Type of the Paper (Article, Review, Communication, etc.)

Genome –Wide Allelic Variation and Population Structure Analysis of Cotton Genotypes using SSR Markers

Abdul Rehman[1*] and Muhammad Iqbal[1]

1 Department of Plant Breeding and Genetics, The Islamia University of Bahawalpur, Pakistan

*Correspondence: abdulrehmantoor@gmail.com muhammadiqbal999@hotmail.com

Abstract: The limited genetic base emphasizing the necessity to study existing cotton germplasm resources. Therefore, this study was conducted to assess genetic diversity of different cotton genotypes. A total of 105 cotton genotypes were analyzed using 208 polymorphic SSR markers; distributed on both A and D genomes. The total alleles in all genomes are 1382, out of which 665 and 717 alleles were recorded in A and D genomes respectively. In A genome, the maximum PIC value (0.89) was observed for markers M92, M82 and M76 located at the chromosomes 12A, 11A and 10A respectively. These markers showed the 0.84 gene diversity (GD) and having the 10 total number of alleles (TNA) per marker which indicating the 100% polymorphism. The highest PIC value 0.89 of D genome was detected for M191, M128 and M117 markers on 24D, 16D and 15D chromosomes having the 0.84 values of gene diversity with 100% polymorphism. Different analysis clustered the studied cotton genotypes into four subpopulations. Phylogenetic tree identified four major groups; Group 1, 2, 3 and 4 consisting of 12, 21, 39 and 69 genotypes, respectively. SSR markers appear to be useful and reliable assets for future genetic and genomic analyses of cotton.

Keywords: Diversity analysis; SSR markers; PIC, gene diversity

Citation: : Rehman, A.; Iqbal, M Genome – Wide Allelic Variations and Population Structure Analysis of Cotton Genotypes using SSR *Pakistan Journal of Biochemistry and Biotechnology*, **2022**, 3 (1), 91-111. https://doi.org/10.52700/pjbb.v3i1.109

Received: 07-02-2022
Accepted: 31-03-2022
Published: 22-04-2022

1. Introduction

Cotton seed is utilized as a source of vegetable oil as well as animal feed [1]. Diverse genetic resources are required for long-term crop improvement, allowing breeders to create unique gene combinations in future cotton cultivars. Commercial genotypes, as well as wild or alien species, may be included in these resources [2]. Previously, genetic diversity research in cotton was primarily based on morphological parameters. However, because of environmental interactions, uncertain genetic control of characteristics, and insufficient genome sampling for phenotypic characterization, morphological markers do not accurately explore genetic relationships among genotypes [3]. Available germplasm of cotton crop have several useful allelic variations that can be used to develop cultivars adapted to emerging environmental condition using traditional and genomic breeding methods

[4].

DNA-based markers are currently widely employed to define germplasm in a variety of crops, including cotton. DNA markers are useful not just for measuring diversity, but

also for fingerprinting genotypes because they are unaffected by the environment [5]. Genotype characterization through molecular markers (DNA Fingerprinting) is the rapid, easy, and most common method to discriminate, identify and characterize various cultivars to protect plant breeder's right and promote marker-assisted breeding [6]. This technique has been revolutionized since the past three decades to distinguish the DNA polymorphism, biological identification, and documentation of species. Through the PCR based approach, individual plant hybrids/varieties can be identified by acquiring a specific pattern of genetic profiles.

Several DNA marker systems have already been designed and used to measure genetic diversity in a variety of crop species, including cotton. Restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD) [7], amplified fragment length polymorphism (AFLP), inter simple sequence repeats (ISSR), and simple sequence repeats (SSRs) are all examples of these [8, 9]. SSRs are effective genetic markers for plant breeding programs because they are plentiful throughout plant genomes, simple to employ, multi-allelic, and polymorphic [9, 10]. SSRs are multi-allelic because the number of repeats of the SSR motif varies widely, resulting in polymorphisms based on the size of the fragments amplified by specific primers to the flanking regions. These markers' loci are also generally reproducible between species, particularly within a genus. SSRs are thus used in a variety of research, including genetic diversity, cultivar fingerprinting, molecular mapping, and marker-assisted selection in crop plants [7, 11].

Furthermore, SSRs exhibited moderate levels of genetic variation at the DNA level and identified three organized subpopulations, implying that these markers might be used for genome-wide association mapping studies for the identification and conservation of valuable alleles in upland cotton germplasm. [2]. Previously, cotton scientists [12] used SSR markers for DNA fingerprinting among cotton accessions however all cotton genotypes in their study were primitive and they only used 104 SSR markers which are unable to reveal genetic diversity in Pakistani cotton genotypes having narrow genetic base. Therefore, the current study was designed the genetic diversity and genome-wide variation of 105 cotton genotypes using 208 SSR markers. UPGMA Cluster analysis was conducted for estimation of genetic distance and to provide a reliable picture of a diverse grouping of genotypes for their effective utilization in future cotton breeding programs. In addition STRUCTURE analysis was used to an insight into different sets of allelic variations in studied cotton genotypes. Present study has provided a comprehensive genetic assessment of 105 cotton genotypes. In addition to this it is conferred that SSR markers are efficient tool for expressing genetic diversity and relatedness of cotton genotypes for drought tolerance.

2. Materials and Methods

2.1 Experiment Material

A representative collection of 105 cotton genotypes were procured from the department of Plant Breeding and Genetics, The Islamia University of Bahawalpur. The name, pedigree (if available) and origin of these 105 cotton genotypes are listed in Table 3. The total 10 genotypes (G1-G10) developed in the Department of Plant Breeding and Genetics, The Islamia University of Bahawalpur, Pakistan. Genotypes 11-55 widely cultivated varieties which were developed from different research Institutes of Pakistan (CCRI, CRI Multan, NIBGE and NIAB) and 40 exotic genotypes (G66-G105) were used in this experiment.

2.2. Plant Growing Condition and DNA Extraction

Seeds were sown in small plastic pots under greenhouse conditions where normal agronomical practices were followed for raising healthy cotton plants in 2019. After three weeks, leaves were collected for DNA extraction. Genomic DNA extracted from approximately ~ 0.2 g fresh leaf tissue using the modified cetyl-trimethyl ammonium bromide (CTAB) method [13]. The quality and concentration of extracted DNA was estimated by Nano-drop (ND1000, Thermo Scientific, USA), followed by gel electrophoresis.

2.3. Molecular Marker and PCR Profile

A total of 300 polymorphic SSR markers, which were distributed evenly over the 26 (13 from A genome and 13 from D genome) cotton chromosomes, were selected according to the information available on the AD-genome wide Reference Map

(http://www.cottondb.org/cgi-bin/cmap/view). The polymorphic 208 SSR markers were selected and analyzed which are distributed across the 26 chromosomes (AD genome) of 105 cotton genotypes with a mean of 8 markers per each chromosome. These 208 polymorphic SSRs comprise of 8 markers at each chromosome of the both A and D genomes of cotton genotypes for determination of genome-wide allelic variation and genetic diversity. Total reaction volume was 20µL for PCR amplifications, having the 2µL genomic DNA as template, 2µL 10X buffer containing 1.5mg MgCl2, 0.1µL Taq polymerase of 5 Units µL-1, 0.4µL 10mM dNTPs, and 2µL (forward and revere) of 2.5µM primer. PCR profile contained a total of 35 cycles at 94ºC for 45s, specific primer pairing at optimum annealing temperature for 45s, and 72ºC for 60s, and the extension at 72ºC for 10min. Amplified PCR products were examined by using 2% agarose gel electrophoresis (AGE) and visualized under Ultraviolet light.

2.4. Molecular Data Analysis

Visualized polymorphic bands were calculated in numeric format like, 1 for presence and 0 for absence and data further aligned for the genome wide allelic pattern and genetic diversity analysis. Total number of alleles per locus and allele frequency were determined using the statistical software GenAlEx version 6.5 [14] and UPGMA (Unweighted pair group method with arithmetic mean, or unweighted neighbor joining tree) was constructed using statistical software DARWIN version 6 [15] for the classification of population into sub population. POWER MARKER software version 3.23 [16] used for calculation of polymorphism information content (PIC) values and gene diversity, Bayesian clustering method was applied to identify clusters of genetically similar individuals using the statistical software STRUCTURE v.2.3 [17] . A burn-in length of 104 cycles (to minimize the effect of starting configuration), a simulation run of 106 cycles, and the admixture model option were applied in the STRUCTURE program. Web-based software package "Structure Harvester v0.6.93" was used [18] to derive the peak or optimal number of clusters ''K.'' which permits the visualization of the STRUCTURE results to understand the number of clusters based on ad-hoc techniques. We chose cluster values (K) ranging from 1 to 10 and six independent runs for each value to obtain consistent results.

3. Results

Genome-wide allelic variation

Among 300 SSR markers, 30 markers were not amplified whereas 32 were monomorphic and the remaining 208 were polymorphic in this experiment. We found only 69% of the markers revealing polymorphism among studied genotypes which is comparable to previous reports [3, 5]. A total 208 polymorphic SSR markers were selected for genomewide allelic variation, among which 104 were present at the A and D, genomes, respectively. From each genome 8 polymorphic SSR markers located on each chromosome as displayed in table 1. A total 1382 of alleles 208 polymorphic SSR markers were observed in all genomes. Out of these 665 and 717 alleles were recorded in A and D genomes of cotton respectively. The average number of allele per locus was 6.64 and 0.71 mean value of polymorphic information content (PIC) 0.71 and 0.67 gene diversity value was calculated among the 208 polymorphic SSR markers.

The total number of alleles (TNA) per marker ranged from 2 to 10 in A genome with the 6.39 average value while D genome showed 3 to 10 total number of alleles having the mean value of 6.89. Polymorphic information contents (PIC) range from 0.33 to 0.89 in A genome with the average value 0.70, while D genome had PIC values ranging from 0.37 to 0.89 with the 0.73 average value. Gene diversity (GD) ranged from 0.11 to 0.96 across the both A and D genomes with averaged values of 0.67and 0.68 respectively (Table 1).

The mean value of allele/locus was 6.644 having the standard error (SE) value is 0.247. The mean number of different alleles with a frequency >= 5% is 5.125 with SE value of 0.117. The average number of effective alleles was 4.842 having SE value of 0.140. Shannon's Information Index (I) value is 1.633 with the 0.032 SE values as shown in table 2. The mean of heterozygosity (He) value is 0.751 in allelic variation, while the mean of observed heterozygosity (Ho) and unbiased expected heterozygosity (uHe) was 0.673 and 0.775 respectively In addition the mean of fixation index (F) was 0.101 with SE value of 0.018 (Table 2). SSR markers with the minimum TNA (total number of alleles) displayed the minimum genetic variation as compared to those markers which perceived more TNA, thereby showing maximum genetic variation.

Further, among 104 polymorphic SSR markers in A genome, 43 markers showed 100% polymorphism in 105 studied cotton genotypes, while 11 markers showed 99% followed by 17 markers showed 98%, 6 markers had 97% and the remaining markers showed 96 to 84% polymorphism (table 1). In A genome the maximum PIC value, 0.89 was recorded for M92, M82 and M76 markers with the 10 total numbers of alleles exhibiting the 100% polymorphism. These markers (M92, M82 and M76) were located on the chromosomes 12A, 11A and 10A respectively with the estimation of 0.84 gene diversity. In A-genome, the lowest PIC value of 0.33 and 0.37 were detected for M43 and M35 markers; which were found on chromosome 6A and 5A with the 98.10% polymorphism, 0.46 and 0.38 gene diversity showing the 2 and 3 total numbers of alleles, respectively. In D Genome, among 104 polymorphic SSR markers, the 57 markers showed 100% polymorphism in 105 studied cotton genotypes followed by 17 markers showed 98% and 6 markers showed 96% polymorphism, while the remaining markers showed 94 to 84% polymorphism as displayed in Table 1. The highest 0.89 PIC value was in D genome was detected for M191,

M128 an M117 markers on 24D, 16D and 15D chromosome with 0.84 gene diversity and 100% polymorphism. The lowest PIC values 0.37 was identified for M155, M169, M175 and M189 markers which were located on chromosomes 20D, 22D and 24D respectively. These markers have the values of 0.38 gene diversity showing the 98.10% polymorphism with the 3 total numbers of alleles in D-genome. Among the all tested SSR markers, 208 were found polymorphic, which produced a total of 1382 loci.

Genetic Diversity

Availability of dense and evenly distributed markers is essential for marker-assisted-breeding in cotton. During the 105 cotton cultivars evaluation, 208 SSR markers were amplified out of 300 SSR markers. In addition Bayesian technique implemented in statistical software STRUCTURE to access the genetic structure of studied germplasm, and the outcomes showed the highest (peak) number of K=4 demonstrating the germplasm distributed into 4 sub-population (Fig 1). Different colored in fig 2 exhibits the distinct group and overall germplasm allocated into four sub-groups.

Molecular UPGMA cluster DARWIN tree analyses and STRUCTURE Bayesian results exhibited that genotypes from department of PBG-IUB contained genetic variations and were not present in the similar cluster, showing that these genotypes derived from diverse forefathers. Additionally, evaluation of each group exposed that genotypes G-1 to G-10 and G-27 to G-28 located in the similar cluster and considered as group first; while the G11 to G-26 and G-29 to G-33 genotypes were presented in the 2nd cluster. Cluster (group) number 3 had G-34 to G-72 genotypes; the cluster fourth contained the G-73 to G-105 genotypes. Similar results obtained from the STRUCTURE Bayesian and DARWIN tree analyses using 302 polymorphic SSR markers in 105 cotton genotypes. Bayesian STRUCTURE analysis showed that germplasm was clustered into four subgroups. Molecular DARWIN cluster analysis (Figure 4) and STRUCTURE Bayesian analysis exhibited that 105 cotton genotypes are genetically distant. Revealing genetic diversity among the tested cultivars of cotton

In this experiment distance among clusters identified the dissimilarities within 105 cotton varieties and all clusters are genetically diverse to each other. The maximum

genetic distance between clusters exhibited that genetically they are different. Mainly, this is an indication of genetic dissimilarity between the clusters and genetic similarity within each cluster.

Figure 1. This result was achieved for 105 cotton genotypes using 208 polymorphic SSR markers from Structure Harvester analysis. It's based on the second order derivation on the variance of the maximum likelihood estimation. Delta K shows only the uppermost clustering level and number of subpopulations in main population

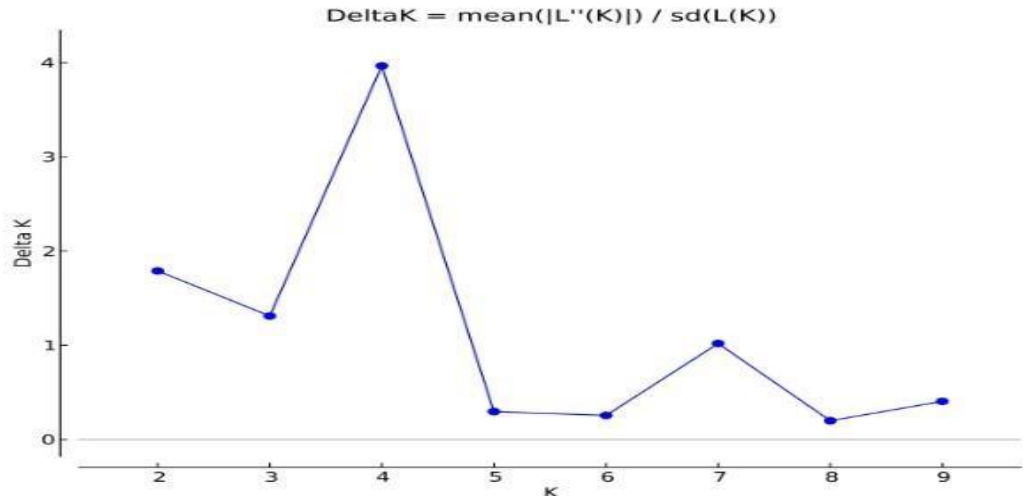


Figure 2 Population structure of 105 cotton genotypes based on Bayesian approach analyzed with 208 polymorphic SSR markers observing 4 clusters, K=4. In this Figure, four different colors represented the different groups.

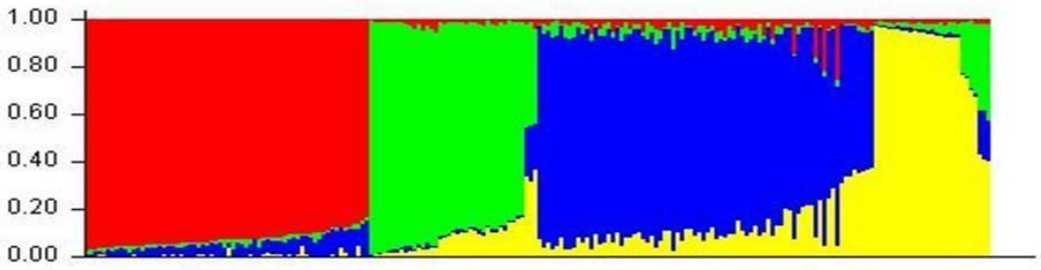


Figure 3a Genome-Wide Allelic Patterns using 208 polymorphic SSR markers across 105 cotton genotypes

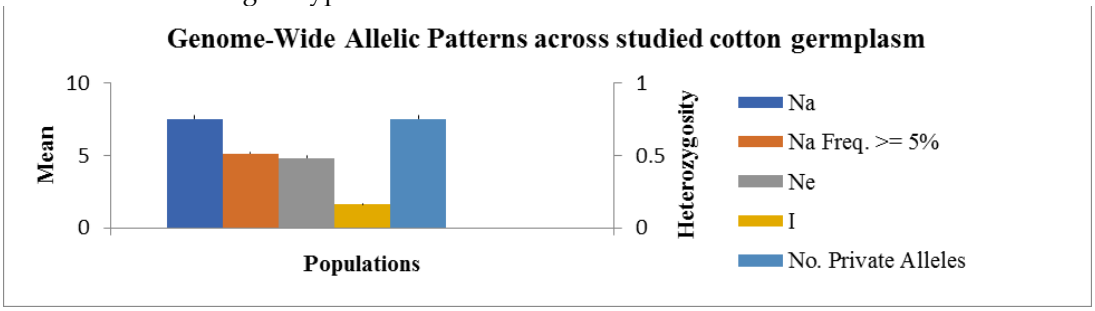


Table 1. List of polymorphic 208 SSR markers used to evaluate 105 cotton genotypes

Marker	genome	CHL	PIC	GD	AF	PP	TNA
CGR6111	A	1	0.88	0.47	97.00	92.38	10
BNL2448	A	1	0.76	0.46	105.00	100.00	9
BNL3441	A	1	0.72	0.70	99.00	94.29	5
BNL3442	A	1	0.67	0.65	99.00	94.29	5
BNL3261	A	1	0.79	0.77	105.00	100.00	7
CGR5111	A	1	0.54	0.15	105.00	100.00	3
BNL3501	A	1	0.37	0.17	105.00	100.00	2
BNL2920	A	1	0.77	0.59	104.00	99.05	7
BNL1379	A	2	0.79	0.54	104.00	99.05	7
BNL1606	A	2	0.61	0.69	105.00	100.00	4
BNL2486	A	2	0.58	0.75	105.00	100.00	5
BNL2496B	A	2	0.58	0.72	103.00	98.10	6
BNL2599	A	2	0.78	0.77	105.00	100.00	7
BNL1693	A	2	0.54	0.64	104.00	99.05	6
BNL2921	A	2	0.80	0.92	105.00	100.00	8
BNL0341	A	2	0.81	0.94	96.00	91.43	7
BNL2837	A	3	0.66	0.97	104.00	99.05	9
BNL3145	A	3	0.84	0.58	104.00	99.05	9
CGR5158	A	3	0.74	0.88	103.00	98.10	9
BNL0786	A	3	0.75	0.55	105.00	100.00	7
CGR5056	A	3	0.80	0.99	105.00	100.00	8
BNL3400	A	3	0.67	0.61	92.00	87.62	4

BNL3257	A	3	0.79	0.98	104.00	99.05	8
BNL0852	A	3	0.53	0.63	100.00	95.24	5
BNL1423	A	4	0.68	0.58	102.00	97.14	6
BNL4003	A	4	0.62	0.31	105.00	100.00	5
BNL0663	A	4	0.57	0.85	105.00	100.00	7
BNL1528	A	4	0.44	0.36	102.00	97.14	3
BNL3030	A	4	0.65	0.88	104.00	99.05	5
BNL2993	A	4	0.72	0.22	104.00	99.05	5
BNL1044	A	4	0.78	0.58	88.00	83.81	8
CGR5098	A	4	0.63	0.96	105.00	100.00	5
BNL3281	A	5	0.73	0.38	103.00	98.10	7
BNL1688	A	5	0.74	0.96	105.00	100.00	5
BNL3792	A	5	0.37	0.38	103.00	98.10	3
BNL2992	A	5	0.53	0.52	103.00	98.10	3
BNL3800	A	5	0.86	0.78	104.00	99.05	8
CER0099	A	5	0.75	0.68	102.00	97.14	8
BNL1017	A	5	0.39	0.36	105.00	100.00	3
BNL0519	A	5	0.81	0.89	100.00	95.24	8
BNL0244	A	6	0.86	0.91	105.00	100.00	10
BNL1080	A	6	0.69	0.96	105.00	100.00	4
BNL4034	A	6	0.33	0.46	103.00	98.10	2
BNL3033	A	6	0.58	0.95	104.00	99.05	4

BNL3463	A	6	0.68	0.78	104.00	99.05	4
BNL0226	A	6	0.37	0.06	103.00	98.10	2
BNL0686	A	6	0.66	0.95	105.00	100.00	5
BNL1707	A	6	0.77	0.90	102.00	97.14	9
BNL3874	A	7	0.40	0.47	103.00	98.10	4
CGR5110	A	7	0.42	0.46	98.00	93.33	3
CGR5135	A	7	0.78	0.72	102.00	97.14	10
BNL1404	A	7	0.76	0.83	103.00	98.10	7
C2-00036	A	7	0.64	0.84	103.00	98.10	4
C2-00114	A	7	0.75	0.87	101.00	96.19	8
BNL1122	A	7	0.59	0.78	102.00	97.14	3
BNL1395	A	7	0.74	0.88	98.00	93.33	5
BNL3171	A	8	0.53	0.11	105.00	100.00	3
BNL1681	A	8	0.71	0.84	99.00	94.29	5
BNL2578	A	8	0.46	0.47	101.00	96.19	3
BNL2768	A	8	0.74	0.93	105.00	100.00	6
BNL3835	A	8	0.81	0.95	105.00	100.00	9
C2-00120	A	8	0.68	0.96	105.00	100.00	6
CGR5015	A	8	0.81	0.80	103.00	98.10	8
BNL4108	A	8	0.87	0.80	105.00	100.00	8
BNL0448	A	9	0.79	0.74	103.00	98.10	8
BNL3241	A	9	0.45	0.42	101.00	96.19	3

BNL3569	A	9	0.76	0.96	105.00	100.00	8
BNL1163	A	9	0.84	0.66	105.00	100.00	9
BNL3992	A	9	0.72	0.77	105.00	100.00	7
BNL1438	A	9	0.81	0.51	90.00	85.71	7
BNL3594	A	9	0.82	0.69	105.00	100.00	8
BNL2589	A	9	0.88	0.47	97.00	92.38	10
CER0086	A	10	0.76	0.46	105.00	100.00	9
BNL2449	A	10	0.72	0.70	99.00	94.29	5
BNL1513	A	10	0.67	0.65	99.00	94.29	5
BNL1064	A	10	0.89	0.84	105.00	100.00	10
BNL1440	A	10	0.76	0.96	105.00	100.00	8
BNL1169	A	10	0.84	0.66	105.00	100.00	9
BNL3472	A	10	0.72	0.77	105.00	100.00	7
BNL1408	A	10	0.81	0.51	90.00	85.71	7
CER0153	A	11	0.82	0.69	105.00	100.00	8
BNL0836	A	11	0.89	0.84	105.00	100.00	10
BNL3144	A	11	0.87	0.80	105.00	100.00	8
BNL3592	A	11	0.79	0.74	103.00	98.10	8
BNL2650	A	11	0.45	0.42	101.00	96.19	3
BNL3649	A	11	0.78	0.58	88.00	83.81	8
BNL1231	A	11	0.63	0.96	105.00	100.00	5
BNL3279	A	11	0.73	0.38	103.00	98.10	7

BNL2709	A	12	0.74	0.96	105.00	100.00	5
BNL0850	A	12	0.37	0.38	103.00	98.10	3
BNL3423	A	12	0.53	0.52	103.00	98.10	3
BNL3537	A	12	0.89	0.84	105.00	100.00	10
BNL1673	A	12	0.76	0.96	105.00	100.00	8
C2-00024	A	12	0.84	0.66	105.00	100.00	9
CER0144	A	12	0.72	0.77	105.00	100.00	7
CGR5152	A	12	0.81	0.51	90.00	85.71	7
BNL0387	A	13	0.82	0.69	105.00	100.00	8
BNL1040	A	13	0.88	0.47	97.00	92.38	10
BNL1495	A	13	0.76	0.46	105.00	100.00	9
BNL0645	A	13	0.72	0.70	99.00	94.29	5
BNL2571	A	13	0.67	0.65	99.00	94.29	5
BNL2906	A	13	0.79	0.74	103.00	98.10	8
BNL4029	A	13	0.45	0.42	101.00	96.19	3
CER0122	A	13	0.88	0.47	97.00	92.38	10
BNL3085	D	14	0.76	0.46	105.00	100.00	9
BNL3090	D	14	0.72	0.70	99.00	94.29	5
BNL3345	D	14	0.67	0.65	99.00	94.29	5
BNL3510	D	14	0.76	0.96	105.00	100.00	8
BNL4082	D	14	0.84	0.66	105.00	100.00	9
C2-00108	D	14	0.72	0.77	105.00	100.00	7

CER0121	D	14	0.81	0.51	90.00	85.71	7
CGR5106	D	14	0.45	0.42	101.00	96.19	3
BNL1350	D	15	0.88	0.47	97.00	92.38	10
BNL3902	D	15	0.76	0.46	105.00	100.00	9
BNL4080	D	15	0.72	0.70	99.00	94.29	5
BNL1454	D	15	0.67	0.65	99.00	94.29	5
BNL0830	D	15	0.89	0.84	105.00	100.00	10
BNL2827	D	15	0.87	0.80	105.00	100.00	9
C2-00135	D	15	0.79	0.74	103.00	98.10	8
CGR5022	D	15	0.45	0.42	101.00	96.19	3
BNL2742	D	16	0.76	0.96	105.00	100.00	8
BNL3590	D	16	0.84	0.66	105.00	100.00	9
CGR5165	D	16	0.72	0.77	105.00	100.00	7
BNL3474	D	16	0.81	0.51	90.00	85.71	7
BNL3638	D	16	0.82	0.69	105.00	100.00	8
BNL3408	D	16	0.76	0.96	105.00	100.00	8
BNL3627	D	16	0.84	0.66	105.00	100.00	9
BNL3255	D	16	0.89	0.84	105.00	100.00	10
CGR5145	D	17	0.87	0.80	105.00	100.00	9
CER0152	D	17	0.79	0.74	103.00	98.10	8
CGR5120	D	17	0.45	0.42	101.00	96.19	3
BNL0252	D	17	0.76	0.96	105.00	100.00	8

BNL3103	D	17	0.84	0.66	105.00	100.00	9
BNL0500	D	17	0.72	0.77	105.00	100.00	7
BNL0269	D	17	0.81	0.51	90.00	85.71	7
BNL1664	D	17	0.82	0.69	105.00	100.00	8
BNL3502	D	18	0.88	0.47	97.00	92.38	10
BNL3932	D	18	0.76	0.46	105.00	100.00	9
BNL3449	D	18	0.72	0.70	99.00	94.29	5
BNL3424	D	18	0.67	0.65	99.00	94.29	5
BNL2651	D	18	0.76	0.96	105.00	100.00	8
BNL3034	D	18	0.84	0.66	105.00	100.00	9
BNL0666	D	18	0.87	0.80	105.00	100.00	9
BNL3259	D	18	0.79	0.74	103.00	98.10	8
BNL2590	D	19	0.45	0.42	101.00	96.19	3
BNL1317	D	19	0.76	0.96	105.00	100.00	8
BNL3140	D	19	0.84	0.66	105.00	100.00	9
BNL1672	D	19	0.72	0.77	105.00	100.00	7
BNL3031	D	19	0.81	0.51	90.00	85.71	7
BNL1878	D	19	0.82	0.69	105.00	100.00	8
BNL3410	D	19	0.78	0.58	88.00	83.81	8
BNL1053	D	19	0.63	0.96	105.00	100.00	5
BNL3418	D	20	0.73	0.38	103.00	98.10	7
BNL0193	D	20	0.74	0.96	105.00	100.00	5

BNL2805	D	20	0.37	0.38	103.00	98.10	3
BNL3976	D	20	0.53	0.52	103.00	98.10	3
BNL1705	D	20	0.87	0.80	105.00	100.00	8
BNL2589	D	20	0.79	0.74	103.00	98.10	8
CGR5167	D	20	0.45	0.42	101.00	96.19	3
BNL4094	D	20	0.76	0.96	105.00	100.00	8
BNL2495	D	21	0.84	0.66	105.00	100.00	9
BNL1115	D	21	0.72	0.77	105.00	100.00	7
BNL0116	D	21	0.81	0.51	90.00	85.71	7
BNL1441	D	21	0.82	0.69	105.00	100.00	8
BNL1669	D	21	0.78	0.58	88.00	83.81	8
BNL2557	D	21	0.63	0.96	105.00	100.00	5
BNL1551	D	21	0.73	0.38	103.00	98.10	7
BNL1402	D	21	0.74	0.96	105.00	100.00	5
BNL3895	D	22	0.37	0.38	103.00	98.10	3
BNL2786	D	22	0.53	0.52	103.00	98.10	3
BNL4069	D	22	0.78	0.58	88.00	83.81	8
BNL2440	D	22	0.63	0.96	105.00	100.00	5
CGR5025	D	22	0.73	0.38	103.00	98.10	7
BNL2749	D	22	0.74	0.96	105.00	100.00	5
BNL3347	D	22	0.37	0.38	103.00	98.10	3
BNL3426	D	22	0.53	0.52	103.00	98.10	3

BNL1417	D	23	0.87	0.80	105.00	100.00	8
BNL1035	D	23	0.79	0.74	103.00	98.10	8
BNL0272	D	23	0.45	0.42	101.00	96.19	3
BNL3098	D	23	0.76	0.96	105.00	100.00	8
CGR5124	D	23	0.84	0.66	105.00	100.00	9
BNL3655	D	23	0.72	0.77	105.00	100.00	7
BNL3806	D	23	0.81	0.51	90.00	85.71	7
BNL3937	D	23	0.82	0.69	105.00	100.00	8
BNL0169	D	24	0.78	0.58	88.00	83.81	8
BNL0390	D	24	0.63	0.96	105.00	100.00	5
BNL0511	D	24	0.73	0.38	103.00	98.10	7
BNL0946	D	24	0.74	0.96	105.00	100.00	5
BNL3660	D	24	0.37	0.38	103.00	98.10	3
BNL3071	D	24	0.53	0.52	103.00	98.10	3
BNL3977	D	24	0.89	0.84	105.00	100.00	10
C2-00133	D	24	0.76	0.96	105.00	100.00	8
BNL0358	D	25	0.84	0.66	105.00	100.00	9
BNL0530	D	25	0.72	0.77	105.00	100.00	7
BNL2527	D	25	0.81	0.51	90.00	85.71	7
BNL2572	D	25	0.82	0.69	105.00	100.00	8
BNL2609	D	25	0.88	0.47	97.00	92.38	10
C2-00076	D	25	0.76	0.46	105.00	100.00	9

C2-00088	D	25	0.72	0.70	99.00	94.29	5
CER0098	D	25	0.67	0.65	99.00	94.29	5
BNL0580	D	26	0.84	0.66	105.00	100.00	9
BNL0569	D	26	0.72	0.77	105.00	100.00	7
BNL2667	D	26	0.81	0.51	90.00	85.71	7
BNL3280	D	26	0.82	0.69	105.00	100.00	8
BNL3473	D	26	0.76	0.96	105.00	100.00	8
BNL3558	D	26	0.78	0.58	88.00	83.81	8
BNL3482	D	26	0.63	0.96	105.00	100.00	5
BNL3479	D	26	0.74	0.96	105.00	100.00	5

TNA= Total Number of Alleles, PIC= Polymorphic Information Contents, GD=Gene Diversity, AF= Allele Frequency, PP= Polymorphism Percentage, CHL= Chromosome Location and SRBP=Size range in Base Pairs.

Table 2. Genome-wide Mean Allelic variations across 208 polymorphic SSRs in 105 cotton genotypes

No. of Different Alleles	Na	Mean	6.644
		SE	0.247
No. of Effective Alleles = 1 / (Sum pi^2)	Ne	Mean	4.842
		SE	0.140
No. of Different Alleles with a Frequency >= 5%	Na Freq. >= 5%	Mean	5.125
		SE	0.117
Shannon's Information Index = -1* Sum (pi * Ln (pi))	I	Mean	1.633
		SE	0.032
Observed Heterozygosity = No. of Hets / N	Ho	Mean	0.673

		SE	0.014
Expected Heterozygosity = 1 - Sum pi^2	He	Mean	0.751
		SE	0.008
Unbiased Expected Heterozygosity = (2N / (2N-1)) * He	uHe	Mean	0.755
		SE	0.008
Fixation Index = (He - Ho) / He = 1 - (Ho / He)	F	Mean	0.101
		SE	0.018
Where pi is the frequency of the ith allele for the population & Sum pi^2 is the sum of the squared population allele frequencies.			

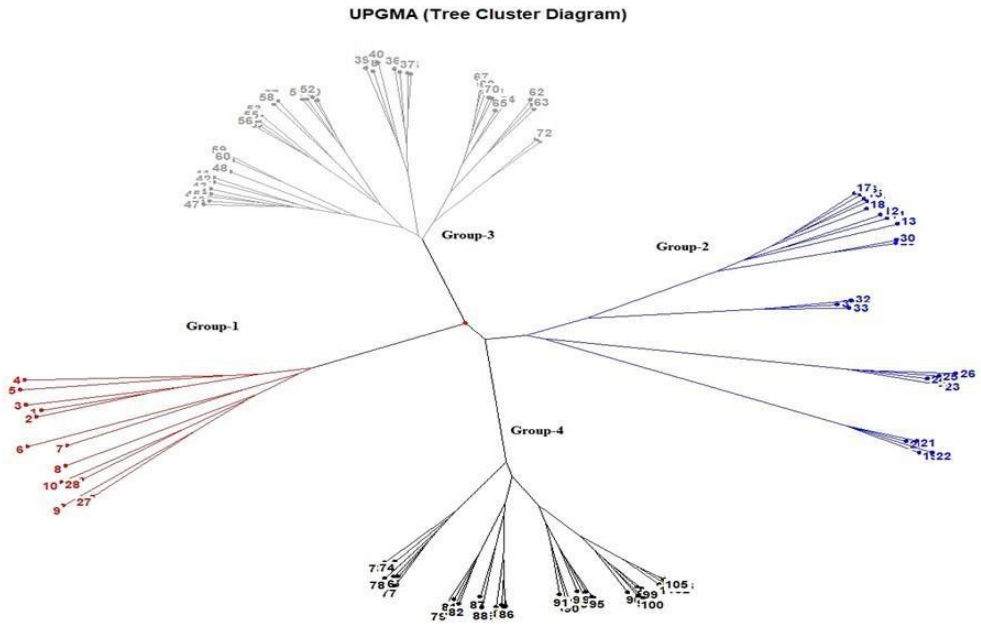
Table 3: List of 105 cotton genotypes used in current study

Sr. No.	Genotype	Sr. No.	Genotype	Sr. No.	Genotype
G1	IUB-09	G36	CYTO-124	G71	H-3
G2	IUB-111	G37	CYTO-177	G72	H-666
G3	IUB-13	G38	CYTO-179	G73	M-8
G4	IUB-222	G39	FH-113	G74	MAC-7
G5	IUB-2233	G40	FH-142	G75	ZQ1P1
G6	IUB-2248	G41	FH-207	G76	UGC-09
G7	IUB-52	G42	FH-326	G77	V-259
G8	IUB-63	G43	FH4243	G78	VGC-09
G9	IUB-75	G44	FH-900	G79	R-88
G10	MM-58	G45	IR-3701	G80	RH-647
G11	CIM-109	G46	IR-NIBGE-7	G81	MOON SOON INDIAN

G12	CIM-1100	G47	IR-NIBGE-8	G82	IRS-1
G13	CIM-240	G48	BH-180	G83	GP-3746-USA
G14	CIM-443	G49	BH-300	G84	GP-CHINA
G15	CIM-446	G50	CCRI-130	G85	CHINESE COTTON
G16	CIM-448	G51	CCRI-136	G86	ARS-703
G17	CIM-473	G52	CEMB-4444	G87	BG-1
G18	CIM-482	G53	CEMB-66	G88	BG-3
G19	CIM-496	G54	CEMB-777	G89	BG-391
G20	CIM-499	G55	CH-111	G90	A-1
G21	CIM-506	G56	CH-41	G91	PIMA-S3
G22	CIM-534	G57	CH-49	G92	VPE-7
G23	CIM-554	G58	CH-58	G93	IS-2-4
G24	CIM-573	G59	MNH-6070	G94	KARNAK-55
G25	CIM-598	G60	MNH-886	G95	KARNAK-65
G26	CIM-599	G61	MNH-986	G96	TADLA-16
G27	CIM-600	G62	NIAB-878	G97	GIZA-26
G28	CIM-602	G63	NIAB-KARISHMA	G98	C-132
G29	CIM-610	G64	NIBGE-115	G99	SKD-10/19
G30	CIM-616	G65	NIBGE-314	G100	GOA-18
G31	CIM-620	G66	BMWC-1534 USA	G101	FDH-170
G32	CIM-622	G67	BT-06	G102	C-50212
G33	CIM-632	G68	BT-081	G103	RA-3364

G34	CIM-70	G69	BT-121	G104	COMILA
G35	CIM-707	G70	BT-555	G105	COTICPHAC

Figure 3b The UPGAMA DARWIN tree displaying the distribution of the 105 cotton genotypes in four groups, and presenting the genetic similarities and dissimilarities within and between the groups



4. Discussion

The availability of a wide set of upland cotton genetic resources is critical for ensuring a steady supply of allelic variants for the various cotton breeding initiatives. Furthermore, a thorough assessment of genetic resources using molecular markers will aid in reducing redundancy and establishing a core germplasm collection that can be used to identify and conserve useful traits as well as diversify the narrow genetic base that is currently limiting cotton productivity. Among the different molecular markers, SSRs have been widely applied to perceive gene variability in cotton germplasm and to determine the genetic diversity.

In this study, the selected polymorphic 208 SSR markers were analyzed which are distributed across the 26 chromosomes (AD genome) of 105 cotton genotypes with a mean of 8 markers per each chromosome. In addition primers amplified a total of 1382 alleles to give an average of 6.64 alleles per microsatellite locus. Previously, a study was conducted [19] they use 179 SSR markers which showed polymorphism and gave 426 allele loci; and the cotton population was classified into seven subpopulations corresponding to pedigree origin, ecological and geographical distribution.

According to the cotton researchers, the number of alleles amplified per marker is determined by the markers utilized, the genotypes being studied, and the procedures used to resolve amplified products. [5, 11]. The mean allele number obtained in this study

was comparable to that obtained by other researchers. PIC values are widely used to determine the degree of SSR marker polymorphism. (Botstein et al. 1980; Gupta and Varshney 2000; Shete et al. 2000), which is mainly used to assess the amount of genetic variability in plant sciences. Because if, the PIC value is greater than 0.5 the marker is suggested to be of maximum diversity, while if the PIC value is less than 0.25 the marker is suggested to be of minimum diversity [19, 20]. According to Botstein et al. (1980), the SSR markers' in formativeness or discriminatory power can be classified as high for PIC >0.50, moderate for PIC values between 0.25 and 0.50, and low or slightly informative for PIC < 0.25. As a result, the 208 polymorphic SSR markers examined across 105 cotton genotypes were shown to be slightly (54.5%) and moderately (45.5%) informative, respectively, in the current study. As a result of the genetic bottleneck effect of domestication resulting from the a small founder population and high selection pressure, the current results from the PIC value analysis revealed that the genetic basis of upland cotton genotypes existed

In this experiment most of the markers having PIC values greater than 0.5 which indicate presence of more allelic diversity in studied germplasm. Our research results indicated that the selected markers had sufficient polymorphic information to reveal the genetic relationship between these upland cotton genotypes [21]. The mean values and SE values propose that there is great genetic diversity at SSR loci among studied germplasm. In current study of private alleles were found in this study (number of alleles unique to a single population), which is similar to the previous study [5, 22].The maximum mean values of gene diversity was identified in the D genome followed by A-genome, suggesting that the D-genome showed more variation and the existence of genetic diversity in cotton germplasm. Further the mean values of total number of alleles per marker in the B-genome exhibited the maximum genetic diversity as compared to the Dgenome followed by A-genome which showed the minimum genetic diversity.

The total numbers of allele (TNA) per locus ranged from 3-22 having 7.8 the mean values which was consistent with the previous reports. For instance, [23] reported 3.72 alleles per locus in genetic diversity study of 22 cotton genotypes using 30 SSR markers. Similarly,[24] reported 3.8 alleles per locus with 0.50 PIC value. In contrast [25] reported a high number of alleles (7.9) per locus. This is expected because they used landraces, and diversity in landraces have more genetic variations than the cultivated varieties. In current study, PIC values revealed a significant and positive association with the gene diversity (GD) and total numbers of allele (TNA) for SSR markers. In addition, these results (GD, TNA and PIC values) suggested the existence of genetic diversity which is higher in Agenome than D-genome. However, average PIC value (0.89) reported in our study is highest among all the previously published reports. High number of alleles in our study and high PIC value corresponds to a large set of SSR markers used in recent research [5]. Current study on genetic diversity and genome-wide allelic variation in cotton genotypes may be favorable for planning the future strategies on cotton genetic resources and cotton breeding scheme for development of novel cotton genotypes.

The evaluation of the population genetic structure is a prerequisite of genome-wide association studies, because false associations are usually caused by population structure. The exploration of accurate population structure can lead to more genetic similarity within each group, and higher genetic differences among groups which ultimately reduce the defectives in association analysis [19] The UPGMA cluster DARWIN tree and

STRUCTURE analysis distribute the 105 cotton genotypes into 4 subgroups or clusters. In this study, distances among clusters or groups clearly showed the variations among 105 cotton genotypes, and all subgroups also genetically diverse to one another. Further, there was a low genetic variation within each cluster or group, the genetic similarity among 105 genotypes. Development of novel cotton genotypes should be attaining the significance level of genetic diversity. Presence of less variation in 105 cotton genotypes which indicate the low genetic diversity, fearlessly, that the studied germplasm introduced from different sources or assumable mechanical mixing.

A phylogenetic tree constructed for the whole germplasm population identified four major clusters (groups), Group1, Group 2, Group 3 and Group 4 consisting of 12, 21, 39 and 69 genotypes, respectively. Among the four major clusters, the largest number of genotypes was categorized in group 4. This cluster also contained cotton genotypes collected from varied geographic origins with a diverse genetic makeup.

According to the provided pedigree record there are three groups of 105 cotton genotypes as shown in Supplementary 1. In first group genotypes G-1 to G-10 (10 genotypes) which developed in PBG-IUB, while group 2 had genotypes G-11 to G-65 (55 genotypes) from indigenous source and group three genotypes G-66 to G-105 (40 genotypes) from exotic source. But according to molecular analysis these genotypes are divided into four clusters or groups. Cotton genotypes developed in PBG-IUB presented in cluster 1, genotypes number G-27 and G-28 are also included in this cluster which demonstrated that these genotypes originated from the similar genetic makeup due to shared progenitors. Total 21 genotypes constituted in cluster 2, which exhibited the genetic similarity with each other. Total 72 cotton genotypes fell into cluster 3 and 4, out of these 33 genotypes from cluster 3 and 32 genotypes from cluster 4. These genotypes are created by a mixture of the diverse genetic constitutions which suggest the diverse pedigree of these genotypes. Partially, outcomes were usable conferring to the previously known pedigree record and origin. Genetic diversity evaluation could be helpful to identify the different genotypes for the advancement and improve the future cotton breeding scheme [26-28]. Genotypes with different genetic makeup can be selected to generate significant combinations of suitable traits to obtaining high yield cotton genotypes

5. Conclusions

Genetic diversity is the base of any genetic improvement breeding program. Therefore, it is necessary to investigate genetic diversity in cotton germplasm to broaden the genetic base in future cotton breeding. Gene diversity and PIC value ranged from 0.41 - 0.96 and 0.37 - 0.89 and 0.89 respectively indicating the significant level of genetic diversity among 105 cotton genotypes. The maximum PIC value 0.89 and lowest PIC value 0.37 among the all genomes precise to A and D genome respectively. Genetic diversity among genome ranked as genome D> genome A genome based on PIC values. The studied genotypes were divided into four genetically different subgroups based on STRUCTURE Bayesian and DARWIN analysis. The documentation of genetic diversity would be a good tool for choosing the desirable genotypes in cotton breeding programs. The present investigations also showed clear-cut identity of studied genotypes which will be of great utility for the protection of Plant Breeder's Rights.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Saeed, A. and E. Elçi, Microsatellite-based characterization of cotton genotypes for verticillium wilt and fiber quality traits. Turkish Journal of Biochemistry, 2018. 43(3): p. 277-288.
2. Seyoum, M., et al., Analysis of genetic diversity and population structure in upland cotton (Gossypium hirsutum L.) germplasm using simple sequence repeats. Journal of genetics, 2018. 97(2): p. 513-522.
3. Rakshit, A., et al., Evaluation of SSR markers for the assessment of genetic diversity and fingerprinting of Gossypium hirsutum accessions. Journal of plant biochemistry and biotechnology, 2010. 19(2): p. 153-160.
4. Hinze, L.L., et al., Diversity analysis of cotton (Gossypium hirsutum L.) germplasm using the CottonSNP63K Array. BMC Plant Biology, 2017. 17(1): p. 1-20.
5. Jamil, S., et al., DNA fingerprinting and genetic diversity assessment of GM cotton genotypes for protection of plant breeders rights. Int. J. Agric. Biol, 2021.
6. Kalia, R.K., et al., Microsatellite markers: an overview of the recent progress in plants. Euphytica, 2011. 177(3): p. 309-334.
7. Sapkal, D.R., et al., Genetic diversity analysis of maintainer and restorer accessions in upland cotton (Gossypium hirsutum L.). Journal of plant biochemistry and biotechnology, 2011. 20(1): p. 20-28.

8. Zhao, Y., et al., Genetic structure, linkage disequilibrium and association mapping of Verticillium wilt resistance in elite cotton (Gossypium hirsutum L.) germplasm population. PloS one, 2014. 9(1): p. e86308.
9. Badigannavar, A., G.O. Myers, and D.C. Jones, Molecular diversity revealed by AFLP markers in upland cotton genotypes. Journal of Crop Improvement, 2012. 26(5): p. 627-640.
10. Ahmed, H., et al., Genome wide allelic pattern and genetic diversity of spring wheat genotypes through SSR markers. Int. J. Agric. Biol, 2017. 19: p. 1559-1565.
11. Ahmed, H.G.M.-D., et al., Genome Wide Diversity in Bread Wheat Evaluated by SSR Markers. International Journal of Agriculture and Biology, 2020. 24(2): p. 263-272.
12. Ullah, I., et al., Genetic diversity analysis of Bt cotton genotypes in Pakistan using simple sequence repeat markers. Genetics and Molecular Research, 2012. 11(1): p. 597-605.
13. Saghai-Maroof, M.A., et al., Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Sciences, 1984. 81(24): p. 8014-8018.
14. Peakall, R. and P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. BIOINFORMATICS, 2012. 28(19): p. 2537-2539.
15. Perrier, X., A. Flori, and F. Bonnot, Data analysis methods in Genetic diversity of cultivated tropical plants. 43-76. Enfield, Science Publishers. Plymouth, UK, 2003.
16. Liu, K. and S.V. Muse, PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics, 2005. 21(9): p. 2128-2129.
17. Pritchard, J.K., et al., Association mapping in structured populations. The American Journal of Human Genetics, 2000. 67(1): p. 170-181.
18. Earl, D.A. and B.M. VonHoldt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources, 2012. 4(2): p. 359-361.
19. Nie, X., et al., Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. BMC genomics, 2016. 17(1): p. 1-16.
20. Wu, D., et al., Genetic relationship and diversity of upland cotton germplasms from different cotton producing countries using SSR markers. Cotton Science, 2011. 23(4): p. 291-299.
21. Qin, H., 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): p. e0118073.
22. Sajjad, M., S.H. Khan, and M. Shahzad, Patterns of allelic diversity in spring wheat populations by SSR-markers. Cytology and Genetics, 2018. 52(2): p. 155-160.
23. Javaid, A., et al., Assessment of allelic diversity among drought-resistant cotton genotypes using microsatellite markers. Genetics and Molecular Research, 2017. 16(2).
24. Gurmessa, D., Genetic diversity study of improved cotton (G. hirsutum L.) varieties in Ethiopia using simple sequence repeats markers. Journal of Biotechnology, 2019. 7(2): p. 6-14.
25. McCarty, J.C., et al., Genetic diversity of day-neutral converted landrace Gossypium hirsutum L. accessions. Euphytica, 2018. 214(10): p. 1-14.
26. Ahmed, H.G.M.-D., et al., Selection Criteria for Drought-Tolerant Bread Wheat Genotypes at Seedling Stage. Sustainability, 2019. 11(9): p. 2584.
27. Zhang, L., et al., Investigation of genetic diversity and population structure of common wheat cultivars in northern China using DArT markers. BMC genetics, 2011. 12(1): p. 42.
28. Lazzaro, M., et al., Unraveling diversity in wheat competitive ability traits can improve integrated weed management. Agronomy for sustainable development, 2019. 39(1): p. 6.