10.4%), 29 (27.4%), 31 (29.2%), and 23 (21.7%) genotypes, respectively (Fig. 3, Supplementary Table S9).

DAPC identified clear genetic structure within the panel. Bayesian Information Criterion (BIC) values supported the presence of four genetic clusters (K = 4) with the lowest BIC observed at this level (Fig. 4, Supplementary Fig. S6 and Supplementary Fig. S7). Cross-validation analysis indicated high and stable predictive success across a wide range of retained PCA axes supporting the robustness of the selected model and the number of principal components used for discrimination. The DAPC scatter plot based on the first two discriminant functions showed clear separation among the four groups with some overlap between clusters DAPC1 and DAPC3 (Fig. 4). Groups DAPC1, DAPC2, DAPC3, and DAPC4 were composed of 41, 48, 74, and 37 individuals, respectively (Supplementary Table S10). The first discriminant axis accounted for the largest proportion of variation (47.2%) among groups (Supplementary Fig. S7) as indicated by the discriminant eigenvalues while subsequent axes contributed progressively less to group separation.

Population structure clusters C1, C2, C4, and C5 showed complete (100%) correspondence with DAPC groups DAPC3, DAPC1, DAPC2 and DAPC4, respectively (Supplementary Fig. S8). However, C3 exhibited partial correspondence with the DAPC analysis: 89.7% of individuals were assigned to DAPC3 while smaller proportions were distributed to DAPC1 (6.9%) and DAPC4 (3.4%). Admixed individuals were distributed across multiple DAPC groups with 36 individuals (38.3%) assigned to DAPC3, 28 (29.8%) to DAPC1, 17 (18.1%) to DAPC2, and 13 (13.8%) to DAPC4 (Supplementary Fig. S8). Although the STRUCTURE-inferred clusters represented well-defined genetic patterns, the multivariate DAPC approach separated admixed individuals into all major genetic groups, thereby resolving additional structure beyond Bayesian assignment alone.

The geographic composition of STRUCTURE-based genetic clusters revealed a weak association between cluster membership and country of origin (Supplementary Fig. S9). Across all clusters, genotypes

from Türkiye constituted the largest proportion accounting for more than half of the individuals in the admixed group (53.2%) and the majority in clusters C2 (72.7%), C4 (67.7%), and C5 (65.2%). Cluster C1 was predominantly composed of genotypes from the USA (83.3%) with only minor representation from Türkiye and Greece. Cluster C3 exhibited the highest geographic heterogeneity with genotypes originating from multiple regions including Türkiye (41.4%), USA (17.2%), Uzbekistan (10.3%), Greece (6.9%), and Pakistan (6.9%), with several additional countries each contributing smaller proportions. The admixed group also displayed substantial international representation with notable contributions from the USA (34.0%) and Australia (7.5%) in addition to Türkiye.

Core collection

Using a multi-objective optimization approach, a core collection of 50 genotypes was generated from the original 200-genotype panel (Supplementary Table S11). This approach ensured simultaneous maximization of allelic richness and genetic distance within the subset. The resulting core set retained broad genetic representation (average diversity of 0.28) across the panel while reducing population size by 75%. The core collection showed a similar distribution pattern to the full panel indicating that the subset retained the overall genetic diversity of the complete dataset (Supplementary Fig. S10). Moreover, the core collection preserveed the overall geographic representation of the full panel with Türkiye and the USA remaining the dominant contributors. Minor origins were also retained in the core subset indicating that the core collection retained the major geographic diversity present in the complete germplasm panel (Supplementary Fig. S11).

Fingerprinting

Application of the minimal marker based fingerprinting approach to the cotton panel identified a set of 20 SNPs that was sufficient to discriminate all 200 genotypes (Fig. 5). This reduced marker set resolved all pairwise genotype combinations without redundancy confirming its full discriminatory capacity within the panel. The complete list of fingerprinting SNPs is provided in Supplementary Table S12. Pairwise IBS distances were calculated using this minimal SNP set to produce a distance matrix. The resulting neighbor-joining dendrogram clearly differentiated all genotypes (Supplementary Fig. S12) demonstrating that a small, strategically selected marker subset can effectively determine genetic relationships within the panel.

Discussion

In this study, we performed a genome-wide assessment of genetic diversity, population structure and linkage disequilibrium in Upland cotton germplasm using high-density SNP markers. The germplasm revealed moderate genetic diversity with slow LD decay and a well-defined population structure. Further, we demonstrated that a reduced core collection and a minimal SNP marker set can efficiently represent and discriminate the germplasm panel.

Genome-wide SNP distribution showing higher marker abundance in the A subgenome than D subgenome

The results showed that our genomic data was suitable for downstream molecular genetic analyses as indicated by transition : transversion ratio of 2.78. This value is consistent with the range reported for cotton SNP datasets including ratios of 2.13–2.17 in Upland cotton lines (Song et al. 2018) and 3.04 in a large tetraploid cotton diversity panel (Rabieyan et al. 2023). In our study, chromosomes in the A subgenome harbored higher SNP counts and denser marker spacing compared to those in the D subgenome reflecting a more uniform and comprehensive SNP distribution across the A subgenome. This uneven chromosomal distribution of markers is also consistent with several previous cotton studies. For example, Song et al. (2018) reported more polymorphisms in the A than the D subgenome in *G. hirsutum* lines, and Guo et al. (2022) similarly found 54% of SNPs in the A subgenome versus 46% in the D subgenome in their cotton panel.

Allotetraploid cotton (*G. hirsutum*) originated approximately 1–2 million years ago through hybridization between an ancestral A genome lineage (A_0), closely related to extant A-genome diploids (*G. herbaceum and G. arboreum*) and a D genome ancestor similar to *G. raimondii* (Huang et al. 2020). Subsequent genome evolution has been characterized by transposable element activity causing expansion of the A subgenome and asymmetric evolution between the A and D subgenomes giving rise to differences in gene expression and selection pressure (Wendel and Grover 2015; Jazayeri et al. 2020; Chen et al. 2020; Huang et al. 2020). Genome-wide resequencing studies have revealed that pronounced genetic bottlenecks were associated with divergence and subsequent domestication of allotetraploid cotton (Fang et al. 2017). Therefore, the observed enrichment of SNPs in the A subgenome is consistent with its distinct evolutionary history and structural expansion relative to the D subgenome.

Extended LD decay reflecting breeding history and reduced genetic variation in *G. hirsutum*

The observed LD decay indicated the presence of long-range marker associations, suggesting that the inferred genetic clusters may be influenced by linkage as well as allele frequency differences. When compared with prior studies, our half-decay estimate (~2.0–2.5 Mb, $r^2 \approx 0.35$) is consistent with reports of multi-megabase LD decay in cultivated Upland cotton GWAS panels. For example, Gapare et al. (2017) reported that genome-wide LD in a panel of 90 *G. hirsutum* accessions decayed to 2 Mb at $r^2 = 0.2$ and reached approximately 5 Mb at $r^2 = 0.1$ showing extensive long-range LD in breeding germplasm. In another study, Yuan et al. (2018) estimated LD decay around 3.2–3.3 Mb at an r^2 of 0.1 in a 196-accession Upland cotton panel, reflecting a similarly long LD decay distance. In contrast, other studies have reported shorter LD decay in Upland cotton populations. For example, P. Xu et al. (2021) reported genome-wide LD decay of ~0.87 Mb ($r^2 = 0.2$) and ~1.12 Mb ($r^2 = 0.1$) across 217 cultivars. Zhao et al. (2022) reported a

half-maximum ($r^2 \approx 0.4$) LD decay around 388 kb in a larger panel (336 accessions) emphasizing that LD extent can shrink in more diverse or less-related collections and when marker density increases.

LD analysis is useful for determining the marker density that is required for further genome wide analysis (DoVale et al. 2022; Hong et al. 2024). When the LD decay distance increases, a SNP represents a larger segment of the genome in GWAS and is less likely to represent a single gene. Thus, mapping resolution will be limited to larger genomic intervals, making candidate gene identification more difficult. The LD decay value observed in this study may reflect a reduction in recombination driven by breeding-related methods including partial inbreeding and founder effects which can lead to the loss of genetic variation in cultivated Upland cotton (Iqbal et al. 2001; Fang et al. 2017). Such patterns are consistent with the historical use of a limited number of elite parental lines and recurrent selection resulting in large haplotype blocks maintained across the genome.

Moderate genetic variation within the Upland cotton panel, consistent with intensive artificial selection.

The level of genetic diversity (0.06–0.37; mean = 0.27) observed in this study, with country-level means ranging from 0.24 to 0.30, indicated a moderate degree of variation within the Upland cotton panel as reported previously for cultivated *G. hirsutum* (Tyagi et al. 2014; Billings et al. 2021; Gowda et al. 2023). Slightly higher diversity levels have been reported in large SNP-based panels. For example, 273 elite cultivars from China and the rest of the world had a genetic diversity of approximately 0.31 (Wang et al. 2022a) and a GWAS panel of 196 accessions had mean genetic diversity around 0.33 probably reflecting the inclusion of exotic accessions with the predominantly Chinese cultivars (Yuan et al. 2018). In contrast, lower diversity values have been reported for more narrowly defined breeding materials including diversity estimates of 0.23 for landraces, 0.20 for elite germplasm and 0.12 for a public breeding panel (Gowda et al. 2023). Together, these results show that genetic diversity in modern cultivars is reduced, due to domestication bottlenecks, founder effects and intensive selection during breeding (Tyagi et al. 2014; Billings et al. 2021; Gowda et al. 2023).

In our study, the slightly higher diversity observed in the USA subset (0.26; n = 60) likely reflected the mixture of materials from different breeding regions and admixed lineages, as previously reported in US cotton panels (Tyagi et al. 2014). The Turkish subset (0.25; n = 107) showed moderate but slightly lower diversity which may indicate shared ancestry among elite cultivars developed within national breeding programs. Therefore, future cotton breeding in Türkiye may benefit from incorporating more genetically diverse germplasm from international collections to broaden the breeding pool and increase the potential for genetic gain. The higher mean values observed in Greece (0.30, n = 4) and Uzbekistan (0.26, n = 3) should be interpreted with caution as small sample sizes can inflate average pairwise distances.

The most divergent and closest related genotype pairs were identified. This information can guide germplasm utilization by allowing selection of parental combinations for maximizing genetic divergence. It also allows detection of nearly duplicate materials that may be redundant in breeding programs. Therefore, this information can be used for devising more efficient parental selection and hybridization strategies and the rational management of cotton genetic resources.

Optimized core collection preserving genetic diversity while reducing population size

Core collection strategies have been widely applied in cotton to reduce large germplasm panels into representative subsets that retain genetic diversity while improving manageability for breeding and genomic studies (Xu et al. 2006; Li et al. 2022a). H. Xu et al. (2006) developed a core collection of *G. barbadense* using a clustering-based approach to capture fiber trait variation. This work demonstrated that reduced subsets can preserve phenotypic and genetic patterns. In Upland cotton, C. Li et al. (2022) constructed a core collection (122) from 491 *G. hirsutum* accessions genotyped with a high-density SNP array to support fiber quality improvement and parental selection. Moreover, their resequencing-based studies in China have employed structured Upland cotton core collections to investigate genome-wide variation and breeding signatures, demonstrating the value of such subsets for association mapping and trait dissection (Li et al. 2022a).

In this context, we generated a 50-genotype core set derived from our Upland cotton panel using established cotton core collection practices. By optimizing both genetic distance and heterozygosity criteria, the selected subset maintains broad genetic representation (0.28) while reducing panel size by 75%. Therefore, the subset may lower phenotyping and multi-environment trial costs without substantially limiting genetic variability.

Population structure analysis provides a framework for reconstructing the genetic composition of a population by inferring shared ancestry and describing genetically distinct subgroups. The population structure inferred by DAPC was highly concordant with Bayesian clustering results, with both approaches supporting major genetic groups. The complete correspondence between population structure clusters and DAPC groups indicated a robust and consistent population structure. The partitioning of admixed individuals across multiple DAPC groups reflected additional structure unresolved by Bayesian assignment.

A robust and cost-effective SNP fingerprinting system for cotton

Application of a minimal marker strategy resulted in a set of 20 SNPs that uniquely discriminated all genotypes in the cotton panel. The minimal marker concept originally described by Fujii et al. (2013) was designed to maximize genotype resolution with a minimum marker number. Our results are consistent with this, showing that complete genotype differentiation can be achieved with a small, information-dense SNP set.

In cotton, fingerprinting was previously described based on SSR markers often requiring 20–30 loci to reliably distinguish cultivars (Wang et al. 2017; Kuang et al. 2022). However, SSR markers often require multi-allelic scoring and can be challenging to standardize in the laboratory. Compared to SSR-based systems, SNP markers offer superior reproducibility, easier assay conversion and scalability across platforms. In this study, the successful resolution of all genotypes using 20 SNPs indicated that high-throughput SNP discovery datasets can be converted into efficient markers for use with low-cost genotyping platforms, such as KASP assays (Hulse-Kemp et al. 2014; Li et al. 2022b). The markers can then be used for seed purity testing, accession verification and varietal protection, offering a practical molecular identity system for cotton breeding and germplasm management.

Conclusion

The long-term productivity and resilience of cotton breeding programs are fundamentally dependent on maintaining sufficient genetic diversity within the breeding pool. However, the continued narrowing of the genetic base in modern Upland cotton has constrained breeding progress. The increasing complexity of breeding targets is not matched by effective genomic strategies for accessing and incorporating beneficial variation from diverse and exotic *Gossypium* germplasm.

The moderate level of genetic diversity detected in the present study reflects both the legacy of intensive selection and the persistence of distinct genetic patterns within modern cotton germplasm. By resolving population structure, identifying divergent and closely related genotype pairs, and characterizing genome-wide LD patterns, our results provide a useful genomic framework for more effective utilization of existing diversity and for guiding the introduction of novel variation into cotton breeding programs. Furthermore, the development of a minimal SNP fingerprinting set capable of uniquely discriminating all genotypes offers an efficient tool for germplasm authentication, varietal protection and routine breeding program quality control. Overall, the genomic resources and insights generated here may provide a foundation for more informed material selection and more efficient management of cotton breeding populations.

Funding

Funding was provided by The Scientific and Technological Research Council of Türkiye, project no. 123O986.

Conflict of interest/competing interests

The authors declare that they have no conflict of interest or competing interests.

Availability of data and materials

Data is available on request.

Code availability

Not applicable.

Authors' contributions

AAB: molecular characterization, data analysis, interpretation of data, manuscript drafting and revision; EGY: data analysis, interpretation of data and manuscript revision; VS: resources and manuscript revision; AF: conception and design, interpretation of data, manuscript revision; SD: conception and design, manuscript revision; All: final approval of the version to be published.

Ethics approval and consent to participate

The authors declare that they have followed all necessary ethical guidelines. Human and animal subjects were not used in this work

Consent for publication

All authors have consented to publication of this work.

References

Ai X, Liang Y, Wang J, et al (2017) Genetic diversity and structure of elite cotton germplasm (*Gossypium hirsutum* L.) using genome-wide SNP data. Genetica 145:409–416. https://doi.org/10.1007/s10709-017-9976-8

Beasley ED (2024) Evaluation of cotton germplasm (*Gossypium* spp.) for resistance to endemic diseases in the southeast United States. Doctoral Dissertation, The University of Georgia

Billings GT, Jones MA, Rustgi S, et al (2021) Population structure and genetic diversity of the Pee Dee cotton breeding program. G3 GenesGenomesGenetics 11:jkab145. https://doi.org/10.1093/g3journal/jkab145

Brubaker CL, Wendel JF (1994) Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Am J Bot 81:1309–1326. https://doi.org/10.1002/j.1537-2197.1994.tb11453.x

Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7. https://doi.org/10.1186/s13742-015-0047-8

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen ZJ, Sreedasyam A, Ando A, et al (2020) Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet 52:525–533. https://doi.org/10.1038/s41588-020-0614-5

Cook OF (1912) Cotton Improvement Under Weevil Conditions. U.S. Department of Agriculture, Washington, D.C.

Coppens d'Eeckenbrugge G, Lacape J-M (2014) Distribution and Differentiation of Wild, Feral, and Cultivated Populations of Perennial Upland Cotton (*Gossypium hirsutum* L.) in

Mesoamerica and the Caribbean. PLoS ONE 9:e107458. https://doi.org/10.1371/journal.pone.0107458

Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De Beukelaer H, Davenport GF, Fack V (2018) Core Hunter 3: flexible core subset selection. BMC Bioinformatics 19:203. https://doi.org/10.1186/s12859-018-2198-2

DoVale JC, Carvalho HF, Sabadin F, Fritsche-Neto R (2022) Genotyping marker density and prediction models effects in long-term breeding schemes of cross-pollinated crops. Theor Appl Genet 135:4523–4539. https://doi.org/10.1007/s00122-022-04236-3

Duggar JF (1907) Descriptions and classification of varieties of American upland cotton. Agricultural Experiment Station of The Alabama Polytechnic Institute, Auburn, Alabama

Fang L, Gong H, Hu Y, et al (2017) Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. Genome Biol 18:33. https://doi.org/10.1186/s13059-017-1167-5

FAO (2026) FAOSTAT production: crops and livestock products

Fryxell PA (1986) Ecological adaptations of *Gossypium* species. In: Mauney JR, Stewart JMcD (eds) Cotton Physiology. The Cotton Foundation, Memphis, TN, pp 1–7

Fujii H, Ogata T, Shimada T, et al (2013) Minimal Marker: An algorithm and computer program for the identification of minimal sets of discriminating dna markers for efficient variety identification. J Bioinform Comput Biol 11:1250022. https://doi.org/10.1142/S0219720012500229

Gapare W, Conaty W, Zhu Q-H, et al (2017) Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. Euphytica 213:66. https://doi.org/10.1007/s10681-017-1855-y

Garrison E (2016) Vcflib, a simple C++ library for parsing and manipulating VCF files

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing

Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9:e90346. https://doi.org/10.1371/journal.pone.0090346

Gowda SA, Bourland FM, Kaur B, et al (2023) Genetic diversity and population structure analyses and genome-wide association studies of photoperiod sensitivity in cotton (*Gossypium hirsutum* L.). Theor Appl Genet 136:230. https://doi.org/10.1007/s00122-023-04477-w

Gruber B, Unmack PJ, Berry OF, Georges A (2018) dartr: an R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Mol Ecol Resour 18:691–699. https://doi.org/10.1111/1755-0998.12745

Guo X, Wang Y, Hou Y, et al (2022) Genome-Wide Dissection of the Genetic Basis for Drought Tolerance in *Gossypium hirsutum* L. Races. Front Plant Sci 13:876095. https://doi.org/10.3389/fpls.2022.876095

Han P, Tian X, Wang Y, et al (2022) Construction of a core germplasm bank of upland cotton (*Gossypium hirsutum* L.) based on phenotype, genotype and favorable alleles. Genet Resour Crop Evol 69:2399–2411. https://doi.org/10.1007/s10722-022-01379-6

Hinze LL, Hulse-Kemp AM, Wilson IW, et al (2017) Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. BMC Plant Biol 17:37. https://doi.org/10.1186/s12870-017-0981-y

Hinze LL, Udall JA, Yu JZ, Frelichowski JE (2025) The National Plant Germplasm System cotton collection—a review of germplasm resources, phenotypic characterization, and genomic variation. Theor Appl Genet 138:220. https://doi.org/10.1007/s00122-025-05009-4

Hong MJ, Ko CS, Kim DY (2024) Genome-Wide Association Study to Identify Marker–Trait Associations for Seed Color in Colored Wheat (*Triticum aestivum* L.). Int J Mol Sci 25:3600. https://doi.org/10.3390/ijms25073600

Hosoya S, Hirase S, Kikuchi K, et al (2019) Random PCR-based genotyping-by-sequencing technology GRAS-Di (Genotyping by Random Amplicon Sequencing, Direct) reveals genetic structure of mangrove fishes. Mol Ecol Resour 19:1153–1163. https://doi.org/10.1111/1755-0998.13025

Huang G, Wu Z, Percy RG, et al (2020) Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. Nat Genet 52:516–524. https://doi.org/10.1038/s41588-020-0607-4

Hulse-Kemp AM, Ashrafi H, Zheng X, et al (2014) Development and bin mapping of gene-associated interspecific SNPs for cotton (*Gossypium hirsutum* L.) introgression breeding efforts. BMC Genomics 15:945. https://doi.org/10.1186/1471-2164-15-945

Iqbal MJ, Reddy OUK, El-Zik KM, Pepper AE (2001) A genetic bottleneck in the 'evolution under domestication' of upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting: Theor Appl Genet 103:547–554. https://doi.org/10.1007/PL00002908

Jazayeri SM, Villamar-Torres RO, Zambrano-Vega C, et al (2020) Transcription factors and molecular markers revealed asymmetric contributions between allotetraploid Upland cotton and its two diploid ancestors. Bragantia 79:30–46. https://doi.org/10.1590/1678-4499.20190161

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129

606 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new
607 method for the analysis of genetically structured populations. BMC Genet 11:94.
608 https://doi.org/10.1186/1471-2156-11-94

609 Kaeuffer R, Réale D, Coltman DW, Pontier D (2007) Detecting population structure using
610 STRUCTURE software: effect of background linkage disequilibrium. Heredity 99:374–
611 380. https://doi.org/10.1038/sj.hdy.6801010

612 Kuang Z, Xiao C, Ilyas MK, et al (2022) Use of SSR Markers for the Exploration of Genetic
613 Diversity and DNA Finger-Printing in Early-Maturing Upland Cotton (*Gossypium*
614 *hirsutum* L.) for Future Breeding Program. Agronomy 12:1513.
615 https://doi.org/10.3390/agronomy12071513

616 Lacape J-M, Dessauw D, Rajab M, et al (2006) Microsatellite diversity in tetraploid *Gossypium*
617 germplasm: assembling a highly informative genotyping set of cotton SSRs. Mol Breed
618 19:45–58. https://doi.org/10.1007/s11032-006-9042-1

619 Lee JA (1984) Cotton as a World Crop. In: Kohel RJ, Lewis CF (eds) Agronomy Monographs, 1st
620 edn. Wiley, pp 1–25

621 Li C, Dong C, Zhao H, et al (2022a) Identification of superior parents with high fiber quality using
622 molecular markers and phenotypes based on a core collection of upland cotton (*Gossypium*
623 *hirsutum* L.). Mol Breed 42:30. https://doi.org/10.1007/s11032-022-01300-0

624 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

625 Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage
626 samples. Bioinformatics 30:2843–2851. https://doi.org/10.1093/bioinformatics/btu356

627 Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools.
628 Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

629 Li L, Sun Z, Zhang Y, et al (2022b) Development and Utilization of Functional Kompetitive Allele-
630 Specific PCR Markers for Key Genes Underpinning Fiber Length and Strength in
631 *Gossypium hirsutum* L. Front Plant Sci 13:853827.
632 https://doi.org/10.3389/fpls.2022.853827

633 Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting
634 population genetics and genomics programs. Bioinformatics 28:298–299.
635 https://doi.org/10.1093/bioinformatics/btr642

636 Lubbers EL, Chee PW, Gannaway JR, et al (2005) Genetic relationships of historically important
637 eastern US Upland cotton. In: Proceedings of the Beltwide Cotton Conferences. New
638 Orleans, Louisiana, pp 1027–1030

639 Lubbers EL, Chee PW, Paterson AH, Smith CW (2006) Fiber quality of a near-isogenic
640 introgression line series from an upland by pima interspecific cross. In: Proceedings of the
641 Beltwide Cotton Conferences. San Antonio, Texas, pp 845–852

Ning W, Rogers KM, Hsu C-Y, et al (2024) Origin and diversity of the wild cottons (*Gossypium hirsutum*) of Mound Key, Florida. Sci Rep 14:14046. https://doi.org/10.1038/s41598-024-64887-8

O'Leary SJ, Puritz JB, Willis SC, et al (2018) These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. Mol Ecol 27:3193–3206. https://doi.org/10.1111/mec.14792

Oliveira Da Silva AC, Morais De Oliveira AF, Alves Cursino Dos Santos DY, Izídio Da Silva S (2010) An approach to chemotaxonomy to the fatty acid content of some Malvaceae species. Biochem Syst Ecol 38:1035–1038. https://doi.org/10.1016/j.bse.2010.10.006

Perrier X, Jacquemoud-Collet JP (2006) DARwin software

Purcell S, Chang C (2019) PLINK 1.9

Puritz JB, Hollenbeck CM, Gold JR (2014) *dDocent* : a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431. https://doi.org/10.7717/peerj.431

Rabieyan E, Darvishzadeh R, Mohammadi R, et al (2023) Genetic diversity, linkage disequilibrium, and population structure of tetraploid wheat landraces originating from Europe and Asia. BMC Genomics 24:682. https://doi.org/10.1186/s12864-023-09768-6

Safiullina A, Ernazarova D, Turaev O, et al (2024) Genetic Diversity and Subspecific Races of Upland Cotton (*Gossypium hirsutum* L.). Genes 15:1533. https://doi.org/10.3390/genes15121533

Song C, Li W, Wang Z, et al (2018) Genome resequencing reveals genetic variation between the parents of an elite hybrid upland cotton. Agronomy 8:305. https://doi.org/10.3390/agronomy8120305

Tanksley SD, McCouch SR (1997) Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. Science 277:1063–1066. https://doi.org/10.1126/science.277.5329.1063

TinHan TC, O'Leary SJ, Portnoy DS, et al (2020) Natural tags identify nursery origin of a coastal elasmobranch *Carcharhinus leucas*. J Appl Ecol 57:1222–1232. https://doi.org/10.1111/1365-2664.13627

Tyagi P, Gore MA, Bowman DT, et al (2014) Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet 127:283–295. https://doi.org/10.1007/s00122-013-2217-3

Tyler FJ (1910) Varieties of American upland cotton. Govt. print. off, Washington

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

Vega M, Quintero-Corrales C, Mastretta-Yanes A, et al (2023) Multiple domestication events explain the origin of *Gossypium hirsutum* landraces in Mexico. Ecol Evol 13:e9838. https://doi.org/10.1002/ece3.9838

Wang G-L, Dong J-M, Paterson AH (1995) The distribution of *Gossypium hirsutum* chromatin in *G. barbadense* germ plasm: molecular analysis of introgressive plant breeding. Theor Appl Genet 91:1153–1161. https://doi.org/10.1007/BF00223934

Wang J, Zhang Z, Gong Z, et al (2022a) Analysis of the genetic structure and diversity of upland cotton groups in different planting areas based on SNP markers. Gene 809:146042. https://doi.org/10.1016/j.gene.2021.146042

Wang P, Dong N, Wang M, et al (2022b) Introgression from *Gossypium hirsutum* is a driver for population divergence and genetic diversity in *Gossypium barbadense*. Plant J 110:764–780. https://doi.org/10.1111/tpj.15702

Wang X-Y, Ai X-T, Wang J-D, et al (2017) Rapid Identification System of Purity and Authenticity in Cotton Varieties by SSR Markers. Acta Agron Sin 43:1565. https://doi.org/10.3724/SP.J.1006.2017.01565

Ware JO (2023) Origin, Rise and Development of American Upland Cotton and Their Status at Present. AAES Research Reports and Research Bulletins

Wendel JF, Brubaker C, Alvarez I, et al (2009) Evolution and Natural History of the Cotton Genus. In: Paterson AH (ed) Genetics and Genomics of Cotton. Springer US, New York, NY, pp 3–22

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. In: Advances in Agronomy. Elsevier, pp 139–186

Wendel JF, Grover CE (2015) Taxonomy and Evolution of the Cotton Genus, *Gossypium*. In: Fang DD, Percy RG (eds) Cotton, 2nd edn. Wiley, pp 25–44

Westengen OT, Huamán Z, Heun M (2005) Genetic diversity and geographic pattern in early South American cotton domestication. Theor Appl Genet 110:392–402. https://doi.org/10.1007/s00122-004-1850-2

Wickham H (2016) ggplot2: elegant graphics for data analysis, Second edition. Springer-Verlag New York

Xu H, Mei Y, Hu J, et al (2006) Sampling a Core Collection of Island Cotton (*Gossypium barbadense* L.) Based on the Genotypic Values of Fiber Traits. Genet Resour Crop Evol 53:515–521. https://doi.org/10.1007/s10722-004-2032-4

Xu P, Guo Q, Meng S, et al (2021) Genome-wide association analysis reveals genetic variations and candidate genes associated with salt tolerance related traits in *Gossypium hirsutum*. BMC Genomics 22:26. https://doi.org/10.1186/s12864-020-07321-3

Yanar EG, Frary A, Doğanlar S (2023) GRAS-Di SNP-based molecular characterization and fingerprinting of a Turkish *Corylus avellana* core set provide insights into the cultivation and breeding of hazelnut in Turkey. Mol Genet Genomics 298:413–426. https://doi.org/10.1007/s00438-022-01990-3

Yuan Y, Wang X, Wang L, et al (2018) Genome-Wide Association Study Identifies Candidate Genes Related to Seed Oil Composition and Protein Content in *Gossypium hirsutum* L. Front Plant Sci 9:1359. https://doi.org/10.3389/fpls.2018.01359

Zhao N, Wang W, Grover CE, et al (2022) Genomic and GWAS analyses demonstrate phylogenomic relationships of *Gossypium barbadense* in China and selection for fibre length, lint percentage and *Fusarium wilt* resistance. Plant Biotechnol J 20:691–710. https://doi.org/10.1111/pbi.13747

Zheng X, Levine D, Shen J, et al (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Tables

Table 1 Chromosome wide distribution, density, and genome coverage of SNP markers across the *Gossypium hirsutum* reference genome (GCF_007990345.1).

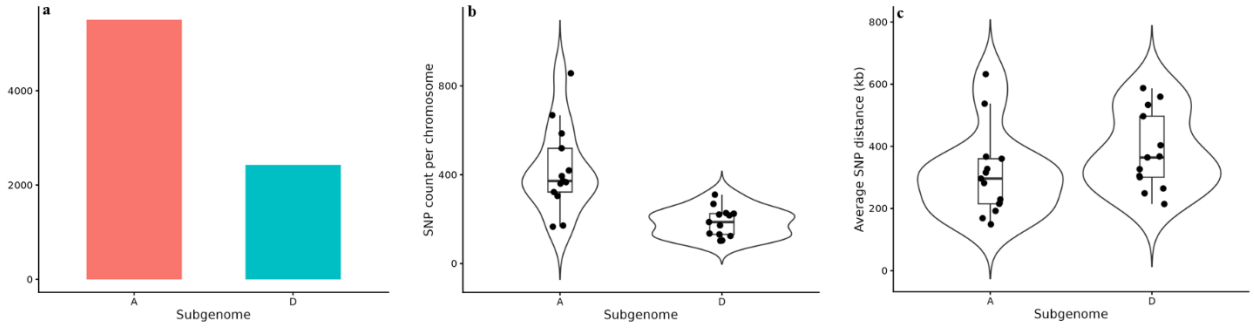
Chromosome	Length (Mb)	Number of SNPs	Average SNP distance (Kb)	Coverage (%)
A01	119.76	366	327.22	99.16
A02	108.14	171	632.40	98.45
A03	113.69	360	315.81	98.97
A04	89.18	166	537.23	97.75
A05	111.10	394	281.98	99.47
A06	128.20	668	191.91	99.09
A07	98.90	586	168.78	96.06
A08	127.50	857	148.77	99.94
A09	85.34	372	229.40	98.66
A10	118.18	322	367.03	99.49
A11	124.18	419	296.38	98.81
A12	109.47	304	360.11	97.54
A13	111.65	519	215.12	99.33
Total		**5,504**	**312.08**	
D01	65.21	217	300.48	92.29
D02	72.19	221	326.64	99.38
D03	54.96	103	533.56	98.14
D04	58.23	104	559.90	96.38
D05	66.48	310	214.47	98.22
D06	66.68	268	248.82	96.06
D07	59.44	225	264.18	95.67
D08	69.43	228	304.51	97.80
D09	54.45	135	403.30	95.52
D10	68.09	187	364.11	97.01
D11	72.82	124	587.29	97.55
D12	63.26	172	367.76	99.22
D13	65.10	131	496.95	99.68
Total		**2,425**	**382.45**	

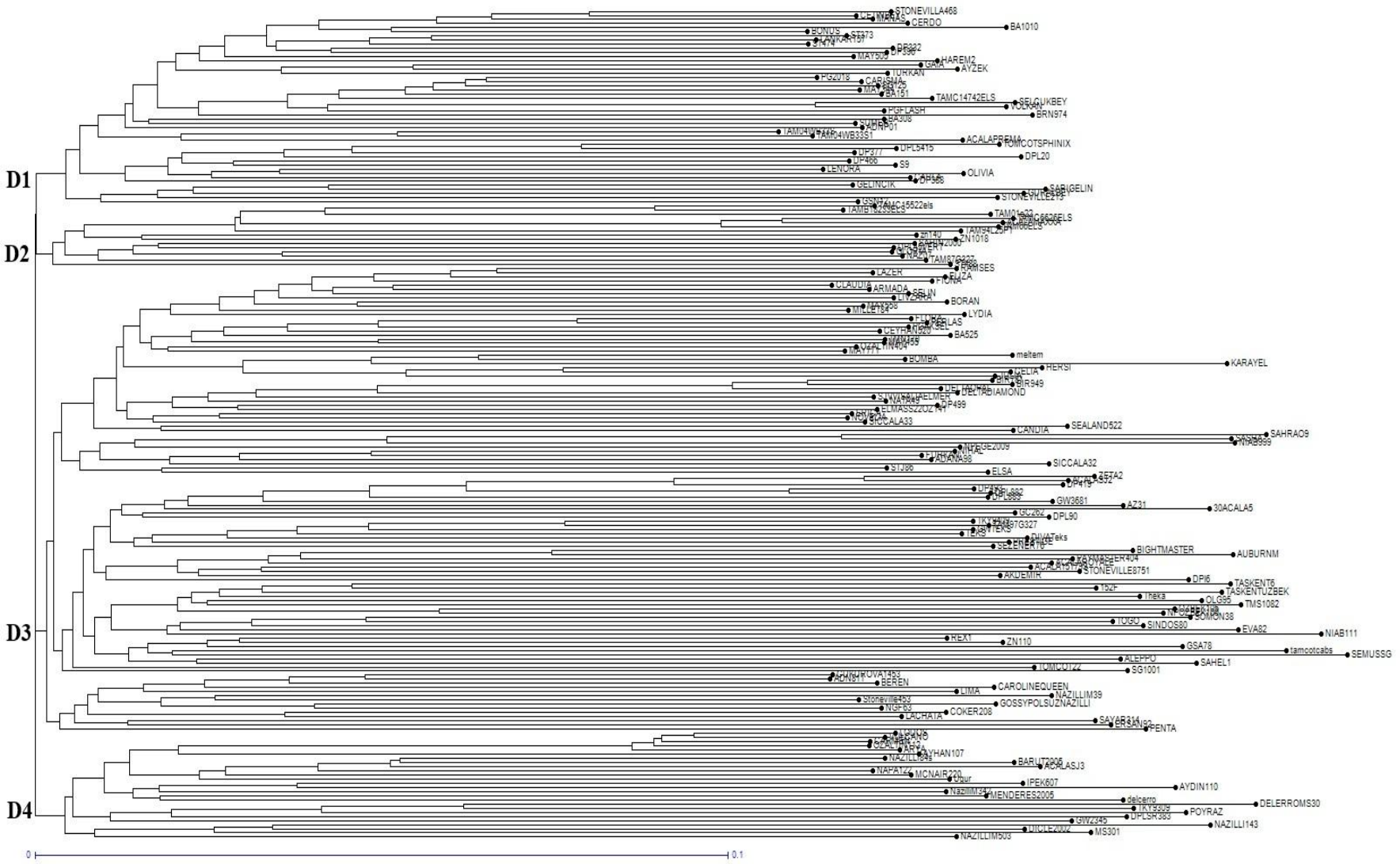
734 **Table 2** Distribution of SNPs across genomic regions in Upland cotton germplasm panel.

Genomic region	Number of SNPs	Number of genes
Exon	574 (7.2%)	382
Intron	203 (2.5%)	174
Splice region	9 (0.1%)	7
Intergenic region	7231 (90.2%)	NA
Total	**8017 (100%)**	**535**

735

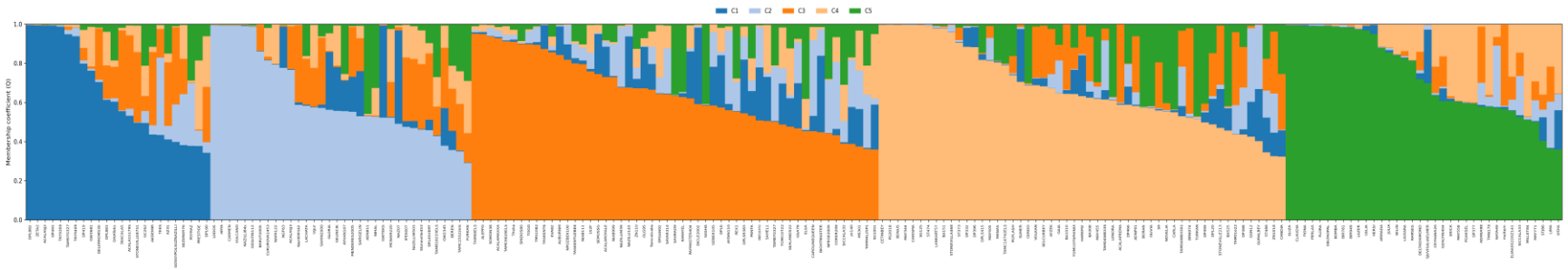
736


a
4000
2000
0
A
D
Subgenome
b
SNP count per chromosome
800
400
0
A
D
Subgenome
c
Average SNP distance (kb)
800
600
400
200
0
A
D
Subgenome



746

747



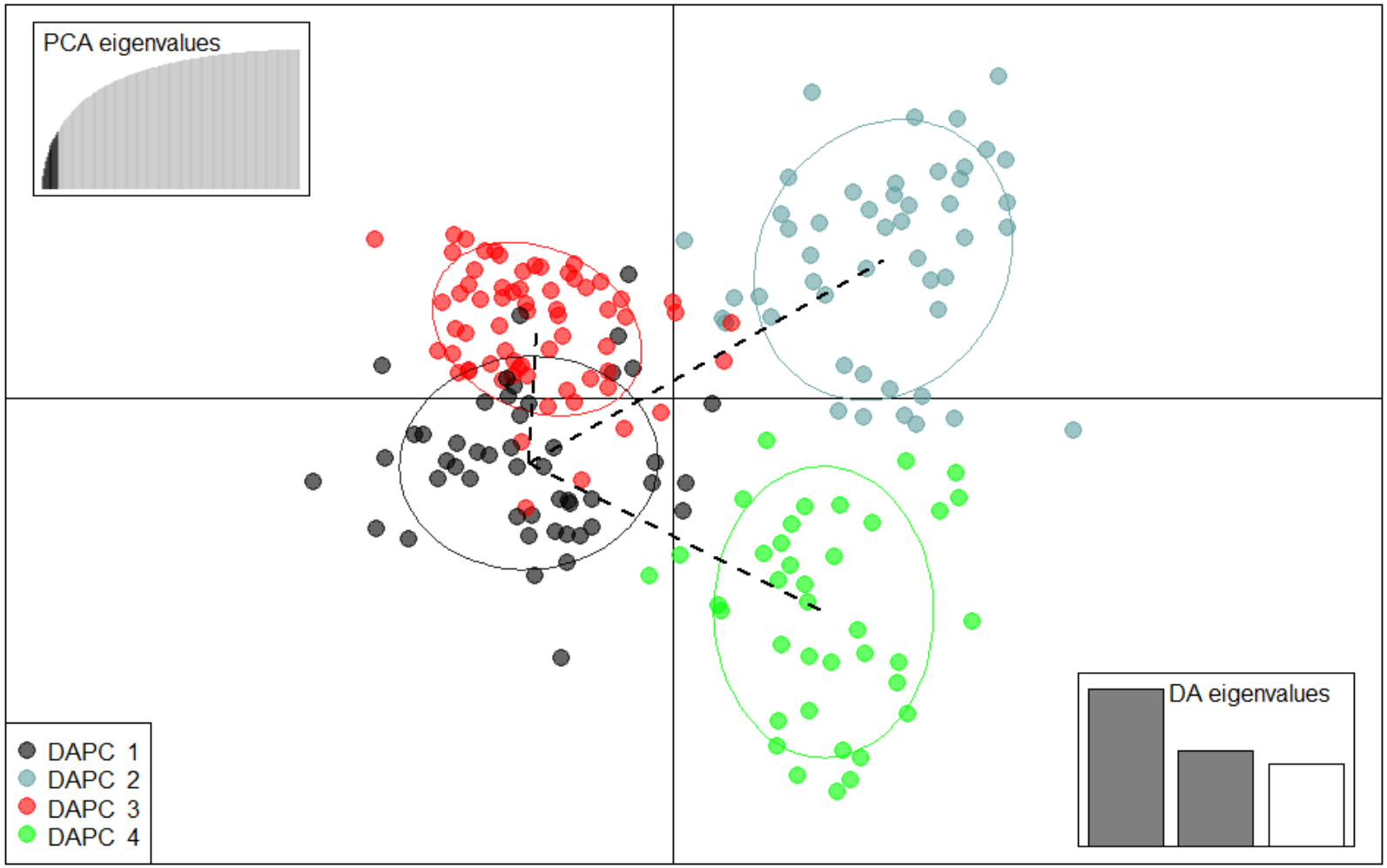
748

749

750

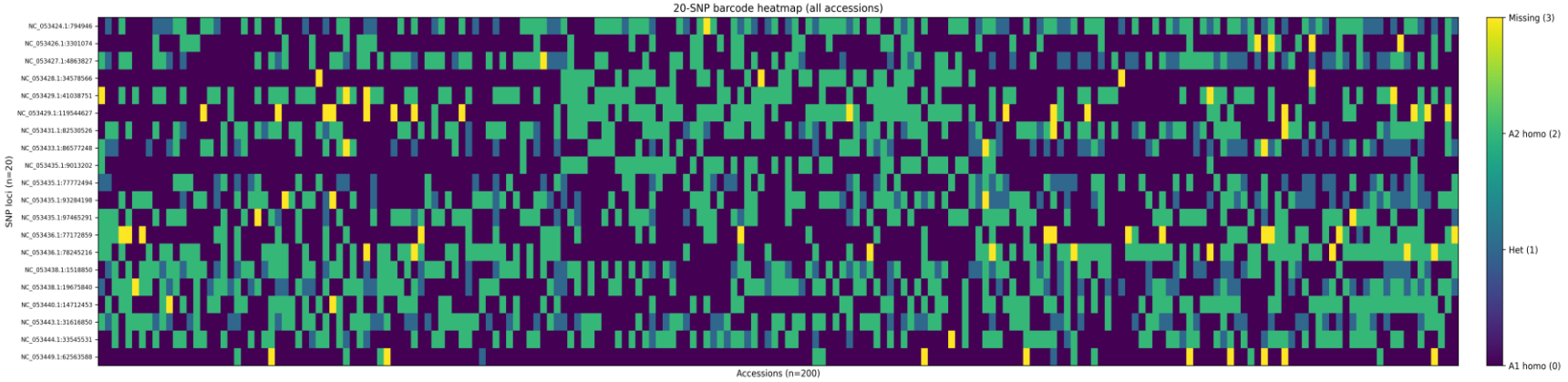
751

752



753

754


20-SNP barcode heatmap (all accessions)
SNP loci (n=20)
NC_053424.1:794946
NC_053426.1:3301074
NC_053427.1:4863827
NC_053428.1:34578566
NC_053429.1:41038751
NC_053429.1:119544627
NC_053431.1:82530526
NC_053433.1:86577248
NC_053435.1:9013202
NC_053435.1:77772494
NC_053435.1:93284198
NC_053435.1:97465291
NC_053436.1:77172859
NC_053436.1:78245216
NC_053438.1:1518850
NC_053438.1:19675840
NC_053440.1:14712453
NC_053443.1:31616850
NC_053444.1:33545531
NC_053449.1:62563588
Accessions (n=200)
Missing (3)
A2 homo (2)
Het (1)
A1 homo (0)

755

756

757

Figure legends

Fig. 1 (a) SNP distribution in cotton A and D subgenomes. The x-axis represents the subgenome whereas the y-axis represents the total number of SNPs aggregated across all chromosomes of that subgenome. (b) Distribution of SNPs on individual chromosomes of the A and D subgenomes. The x-axis indicates the subgenome grouping of chromosomes (A01–A13 and D01–D13) while the y-axis corresponds to the number of SNP markers identified on each chromosome. (c) Density of SNPs on the chromosomes in each subgenome. The x-axis represents the subgenome category of chromosomes belonging to the subgenomes while the y-axis shows the average SNP distance (kb) calculated for each chromosome.

Fig. 2 Unweighted neighbor-joining dendrogram showing genetic relationships among 200 upland cotton genotypes based on genome-wide SNP data. The dendrogram shows four main groups comprising 47, 15, 112, and 26 genotypes assigned to D1, D2, D3, and D4, respectively.

Fig. 3 Population structure of the Upland cotton panel (K = 5). Each vertical bar represents an individual genotype. The colored segments indicate the estimated membership coefficient of that genotype in each genetic cluster. The five inferred clusters are represented by colors: C1 (dark blue), C2 (light blue), C3 (orange), C4 (peach), and C5 (green). Individuals with multiple colors indicate admixed ancestry showing shared genetic background among clusters. The x-axis shows genotypes and the y-axis represents the proportion of ancestry (Q value) assigned to each cluster.

Fig. 4 Discriminant analysis of principal components (DAPC) clustering. Scatter plot of the DAPC showing genetic relationships among individuals based on the first two discriminant functions. Each point represents an individual color-coded according to its assigned genetic cluster (K = 4). Ellipses indicate the dispersion of individuals within each cluster. The plot reveals clear separation among the four genetic groups with limited overlap.

Fig. 5 Barcode heatmap of the 20-SNP fingerprint panel across 200 cotton genotypes. Rows represent SNP loci and columns represent individual genotypes. Genotype classes are color-coded as follows: allele 1 homozygote (navy blue), heterozygote (medium blue), allele 2 homozygote (green) and missing data (N) (yellow).

Supplementary file information

Supplementary Tables S1-S12.xlsx

Description: Supplementary tables including cotton panel information (Supplementary Table S1), primer sequences used for library preparation (Supplementary Table S2), SNP substitution types (Supplementary Table S3), functional annotation of SNPs (Supplementary Table S4), distributions of SNP annotations and impacts (Supplementary Tables S5–S7), pairwise genetic distance information for the most divergent genotype pairs (Supplementary Table S8), STRUCTURE membership coefficients (Supplementary Table S9), DAPC membership matrix (Supplementary Table S10), core collection composition (Supplementary Table S11), and the minimal SNP fingerprinting marker set (Supplementary Table S12).

Supplementary Figures S1-S12

Figure legends and descriptions

Supplementary Fig. S1 Genome-wide SNP density distribution across 26 chromosomes. SNP density was calculated in 500 kb sliding windows along each chromosome based on the *Gossypium hirsutum* reference genome (v2.1). Chromosomes are labeled according to the A and D subgenomes (A01–A13 and D01–D13). The color scale represents the number of SNPs per window, ranging from low (orange) to high (black).

Supplementary Fig. S2 Linkage disequilibrium (LD) decay across the fig genome. Genome-wide linkage disequilibrium decay was assessed by calculating pairwise LD (r^2) values between SNPs as a function of physical distance. Each point represents an individual SNP pair while the solid line indicates the fitted nonlinear decay curve. LD declined rapidly with increasing distance with r^2 values decreasing sharply within the first few kilobases and approaching a plateau at approximately $r^2 \approx 0.15$ at larger distances. The estimated half-decay distance (~2.0–2.5 Mb, $r^2 \approx 0.35$) was used to guide SNP thinning for downstream analyses.

Supplementary Fig. S3 Distribution of genetic dissimilarity values among Upland cotton panel. Dissimilarity values ranged from 0.06 to 0.37 with a mean of 0.27 indicating moderate to high levels of genetic variation across the panel. The red line indicates a genetic distance threshold of 0.1.

Supplementary Fig. S4 Mean within-country IBS genetic distance across cotton accessions grouped by country of origin. Bars represent the mean pairwise IBS genetic distance calculated from SNP data within each origin group. Error bars indicate the minimum–maximum range of pairwise distances observed within each country. Sample sizes are shown in parentheses below each country name. The highest within-country genetic divergence was observed in the Greek subset (mean = 0.301; n = 4), followed by Uzbekistan (0.264; n = 3) and USA (0.262; n = 60), whereas Türkiye (0.246; n = 107), Spain (0.244; n = 5), and Australia (0.240; n = 13) exhibited comparatively lower internal genetic distances.

Supplementary Fig. S5 Determination of the optimal number of genetic clusters using the ΔK method. The plot shows ΔK values calculated according to the Evanno method across tested K values.

Supplementary Fig. S6 Bayesian Information Criterion (BIC) values for DAPC clustering. BIC values plotted against the number of inferred genetic clusters (K) used in DAPC. The lowest BIC value was observed at K = 4, indicating this clustering level as the most parsimonious model for describing the genetic structure of the panel.

Supplementary Fig. S7 Discriminant analysis eigenvalues. Bar plot of discriminant analysis eigenvalues illustrating the relative discriminatory power of each linear discriminant axis. The first discriminant axis exhibits the highest eigenvalue indicating the strongest contribution to among-group separation while subsequent axes show progressively reduced discriminatory power.

Supplementary Fig. S8 Neighbor-joining dendrogram illustrating the genetic relationships among 200 Upland cotton genotypes based on genome-wide SNP data and IBS distances. Branch colors indicate STRUCTURE clusters as shown circular symbols: C1 (magenta), C2 (green), C3 (blue), C4 (red), C5 (yellow), and admixed individuals (black).

Supplementary Fig. S9 Geographic origin composition of STRUCTURE clusters. Stacked bar plot showing the proportional contribution (%) of each geographic origin within STRUCTURE clusters (Admixed and C1–C5). Each bar represents a STRUCTURE group, and colored segments indicate the relative frequency of genotypes originating from different countries. Percentages were calculated based on the total number of genotypes assigned to each cluster.

Supplementary Fig. S10 Genetic distance distribution in the full cotton panel and the core collection. (a) Frequency distribution of pairwise genetic distances calculated among genotypes in the full panel (n = 200) and the core collection (n = 50). (b) Mean genetic distance for the full panel and the core collection.

Supplementary Fig. S11 Geographic origin composition of the full panel and the core collection. Donut charts show the proportional distribution of genotypes according to their country of origin in the full panel (n = 200) and the core collection (n = 50).

Supplementary Fig. S12 Neighbor-Joining dendrogram based on the minimal SNP fingerprinting panel.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GAbstractedited.jpg
- SupplementaryFigureS1.png
- SupplementaryFigureS10.png
- SupplementaryFigureS11.png
- SupplementaryFigureS12.jpg
- SupplementaryFigureS2.png
- SupplementaryFigureS3.png
- SupplementaryFigureS4.png
- SupplementaryFigureS5.png
- SupplementaryFigureS6.png
- SupplementaryFigureS7.png
- SupplementaryFigureS8.jpg
- SupplementaryFigureS9.png
- SupplementaryfilesTableS1S12.xlsx