BMC Genomics

RESEARCH

Open Access

Genetic diversity and population structure analysis of cotton (*Gossypium hirsutum* L.) genotypes using DArTseq technology

Kamil Haliloğlu[1], Batuhan Akgöl[2], Tuğçe Hançer[2], Hadi Alipour[3*] and Aras Türkoğlu[4*]

Abstract

Cotton is a globally cultivated crop that is exposed to various biotic and abiotic stress conditions. The most effective strategy to mitigate these stresses is the development of tolerant or resistant varieties through plant breeding programs. Appropriate breeding strategies such as conventional breeding are crucial for generating beneficial genetic variations and identifying desirable traits. Integrating conventional breeding with molecular breeding is a key approach to addressing the challenges of sustainable cotton production. We aimed to assess the genetic diversity and the population structure of cotton genotypes, therefore evaluating genotypes value in terms of use for breeding purposes. A total of 913 genotypes, including advanced breeding lines from the *Gossypium hirsutum* L., were genotyped using the diversity array technology sequencing (DArTseq array) with a high throughput of the single-nucleotide polymorphisms (SNPs). Out of 5,986 SNPs, 1,431 high-quality SNPs were selected for genome diversity analysis. The genotypes investigated were classified into subgroups based on kinship relationships, UPMGA clustering, discriminant analysis of principal components, and principal coordinate analysis. Population structure analysis using the model-based ΔK method suggested an optimal K = 2, representing the highest level of population differentiation. However, the multivariate DAPC method revealed a finer substructure of K = 3, providing a more detailed view of the genetic relationships within the breeding lines. The highest fixation index (F_{ST}) and lowest gene flow (Nm) were observed between subpopulation I and subpopulation II. The mean values of minor allele frequency (MAF), genetic diversity (GD), polymorphism information content (PIC), observation heterozygosity (Ho), and the highest F_{ST} across the whole genome were estimated at 0.2328, 0.3208, 0.2607, 0.0210, and 0.0830, respectively. The findings of this study will provide valuable insights into breeders in selecting parents for cultivar development breeding programs through crossbreeding.

Keywords Cotton, DArTseq, Genetic diversity, Polymorphism, Population structure

*Correspondence:
Hadi Alipour
ha.alipour@urmia.ac.ir
Aras Türkoğlu
aras.turkoglu@erbakan.edu.tr
[1]Department of Biology, Faculty of Science, Gazi University, Ankara 06170, Türkiye
[2]ProGen Tohum A.Ş., Büyükdalyan Mh. 2. Küme Evler Sk. No:49, Antakya/HATAY 31060, Türkiye
[3]Department of Plant Production and Genetics, Urmia University, Urmia, Iran
[4]Department of Field Crops, Faculty of Agriculture, Necmettin Erbakan University, 42310 Konya, Türkiye

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

For the 2023–2024 growing season, global cotton production is estimated at approximately 113.0 million bales, cultivated on a harvested area of roughly 31.2 million hectares. Within this global landscape, Türkiye plays a significant role as both a major producer and a key consumer, driven by its robust textile industry. In the same period, Türkiye's cotton production was estimated at 3.2 to 3.5 million bales (approximately 755,000 metric tons) from a harvested area of between 335,000 and 417,000 hectares. This positions Türkiye as one of the top ten global producers, contributing approximately 3% of the world's total cotton supply. Significantly, Türkiye's domestic cotton consumption, projected at 1.65 million metric tons (7.6 million bales) for MY 2023/24, significantly outstrips its production. This production deficit makes Türkiye one of the world's largest cotton importers, highlighting the strategic economic importance of enhancing the yield, quality, and local adaptation of domestic cultivars. The advanced breeding lines analyzed in this study represent a direct effort to address this challenge, making an understanding of their genetic architecture essential for national agricultural strategy [1].

Cotton (*Gossypium* spp.) is a vital crop widely cultivated for its natural fiber and oil, contributing to 35% of global fiber consumption. The genus consists of approximately 50 species, of which 45 are diploid (2n = 2x = 26) and five are allotetraploid (2n = 4x = 52) [2]. Among the diploid species, *G. arboreum* L. (A_2A_2) and *G. herbaceum* L. (A_1A_1) belong to the AA genome group (2n = 2x = 26), while the tetraploid species *G. hirsutum* L. and *G. barbadense* L. fall under the AADD genome group (2n = 4x = 52) [3–5]. The tetraploid cotton genome is relatively large, encompassing approximately 2,200–3,000 Mb of DNA [6]. Cultivated tetraploid cotton evolved approximately 1–2 million years ago through the hybridization of an A genome donor species (*G. herbaceum* and *G. arboreum*) with a D genome species (*G. raimondii* and *G. gossipioides*), followed by polyploidization [7]. The progenitor allotetraploid "AD" diverged, giving rise to the "AD" tetraploid species, *G. hirsutum* L. and *G. barbadense* L. [8]. *G. hirsutum* contributes 90% of global cotton production, while *G. barbadense* accounts for 8% [9]. *G. herbaceum* and *G. arboreum* together contribute 2% of the world's cotton production [10].

Most of the global cotton production, estimated at over 95%, is derived from Upland cotton (*G. hirsutum*) due to its high yield and broad adaptability. Extra-long staple cotton (*G. barbadense*), prized for its superior fiber quality, accounts for approximately 3–5% of world production. The two cultivated diploid species, *G. arboreum* and *G. herbaceum*, collectively contribute the remaining ~ 2%, and are grown primarily in specific regions of Asia and Africa. As a result, many global breeding efforts are concentrated on improving Upland cotton varieties [11]. Globally, around 20 diseases, pests, and environmental stress factors negatively impact cotton cultivation by reducing yield and quality. Notably, biotic factors alone account for over 41% of yield losses, with insects contributing 15%, weeds 13%, and various harmful pathogens making up the remainder [4]. To mitigate these stress factors, Cook [12] suggested four key strategies: (i) breeding new tolerant or resistant varieties, (ii) improving root health, (iii) enhancing water quality, and (iv) safeguarding plants from airborne threats. Developing stress-tolerant or resistant genotypes is considered a more environmentally friendly approach than chemical applications, which contribute to atmospheric pollution.

Cotton production worldwide is often limited by biotic and abiotic stresses, leading to reduced yields. This challenge is primarily attributed to the narrow genetic base and the absence of desirable traits in cultivated genotypes. To overcome these limitations, incorporating diverse genotypes from various sources into breeding programs is crucial. This approach enhances the frequency of beneficial alleles and promotes transgressive segregation, ultimately improving stress tolerance and maximizing seed yield and quality. The selection of parental lines is crucial for determining the level of variation, heterosis, hybrid strength, and molecular breeding efficiency, ultimately aiding in the development of cotton genotypes that are tolerant or resistant to diseases and pests, as well as early-maturing and high-yielding. Achieving these breeding goals requires the use of molecular markers to assess genetic diversity at the DNA level [13]. Genetic diversity arises from allelic variations in the genome, such as insertions and deletions in DNA, and serves as the foundation for marker-assisted breeding. Molecular markers have been widely applied in studying the genetic and structural heterogeneity of collected or natural germplasm [14]. They play a crucial role in evaluating variation-related indices, which helps streamline the screening process in breeding programs [15].

The potential of plant breeding to enhance crop yield and adaptation to environmental stresses relies on genetic variability within the germplasm [16]. While some studies have explored genetic variation and diversity in cotton germplasm, the genetic diversity of cotton continues to expand with the development of new breeding lines and ongoing germplasm collections. Therefore, it is essential to continually assess genetic diversity to support cotton breeding programs and population development. Characterizing genetic diversity in crop species is essential for the effective conservation and utilization of germplasm, as well as for developing successful breeding programs. So far, various molecular markers have been used to study genetic diversity in cotton such as restriction fragment length polymorphism (RFLP) [17],

randomly amplified polymorphic DNA (RAPD) [18], sequence characterized amplified region (SCAR) [19] cleaved amplified polymorphic sequence (CAPS) [20], amplified fragment length polymorphism (AFLP) [21, 22], sequence related amplified polymorphism (SRAP) [23, 24], sequence tagged site (STS) [25], simple sequence repeat or microsatellites (SSR) [26, 27], expressed sequence tags (EST-SSRs) [28, 29], inter-simple sequence repeat (ISSR) [30–32] and single nucleotide polymorphism (SNP) [33–36].

SNPs are high-throughput molecular markers that cover the entire plant genome, making them among the most widely used in plant breeding [37]. Their abundance allows for high-density mapping, and despite having only two alleles per locus, the accuracy of genetic diversity assessments depends more on the number of loci analyzed rather than allele variation [33]. The main advantage of SNP markers lies in their ability to enable simultaneous genotyping of numerous loci in a high-throughput and cost-efficient manner due to their bi-allelic nature [38]. SNP markers are primarily developed using next-generation sequencing (NGS) technologies, with genotyping-by-sequencing (GBS) playing a key role in accelerating their application in genomics-assisted breeding. The rapid advancement of GBS has facilitated quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) across various crops [39]. In cotton, research has primarily focused on SNP discovery and validation, leading to the development of high-throughput genotyping platforms such as the CottonSNP63K array, which enables efficient, large-scale SNP analysis with high reproducibility [40]. More recently, Diversity Arrays Technology (DArT) introduced a GBS-based approach known as "DArTseq," which enables high-density SNP genotyping across multiple crop species. DArTseq-derived SNP markers have been widely applied in linkage mapping, QTL identification, GWAS, and genetic diversity analyses in different crops [41].

The global demand for cotton products is projected to increase by 102% by 2030, despite significant challenges such as declining arable land, diminishing water resources, and the uncertain impacts of climate change on production. To meet this growing demand, yield improvements must accelerate while minimizing input costs and ensuring scalability across diverse agricultural systems. Currently, genetic gains in yield under standard plant densities are primarily driven by conventional breeding, selection, and biotechnology [17]. However, sustaining and enhancing genetic progress necessitates the integration of advanced genomic tools with traditional breeding approaches. Studies on the population structure and genetic diversity of cotton indicate that breeding programs have exerted distinct effects on their genome [42]. However, identifying SNPs in allotetraploid species such as cultivated cotton remains challenging due to the complexity of its genome [43]. Genotyping-by-sequencing (GBS) has emerged as an efficient approach for whole-genome SNP detection, genotyping, and genetic variation studies in various plant species [44, 45]. Although the availability of DArT and SNP genotyping technologies facilitates marker-assisted selection and accelerates genetic improvement efforts in cotton [46], the population structure and genome-wide diversity of advanced tetraploid cotton lines have yet to be comprehensively examined using high-throughput SNP genotyping. The application of DArTseq SNP markers in cotton remains limited. Addressing this gap through high-density SNP arrays will enhance genetic analyses and provide valuable insights for conservation and hybridization strategies in breeding programs. In this study, we aim to utilize SNP arrays to investigate genetic diversity, genetic differentiation, and population structure of advanced *Gossypium hirsutum* breeding lines, contributing to the development of more efficient breeding strategies.

Materials and methods

Plant materials

The plant material consisted of a collection of 913 advanced breeding lines of *Gossypium hirsutum* L. from the commercial breeding program of ProGen Seed Company, Türkiye. These lines are primarily in advanced generations of development (F_6 to F_8), derived from a series of targeted crosses designed to pyramid desirable traits for yield, fiber quality, and adaptation to the Eastern Mediterranean agricultural environment. The breeding program has historically incorporated diverse parental material, including elite Turkish cultivars, high-performing U.S. varieties, and exotic germplasm, which is expected to contribute to the substantial genetic diversity observed in this panel. All lines have undergone multiple cycles of phenotypic selection for agronomic performance and uniformity under local conditions. Ten plants per genotype were grown in a controlled greenhouse, and fresh pooled leaves from each genotype were used for DNA extraction.

DNA extraction and genotype-by-sequencing (GBS), and SNP calling

Genotyping-by-sequencing (GBS) analysis of 913 advanced breeding lines was conducted through a whole-genome profiling service for SNP and DArTseq markers. Detailed information on GBS analysis is available on the official website of Diversity Arrays Technology Pty Ltd, Australia (https://www.diversityarrays.com/). The DNA samples were processed at Diversity Array Technology Pty, Ltd., Australia (available at https://www. diversityarrays.com/orderinstructions/

plantDNAextractionprotocolfordart/) for DArTseq analysis using the genotyping-by-sequencing platform, as previously described by Kilian et al. [47]. DNA concentration was determined using the NanoDrop® ND-1000 UV/Vis spectrophotometer (Thermo Scientific, Wilmington, USA) and a total of 50 µl of DNA was submitted to Diversity Array Technology for SNP and DArTseq marker analysis, following the protocol outlined by Akbari et al. [48] for library preparation. The library was purified and quantified for cluster generation using an automated clonal amplification system (cBOT Illumina). Subsequently, next-generation sequencing was carried out on the Illumina NovaSeq 6000/X platform. To ensure a high-quality dataset for downstream analysis, the raw SNP calls were subjected to stringent filtering. Markers were retained only if they met the following criteria: a minor allele frequency (MAF) greater than 5%, a call rate greater than 90% (i.e., less than 10% of missing data), and an observed heterozygosity of less than 10%. These thresholds were chosen to remove rare alleles with low informative content, markers with excessive missing data that could bias analyses, and potentially erroneous SNP calls arising from paralogous sequence alignment in the complex allotetraploid cotton genome, respectively. This filtering process ensures that the final set of 1,431 SNPs is robust and reliable for population genetic analysis.

Data analysis

Genetic diversity analysis and analysis of molecular variance (AMOVA)

Basic genetic statistics including observed heterozygosity (H_o) and genetic diversity (H_s), overall gene diversity (H_t), gene diversity among samples (D_{st}), corrected overall gene diversity (H_{tp}), corrected gene diversity among samples (D_{stp}), genetic differentiation (fixation index, F_{st}), corrected fixation index (F_{stp}), inbreeding coefficient (F_{is}), and haplotypic differentiation (D_{est}) were calculated using the adegenet [49] and hierfstat [50] R-packages. The polymorphic information content (PIC), minor allele frequency (MAF), heterozygosity percentage, and gene diversity of all 1,431 SNP markers were analyzed using PowerMarker software v3.25 [51]. The PIC values were computed using the following formula [52]. Genetic variation was analyzed using DARwin version 6.010 software [53] based on the Jaccard index. Analysis of molecular variance (AMOVA) was conducted using the pegas package in R [54] to partition genetic variation into intra- and inter-subgroups. The level of gene flow among subgroups was estimated as the number of migrants per generation (Nm), calculated from pairwise Fst values using the formula $Nm = (1 - Fst)/(4 \times Fst)$. A higher Nm indicates stronger gene flow and reduces genetic differentiation among populations, whereas an Nm value lower than 1 suggests restricted gene flow and greater population differentiation.

Analysis of population structure

STRUCTURE version 2.3.4 was utilized for Bayesian cluster analysis to assess population structure [55], with all parameters set to default. The analysis was performed 10 times for each K value (K = 1 to 10) using 30,000 steps for both the Markov Chain and burn-in period, applying an admixture model [56]. The optimal number of clusters was determined using the ad hoc statistic ΔK in StructureSelector [57], which evaluates the rate of change in the log probability of data across successive K values [58]. Samples with a membership probability of ≥ 0.50 were assigned to their respective clusters [59]. Heatmap of the genomic kinship matrix based on VanRaden method was performed in GAPIT3 R-package. The Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) clustering was also constrcuted using factoextra R-package to determine the relationships among the investigated samples. Discriminate analysis of principal components (DAPC) was used alongside STRUCTURE and UPMGA clsuering analysis to gain deeper insights into the population structure. The optimal K value was determined as the smallest number of clusters beyond which the Bayesian Information Criterion (BIC) showed minimal change [60, 61]. DAPC analysis was conducted using the Adegenet package in R [60]. Principal Coordinate Analysis (PCoA) is a distance-based model that utilizes a dissimilarity matrix calculated with a simple-matching index. PCoA of the DArT-seq markers was conducted using TASSEL version 5 software [62] and visualized in Excel. Pairwise linkage disequilibrium (LD) values (r^2) were also calculated in TASSEL version 5 software and plotted against their genetic distance in R 4.4.1.

Results

Before presenting the analysis of molecular diversity, a visual representation of the germplasm in the field is provided to illustrate the context of the breeding program and the phenotypic variation present in the material. The advanced breeding lines were cultivated in extensive field trials, characterized by organized, replicated plots to facilitate rigorous agronomic evaluation (Fig. 1a). At maturity, the lines exhibited clear phenotypic diversity in traits such as plant architecture, boll size, and boll density, which are key targets of selection in the breeding program (Fig. 1b). This observed phenotypic variation provides the physical basis for the genetic diversity that is subsequently explored using molecular markers.

(a)



(b)



Fig. 1 Field view and phenotypic variation of the advanced cotton breeding lines. **a** Aerial view of the extensive field trials (**b**) Ground-level view of mature cotton plants at the boll-setting stage, illustrating phenotypic variation in plant architecture and boll density

The genome wide SNP distribution in tetraploid cotton genotypes

A total of 1,431 polymorphic SNPs were identified in 913 tetraploid cotton advanced breeding lines using reference genomes. Of these, 1,059 SNPs (74%) were distributed across 13 cotton chromosomes, and 372 SNPs (26%) were unmapped. The number of identified SNP substitutions varied among chromosomes, ranging from 28 on chromosome 03 to 161 on chromosome 09. Although the distribution of transition- and transversion-type SNPs varied across chromosomes, transition SNPs (A↔G and C↔T) were more prevalent (772) than transversion SNPs (A↔T, A↔C, T↔/G, and C↔G), which totaled 659. The transition-to-transversion (Ts/Tv) ratio, a key metric for assessing the specificity of newly identified SNPs, ranged from 0.98 on chromosome 04 to 1.61 on chromosome 13. The most frequent substitution was detected for G/and C/TA (203 occurrences), while the least frequent substitution were observed for C/G (59 occurrences) (Table 1).

Genetic diversity

The highest minor allele frequency (MAF) for SNPs was observed on chromosome 12 (0.311), while the lowest was on chromosome 10 (0.1885), with an overall average of 0.2328. All SNPs exhibited a low allele frequency greater than 0.1. Gene diversity (GD) was highest on chromosome 12 (0.3888) and lowest on chromosome 04 (0.2779), with an average of 0.3208. Similarly, polymorphism information content (PIC) values were highest on chromosome 12 (0.3054) and lowest on chromosome 04 (0.2317), averaging 0.2607. Notably, 57.14% of SNPs across all chromosomes had PIC values exceeding 0.25, indicating a high level of polymorphism for many markers. Overall, MAF, GD, and PIC values were closely aligned across cotton chromosomes, with chromosome 12 exhibiting the highest values. The highest observed heterozygosity (H_o) was recorded on chromosome 03 (0.0337), while the lowest was on chromosome 02 (0.0160), with an average of 0.0210. Genetic diversity (H_s) was highest on chromosome 12 (0.3290) and lowest on chromosome 08 (0.2615), with an average of 0.2933. The highest overall gene diversity (H_t), gene diversity among samples (D_{st}), corrected overall gene diversity (H_{tp}), corrected gene diversity among samples (D_{stp}), genetic differentiation (F_{st}), and corrected Fst (F_{stp}) were observed on chromosome 12 (0.3834, 0.0544, 0.4106, 0.0816, 0.1209, and 0.1662, respectively). In contrast, the lowest values were recorded on chromosome 06 (0.2870, 0.0157, 0.2948, 0.0235, 0.0483, and 0.0695, respectively), with overall averages of 0.3259, 0.0326, 0.3421, 0.0489, 0.0830, and 0.1162, respectively. The inbreeding coefficient (F_{is}) ranged from 0.8972 on chromosome 02 to 0.9249 on chromosome 06, with an average of 0.9251, indicating

Table 1 Summary of single nucleotide substitutions identified in cotton chromosomes and genomes using 1,431 DArT markers

Allele	Chr01	Chr02	Chr03	Chr04	Chr05	Chr06	Chr07	Chr 08	Chr 09	Chr 10	Chr 11	Chr 12	Chr13	ChrNA	Total
A/G	14	13	5	12	11	4	15	8	18	9	17	11	8	43	188
T/C	14	7	2	11	10	10	10	9	23	10	9	5	12	46	178
G/A	8	5	3	8	18	9	9	14	20	16	9	8	12	64	203
C/T	8	10	7	18	15	9	9	15	19	11	10	9	13	50	203
Ts	44	35	17	49	54	32	43	46	80	46	45	33	45	203	772
A/T	5	1	2	6	4	1	3	12	11	5	6	4	5	22	87
A/C	2	1	1	6	3	4	4	6	9	4	6	4	3	18	71
T/A	4	5	1	2	4	7	6	10	9	8	7	4	3	30	100
T/G	4	3	1	7	5	5	6	5	6	3	3	2	2	13	65
C/A	6	3	2	10	5	4	7	6	14	6	4	4	3	22	96
C/G	3	1	1	8	1	2	1	3	6	2	7	1	6	17	59
G/T	4	5	2	7	13	8	4	11	15	6	5	1	2	31	114
G/C	4	4	1	4	1	1	8	3	11	6	3	1	4	16	67
Tv	32	23	11	50	36	32	39	56	81	40	41	21	28	169	659
Ts/Tv	1.38	1.52	1.55	0.98	1.50	1.00	1.10	0.82	0.99	1.15	1.10	1.57	1.61	1.20	1.17
Total	76	58	28	99	90	64	82	102	161	86	86	54	73	372	1431

variation among chromosomes. Population differentiation (Dest) was highest on chromosome 12 (0.1304) and lowest on chromosome 06 (0.0357), with an average of 0.0768. Elevated and positive F_{is} values suggest a high level of inbreeding within the populations (Table 2).

The relationship and structure of the population

The population structure was analyzed using the Bayesian Information Criterion (BIC), supported by discriminant analysis of principal components (DAPC) and principal coordinate analysis (PCoA). The determination of populations at each K value and the membership coefficients in the STRUCTURE analysis provided highly informative insights. Delta K (ΔK) and log-likelihood were used to determine the population structure of 913 tetraploid cotton advanced breeding lines and classify them into subgroups (K) (Fig. 2a). The log-likelihood analysis identified K = 2 as the optimal value, suggesting that the 913 breeding lines could be divided into two distinct subgroups with the highest probability. The first subgroup consisted of 292 breeding lines (31.98%), and the second subgroup comprised 621 breeding lines (68.02%) (Supplementary Table 1 and Fig. 2b). Clustering genetic diversity using kinship matrix also revealed that the investigated samples were composed of three classes (i.e., red to yellow in the heat map clustering output). Given that the results of cluster analysis and kinship relationships revealed the presence of three distinct subpopulations, the number of samples within each subpopulation was determined and reported to provide a clearer understanding of the population structure.

The DAPC approach was employed to further investigate population structure and identify subclusters. The Bayesian Information Criterion (BIC) curve in the DAPC analysis exhibited a sharp decline from one to three clusters, followed by a more gradual reduction in BIC values (Fig. 4a). Consequently, K = 3 was determined as the optimal number of clusters. While both methods confirmed the presence of significant genetic stratification, they provided different levels of resolution. The DAPC analysis, by identifying a finer structure (K = 3) and a substantial admixed group (21.79% of lines), offers a more nuanced view of the genetic inter-relationships shaped by the ongoing breeding process. Specifically, subgroup I comprised 419 individuals (45.89%), subgroup II included 154 (16.86%), and subgroup III contained 141 (10.95%), while 199 individuals (21.79%) formed a mixed subgroup, reflecting a diverse population structure (Supplementary Table 2). Heatmap clustering of kinship matrix illustrates relatively high relationship among the first subgroup samples while the samples classified into the third subgroup showed lower relatedness (Fig. 2c).

Cluster analysis was performed using the Weighted Pair Group Method with Arithmetic Mean (WPGMA) to construct a dendrogram based on a pairwise similarity matrix (Fig. 3). This approach classified the panel into three distinct subgroups, aligning with the results of the STRUCTURE analysis. The first subgroup (I) comprised 36 samples (red color), the second main subgroup (II) included 514 samples (green color), and the third main subgroup (III) consisted of 358 samples (blue color). The results of PCoA were adopted with WPGMA-based clustering results which divided the 913 advanced breeding lines into three subgroups (Fig. 4a, b). The first and second coordinates respectively described 9.11% and 2.94% of the total diversity. PCoA1 separated the two subgroups well so that subgroup I was near the origin of the biplot, and subgroup II had high negative values (Fig. 4c).

Genetic differentiation of populations

The AMOVA analysis partitioned genetic variance into two levels: among and within subgroups, each contributing differently to the total genetic variation. This analysis was conducted based on the DAPC and STRUCTURE results. In both cases, genetic variance among and within subpopulations was found to be significant. For advanced breeding lines, the variation among subgroups accounted for 12% of the total genetic variation, whereas the variation between populations contributed 82%, and within individuals contributed 6% (Table 3).

The fixation index (F_{st}) was estimated among the subgroups identified by DAPC and STRUCTURE (Fig. 4b) to assess population differentiation arising from genetic structure. An F_{st} value close to zero indicates low genetic differentiation, making it difficult to distinguish one subgroup from another. Based on genetic distance, F_{st} values indicated lower differentiation between subgroup I and subgroup III, whereas subgroup II appeared to be genetically distinct from the other two subgroups. Pairwise genetic differentiation and gene flow among the three subgroups revealed the highest genetic differentiation and lowest gene flow between subgroup I and subgroup II ($F_{st} = 0.190$, $N_m = 1.066$) while the lowest genetic differentiation and highest gene flow was detected between subgroup I and subgroup III ($F_{st} = 0.069$, $N_m = 3.381$) (Table 4).

The allelic pattern across the populations

Genetic variation within the subgroups were assessed using 913 advanced breeding lines. The highest values for the average number of alleles per locus (Na = 2.00), effective number of alleles (Ne = 1.62), Shannon diversity index (I = 0.543), and percentage of polymorphic loci (PPL = 100) were observed in population III. In contrast, the lowest values (Na = 1.99, Ne = 1.35, I = 0.375, and PPL = 99.03) were recorded in subgroup II. The overall mean values for Na, Ne, I, and PPL were 2.00, 1.48, 0.454, and 99.68, respectively (Table 5).

Table 2 Profile of single nucleotide polymorphism (SNP) in the 913 tetraploid cotton advanced breeding lines

Chromosome	MAF[1]	GD	PIC	Ho	Hs	Ht	Dst	Htp	Dstp	Fst	Fstp	Fis	Dest
Chr01	0.2312	0.3188	0.2597	0.0238	0.2965	0.3248	0.0283	0.3390	0.0425	0.0766	0.1096	0.9111	0.0681
Chr02	0.2484	0.3417	0.2768	0.0337	0.3164	0.3572	0.0408	0.3775	0.0612	0.1002	0.1388	0.8972	0.0953
Chr03	0.2054	0.2933	0.2421	0.0160	0.2812	0.3058	0.0246	0.3181	0.0369	0.0593	0.0833	0.9390	0.0591
Chr04	0.1918	0.2779	0.2317	0.0207	0.2705	0.2939	0.0235	0.3056	0.0352	0.0636	0.0903	0.9188	0.0548
Chr05	0.2799	0.3611	0.2883	0.0221	0.3001	0.3414	0.0413	0.3620	0.0619	0.1047	0.1467	0.9204	0.0980
Chr06	0.1988	0.2898	0.2386	0.0182	0.2713	0.2870	0.0157	0.2948	0.0235	0.0483	0.0695	0.9249	0.0357
Chr07	0.2202	0.3087	0.2525	0.0180	0.2844	0.3092	0.0248	0.3216	0.0372	0.0658	0.0937	0.9347	0.0588
Chr08	0.1933	0.2860	0.2370	0.0180	0.2615	0.2929	0.0314	0.3086	0.0471	0.0880	0.1226	0.9252	0.0705
Chr09	0.2334	0.3123	0.2529	0.0207	0.2832	0.3240	0.0407	0.3443	0.0611	0.1017	0.1409	0.9245	0.0959
Chr10	0.1885	0.2908	0.2438	0.0193	0.2716	0.3010	0.0294	0.3157	0.0442	0.0798	0.1109	0.9290	0.0660
Chr11	0.1969	0.2909	0.2399	0.0182	0.2684	0.2948	0.0263	0.3079	0.0395	0.0656	0.0920	0.9271	0.0621
Chr12	0.3110	0.3888	0.3054	0.0242	0.3290	0.3834	0.0544	0.4106	0.0816	0.1209	0.1662	0.9249	0.1304
Chr13	0.2653	0.3513	0.2811	0.0220	0.3210	0.3511	0.0301	0.3662	0.0452	0.0800	0.1141	0.9291	0.0724
ChrNA	0.2524	0.3408	0.2745	0.0210	0.3132	0.3473	0.0340	0.3643	0.0511	0.0834	0.1165	0.9301	0.0830
Whole genome	0.2328	0.3208	0.2607	0.0210	0.2933	0.3259	0.0326	0.3421	0.0489	0.0830	0.1162	0.9251	0.0768

[1]MAF Minor allele frequency, GD Gene diversity, PIC Polymorphic information content, Ho Heterozygosity, genetic diversity (Hs) among chromosomes and genomes; Ht, overall gene, diversity (Ht), corrected Ht (HTp), gene diversity among samples (Dst), corrected Dst (Dstp), genetic differentiation (Fst), corrected Fst (Fstp), inbreeding coefficient (Fis), population differentiation (Dest)

Evaluation of linkage disequilibrium

Based on the analysis of linkage disequilibrium, it was found that LD decayed with genetic distance. The 19,059 pairs of SNPs in the tested genotypes showed an average R^2 value of 0.0353, suggesting no high LD. We found that finding the average of the LD in each genome, rather than measuring the LD between two SNPs located on the same chromosome, was more useful for identifying the pattern of LD across the two genomes. Table 6 summarizes the average LD per chromosome, the total number of SNP pairs, and the number of significant SNP pairs located on the same chromosome. The total number of SNP pairs varied from 220 on Chr03 to 3,410 on Chr09. Linkage disequilibrium (LD) within each chromosome ranged from 0.0156 (Chr07) to 0.0547 (Chr05). The highest number of significant marker pairs (740) was found on Chr09, while the lowest (36) was observed on Chr03 (Table 6). Figure 5 illustrates LD decay across each chromosome and the entire genome.

Discussion

This study focused on identifying the genetic diversity and population structure of advanced cotton breeding lines. Molecular marker- based methods of genetic diversity and population genetic studies are not only reliable and efficient but genotypic information can be used for genetic mapping studies, introgressive breeding and genomics-assisted breeding. Several studies have investigated genetic diversity in cotton, with most focusing on germplasm from specific geographic regions, including USA [63], Brazil [64], China [65], Australia [66] and India [67]. Some studies, except for Gowda et al. [63], focused on specific breeding programs, such as the Pee Dee program [68] and the Acala program [9]. Additionally, many studies primarily investigated cultivated types [69] or landraces [70]. To the best of our knowledge, this is the first attempt to assemble the population and investigate the genetic structure of cotton advanced breeding lines in Türkiye. In this study, next-generation sequencing (NGS) techniques were employed based on genotyping-by-sequencing (GBS). This approach is a widely used NGS technique that enables efficient marker discovery and high-density genotyping at a relatively low cost. [71, 72]. In this study, a total of 1,431 SNPs derived from GBS analysis using DArTseq technology on 913 advanced breeding lines were utilized for population diversity and structure analysis. Continuous evaluation of genetic diversity in breeding programs is a fundamental step toward systematically improving the genetic base of cotton through introgressive breeding. The effectiveness of SNP markers in assessing genetic diversity and population structure in cotton has been well established [63, 73–75].

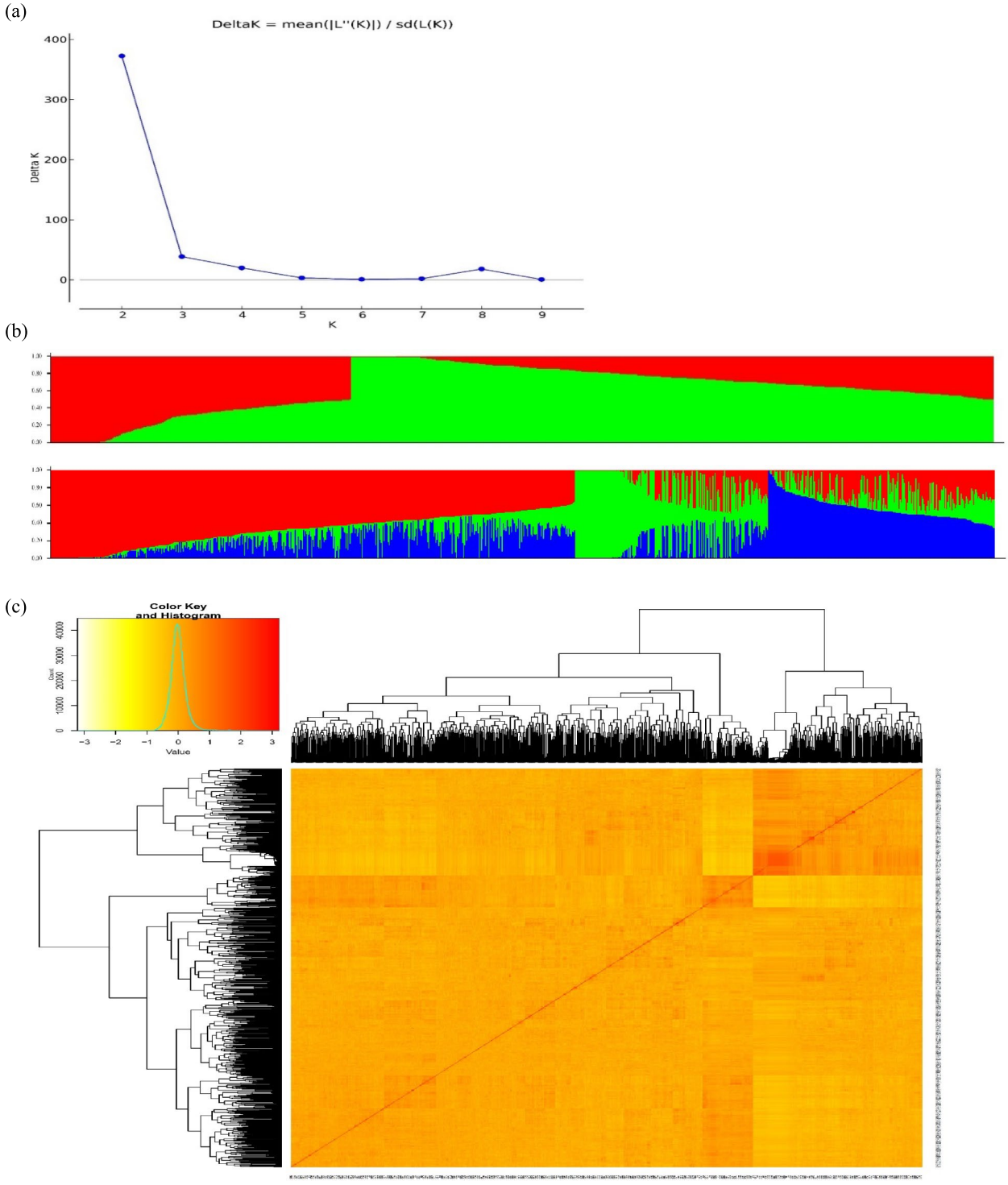

Fig. 2 Determination of subgroups number in cotton advanced breeding lines based on ΔK values (**a**), a structure plot of the 913 advanced cotton breeding lines determined by K = 2 and K = 3 (**b**), cluster analysis using kinship matrix for cotton advanced breeding lines (**c**) using 1,431 DArT markers for 913 advanced cotton breeding lines reflecting three subgroups

In this study, the molecular markers employed successfully detected substantial genetic variability among the cotton breeding lines. A total of 1,431 polymorphic SNPs were identified in 913 tetraploid cotton breeding lines. These SNPs substitution was distributed across 13 cotton chromosomes, with the number of SNPs ranging from 28 on chromosome 03 to 161 on chromosome 09. Transition SNPs (A ↔ G and C ↔ T) were more common (772) than

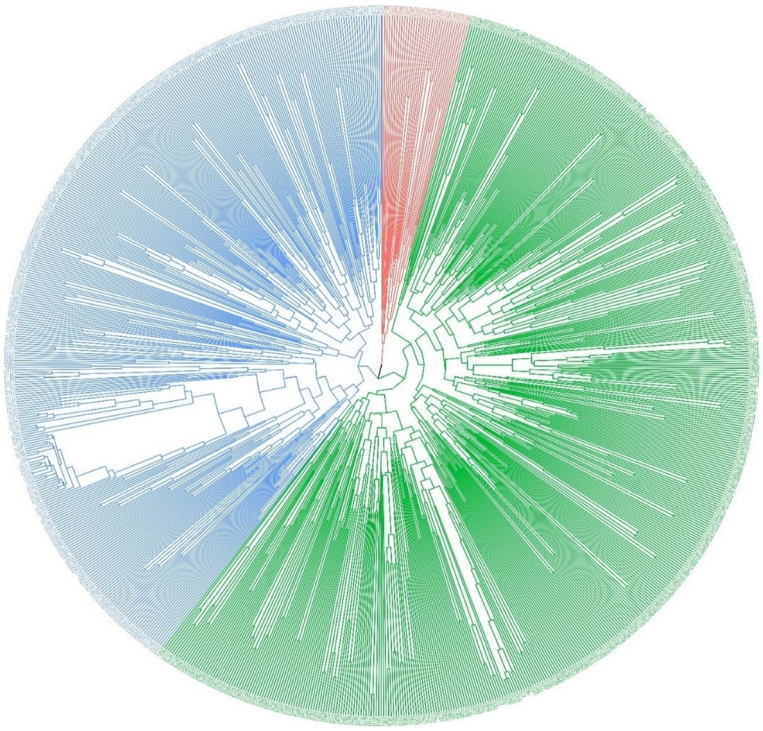

Fig. 3 The dendrogram of UPGMA clustering constructed using 1,431 DArT markers for 913 advanced breeding lines

transversion SNPs (A↔T, A↔C, T↔G, and C↔G), which totaled 659. The Ts/Tv ratio varied between chromosomes, ranging from 0.82 on chromosome 04 to 1.61 on chromosome 13. G/A and C/T were the most frequent substitution (203 occurrences), while C/G was the least common (22 occurrences). The abundance of transition-type SNPs is commonly observed in true SNPs and reflects the transition of cytosine to thymine through the determination of 5-methylcytosine following cytosine methylation [76]. Additionally, the number of A/G transitions was slightly higher than that of C/T transitions, with A/T transversions being the most common. Similar findings have been reported by Wang et al. [28] and Song et al. [77]. The Ts/Tv ratio provides a high level of confidence in the authenticity of the SNPs, suggesting that the observed SNPs are not the result of sequencing and mapping errors [78]. The Ts/Tv ratio serves as a benchmark for evaluating the specificity of new SNP calls. The Ts/Tv ratio observed in this study is significantly lower than that reported by Joshi et al. [79]. The SNP distribution across the genome was uneven, varying according to genome content and gene density.

PIC values are crucial parameters for assessing polymorphism among genotypes in breeding programs. In addition to the PIC values, the GD and MAF of each marker in the diversity panel were also evaluated, which may reflect the impact of breeding programs and selection pressures. The average PIC value was 0.2607, indicating a relatively low degree of genetic diversity, a finding consistent with those of Fang et al. [25] and Ai et al. [80]. Variations in genetic diversity values may be attributed to the types of germplasm used and the methods employed to reveal PCR products. The MAF for SNPs averaged 0.2328, while the GD had a mean value of 0.3208. Our results revealed significant differences in the GD and MAF values across cotton chromosomes. These findings suggest that the markers successfully captured genetic diversity within the advanced cotton lines, as evidenced by their PIC values and the well-distributed SNP markers studied. In addition, MAF is widely recognized as a critical threshold for identifying high-quality and effective markers [81]. A higher number of SNPs with MAF > 0.5% further supports the quality of these markers for subsequent genetic variation analyses. The high MAF and PIC values observed suggest considerable polymorphism in the SNPs across the genomes. It is noteworthy that increasing the number of SNPs by enhancing polymorphism contributes to higher genetic diversity and enhances the likelihood of detecting rare alleles [82]. These results emphasize the importance of incorporating MAF and PIC values in genome-related analyses, such as genome-wide mapping studies. From a practical perspective, assessing the genetic diversity and population structure of advanced cotton lines provides valuable insights for improving cotton traits. The estimated genetic parameters, particularly the high level of allelic diversity, indicate that the investigated cotton line is a rich genetic resource that can be utilized by breeders.

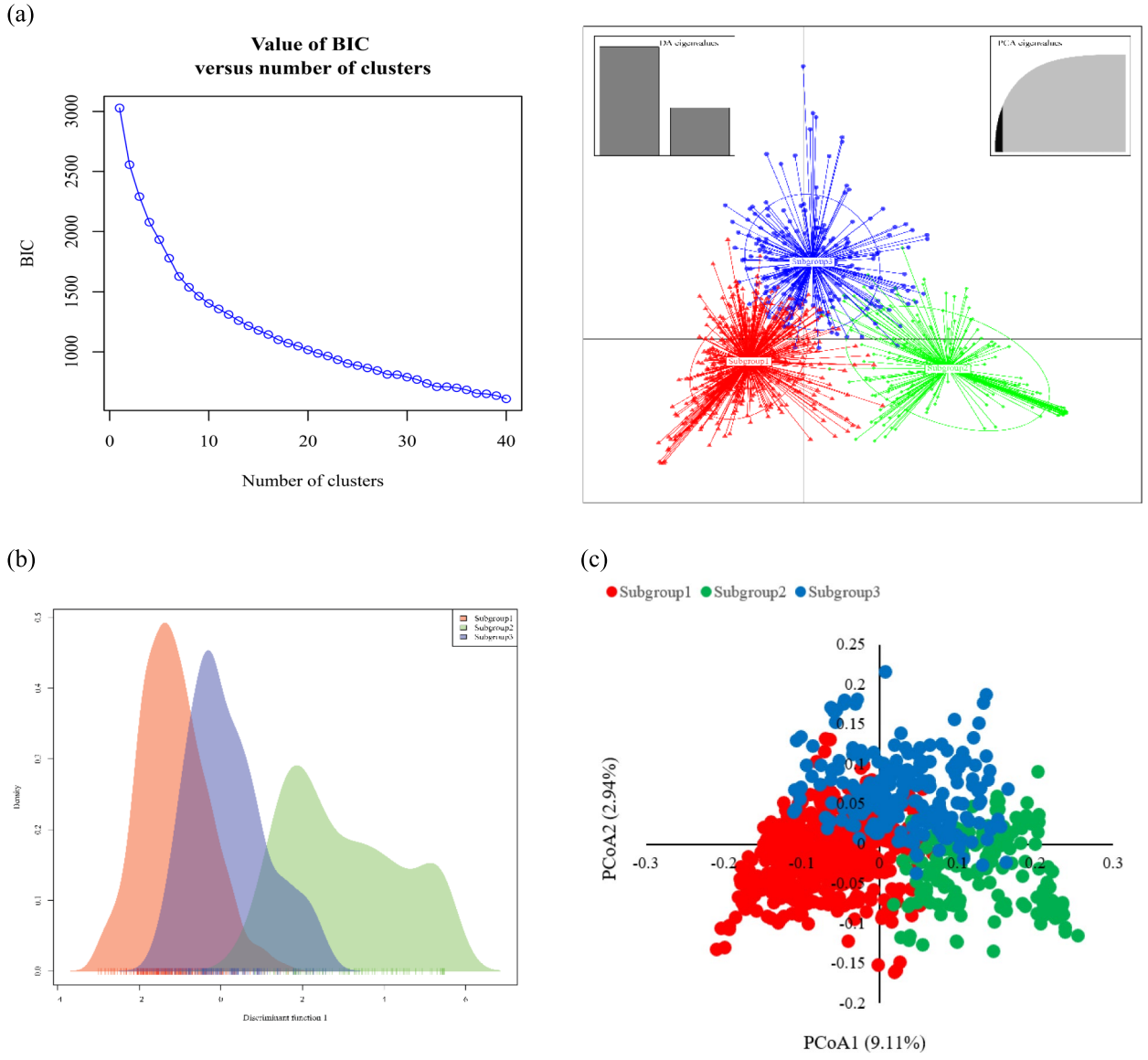

Fig. 4 Bayesian information criterion (BIC) showing a rapid decline from 1 to 3 followed by BIC declining value reduction at 3 (**a**), density plot of individuals, partitioned into 3 subgroups by DAPC. Subgroup II is further apart from Subgroup I and III. Subgroups I and III present an intersected area, (**b**) DAPC scatters plot of 913 advanced breeding lines shows three subgroups (**c**) principal coordinate analysis of the 913 advanced breeding lines

Table 3 Analysis of molecular variance showing the partitioning of genetic variation within 913 advanced breeding lines based on DAPC analysis

Source	df	SS	MS	Est. Var	%	Phi-statistic	*P*-value
Among Population	2	13096.851	6548.426	11.855	12%	0.122	< 0.001
Among Individual	910	149808.958	164.625	79.542	82%		
Within Individual	913	5060.000	5.542	5.542	6%		
Total	1825	167965.809		96.938	100%		

In our study, the mean values of H_o, H_s, H_t, D_{st}, H_{tp}, D_{stp}, F_{st}, F_{stp}, F_{is}, and D_{est} were 0.0210, 0.2933, 0.3259, 0.0326, 0.3421, 0.0489, 0.0830, 0.1162, 0.9251, and 0.0768, respectively. In our study, the F_{st} in chromosomes 02, 05, 09, and 12 were more than 0.10, which could be due to the high levels of differentiation among the subgroups acrosss those chromosomes [83]. F_{st} is a measure of population differentiation resulting from the genetic structure. Higher observed heterozygosity (H_o), overall gene diversity (H_t), and genetic differentiation (Fst) indicate a lack of uniformity and limited gene flow among the three subgroups, leading to increased genetic diversity.

Table 4 Pairwise comparison of genetic differentiation (Fst, below diagonal) and gene flow (Nm, upper diagonal) between subgroups of cotton advanced breeding lines

Population	Subgroup I	Subgroup II	Subgroup III
Subgroup I	-	1.066	3.381
Subgroup II	0.190	-	2.050
Subgroup III	0.069	0.109	-

Table 5 Genetic variation of the 913 advanced breeding lines

Population	Na[1]	Ne ± SE	I ± SE	PPL
Subgroup I	2.00	1.46 ± 0.013	0.447 ± 0.008	100
Subgroup II	1.99	1.35 ± 0.012	0.374 ± 0.008	99.03
Subgroup III	2.00	1.62 ± 0.012	0.543 ± 0.006	100
Mean	2.00	1.48 ± 0.008	0.454 ± 0.005	99.68

Values in tables are estimates ± S.E

[1]*Na* Average number of alleles per locus, *Ne* Effective number of alleles, *I* Shannon diversity index, *PPL* Percentage of polymorphic loci

Table 6 Linkage disequilibrium between 1,431 DArT markers in 913 advanced breeding lines

Chromosome	TNSP[a]	Dist	R^2	NSSP[b]
Chr01	1540	15758393	0.0277	313
Chr02	1063	27387556	0.0350	195
Chr03	220	19885014	0.0120	36
Chr04	1499	16763533	0.0511	246
Chr05	1694	20851536	0.0547	382
Chr06	784	15705374	0.0356	102
Chr07	1655	21199490	0.0156	233
Chr08	1820	10330201	0.0419	264
Chr09	3410	8515466	0.0336	740
Chr10	1498	20674289	0.0516	197
Chr11	1577	19916770	0.0210	190
Chr12	996	14660255	0.0312	178
Chr13	1303	21476734	0.0266	200
Whole genome	19059	16706445	0.0353	3276

[a]*TNSP* Total number of SNP pairs, [b]*NSSP* Number of significant SNP pairs ($P < 0.001$)

Additionally, the low variation among individuals within each subgroup suggests a reduced rate of pollination and genetic exchange [84]. The high levels of observed heterozygosity and genetic diversity further support the Fst results, clearly distinguishing the three subgroups. These findings highlight the strong capacity of SNP markers to differentiate the subgroups investigated. Chang et al. [85] reported that the accuracy of genomic relationship matrices could be improved by prioritizing SNPs based on Fst scores. While these scores are effective for prioritization, they primarily measure fixation. Therefore, incorporating heterozygosity and other genetic differentiation indices may enhance the understanding of population structure and improve SNP marker selection.

In applied breeding, understanding population structure in germplasm is crucial for identifying genes and quantitative trait loci (QTL) that control phenotypic variation [86]. The results of this study revealed the existence of a genetic structure within the cotton germplasm. Structure analysis classified the landraces into two main subgroups (K = 2). This result was supported by Joshi et al. [79]. The DAPC approach was employed to further investigate population structure and identify subclusters. The Bayesian Information Criterion (BIC) curve in the DAPC analysis exhibited a sharp decline from one to three clusters, followed by a more gradual reduction in BIC values. Consequently, K = 3 was determined as the optimal number of clusters. This result is consistent with [87]. A key finding of our population structure analysis was the identification of different optimal cluster numbers by the STRUCTURE (K = 2) and DAPC (K = 3) methods. This is not a contradiction but rather a reflection of the different sensitivities of the algorithms and the hierarchical nature of the genetic structure within this breeding panel. The Evanno ΔK method used with STRUCTURE is known to robustly detect the highest level of population subdivision, which in our case points to two primary ancestral pools. In contrast, DAPC, a non-model-based approach, is adept at uncovering more subtle subdivisions. The K = 3 result from DAPC likely represents a significant secondary division within one of the two major groups, reflecting the creation of distinct breeding populations through targeted crosses. This hierarchical structure provides breeders with a multi-level map of genetic diversity, allowing for strategic parental selection at both broad and fine scales. The identification of a large proportion of admitted individuals (21.79%) by DAPC is a hallmark of an active and dynamic breeding program. These individuals are not unclassified artifacts but rather represent the tangible outcomes of strategic hybridization between distinct genetic pools. From a breeding perspective, this admixed group is arguably the most valuable genetic reservoir within the panel. It contains novel combinations of alleles from different parental backgrounds, providing the raw material for selection. These lines are the most likely source for identifying transgressive segregants—progeny that exhibit phenotypes superior to both parents—for key traits such as yield, fiber quality, and stress resilience. Therefore, this analysis provides breeders with a powerful tool to prioritize these admixed lines for intensive phenotypic evaluation and advancement. A notable finding of our genotyping effort was that a significant portion (26%) of the high-quality polymorphic SNPs could not be mapped to the current *G. hirsutum* reference genome. This is likely attributable to several factors, including the inherent complexity and repetitive nature of the allotetraploid cotton genome, and the presence of novel genomic regions or structural variants in our Turkish breeding material that are not captured in the reference assembly. While the exclusion of these markers from chromosomal analyses means our genome-wide

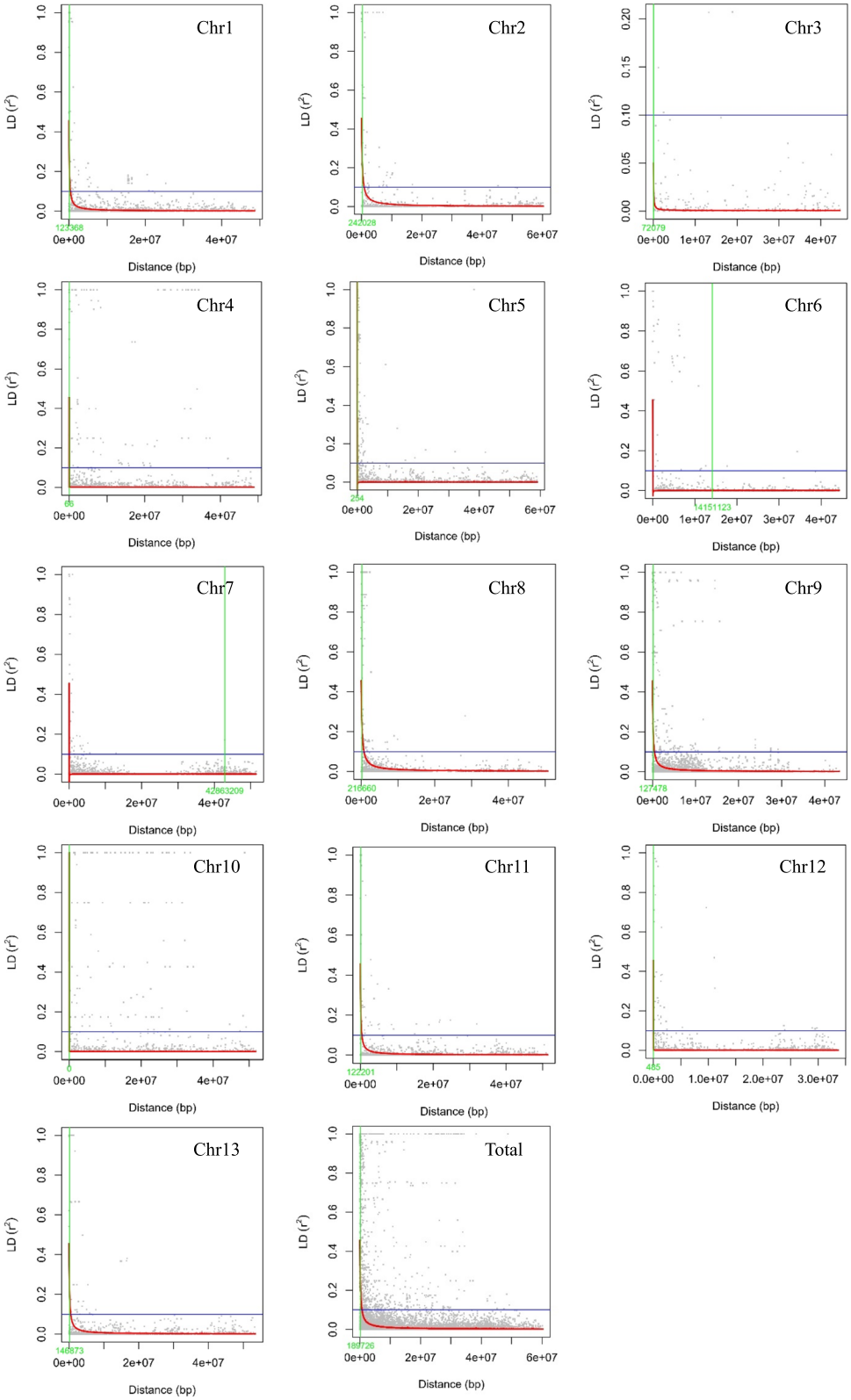

Fig. 5 The rate of linkage disequilibrium (LD) in 913 advanced breeding lines based on the 1,431 DArT markers into two and three clusters

diversity estimates may be conservative, their existence is highly significant. It strongly suggests that the ProGen germplasm harbors a considerable reservoir of genetic diversity beyond that which is characterized by the reference sequence. This unmapped variation could be linked to important traits for regional adaptation and represent a valuable, untapped resource for future breeding efforts. This finding also underscores the critical need for the development of more comprehensive genomic resources for cotton, such as a pangenome, to fully capture the species' genetic diversity.

A notable finding of our genotyping effort was that a significant portion (26%) of the high-quality polymorphic SNPs could not be mapped to the current *G. hirsutum* reference genome. While unmapped markers are often excluded from downstream analysis, their high prevalence here is scientifically significant. It strongly suggests that the ProGen germplasm, shaped by decades of selection in the Turkish environment and introgression from diverse sources, harbors a considerable reservoir of genetic variation not captured in the reference sequence. This unmapped allelic diversity could be linked to important loci for regional adaptation and represents a valuable, untapped resource for future breeding. This finding underscores the need for developing more comprehensive genomic resources, such as a pangenome for cotton, to fully harness the genetic potential present in diverse breeding programs globally. While this study provides a comprehensive overview of the genetic landscape of these advanced breeding lines, its full potential will be realized when this genomic information is linked to phenotypic performance. A thorough understanding of population structure and linkage disequilibrium, as established herein, is an essential prerequisite for conducting a robust genome-wide association study (GWAS) and mitigating the risk of spurious marker-trait associations. Extensive multi-location field trials to phenotype this panel for key yield, fiber quality, and stress tolerance traits are currently in progress. The results of the present study provide the foundational genomic framework for a forthcoming GWAS that will aim to identify specific loci and candidate genes for marker-assisted selection in Turkish cotton breeding.

In applied breeding, understanding population structure in germplasm is crucial for identifying genes and quantitative trait loci (QTL) that control phenotypic variation. The results of this study revealed the existence of a genetic structure within the cotton germplasm. Structure analysis classified the landraces into two main subgroups (K = 2). This result was supported by [79] The DAPC approach was employed to further investigate population structure and identify subclusters. The Bayesian Information Criterion (BIC) curve in the DAPC analysis exhibited a sharp decline from one to three clusters, followed by a more gradual reduction in BIC values. Consequently, K = 3 was determined as the optimal number of clusters. This result is consistent with the multivariate methods, including WPGMA clustering, PCoA, and the Bayesian model-based clustering approach implemented in STRUCTURE software, were successful in assigning genotypes to one of the two primary subgroups. The results of PCoA were adopted with WPGMA-based clustering results which divided the 913 advanced breeding lines into three subgroups. The clustering of the germplasm into 2 or 3 subgroups did not qualitatively affect our conclusion, even though other researchers have reported the soybean population structure with more than 3 ($\Delta K \geq 3$) subgroups [88].

Genetic distance plays a crucial role in the selection of lines for breeding programs, as it directly impacts the efficiency of genetic improvement. The use of lines with minimal genetic diversity may not be advisable, as it could limit the potential for genetic gain. Therefore, a comprehensive understanding of the population structure of the 913 advanced cotton lines is essential to elucidate their genetic relationships and facilitate the selection of genetically diverse lines for breeding efforts. In this study, genetic differentiation among subgroups was assessed using AMOVA, incorporating the results of structure and DAPC analyses. AMOVA revealed that the two subgroups exhibited highly significant genetic diversity, with a significant difference between the evaluated subgroups ($P < 0.001$), confirming genetic differentiation among individuals. This result is consistent with Liang et al. [80]. The low genetic diversity observed within the populations may be attributed to the nature of the advanced cotton lines. Consequently, selecting advanced lines from different subgroups for the improvement of target traits may be more effective than selecting genotypes from the same subgroup.

To assist breeders in selecting parental accessions for cotton improvement, we quantified the differentiation between identified subgroups using F_{st} statistics. In this study, F_{st} values indicated lower differentiation between subgroup II and subgroup III, whereas subgroup I appeared to be genetically distinct from the other two subgroups. Pairwise genetic differentiation and gene flow among the three subgroups revealed the highest genetic differentiation between subgroup I and subgroup II. Conversely, the lowest gene flow was also observed between these two subgroups, while the highest gene flow was detected between subgroup I and subgroup III. Our results showed that the F_{st} value ranged from 0.069 to 0.190, which is lower than the values reported by Wang et al. [89], which ranged from 0.04 to 0.61. Similarly, Gowda et al. [90] observed F_{st} values ranging from 0.22 to 0.65. The level of gene flow between subgroups was assessed by calculating the haploid number of migrants

(N_m). In this study, however, a high level of gene flow was observed among the tested materials, supporting the distribution of genotypes across subgroups within the plant material.

The overall mean values for Na, Ne, I, and PPL were 2.00, 1.48, 0.454, and 99.68, respectively. Among the three subgroups, subgroup III exhibited the highest genetic diversity, as indicated by all allelic pattern indices (Na, Ne, I, He, F, and PPL). Determining the magnitude and decay of linkage disequilibrium (LD) is crucial, as these factors influence SNP marker efficiency, and the resolution of association mapping required for genome-wide association studies [91]. The extent of LD varies significantly across genomes and species. In wheat, LD decay was analyzed separately for each of its two genomes. Based on the nonlinear logarithmic trend line, LD decay was estimated when LD values dropped below 0.1. Based on the analysis of linkage disequilibrium, it was found that LD decayed with genetic distance. The 19,059 pairs of SNPs in the tested genotypes showed an average R^2 value of 0.0353, suggesting no high LD. We found that finding the average of the LD in each genome, rather than measuring the LD between two SNPs located on the same chromosome, was more useful for identifying the pattern of LD across the two genomes. The total number of SNP pairs varied from 220 to 3,410.

It is acknowledged that the 913 advanced breeding lines analyzed in this study are proprietary to the ProGen Seed Company breeding program, which may limit their direct utilization by the broader public breeding community. However, this does not confine the study's findings to a purely academic context. On the contrary, this collection represents a unique and valuable snapshot of the genetic architecture of an elite, commercially focused germplasm pool adapted to the Eastern Mediterranean environment. The population structure, patterns of genetic diversity, and linkage disequilibrium identified herein serve as a powerful model and a set of guiding principles for other breeding programs, both public and private. By elucidating genetic relationships within a highly selected population, this research provides a roadmap for how to strategically manage and exploit genetic diversity for trait improvement, offering insights that are transferable to other cotton breeding efforts worldwide. Furthermore, the public deposition of the marker data contributes a significant resource to the global cotton research community for comparative genomics and future meta-analyses.

Conclusion

This study presents the first comprehensive SNP-based assessment of genetic diversity and population structure in advanced cotton breeding lines from Türkiye. The results show that most genetic variation resides within populations, and clustering analyses consistently identify two to three major subgroups. These findings highlight the presence of substantial, exploitable diversity within elite germplasm. By pinpointing genetically distant subgroups, the study provides clear guidance for strategic parental selection to enhance heterosis and broaden variation in breeding populations. Overall, the generated genomic insights establish a strong foundation for future GWAS, marker-assisted selection (MAS), and the efficient management and use of cotton genetic resources in breeding programs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12643-9.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

The authors acknowledge the support given by ProGen Seed Company, Türkiye (https://www.progenseed.com/), for providing all the necessary facilities and materials used for experiments.

Germplasm availability and intellectual property

The company is open to establishing formal research collaborations with public institutions for noncommercial research projects. Any such collaboration would be governed by a standard Material Transfer Agreement (MTA) to clearly define the terms of use, publication rights, and intellectual property considerations.

Permission for land study

The authors declare that all land experiments and studies were performedaccording to authorized rules.

Authors' contribution

Conceptualization: K.H. and A.T.; methodology: K.H. and A.T.; software: A.T. and H.A.; validation: K.H.; investigation: A.T. and H.A.; resources: K.H. B.A. and T.H.; data curation: K.H, B.A., T.H., H.A. and A.T.; writing the original draft preparation: A.T; writing the review and editing: K.H., H.A. and A.T.; visualization: K.H. and A.T.; supervision: K.H., A.T, and H.A.; project administration: K.H.; All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Scientific and Technological Research Council of Türkiye (TEYDEP-1501 grant number: 3191998).

Data availability

The dataset generated and analyzed during the current study has been deposited in the European Variation Archive (EVA) at EMBL-EBI under project accession number PRJEB105595. The data is publicly available and can be accessed at: https://www.ebi.ac.uk/eva/?eva-study=PRJEB105595.

Declarations

Ethics approval and consent to participate

The plant material was collected and formally identified by Dr. Batuhan Akgol. A voucher specimen has been deposited in the ProGen gene bank (Antakya/Hatay, Türkiye) and is publicly available to the research community under the accession number [PG_H_0001].
Not applicable. No human participant.

Consent for publication

Not applicable. No human participant.

Competing interests
The authors declare no competing interests.

Received: 30 May 2025 / Accepted: 6 February 2026
Published online: 14 February 2026

References

1. Campbell B, Liu Y. 2024 Cotton Outlook. Athens, GA: University of Georgia Cooperative Extension; 2024. Annual Publication 130-2-04. Available from: B, & Liu, Y. : 2024 Cotton outlook. UGA Cooperative Extension, Annual Publication 130–2–04. Retrieved from 2024.
2. Zhang J, Fang H, Zhou H, Sanogo S, Ma Z. Genetics, breeding, and marker-assisted selection for Verticillium wilt resistance in cotton. Crop Sci. 2014;54(4):1289–303.
3. Fang H. Development of molecular markers and mapping of quantitative trait loci for resistance to Verticillium wilt disease using two inbred line populations in tetraploid cotton. PhD thesis. Las Cruces: New Mexico State University; 2013.
4. Pimentel D. Techniques for reducing pesticide use: environmental and economic benefits. In: Pimentel D, editor. Techniques for Reducing Pesticide Use: Economic and Environmental Benefits. Chichester: Wiley; 1997. pp. 1–11.
5. Shrestha N, Zhang K, Gowda SA, Abdelraheem A, Jones DC, Kuraparthy V. Identification of quantitative trait loci for fiber quality, yield, and plant height traits in upland cotton. Crop Sci. 2023;63(3):1421–41.
6. Paterson AH, Smith RH. Future horizons: biotechnology for cotton improvement. In: Smith CW, Cothren JT, editors. Cotton: Origin, History, Technology, and Production. New York: Wiley; 1999. pp. 415–32.
7. Wendel JF, Brubaker CL, Percival AE. Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. Am J Bot. 1992;79(11):1291–310.
8. Abdalla A, Reddy O, El-Zik K, Pepper A. Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theor Appl Genet. 2001;102:222–9.
9. Zhang Z-S, Xiao Y-H, Luo M, Li X-B, Luo X-Y, Hou L, et al. Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). Euphytica. 2005;144:91–9.
10. Zhang H-B, Li Y, Wang B, Chee PW. Recent advances in cotton genomics. Int J Plant Genomics. 2008;2008(1):742304.
11. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, et al. Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol. 2007;145(4):1303–10.
12. Cook RJ. Advances in plant health management in the twentieth century. Annu Rev Phytopathol. 2000;38(1):95–116.
13. Jia JZ. Molecular germplasm diagnostics and molecular marker-assisted breeding. Sci Agric Sin. 1996;29(4):1–10. Chinese.
14. Hao C, Dong Y, Wang L, You G, Zhang H, Ge H, et al. Genetic diversity and construction of core collection in Chinese wheat genetic resources. Chin Sci Bull. 2008;53(10):1518–26.
15. Astarini IA, Plummer JA, Lancaster RA, Yan G. Fingerprinting of cauliflower cultivars using RAPD markers. Aust J Agric Res. 2004;55(2):117–24.
16. Lenaerts B, Collard BC, Demont M. Improving global food security through accelerated plant breeding. Plant Sci. 2019;287:110207.
17. Malik W, Ashraf J, Iqbal MZ, Ali Khan A, Qayyum A, Ali Abid M, et al. Molecular markers and cotton genetic improvement: current status and future prospects. Sci World J. 2014;2014(1):607091.
18. Chaudhary L, Sindhu A, Kumar M, Kumar R, Saini M. Estimation of genetic divergence among some cotton varieties by RAPD analysis. J Plant Breed Crop Sci. 2010;2(3):039–43.
19. Guo W, Zhang T, Shen X, Yu JZ, Kohel RJ. Development of SCAR marker linked to a major QTL for high fiber strength and its usage in molecular-marker assisted selection in upland cotton. Crop Sci. 2003;43(6):2252–6.
20. Karaca M, İnce AG. New non-redundant microsatellite and CAPS-microsatellite markers for cotton (*Gossypium* L.). Turk J Field Crops. 2011;16(2):172–8.
21. Ismail S, Adawy SS, El-Din A-KYG, Hussein E. Linkage map construction and detection of QTLs associated with earliness, fiber quality and yield in an interspecific cross between *Gossypium hirsutum* L. and *Gossypium barbadense* L. Int J Adv Res. 2015;3(3):637–49.
22. Rakshit A, Rakshit S, Singh J, Chopra S, Balyan H, Gupta P, et al. Association of AFLP and SSR markers with agronomic and fibre quality traits in *Gossypium hirsutum* L. J Genet. 2010;89:155–62.
23. Lin Z-x, He D, Zhang X-l, Nie Y, Guo X, Feng C, et al. Linkage map construction and mapping QTL for cotton fibre quality using SRAP, SSR and RAPD. Plant Breed. 2005;124(2):180–7.
24. Zhang Z-S, Hu M-C, Zhang J, Liu D-J, Zheng J, Zhang K, et al. Construction of a comprehensive PCR-based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*Gossypium hirsutum* L.). Mol Breed. 2009;24:49–61.
25. Fang DD, Hinze LL, Percy RG, Li P, Deng D, Thyssen G. A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in Upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. Euphytica. 2013;191:391–401.
26. Qin H, Chen M, Yi X, Bie S, Zhang C, Zhang Y, et al. Identification of associated SSR markers for yield component and fiber quality traits based on frame map and upland cotton collections. PLoS ONE. 2015;10(1):e0118073.
27. Wang B, Draye X, Zhuang Z, Zhang Z, Liu M, Lubbers EL, et al. QTL analysis of cotton fiber length in advanced backcross populations derived from a cross between *Gossypium hirsutum* and *G. mustelinum*. Theor Appl Genet. 2017;130:1297–308.
28. Wang Q, Fang L, Chen J, Hu Y, Si Z, Wang S, et al. Genome-wide mining, characterization and development of microsatellite markers in *Gossypium* species. Sci Rep. 2015;5(1):10638.
29. Zhang Y, Wang X, Li Z, Zhang G, Ma Z. Assessing genetic diversity of cotton cultivars using genomic and newly developed expressed sequence tag-derived microsatellite markers. Genet Mol Res. 2011;10(3):1462–70.
30. Noormohammadi Z, Sheidai M, Ghasemzadeh-Baraki S, Alishah O. Genetic diversity analysis in Opal cotton hybrids based on SSR, ISSR, and RAPD markers. Genet Mol Res: GMR. 2013;12(1):256–69.
31. Sethi K, Siwach P, Verma SK, Sihag M. Assessing genetic diversity among Gossypium arboreum L. genotypes using ISSR markers. Int J Pharm Bio Sci. 2015;6(1):201–8.
32. Sharma R, Joshi A, Maloo S, Rajaman G. Assessment of genetic finger printing using molecular marker in plants: a review. Sci Res Impact. 2012;1:29–36.
33. Cai C, Zhu G, Zhang T, Guo W. High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. BMC Genomics. 2017;18:1–14.
34. Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD, et al. Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. G3 Genes|Genomes|Genetics. 2015;5(6):1187–209.
35. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014;46(6):567–72.
36. Wang X, Lu X, Wang J, Wang D, Yin Z, Fan W, et al. Mining and analysis of SNP in response to salinity stress in upland cotton (Gossypium hirsutum L.). PLoS ONE. 2016;11(6):e0158142.
37. Batley J, Edwards D. SNP applications in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN, editors. Association mapping in plants. New York: Springer; 2007. pp. 95–102.
38. Hamblin MT, Warburton ML, Buckler ES. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE. 2007;2(12):e1367.
39. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136(2):245–57.
40. Hinze LL, Hulse-Kemp AM, Wilson IW, Zhu Q-H, Llewellyn DJ, Taylor JM, et al. Diversity analysis of cotton (*Gossypium hirsutum* L.) germplasm using the CottonSNP63K Array. BMC Plant Biol. 2017;17:1–20.
41. Baloch FS, Alsaleh A, Shahid MQ, Çiftçi V, de E. SáenzMiera L, Aasim M, et al. A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. PLoS ONE. 2017;12(1):e0167821.
42. Hinze LL, Dever JK, Percy RG. Molecular variation among and within improved cultivars in the US cotton germplasm collection. Crop Sci. 2012;52(1):222–30.
43. Farahani S, Maleki M, Mehrabi R, Kanouni H, Scheben A, Batley J, et al. Whole genome diversity, population structure, and linkage disequilibrium analysis of chickpea (*Cicer arietinum* L.) genotypes using genome-wide DArTseq-based SNP markers. Genes. 2019;10(9):676.
44. Ashrafi H, Hulse-Kemp AM, Wang F, Yang SS, Guan X, Jones DC, et al. A long-read transcriptome assembly of cotton (*Gossypium hirsutum L.*) and intraspecific single nucleotide polymorphism discovery. Plant Genome. 2015;8(2). https://doi.org/10.3835/plantgenome2014.10.0068.
45. Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet. 2012;124:1201–14.

46. Mehrabi AA, Pour-Aboughadareh A, Mansouri S, Hosseini A. Genome-wide association analysis of root system architecture features and agronomic traits in durum wheat. Mol Breed. 2020;40(6):55.
47. Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C, et al. Diversity arrays technology: a generic genome profiling technology on open platforms. In: Pompanon F, Bonin A, editors. Data Production and Analysis in Population Genomics. Methods in Molecular Biology, vol 888. Totowa: Humana Press; 2012. pp. 67–89.
48. Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, et al. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet. 2006;113:1409–20.
49. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:1–15.
50. Goudet J. Hierfstat, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes. 2005;5(1):184–6.
51. Liu K: PowerMarker: new genetic data analysis software. Version 3.0. http://www.powermarkernet 2005.
52. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3):314.
53. Perrier X, Jacquemoud-Collet JP. DARwin software version 6.0. Montpellier: CIRAD; 2006.
54. Paradis E. Pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 2010;26(3):419–20.
55. Pritchard JK, Wen X, Falush D. Documentation for structure software: Version 2.3. Chicago: University of Chicago; 2010. Available from: https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf.
56. Würschum T, Maurer HP, Kraft T, Janssen G, Nilsson C, Reif JC. Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet. 2011;123(7):1121–31.
57. Li YL, Liu JX. StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. Mol Ecol Resour. 2018;18(1):176–7.
58. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–20.
59. Seyoum M, Du XM, He SP, Jia YH, Pan Z, Sun JL. Analysis of genetic diversity and population structure in upland cotton (*Gossypium hirsutum* L.) germplasm using simple sequence repeats. J Genet. 2018;97(2):513–22.
60. Jombart T, Solymos P, Cori A, Calboli F, Kamvar Z, Lustrik R: Package 'adegenet'. Exploratory Analysis of Genetic and Genomic Data. In.; 2021. https://cran.r-project.org/package=adegenet.
61. Sodedji FAK, Agbahoungba S, Agoyi EE, Kafoutchoni KM, Kim HY, Nguetta SPA, Assogbadjo AE. DArT-seq based SNP analysis of diversity, population structure and linkage disequilibrium among 274 cowpea (Vigna unguiculata L. Walp.) accessions. PLoS One. 2020;15(3):e0229167.
62. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.
63. Gowda SA, Bourland FM, Kaur B, Jones DC, Kuraparthy V. Genetic diversity and population structure analyses and genome-wide association studies of photoperiod sensitivity in cotton (Gossypium hirsutum L.). Theor Appl Genet. 2023;136(11):230.
64. Morales-Aranibar L, Rivera MYN, Gonzales HHS, Aranibar CGM, Gutiérrez NL, Gomez FG, et al. Comparative analysis of key fiber characteristics in white Pima cotton (Gossypium barbadense L.): native accessions from the Peruvian Amazon. Agrosyst Geosci Environ. 2024;7(2):e20517.
65. Chen X, Feng S, Qi Z, Sima MW, Zeng F, Li L, et al. Optimizing irrigation strategies to improve water use efficiency of cotton in Northwest China using RZWQM2. Agriculture. 2022;12(3):383.
66. Gapare W, Conaty W, Zhu Q-H, Liu S, Stiller W, Llewellyn D, et al. Genome-wide association study of yield components and fibre quality traits in a cotton germplasm diversity panel. Euphytica. 2017;213:1–22.
67. Handi SS, Katageri IS, Adiger S, Jadhav MP, Lekkala SP, Reddy Lachagari VB. Association mapping for seed cotton yield, yield components and fibre quality traits in upland cotton (*Gossypium hirsutum* L.) genotypes. Plant Breed. 2017;136(6):958–68.
68. Billings GT, Jones MA, Rustgi S, Hulse-Kemp AM, Campbell BT. Population structure and genetic diversity of the Pee Dee cotton breeding program. G3 Genes|Genomes|Genetics. 2021;11(7):jkab145.
69. Tyagi P, Gore MA, Bowman DT, Campbell BT, Udall JA, Kuraparthy V. Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet. 2014;127:283–95.
70. Kaur B, Tyagi P, Kuraparthy V. Genetic diversity and population structure in the landrace accessions of *Gossypium hirsutum*. Crop Sci. 2017;57(5):2457–70.
71. Boakyewaa Adu G, Badu-Apraku B, Akromah R, Garcia-Oliveira AL, Awuku FJ, Gedil M. Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. PLoS ONE. 2019;14(4):e0214810.
72. Larsen B, Gardner K, Pedersen C, Ørgaard M, Migicovsky Z, Myles S, et al. Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. PLoS ONE. 2018;13(8):e0201889.
73. Ayyaz M, Chang Z, Ding S, Han P, Xu L, Abudukeyoumu A, et al. QTL mapping associated with Verticillium wilt resistance in cotton based on MAGIC population. J Cotton Res. 2025;8(1):1–15.
74. Mohammed J, Thyssen GN, Hinze L, Zhang J, Zeng L, Fang DD. A GWAS identified loci and candidate genes associated with fiber quality traits in a new cotton MAGIC population. Theor Appl Genet. 2025;138(1):10.
75. Wang H, Cai X, Umer MJ, Xu Y, Hou Y, Zheng J, et al. Genetic analysis of cotton fiber traits in *Gossypium* hybrid lines. Physiol Plant. 2024;176(4):e14442.
76. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature. 1978;274(5673):775–80.
77. Song C, Li W, Wang Z, Pei X, Liu Y, Ren Z, et al. Genome resequencing reveals genetic variation between the parents of an elite hybrid upland cotton. Agronomy. 2018;8(12):305.
78. Lorenc MT, Hayashi S, Stiller J, Lee H, Manoli S, Ruperao P, et al. Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. Biology. 2012;1(2):370–82.
79. Joshi B, Singh S, Tiwari GJ, Kumar H, Boopathi NM, Jaiswal S, et al. Genome-wide association study of fiber yield-related traits uncovers the novel genomic regions and candidate genes in Indian upland cotton (*Gossypium hirsutum* L.). Front Plant Sci. 2023;14:1252746.
80. Ai X, Liang Y, Wang J, Zheng J, Gong Z, Guo J, et al. Genetic diversity and structure of elite cotton germplasm (*Gossypium hirsutum* L.) using genome-wide SNP data. Genetica. 2017;145:409–16.
81. Mammadov JA, Chen W, Ren R, Pai R, Marchione W, Yalçin F, et al. Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays* L.] genome for use in marker-assisted breeding. Theor Appl Genet. 2010;121:577–88.
82. Blair MW, Cortés AJ, Penmetsa RV, Farmer A, Carrasquilla-Garcia N, Cook DR. A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. 2013;126:535–48.
83. Asmamaw M, Keneni G, Tesfaye K. Genetic diversity of Ethiopian durum wheat (*Triticum durum* Desf.) landrace collections as reveled by SSR markers. Adv Crop Sci Technol. 2019;7(1):413.
84. Pickup M, Young A. Population size, self-incompatibility and genetic rescue in diploid and tetraploid races of *Rutidosis leptorrhynchoides* (Asteraceae). Heredity. 2008;100(3):268–74.
85. Chang L-Y, Toghiani S, Hay EH, Aggrey SE, Rekaya R. A weighted genomic relationship matrix based on fixation index (FST) prioritized SNPs for genomic selection. Genes. 2019;10(11):922.
86. Zavinon F, Adoukonou-Sagbadja H, Keilwagen J, Lehnert H, Ordon F, Perovic D. Genetic diversity and population structure in Beninese pigeon pea [*Cajanus cajan* (L.) Huth] landraces collection revealed by SSR and genome wide SNP markers. Genet Resour Crop Evol. 2020;67:191–208.
87. Lukanda MM, Dramadri IO, Adjei EA, Arusei P, Gitonga HW, Wasswa P, et al. Genetic diversity and population structure of Ugandan Soybean (*Glycine max* L.) germplasm based on DArTseq. Plant Mol Biol Rep. 2023;41(3):417–26.
88. Chander S, Garcia-Oliveira AL, Gedil M, Shah T, Otusanya GO, Asiedu R, et al. Genetic diversity and population structure of soybean lines adapted to sub-Saharan Africa using single nucleotide polymorphism (SNP) markers. Agronomy. 2021;11(3):604.
89. Wang J, Zhang Z, Gong Z, Liang Y, Ai X, Sang Z, et al. Analysis of the genetic structure and diversity of upland cotton groups in different planting areas based on SNP markers. Gene. 2022;809:146042.

90. Gowda SA, Katageri IS, Patil RS, Kumar PS, Tiwari GJ, Jena SN, et al. 63 K and 50 K SNP array based high-density genetic mapping and QTL analysis for productivity and fiber quality traits in cotton. Euphytica. 2022;218(7):93.
91. Flint-Garcia SA, Thornsberry JM, Buckler IVES. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003;54(1):357–74.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.