



Article

Assessment of Genetic Diversity, Population Structure, and Evolutionary Relationship of Uncharacterized Genes in a Novel Germplasm Collection of Diploid and Allotetraploid *Gossypium* Accessions Using EST and Genomic SSR Markers

Allah Ditta ^{1,2,†} , Zhongli Zhou ^{1,†}, Xiaoyan Cai ¹, Xingxing Wang ¹, Kiflom Weldu Okubazghi ^{1,3}, Muhammad Shehzad ¹, Yanchao Xu ¹ , Yuqing Hou ¹, Muhammad Sajid Iqbal ¹, Muhammad Kashif Riaz Khan ², Kunbo Wang ^{1,*} and Fang Liu ^{1,*}

¹ State Key Laboratory of Cotton Biology/Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China; adbotanist@yahoo.com (A.D.); zhonglizhou@163.com (Z.Z.); cxycri@163.com (X.C.); wx1991@126.com (X.W.); wediweldu81@yahoo.com (K.W.O.); mshehzad534@gmail.com (M.S.); xuyanchao2016@163.com (Y.X.); houyuqing18@163.com (Y.H.); sajidises@hotmail.com (M.S.I.)

² Nuclear Institute for Agriculture and Biology (NIAB), Jhang Road, Faisalabad 38000, Punjab, Pakistan; mkrkhan@gmail.com

³ Hamelmalo Agricultural College, P.O. Box 397, Keren, Eritrea

* Correspondence: wkbcric@163.com (K.W.); liufcri@163.com (F.L.)

† These authors contributed equally.

Received: 23 June 2018; Accepted: 13 August 2018; Published: 14 August 2018



Abstract: This study evaluated the genetic diversity and population structures in a novel cotton germplasm collection comprising 132 diploids, including *Glossypium klotzschianum* and allotetraploid cotton accessions, including *Glossypium barbadense*, *Glossypium darwinii*, *Glossypium tomentosum*, *Glossypium ekmanianum*, and *Glossypium stephensii*, from Santa Cruz, Isabella, San Cristobal, Hawaiian, Dominican Republic, and Wake Atoll islands. A total of 111 expressed sequence tag (EST) and genomic simple sequence repeat (gSSR) markers produced 382 polymorphic loci with an average of 3.44 polymorphic alleles per SSR marker. Polymorphism information content values counted 0.08 to 0.82 with an average of 0.56. Analysis of a genetic distance matrix revealed values of 0.003 to 0.53 with an average of 0.33 in the wild cotton collection. Phylogenetic analysis supported the subgroups identified by STRUCTURE and corresponds well with the results of principal coordinate analysis with a cumulative variation of 45.65%. A total of 123 unique alleles were observed among all accessions and 31 identified only in *G. ekmanianum*. Analysis of molecular variance revealed highly significant variation between the six groups identified by structure analysis with 49% of the total variation and 51% of the variation was due to diversity within the groups. The highest genetic differentiation among tetraploid populations was observed between accessions from the Hawaiian and Santa Cruz regions with a pairwise F_{ST} of 0.752 ($p < 0.001$). DUF819 containing an uncharacterized gene named yjcL linked to genomic markers has been found to be highly related to tryptophan-aspartic acid (W-D) repeats in a superfamily of genes. The RNA sequence expression data of the yjcL-linked gene Gh_A09G2500 was found to be upregulated under drought and salt stress conditions. The existence of genetic diversity, characterization of genes and variation in novel germplasm collection will be a landmark addition to the genetic study of cotton germplasm.

Keywords: novel accessions; PIC; PCR; EST-gSSRs; genes; genetic distance

1. Introduction

The leading natural fiber in the world is a product of cotton crops. Cotton is placed in the taxonomic order with the genus *Gossypium* and has broad phenotypic diversity, which includes more than 50 species [1–3]. There are now 7 tetraploid and 46 diploid cotton species after molecular confirmation and taxonomic designated two new tetraploid ones, i.e., *Gossypium ekmanianum* (AD6) and *Gossypium stephensii* (AD7) [3–6]. Among those, four are cultivated throughout the world: two of these species are diploids ($2n = 2x = 26$) and two are allotetraploids ($2n = 4x = 52$). Global cotton production is manifested from the two allotetraploid species *Gossypium hirsutum* and *Gossypium barbadense* [7–9].

Data on allotetraploid cotton evolution indicates that the seven tetraploid cottons evolved about 1.5 million years ago by hybridization of the Old world cotton *Gossypium herbaceum* (A₁ genome) and the New world cotton *Gossypium raimondii* (D₅ genome) as a consequence of subsequent diploidization and domestication [3,4,8,10–12]. *Gossypium hirsutum*, also called “Upland cotton”, represents 90% of global cotton fiber production [13], while *Gossypium barbadense* (also known as Pima) is valued for its extra-long staple fiber source, is domesticated in North-West South America, has its native origin in Egypt, and contributes around 8% of total world lint [9]. Wild *Gossypium darwinii* originated from Galapagos Island and, relative to *G. barbadense*, also has good fiber fineness characteristics and is a rich source of resistance to *fusarium* and *verticillium* wilts [14]. The D-genome *Gossypium klotzschianum*, having glabrous seed coverings, evolved through long-distance dispersals, is endemic to Galapagos Island, and is considered a New-World D-genome diploid along with *G. raimondii*. *Gossypium tomentosum* is drought-tolerant, native to a Hawaiian Island, and has a more diffuse population structure falling typically as scattered individuals and small populations on several islands. *G. tomentosum* (AD3), *G. darwinii* (AD5), *G. ekmanianum* (AD6), and *G. stephensii* (AD7) are wild and are not grown commercially [1,2,15,16]. Wendel and Percy analyzed 58 *G. darwinii* accessions from six islands using 17 isozyme markers and identified a high genetic diversity level within these accessions and relationships with *G. barbadense* and *G. hirsutum* genomes. This classic study suggested that *G. darwinii* and *G. barbadense* are separated and each has a distinct genome [17].

The genetic diversity of different plant species is an essential element for crop production in agriculture, including cotton. Genetic variation in the *Gossypium* species is widespread, covering large geographic and ecological niches. It is a vital source of conserved genetic diversity in situ in Mexico for cotton origin [18,19] and is preserved ex situ within worldwide cotton germplasm collections and materials of breeding programs. The productivity of cotton and future efforts to improve cotton depend to a large extent on the elucidation of genetic diversity in cotton genetic stocks and their effective utilization in cotton improvement programs [20].

The narrow genetic background of Upland cotton has become a major concern as low genetic diversity gives rise to stagnant yield and quality of breeding. The elite breeding programs cannot make robust inferences without using the unexploited standing genetic variation of archaic cultivars typically associated with wild accessions [14,21,22]. The characterization of genetic diversity between and within groups enables us to find heterozygous groups, understand population structures, and isolate a core set of lines for genetic analysis studies in cotton. A multitude of studies indicate the extensive usage of model-based structure analysis for investigating genetic diversity in cotton [22,23]. Genetic diversity estimates have been established using genotypic data and DNA-based molecular markers [24–28]. Molecular markers are more reliable since they can directly determine allelic diversity and give robust estimates of genetic distances.

The DNA-based markers used for determining genetic diversity in cotton include restriction fragment length polymorphisms (RFLPs) [29], random amplified polymorphic DNA (RAPD) [30–32], amplified fragment length polymorphisms (AFLPs) [33], simple sequence repeat (SSR) [9,34–36], expressed sequence tags (ESTs) [37], inter-simple sequence repeat (ISSR) [38,39], and single nucleotide polymorphisms (SNPs) [40]. Compared with other biomarkers, SSR has advantages that include more reproducibility, co-dominant inheritance, distribution throughout the genome, and its being highly transferable, informative, and reliable [41].

Although data from several studies implicates the marker-based estimation of genetic diversity in cotton, the majority of those remain bound to the number of accessions included or the number of markers used to describe genetic diversity [42]. Recently, an effort has been made by Kirungu et al. [43] to explore the important genes linked to SSR markers by constructing a genetic linkage map between *Gossypium davidsonii* and *G. klotzschianum*. Similarly, a study of gene diversity, their functionality, and especially the diagnosis of uncharacterized domains of proteins in developing the evolutionary relationship among cotton accessions will be fruitful for exploring the mystery of cotton evolution. Among all protein domains with a unique structure and functions, nearly more than 20% are currently described as “domains of unknown function” (DUFs). They are often overlooked as irrelevant as many of them are found in only a few genomes. Approximately 2700 DUFs exist in bacteria as compared to eukaryotes, which have only 1500. More than 800 DUFs have been found to be common in bacteria and eukaryotes, and about 300 of these are also present in archaea. Evolutionary conservation suggests that many of these DUFs are important in biology as they mostly represent single-domain proteins, clearly establishing the biological importance of DUFs [44].

The importance of prioritizing DUFs has been recognized in various experimental and/or computational characterization efforts [45–48]. We identified DUF819 (PF005684), which is not only highly conserved but also plays an important role against biotic and abiotic stress, among four sequenced cotton species by using the WDR (PF00400) superfamily as reference-genome-sequenced proteins. Genome-wide characterization of WD-repeats, also known as tryptophan-aspartic acid or the W-D superfamily, has only been conducted in *Arabidopsis* and *Cucumber* [49,50] till now. Therefore, a comprehensive study comprising a wide collection of germplasms, more efficient genotyping, and collective genomic platforms is required to measure the overall genetic diversity in diploid and allotetraploid cotton, which will help overcome the future challenges of the gene pool’s disastrous escape.

The objectives of this study were to explore the genetic diversity and evolutionary relationship among the domains of uncharacterized proteins in natural diploid and allotetraploid cotton germplasm resources and to analyze the population structures to maximize estimations about the accessions of cotton present in a wild nursery of China for their efficient utilization in cotton-breeding programs.

2. Results

2.1. SSR Marker Analysis

Among a total of 853 SSR primer pairs used for genotyping 132 accessions, 205 primer pairs were found to be polymorphic with a polymorphism rate of 24%. Accessions with more than 5% missing data were removed and 94 SSRs were dropped; the selected 111 SSR primer pairs can be scored confidently and read clearly on PCR products. Data for monomorphic loci were also excluded from the analysis. Data generated from the selected 111 SSR primer pairs was analyzed. Among 132 accessions, a total of 382 SSR alleles were detected as marker loci with an average of 3.44 alleles per SSR ranging from 2 to 8. All 382 SSR loci were found to be polymorphic. The average polymorphism information content (PIC) value for SSRs was 0.555 with a range of 0.078 to 0.821, and the major allele frequency was 0.738 ranging from 0.541 to 0.959 for the complete panel. Seventy-six (68.468%) SSR markers in total were found to be highly informative with a PIC value ≥ 0.50 , 29 (26.126%) were moderately informative with PICs value ≥ 0.25 and < 0.50 , and 6 (5.405%) were least informative with a PIC value < 0.25 . A summary of marker statistics for *G. hirsutum* accessions is listed in Supplementary Table S2.

2.2. Unique Alleles

Among the 382 alleles detected in the studied accessions, 123 alleles were found and were termed as unique alleles (Supplementary Table S3). A high percentage (17.51%) of unique alleles was observed in *G. ekmanianum* genotypes (Table 1). Twenty-five unique alleles were found in two accessions of *G. hirsutum*. Nineteen, 13, 18, 13, and 4 unique alleles were observed in *G. barbadense*, *G. tomentosum*,

G. darwinii, *G. klotzschianum*, and *G. stephensii*, respectively. *G. ekmanianum* had the highest number (31) of unique alleles, which were collected from the Dominican Republic, National Plant Germplasm System (NPGS) USA (Supplementary Table S3). These unique alleles are an important genetic resource for cotton and have never been studied before.

Table 1. Summary of unique (present in one accession) and rare alleles (present in <5% accessions) observed in a combined Panel of 132 accessions.

Panel	Total Alleles	Total Lines	Unique Alleles	Rare Alleles (Freq < 5%)
Combined Panel	382	132	123 (32.19%)	108 (28.27%)
<i>Gossypium barbadense</i>	258	20	19 (7.36%)	12 (4.65%)
<i>Gossypium darwinii</i>	309	59	18 (5.83%)	57 (18.44%)
<i>Gossypium tomentosum</i>	205	32	13 (6.34%)	15 (7.31%)
<i>Gossypium hirsutum</i>	143	2	25 (17.48%)	22 (15.38%)
<i>Gossypium ekmanianum</i>	177	10	31 (17.51%)	2 (1.13%)
<i>Gossypium stephensii</i>	125	4	4 (3.2%)	0
<i>Gossypium klotzschianum</i>	90	5	13 (14.44%)	0

2.3. Common Alleles

Common alleles were estimated to understand the phenomenon of cotton evolution and the gene flow mechanism. All six species of cotton considered in this study have common alleles at 114 loci, keeping *G. hirsutum* as fixed. The SSR marker DPL0330-A showed the maximum number of common alleles (124) among all tetraploid cottons except for *G. klotzschianum*, which is diploid, while DPL0249-C showed the minimum number of common alleles, which were only found in *G. barbadense*, *G. klotzschianum*, and *G. hirsutum*. The number of common alleles ranged from 21 to 123 in all six species of *Gossypium*. In this investigation, a total of 459 common loci were observed. Eighty-eight, 101, 82, 91, 67, and 30 loci having common alleles specific to *G. hirsutum* were observed in *G. barbadense*, *G. darwinii*, *G. tomentosum*, *G. ekmanianum*, *G. stephensii*, and *G. klotzschianum*, respectively (Supplementary Table S4). These *G. hirsutum*-specific alleles were amplified by 85 out of 111 SSR markers. The presence of *G. hirsutum*-specific alleles in all six species of *Gossypium* indicated a high level of natural introgression. The level of introgression was found to vary among these wild-type accessions [36,51].

2.4. Analysis of Population Structures

Based on the ΔK value, the analysis of population structures divided 97 out of 132 accessions into six subpopulations (Figures 1–3). Group 1 contained five diploid (D3) accessions collected from Santa Cruz Island. Group 2 had 32 accessions of *G. tomentosum* (AD3) obtained from a Hawaiian Island. Group 3 was composed of 14 accessions and was demarcated with the accession of *G. darwinii* (AD5) collected from Isabella Island but also including one from the China Wild Cotton Germplasm Nursery. Group 4 had 10 accessions of *G. ekmanianum* (AD6) collected from the Dominican Republic, NPGS, USA. Group 5 contained 19 accessions of *G. darwinii* that were collected from San Cristobal. Group 6 had 17 accessions of *G. barbadense* (AD2); of them, two were collected from the China Wild Cotton Germplasm Nursery (Supplementary Table S5). Based on a phylogenetic analysis using the Unweighted Pair-Group Method using Arithmetic average (UPGMA), the same accessions were placed under discriminating subgroups having significant genetic distance in accordance with the geographical locations of the collection. The results were further validated by using Shannon's information index to determine the genetic diversity among six populations. It was found that population 3 (Isabella) had the highest degree of heterozygosity with 55.5% polymorphic loci (Figure 4).

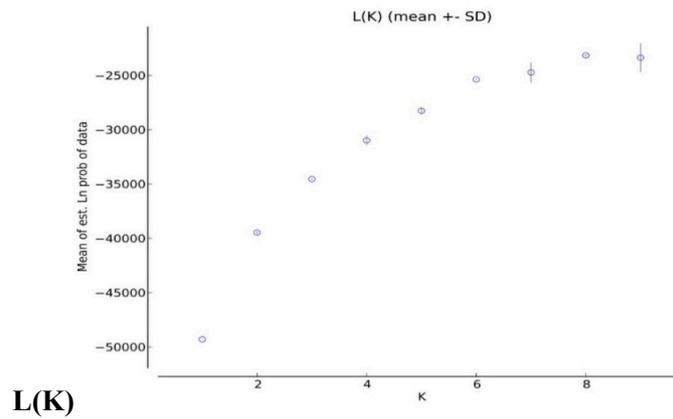


Figure 1. K means for 132 accessions.

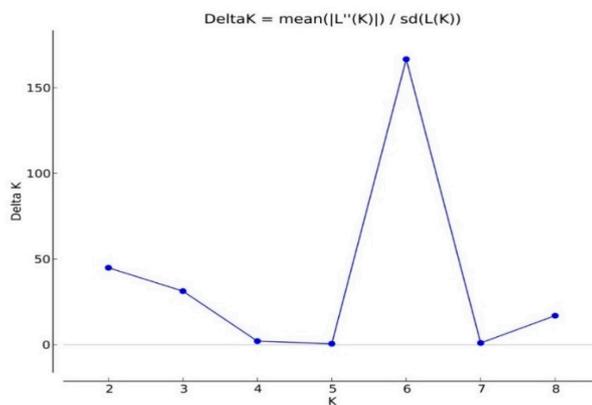


Figure 2. Delta K for 132 accessions.

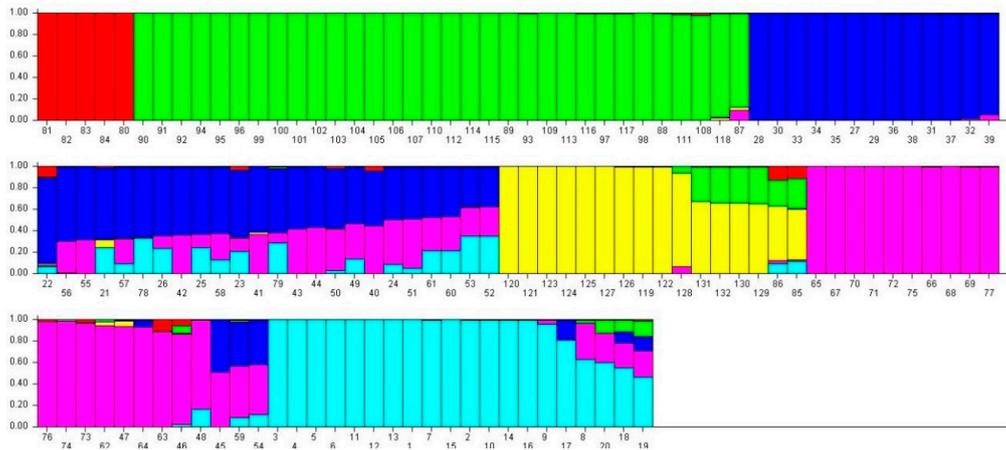


Figure 3. Q plot showing clustering of 132 accessions in 6 subpopulations based on an analysis of genotypic data using STRUCTURE software ver. 2.2. Each accession is indicated by vertical bars. The color subsections within each vertical bar represent the membership coefficient (Q) of the accession to different colors. Six groups were identified. The identified groups are I (red), II (lime), III (Blue), IV (yellow), V (Fuchsia), and VI (Aqua) colors in regular patterns.

Allelic Patterns across Populations

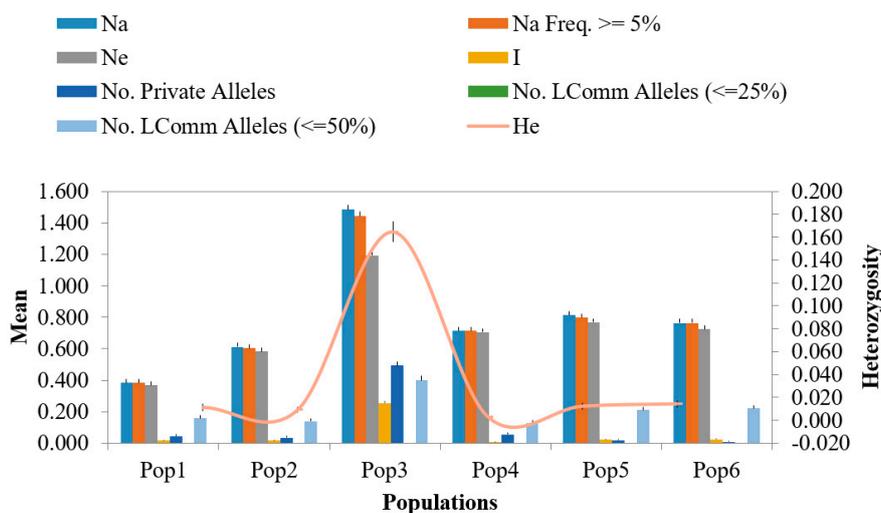


Figure 4. Allelic patterns across populations.

2.5. Genetic Diversity and Cluster Analysis of Phylogenetic Tree

A total of 382 alleles, generated by 111 EST-SSRs, were used to run UPGMA for generating the dendrogram. Based on Nei's criteria [52], the genetic distance among wild cotton accessions ranged from 0.003 to 0.529 with an average of 0.325. The highest genetic distance (0.529) was between D3k-21-3 and AD5-lz. The phylogenetic tree was in agreement with the structure results with the exception that *G. hirsutum* and *G. stephensii* sit in different clusters in the phylogenetic tree but in the structure analysis these were grouped together. In order to see how the results correspond to each other between the STRUCTURE and phylogenetic analyses, the dendrogram was manually edited to show the STRUCTURE grouping (Figure 5 and Supplementary Figure S1). Six groups identified in the structure analysis were also clustered together in the phylogenetic tree analysis. Overall, there was good agreement between the two estimates. The clustering pattern also showed agreement with relationships based on pedigree studies [53]. The first two axes of the principal coordinate analysis (PCoA) accounted for 42.2% of the variation (Figure 6). This indicates a high level of genetic diversity in the *Gossypium* germplasm with continuous variation between and within the subgroups. Analysis of molecular variance (AMOVA) revealed highly significant variation between the six groups identified by the structure analysis, with 49% of the total variation contributing to between-group differences. However, a larger amount of variation (51%) was due to diversity within the groups having different populations (Table 2). Pairwise F_{ST} analysis revealed that accessions from Pop 3 (Isabella region) are closer to accessions from the San Cristobal (Pop 5) and Santa Cruz regions (Pop 6) as compared with the Hawaiian accessions. The highest genetic differentiation was observed among tetraploid populations between accessions from the Hawaiian (Pop 2) and Santa Cruz (Pop 6) regions with a pairwise F_{ST} of 0.752 ($p < 0.001$) (Table 3).

A cluster analysis clearly discriminated diploid wild-type cotton from other tetraploid wild-types. These accessions were collected from different locations, namely the Galapagos Islands, Hawaii, the Dominican Republic, Wake Atoll, and the Wild Cotton Germplasm Nursery of China. The dendrogram was truncated at a genetic distance level of (0.05) and divided 132 cotton genotypes into seven clusters (Supplementary Figure S1).

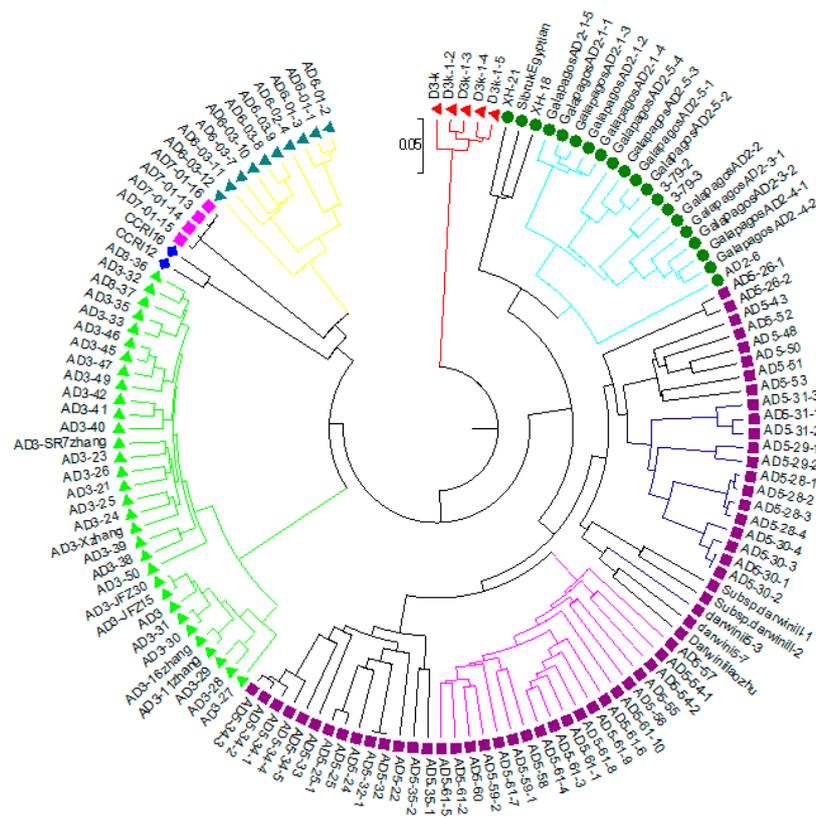


Figure 5. Dendrogram of 132 wild cotton accessions by Unweighted Pair-Group Method using Arithmetic average (UPGMA) analysis. Colors in the dendrogram lines correspond to *Gossypium* accession populations as identified by structure analysis while the colors in the circle represent the seven species. A membership threshold of 70% was used to assign accessions to different clusters in this dendrogram based on structure analysis.

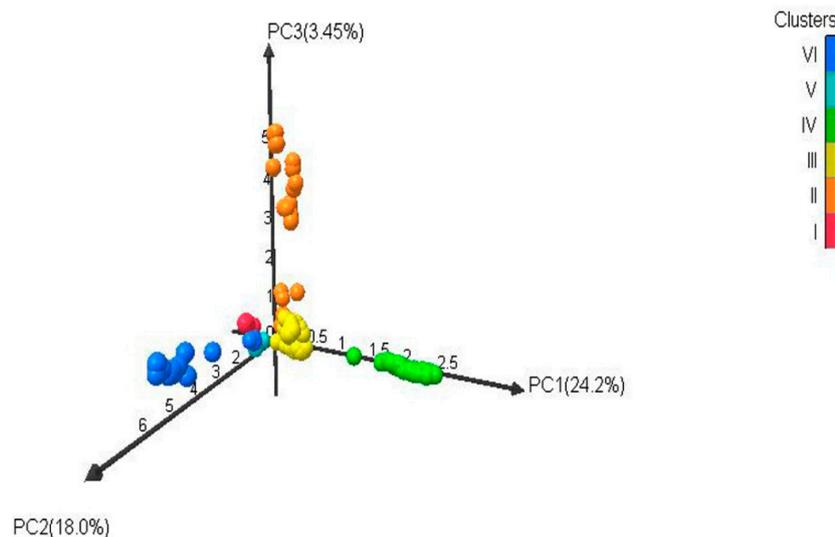


Figure 6. Three-dimensional principal coordinate analysis (PCOA) of a *Gossypium* accessions diversity panel genotyped with expressed sequence tags (EST) and Genomic simple sequence repeats (SSRs). The different colors in the figure correspond to six clusters: Red (Cluster I), orange (Cluster II), yellow (Cluster III), Bright green (Cluster IV), Sky blue (Cluster V), Blue (Cluster VI).

Table 2. Analysis of molecular variance for wild cotton accessions among and within six populations as identified by STRUCTURE.

Source of Variation	df	Sum of Squares	Mean Squares	Estimated Variation	Percentage of Variation
Among Pops	5	4201.563	840.313	38.134 **	49%
Within Pops	126	4932.945	39.150	39.150	51%
Total	131	9134.508		77.284	100%

(PhiPT < 0.493; ** significance at $p < 0.001$).**Table 3.** Pairwise Fst estimates for the five groups corresponding to six regions of accession collections as identified by STRUCTURE.

Populations with Origin	Pop1 (Santa Cruz)	Pop2 (Hawaiian)	Pop3 (Isabella)	Pop4 (Dominican Republic)	Pop5 (San Cristobal)
Pop2 (Hawaiian)	0.869				
Pop3 (Isabella)	0.697	0.651			
Pop4 (Dominican Republic)	0.757	0.689	0.544		
Pop5 (San Cristobal)	0.810	0.749	0.301	0.638	
Pop6 (Santa Cruz)	0.813	0.752 **	0.413	0.626	0.517

(** significance at $p < 0.001$).

2.6. Phylogenetic Analysis of Mined Genes and Functional Annotation of DUF819 (PF005684)

The study was extended to dissect the evolutionary relationship among the uncharacterized genes because these are considered to be highly conserved and have an important role in biology. DUF819 is a family containing proteins (PF005684) found in 532 species with a total of 756 sequences. A total 1517 genes were found among *Gossypium arboreum* (258), *G. raimondii* (258), *G. hirsutum* (513), and *G. barbadense* (488) from the cotton functional genomic database (www.cottonfgd.org). These were the best-fit matched homologue genes having the highest similarity with the four cotton species. Out of 1517, only 116 genes differentially expressed in experiments and belonging to PF00400 and PF005684 were identified among *G. arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*. A total of 24 uncharacterized genes identified to diagnose their evolutionary relationship with these cotton species. A phylogenetic tree consisting of 115 genes out of 116 expressed genes, including uncharacterized genes, in different experiments was constructed for the sorted PF00400 (105) and PF005684 (11) (Figure 7). The protein sequence of one gene remained unaligned during a ClustalW alignment. The PF00400 belonging to the superfamily WDR was used as a reference because 13 out of 24 uncharacterized genes were linked to this protein domain. The remaining 11 uncharacterized genes were named yjcl. The 11 yjcl (PF005684) genes were found to be more closely related to Cotton_A04_RF178 and then to GOBAR_DD_12SPA2, whose functions are known. Cotton_A04_RF178 is a well-known E3 ubiquitous protein having an important role in stress response in plants and animals [54]. It is predicted that, as these genes make a very close cluster with a well-known gene playing a crucial role in plant survival, they may have same function because they can be assigned to proteins by using the bioinformatics tools in comparative genomics [55]. The yjcl of A_ and D_, which are subgenomes of *G. hirsutum* and *G. barbadense*, make close groups with the yjcl of *G. arboreum* and *G. raimondii*, respectively. This indicates that yjcl genes may flow from *G. arboreum* and *G. raimondii* to *G. barbadense* and *G. hirsutum* in equal proportion. Moreover, the uncharacterized genes in *G. hirsutum*, *G. raimondii*, and *G. arboreum* indicated as all2124 are grouped close to Gh_A03ago, which has the known function of a protein related to F-box/WD repeats. Similarly, these results can also be validated by predicting that Gh_A01all2124, Gorai-all2124, Gorai-als3466, and Cott_A_all2124 perform the same function as that determined for Gh_A03ago. GOBAR_AA12SPAC343 is separate from but close to the gene pak1p1 with the known function of p21-activated protein kinase-interacting protein 1-like. The gene SPBC1711 located on the A_ and D_ genome of *G. hirsutum*, *G. arboreum*, and *G. raimondii* makes a close group with the gene named RUP2 in the A03 and A09 chromosomes of *G. hirsutum*. RUP2 is a gene composed of WD protein domains. Meanwhile, another gene, SPAC3H5, originating in *G. hirsutum* and *G. arboreum*

was found to be lying nearest to the DDB2 gene of known function in *G. hirsutum*. These genes were distributed throughout the 26 chromosomes. The maximum number of genes (11) was found on chromosome D11, and the minimum number (4) was found on chromosome 2 (Supplementary Table S6). The coding DNA sequences (CDS) were characterized and the GC content percentage ranged from 38.9 to 52.7 with its length ranging from 417 to 3221 (bp). The maximum number of exons was noted to be 24 in the Gh_D11G0779 homologue of GOBAR_DD21377, indicating the highest intron disruption (Supplementary Table S6). The majority of these mined genes, especially the uncharacterized ones, have a single protein domain, which means that these genes are highly conserved. We analyzed the features of these genes and the results showed several categories related to stress and fiber development in upland cotton. We further analyzed the genes through annotations and Gene Ontology (GO) terms that were associated with the mined genes, which describe the genes in relation to cellular components (CCs), molecular function (MF), and biological process (BP) [56]. In cellular components, functions such as microtubule organizing center (11%), microtubule-associated complex (10%), membrane coat (13%), coated membrane (13%), cytoskeleton part (10%), microtubule cytoskeleton (%), cytoskeleton (10%), and protein complex (23%) were observed. Similarly, 14 molecular functions and 5 biological processes were observed (Figure 8). Finally, we carried out RNA sequence expression to validate our results. The 65 genes with differential expression in *G. hirsutum* were selected to construct a heat map. The genes were both up and downregulated in cold, hot, polyethylene glycol (PEG), and salt treatments and different developmental stages of different tissue organs, such as calyx, leaf, petal, pistil, root, stamen, stem, and torus tissue (Figure 9). The genes were categorized into two main groups. Group 1 comprised 34 genes that were significantly expressed; i.e., with fragments per kilobase of transcript per million mapped reads (FPKM) value of more than 1. Among the 34 upregulated genes, SPA2 (protein SPA1-RELATED 2) with Gene ID Gh_D12G2294 has five Go functions: protein kinase activity (GO:0004672 = MF), protein binding (GO:0005515 = MF), ATP binding (GO:0005524 = MF), protein phosphorylation (GO:0006468 = BP), and transferase activity transferring phosphorus-containing groups (GO:0016772 = MF). CDC40 (Pre-mRNA-processing factor 17) with Gene ID Gh_A05G0018 depicts three GO functions: mRNA splicing via spliceosome (GO:0000398 = BP), protein binding (GO:0005515 = MF), and catalytic step 2 spliceosome (GO:0071013 = CC). Two Guanine nucleotide-binding protein subunit beta-2s with different Gene IDs were found to have two similar GO functions, namely protein binding (GO: 0005515 = MF) and signal transduction (GO:0007165 = BP). All remaining genes were found to be associated in molecular function with protein binding with GO:0005515. The yjL-linked gene ID Gh_A09G2500 showed significant expression against drought and salt stress and fell into group 1. The other two Gene IDs associated with SPAC3H5, an uncharacterized WD-repeat-containing protein (GO:0005515 = MF), also indicated significant expression. Group 2 has 31 genes that exhibited the differential expression of both up and downregulation (Supplementary Table S7). Among these, only the Gh_D07G1711 gene showed three GO functions: mRNA splicing via spliceosome (GO:0000398 = BP), protein binding (GO:0005515 = MF), and catalytic step 2 spliceosome (GO:0071013 = CC). All others were associated with WDR25 (WD-repeat containing protein 25) with the GO function GO:0005515 = MF except for three genes with the IDs Gh_D11G0109, Gh_A11G2961, and Gh_D09G0432, which are linked to the uncharacterized gene yjL and have no GO functions. In the second group, four genes, namely Gh_D07G2259, Gh_A10G2180, Gh_D09G0432, and Gh_D02G1696, were relatively downregulated while all other genes showed differential expression. Gh_D04G1713 showed upregulated expression in petal and stamen tissues but relative downregulation in other tissues and under stress treatments.

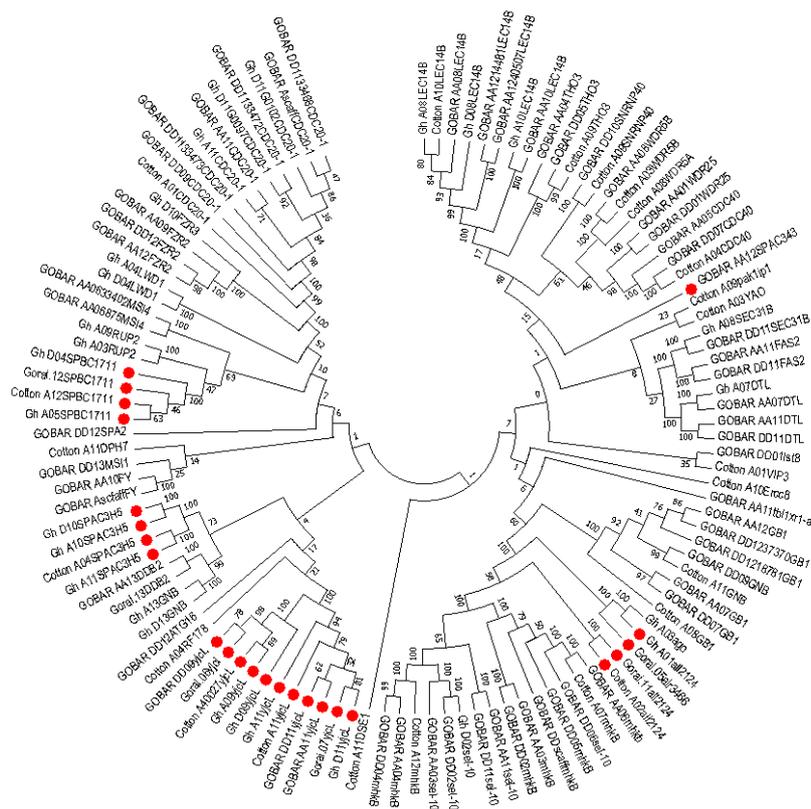


Figure 7. Evolutionary relationship of 115 genes belonging to protein domains of DUF819 (PF005684) and WDR (PF00400) in *Gossypium arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*. The phylogenetic tree was constructed using MEGA software ver. 7.0 by the neighbor-joining method. The parameters were 1000 bootstraps and pairwise deletion. The 24 uncharacterized genes are indicated by red dots in four *Gossypium* species.

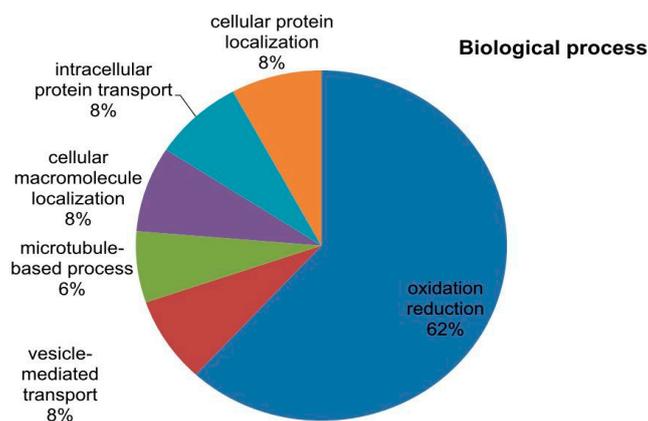


Figure 8. Cont.

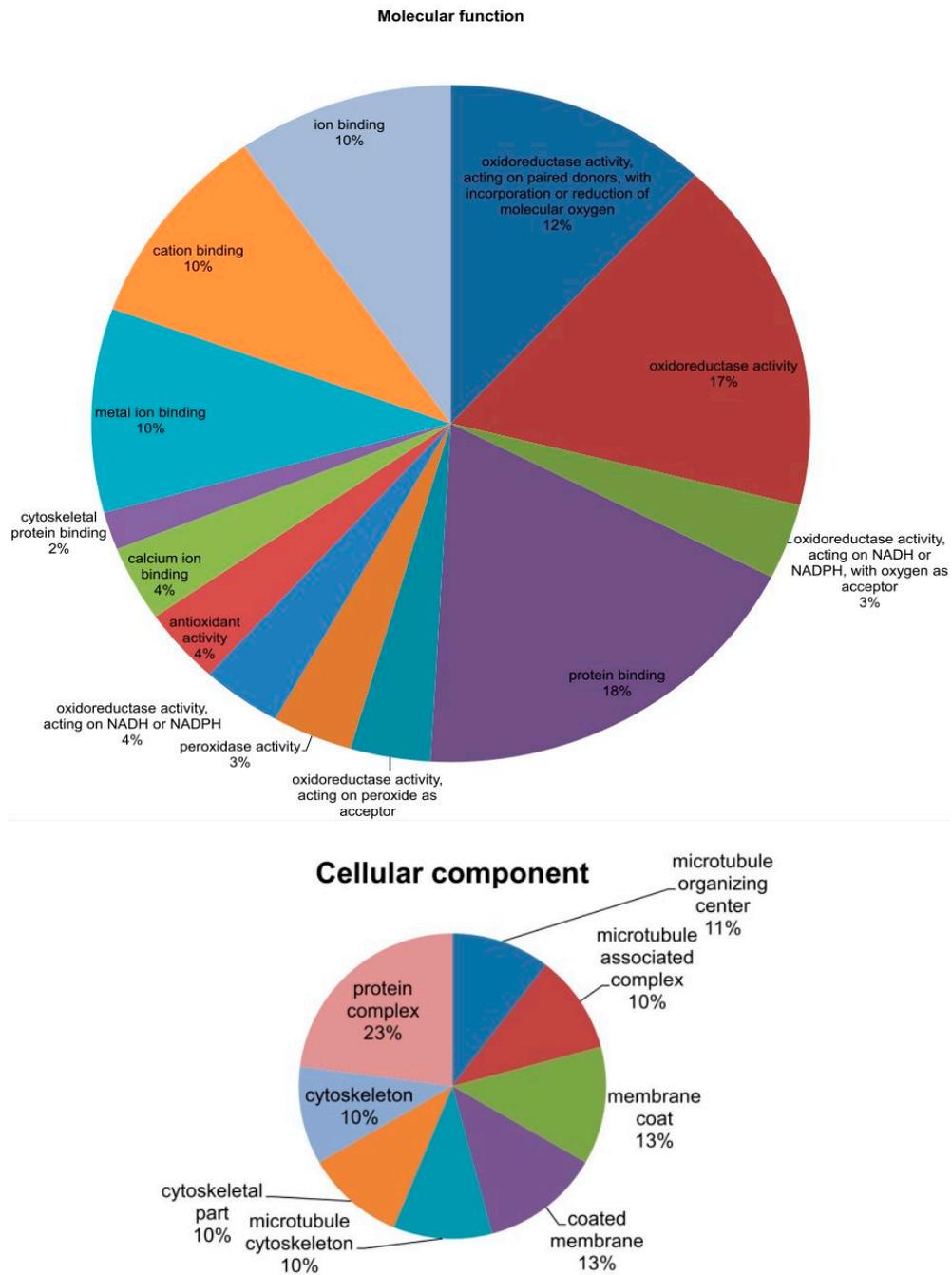


Figure 8. Genes were analyzed using the Agrigo v 2.0 software. Gene Ontology (GO) annotation results for *Gossypium hirsutum* mined genes. GO functional classification of genes mined with protein sequences predicted for their involvement in biological processes (BPs), molecular functions (MFs), and cellular component (CCs).

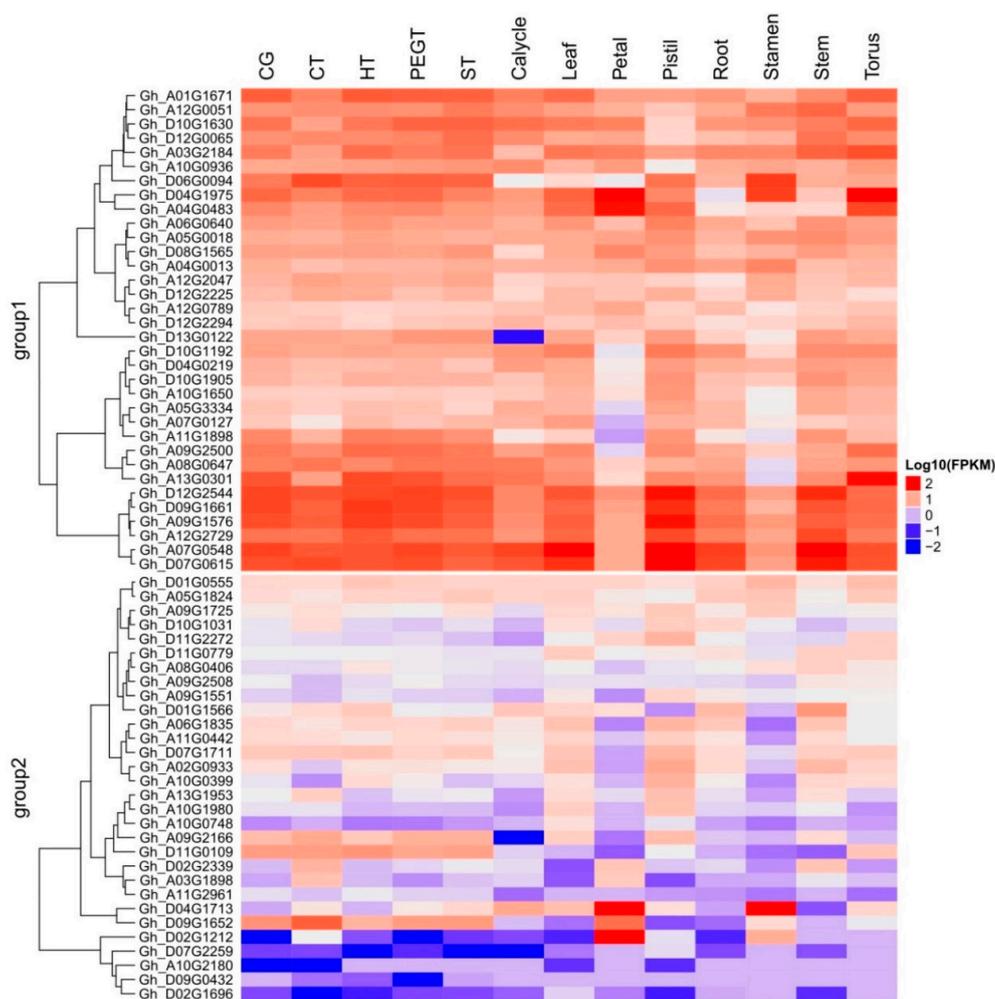


Figure 9. RNA sequence data analysis of 65 differentially expressed genes in eight different cotton tissues with reference to a control group (CG) under different cold (CT), heat (HT), PEG (PEGT), and salt treatments (ST) listed at the top of the figure. The names of genes are listed to the left of the figure. The heat map was generated from the \log_{10} (FPKM) of the expression values by using R software. The Y axis represents the relative expression ($2^{-\Delta\Delta C_t}$).

3. Discussion

In this study, 111 SSR primer pairs generated 382 polymorphic loci in the 132 tested accessions. An average of 3.44 alleles amplified per marker was observed for all accessions, ranging from 2 to 8 alleles. This value is comparable with the findings of Dahab et al. [57] on allele number using 70 SSR markers on *Gossypium hirsutum*. Consistent results were found with 3.93 alleles per locus by Bardak and Bolek [58] for assessing genetic diversity in diploid and tetraploid cottons using Simple Sequence Repeat (SSR) and Inter Simple Sequence Repeat (ISSR) markers. Wendel and Percy [17] detected 3.47 enzymes per locus by studying 17 enzymes encoding 59 loci in a collection of 58 accessions of *darwinii* from six islands. However, other studies showed a variable allele number per locus. For example, 2.13, 2.20, 5.46, and 7.64 alleles per marker have been found in several genetic diversity assessments for the cotton germplasm [26,59–61]. The semi-wild accessions retaining a diverse germplasm showed high allele numbers in the majority of studies consistent with a recent study because these accessions have not yet been exposed to extensive human selection pressure for accumulating a particular type of alleles [26,62,63]. The number of alleles observed per marker is contingent on the selection of markers, the collection of germplasm to be genotyped, and the platform used for the resolution of amplified products [64].

Our study results determined an average PIC value of 0.555 with a range of 0.078 to 0.821, which completely corresponds to the literature-cited average PIC value for cotton SSRs, which ranges from 0.122 [65] to 0.71 [26]. Higher PIC values in cotton as shown in the current study suggest that these accessions can be useful for improving cotton [57]. The unique alleles identified in this study had percentages of 7.36, 5.83, 6.43, 17.36, 17.51, 3.2, and 14.44 in *G. barbadense*, *G. darwinii*, *G. tomentosum*, *G. hirsutum*, *G. ekmanianum*, *G. stephensii* and *G. klotzschianum*, respectively. These are higher than the percentages reported in the earlier report [65]. It is an interesting finding that all the unique alleles were found in the newly collected accessions along with a few from *G. hirsutum*. The unique alleles may be related to unique characteristics, such as extra-long fibers in *G. barbadense* and drought and salt tolerance in *G. tomentosum* and *G. darwinii*, and be similar to other wild accessions.

The common alleles were estimated to understand the gene flow mechanism of the *Gossypium* species during evolution. The results are supported by previous studies and the hypothesis that *G. hirsutum* has a single evolutionary lineage because all of the species in this study have common alleles with reference to the fixed alleles of *G. hirsutum*. *G. tomentosum* is considered to be a sister of *G. hirsutum*, while *G. barbadense* originates from a geographically overlapping region. *G. darwinii* has the same origin, the Galapagos Island, as *G. barbadense*. The *G. klotzschianum* diploid cotton is considered endemic to the New World and has an origin similar to that of *G. raimondii*. The two new species *G. ekmanianum* and *G. stephensii* are sister clades even though they make distinct groups with *G. hirsutum*, but *G. ekmanianum* and *G. stephensii* are monophyletic to *G. hirsutum* which strengthens the hypothesis of gene flow from different species to *G. hirsutum* having a single-lineage evolution [1,13,15]. The present analysis with a high level of natural introgression among the wild accessions shows consistency with the results of Yu et al. [51] and Hinze et al. [36] who described the distribution of introgression within the *G. hirsutum* and *G. barbadense* genomes using the chromosome positions as markers.

Populations from different islands are isolated distinctly, indicating general correspondence to Wedel and Percy's investigations [17]. Due to good agreement with Wendel and Percy [17], an exploration that occurred 30 years prior to the present study, this novel study will also be helpful in understanding the basis of the hybridization and domestication phenomenon in cotton evolution. Our findings also suggest that a wild germplasm has higher genetic diversity than that in cultivated cotton.

A phylogenetic tree constructed based on genotypic data completely validated the distinct clustering of the accessions detected. The results are quite congruent to prior taxonomic studies [2,3,66]. The average genetic distance (GD = 0.325) revealed the overall level of genetic diversity to be high among semi-wild and cultivated accessions; this finding is similar to earlier reports [26,67,68]. However, this estimate may be inflated since data from monomorphic SSR loci were excluded in the current study.

An evolutionary relationship among the genes was also developed. It has been estimated that the majority of genes are linked to responses towards biotic and abiotic stress conditions. For example, the damage-specific DNA binding protein 2 (DDB2), Autophagy-related protein 16 (ATG16), WD repeat-containing protein LWD1, Denticleless protein homolog (DTL), Protein FIZZY-RELATED FZR2, FZR3, Pre-mRNA-processing factor 17 (CDC40), Diphthine methyltransferase (DPH7), Chromatin assembly factor 1 subunit FAS2, DNA excision repair protein ERCC-8, Flowering time control protein FY, Guanine nucleotide-binding protein subunit beta-like protein (GB1), Myosin heavy chain kinase B (mhkB), WD-40 repeat-containing protein MSI1, MSI4, F-box/WD repeat-containing protein sel-10, U5 small nuclear ribonucleoprotein 40 kDa protein (SNRNP40), THO complex subunit 3 (THO3), WD repeat-containing protein VIP3, WDR25, WDR5A, WDR5B, and U3 snoRNP-associated protein-like YAO belong to the superfamily of WDR and are involved in repairing damaged DNA under various stress conditions [69,70]. Similarly, RUP2 has been found to play a very crucial role in vegetative development and flowering in *Arabidopsis* [49]. F-box/WD repeat-containing protein 7, named as "ago" and associated with Gh_A03G1152, supports plants against disease and repairs

damaged DNA [71]. SEC31 homolog B transports proteins and is situated at Golgi-associated endoplasmic reticulum exit sites. CDC20-1 is a known component of the anaphase promoting complex/cyclosome (APC/C), a cell-cycle-regulated E3 ubiquitin–protein ligase complex that controls progression through mitosis and the G1 phase of the cell cycle. This protein is involved in the pathway protein ubiquitination, which is part of protein modification. The intron-containing CDC20 gene copies provide conserved and redundant functions for cell-cycle progression in plants and are required for meristem maintenance, plant growth, and male gametophyte formation [69]. These results also support our hypothesis that *yjL* (PF005684) has the same functions as these WDR family genes.

The current study is highly associated with the pedigree information recently provided by Gallagher et al. [3] after molecular confirmation of newly designated species of *Gossypium*. Genetic diversity within the group was highest for the Isabella group and lowest for the Hawaiian group (Table 2 and Figure 1). The genetic differentiation between groups was further validated by AMOVA, with 49% of the variation among populations and 51% of the variation within populations (** significant $p > 0.001$) being explained by the population structure of the wild cotton germplasm (Table 3). Such higher variation may be due to the complete study of seven different species of diploid and tetraploid cotton. This also indicates the presence of a great genetic difference among tetraploid and diploid cottons as well as a good level of genetic diversity within each group, which can be used in further hybridization breeding programs in cotton to broaden the narrow genetic base of *G. hirsutum*, which is becoming a serious threat due to limited allelic availability [72]. The F_{ST} values for the diploid and tetraploid cottons observed in this study (0.301–0.869) are very high, indicating high genetic distance and diversity. PCoA plots separated tetraploid cottons from diploid plants, supporting the AMOVA results. All these results are in good agreement with Noormohammadi et al.'s genetic diversity analysis [68] between diploid and tetraploid accessions.

Thus, our results could help breeders to determine the selection of appropriate parental combinations in germplasm enhancement programs and conserve genetic diversity and the evolutionary relationship among the genes of uncharacterized functions. The presence of profound population differentiation could pose a challenge to successful Genome-Wide Association Mapping (GWAS) studies in the Upland cotton germplasm for traits that are associated with population structures. The power of structure-based association studies to detect the effects of a single gene would be reduced if a large fraction of variation was explained by the population structures [22,73]. In such cases, alternative association mapping populations would be more useful.

4. Materials and Methods

This study was conducted at the Institute of Cotton Research (ICR, Anyang, China), Chinese Academy of agricultural Sciences (CAAS), Anyang, China. The cotton accessions were obtained from six islands, namely Santa Cruz, San Cristobal, Isabella of Galapagos Island, a Hawaiian Island, Dominican Republic, and Wake Atoll. The screening of this unique collection was carried out using microsatellite markers for the detection of a polymorphism among these accessions.

4.1. Plant Material and DNA Extraction

We sampled a total of 132 accessions belonging to different species, including five *G. klotzschianum* (D3), two *G. hirsutum* (AD1), 20 *G. barbadense* (AD2), 32 *G. tomentosum* (AD3), 59 *G. darwinii* (AD5), 10 *G. ekmanianum* (AD6), and four *G. stephensii* (AD7), from six islands and the Wild Cotton Germplasm Nursery of China. Among the 132 accessions, 32, 25, 16, 32, 10, 4, and 12 were obtained from Santa Cruz, Isabella, San Cristobal, a Hawaiian island, the Dominican Republic, Wake Atoll (NPGS, USA), and the Wild Cotton Germplasm Nursery of China, respectively (Supplementary Table S1). Seedlings of these accessions were grown at the wild cotton germplasm nursery of China, Sanya Hainan during October 2015, 2016, and 2017, respectively. When the plants were about 30–35 days old, fresh leaves were sampled and immediately frozen at -80°C for later DNA extraction. Total genomic DNA was extracted from the frozen leaves by the cetyltrimethylammonium bromide (CTAB) method as described

by Zhang and Stewart [74] with slight modifications. DNA was quantified using Nanodrop at a 260/280 nm absorbance ratio and the quality was checked by 1% (*w/v*) agarose gel electrophoresis.

4.2. SSR Marker Selection and Genotyping

A total of 853 randomly selected SSR markers, including 200 DPL, 310 MonCGR, 48 NAU, 41 MUCS, and 254 SWU, were surveyed for their polymorphisms in 132 genotypes belonging to seven cotton species. Then, 111 EST and genomic SSR (based on D-genome) polymorphic primers from the Cotton Marker Database (CMD; <http://www.cottonmarker.org/>) were used in the SSR analysis. The reaction contained 5 μ L 2 \times Taq Master Mix (containing buffer, dNTPs, and Taq DNA Polymerase), 2 μ L primers, 1 μ L DNA, and 2 μ L H₂O. The PCR reaction was performed using a together TP 600 thermal cycler (TAKARA Bio Inc., Kusatsu, Japan) and then followed by silver staining according to a previous method described by Zhang et al. [75]. The PCR temperature program was two cycles of 95 °C for 3 min pre-denaturing followed by 30 cycles of 94 °C for 45 s denaturing, 57 °C for 36 s annealing, 72 °C for 1 min extension, with a final step of 1 cycle at 72 °C for 5 min extension. To confirm that the observed amplicons were amplified from genomic DNA and not a primer artifact, genome DNA was omitted from the control reaction. No amplification products were detected without genomic DNA in any PCR.

4.3. Analysis of Genotypic Data and Genetic Diversity

Pairwise genetic distances between accessions were calculated using the Powermarker software package ver.3.25 by Nei et al. [52] D_A distance. The dendrogram was constructed on the basis of the distance matrix. We estimated the similarity between genotypes for each accession by awarding a score to each microsatellite (i.e., 0 when an allele was absent, 1 when the allele was present). The cluster analysis was carried out using the unweighted pair group method using arithmetic average (UPGMA) and the dendrogram resulting from these calculations was plotted using MEGA 6.0 to visualize and edit the dendrogram. The basic summary statistics for biallelic data were calculated using the POWERMARKER software package version 3.25 [76]. The polymorphism information content (PIC) of an SSR marker was determined according to the method described by Anderson et al. [77] based on the allele frequency of all genotypes.

$$\text{PIC} = 1 - \sum_{i=1}^n p_i^2 \quad (1)$$

where P_{ij} is the frequency of the allele for locus i and the summation covered n patterns.

A PIC value of 1 indicates that the marker can differentiate each line, and 0 indicates a monomorphic marker. The informative potential of a marker is high if its PIC value is more than 0.5, moderate if its PIC is between 0.5 and 0.25, and only slightly informative if its PIC value is below 0.25. Other statistics calculated were the number of alleles and availability and gene diversity for each marker. Further analysis of genetic structure was done by means of Principal co-ordinate analysis (PCoA) using XLSTAT, 2014 [78] and a three-dimensional diagram was constructed. Dominant data (0, 1 binary data) were used for the PCoA analysis.

4.4. Analysis of Genetic Structure

The STRUCTURE software version 2.3.4 [79] was employed to define 132 accessions into clusters consisting of genotypes by using co-dominant genotypic data. The admixture model was used to estimate a mixed group by using correlated allele frequencies between populations as described by Falush et al. [80]. The optimum number of subpopulations was calculated based on the recommendation of Evanno et al. [81] by defining the values for $K = 2$ to $K = 10$ with a burn length of 10,000 and a run length of 100,000 each in 10 runs. The results were uploaded in a Zip file to the STRUCTURE harvester software for finding the ΔK [82]. Grouping and subgrouping of accessions

was done if the probability of membership was more than 70% [83]. The accessions with membership <70% were placed into the mixed subgroup.

4.5. Gene Mining and Phylogenetic Analysis of DUF819 Proteins

The complete sequence of Markers SWU15000–SWU15194 mapped for chromosome 6 was downloaded from the Genome database of *G. raimondii* [84] and blastx was used to find the homologue similarity of genes in the genome sequence of *G. raimondii*, *G. arboreum*, *G. hirsutum*, and *G. barbadense*. The mining of genes from the marker regions has been done extensively; see for instance Kirungu et al. [43]. Similarly, the same has been applied by Magwanga et al. [85]. The uncharacterized gene named yjcL of DUF819 (PF008654) was selected for the evolutionary study of genes in sequenced *Gossypium* species. The full-length sequences of DUF819 (PF005684) were downloaded from the pfam database (<http://pfam.xfam.org/>). The dendrogram was constructed by using Molecular Evolutionary Genetics version 7.0 [86]. The functional description related to domains of uncharacterized proteins has been predicted using the protein sequence of 116 genes downloaded from the Cotton Functional Genomics database (www.cottonfgd.org) [87]. The evolutionary relationship among all selected genes was summed up to provide a clear picture of functions with reference to upregulated genes of the superfamily WDR.

Thus, 115 genes out of 116 were grouped into 6 major clusters and their evolutionary history was inferred using the Neighbor-Joining method [88]. The optimal tree with the sum of branch length equal to 41.56 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches [89]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [90] and the units for the number of amino acid substitutions per site. The analysis involved 115 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 1466 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Figure 7).

The explored genes were analyzed for their gene features, protein characteristics, and RNA expression using the cotton functional genome database (<https://cottonfgd.org/search/>), While GO functional classification was done using Agrigo ver. 2.0 software acquiring *Gossypium hirsutum* as the reference genome. The analysis of RNA expression data inferred was then carried out to construct a heatmap using the R statistical software package.

5. Conclusions

SSR markers can be used to describe the degree of differentiation between populations and to control the conservation of genetic resources. The study concludes that the evaluated cotton accessions have a broad genetic basis. The recurrent use of these accessions as parents will produce significant results. The genetic diversity and evolutionary relationship recognized among the uncharacterized genes and population structures established in this study would be informative to select parental accessions for breeding and genetic analysis as well as for efficient management and conservation of allotetraploid cotton genetic diversity. We identified DUF819 (PF005684), which is not only highly conserved but also plays an important role against biotic and abiotic stress, among four sequenced cotton species by using the WDR (PF00400) superfamily as reference genome-sequenced proteins. Additionally, the current diversity panel of semi-wild cottons will be invaluable as a community resource for measuring linkage disequilibrium (LD) and for fine-scale mapping of traits through LD mapping or a Genome-Wide Association Study (GWAS) that can be streamlined for genomics-assisted plant breeding programs. Our findings suggest that allotetraploid cotton species, including *G. barbadense* (AD2), *G. tomentosum* (AD3), *G. darwinii* (AD5), *G. ekmanianum* (AD6), and *G. stephensii* (AD7), are a rich source for the creation of genetic diversity in upland cotton.

Supplementary Materials: Supplementary materials can be found at <http://www.mdpi.com/1422-0067/19/8/2401/s1>.

Author Contributions: K.W., A.D., and F.L. designed the experiments. A.D. and Z.Z. conceived the experiments and analyzed the results. A.D. and Z.Z. carried out the majority of the experiments and contributed equally. A.D. carried out all computational analyses. F.L., X.C., X.W., K.W.O., M.S., Y.X., Y.H., M.K.R.K., and M.S.I. participated in the mapping experiments. A.D. drafted the manuscript and K.W. revised the manuscript. All authors read and approved the final manuscript.

Funding: This research program was financially sponsored by the National key research and development plan (2016YFD0100306, 2016YFD0100203) and National Natural Science Foundation of China (31530053, 31671745).

Acknowledgments: We are indebted to give appreciation to State Key Laboratory of Cotton Biology, Institute of Cotton Research (ICR), Chinese Academy of Agricultural Sciences (CAAS); Anyang, China for providing umbrella for research program. We are also grateful to National Plant Germplasm System (NPGS), USA for providing germplasm. We express profound sense of reverence to Kunbo Wang and Fang Liu (Institute of Cotton Research), for timely guidance and provision of material whenever we needed during research work. To the entire research team, friends, and any other person who contributed, we have deep gratitude for you so much.

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

Abbreviations

PIC	Polymorphism information content
MAF	Major Allele frequency
GD	Genetic distance
Pop	population
AMOVA	Analysis of molecular variance
PCA	Principal coordinate analysis
PCR	Polymerase chain reaction
DNA	Deoxyribonucleic acid
NPGS	National Plant Germplasm System
EST	Expressed sequence tags
gSSR	Genomic SSR
SSR	Simple sequence repeats
RFLP	Restriction fragment polymorphism
RAPD	Random amplified polymorphic DNA
AFLP	Amplified fragment length polymorphism
SNP	Single nucleotide polymorphism
G	<i>Gossypium</i>
WCGN	Wild cotton Germplasm Nursery of China

References

- Wendel, J.F.; Brubaker, C.; Alveraz, I.; Cronn, R.; Stewart, J.M. Evolution and natural history of cotton genus. *Genet. Genom.* **2009**, *3*, 3–22. [[CrossRef](#)]
- Wendel, J.F.; Grover, C.E. *Taxonomy and Evolution of Cotton Genus, Gossypium*; Issue Agronomogr; Cotton, American Society of Agronomy, Inc.: Madison, WI, USA; Crop Science Society of America, Inc.: Madison, WI, USA; Soil Science Society of America, Inc.: Madison, WI, USA, 2015; Volume 57. [[CrossRef](#)]
- Gallagher, J.P.; Grover, C.E.; Rex, K.; Moran, M.; Wendel, J.F. A New Species of Cotton from Wake Atoll, *Gossypium stephensii* (Malvaceae). *Syst. Bot.* **2017**, *42*, 115–123. [[CrossRef](#)]
- Grover, C.E.; Zhu, X.; Grupp, K.K.; Jareczek, J.J.; Gallagher, J.P.; Szadkowski, E.; Seijo, J.G.; Wendel, J.F. Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. *Genet. Resour. Crop Evol.* **2015**, *62*, 103–114. [[CrossRef](#)]
- Chandrakanth, K. *Wild Crop Relatives: Genomic and Breeding Resources*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 109–122.
- Krapovickas, A.; Seijo, G. *Gossypium ekmanianum* (Malvaceae), algodon silvestre de la Republica Dominicana. *Bonplandia* **2008**, *17*, 55–63.
- Wendel, J.F.; Brubaker, C.L.; Percival, A.E. Genetic diversity in *Gossypium hirsutum* and the origin of Upland cotton. *Am. J. Bot.* **1992**, *79*, 1291–1310. [[CrossRef](#)]

8. Wendel, J.F.; Cronn, R.C. Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **2003**, *78*, 139–186.
9. Wang, Q.; Fang, L.; Chen, J.; Hu, Y.; Si, Z.; Wang, S.; Chang, L.; Guo, W.; Zhang, T. Genome-Wide Mining, Characterization, and Development of Microsatellite Markers in *Gossypium* Species. *Sci. Rep.* **2015**, *5*, 10638. [[CrossRef](#)] [[PubMed](#)]
10. Wendel, J.F.; Brubaker, C.L. RFLP diversity in *Gossypium hirsutum* L. and new insights into the domestication of cotton. *Am. J. Bot.* **1993**, *80*, 71.
11. Brubaker, C.L.; Paterson, A.H.; Wendel, J.F. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **1999**, *42*, 184–203. [[CrossRef](#)]
12. Senchina, D.S.; Alvarez, I.; Cronn, R.C.; Liu, B.; Rong, J.; Noyes, R.D.; Paterson, A.H.; Wing, R.A.; Wilkins, T.A.; Wendel, J.F. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **2003**, *20*, 633–643. [[CrossRef](#)] [[PubMed](#)]
13. Campbell, B.T.; Saha, S.; Percy, R.; Frelichowski, J.; Jenkins, J.N.; Park, W.; Mayee, C.D.; Gotmare, V.; Dessauw, D.; Giband, M.; et al. Status of the global cotton germplasm resources. *Crop Sci.* **2010**, *50*, 1161–1179. [[CrossRef](#)]
14. Chen, H.; Khan, M.K.R.; Zhou, Z.; Wang, X.; Cai, X.; Ilyas, M.K.; Wang, C.; Wang, Y.; Li, Y.; Liu, F.; et al. A high density SSR genetic map constructed from F₂ population of *Gossypium hirsutum* and *Gossypium darwinii*. *Gene* **2015**, *574*, 273–286. [[CrossRef](#)] [[PubMed](#)]
15. Dejoode, D.; Wendel, J. Genetic diversity and origin of the hawaiian-islands cotton, *Gossypium tomentosum*. *Am. J. Bot.* **1992**, *79*, 1311–1319. [[CrossRef](#)]
16. Wendel, J.F.; Rowley, R.; Stewart, J. Genetic diversity in and phylogenetic-relationships of the brazilian endemic cotton, *Gossypium mustelinum* (Malvaceae). *Plant Syst. Evol.* **1994**, *192*, 49–59. [[CrossRef](#)]
17. Wendel, J.F.; Percy, R. Allozyme diversity and introgression in the galapagos-islands endemic *Gossypium darwinii* and its relationship to continental *Gossypium barbadense*. *Biochem. Syst. Ecol.* **1990**, *18*, 517–528. [[CrossRef](#)]
18. Ulloa, M.; Stewart, J.M.; Garcia-C, E.A.; Goday, A.S.; Gaytan-M, A.; Acosta, N.S. Cotton Genetic Resources in the Western States of Mexico: In Situ Conservation Status and Germplasm Collection for ex situ Preservation. *Genet. Resour. Crop Evol.* **2006**, *53*, 653–668. [[CrossRef](#)]
19. Abdurakhmonov, I.Y.; Buriev, Z.T.; Shermatov, S.E.; Abdullaev, A.A.; Urmonov, K.; Kushanov, F.; Egamberdiev, S.S.; Shapulatov, U.; Abdukarimov, A.; Saha, S.; et al. Genetic Diversity in *Gossypium* genus. In *Genetic Diversity in Plants*; Caliskan, M., Ed.; InTech: Rijeka, Croatia, 2012; pp. 313–338. [[CrossRef](#)]
20. Van Esbroeck, G.A.; Bowman, D.T. Cotton germplasm diversity and its importance to cultivar development. *J. Cotton Sci.* **1998**, *2*, 121–129.
21. Bolek, Y.; El-Zik, K.M.; Pepper, A.E.; Bell, A.A.; Magill, C.W.; Thaxton, P.M.; Reddy, O.U.K. Mapping of verticillium wilt resistance genes in cotton. *Plant Sci.* **2005**, *168*, 1581–1590. [[CrossRef](#)]
22. Tyagi, P.; Gore, M.A.; Bowman, D.T.; Campbell, B.T.; Udall, J.A.; Kuraparthi, V. Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). *Theor. Appl. Genet.* **2014**, *127*, 283–295. [[CrossRef](#)] [[PubMed](#)]
23. Iqbal, M.A.; Rehman, M.U. Identification of Marker-Trait Associations for Lint Traits in Cotton. *Front. Plant Sci.* **2017**, *8*, 86. [[CrossRef](#)] [[PubMed](#)]
24. Kalivas, A.; Xanthopoulos, F.; Kehagia, O.; Tsaftaris, A.S. Agronomic characterization, genetic diversity and association analysis of cotton cultivars using simple sequence repeat molecular markers. *Genet. Mol. Res.* **2011**, *10*, 208–217. [[CrossRef](#)] [[PubMed](#)]
25. Sun, G.; He, S.; Pan, Z.; Du, X. Homologous simple sequence repeats (SSRs) analysis in tetraploid (AD1) and diploid (A2, D5) genomes of *Gossypium*. *Hereditas (Beijing)* **2015**, *37*, 192–203.
26. Kiflom, W.O.; Xiao, L.; Cai, X.; Wang, X.; Chen, H.; Zhou, Z.; Wang, C.; Wang, Y.; Liu, F.; Wang, K. Genome wide assessment of genetic diversity and fiber quality traits characterization in *Gossypium hirsutum* races. *J. Integr. Agric.* **2017**, *16*, 2402–2412. [[CrossRef](#)]
27. Khan, M.K.R.; Haodong, C.; Zhongli, Z.; Ilyas, M.K.; Xingxing, W.; Cai, X.; Chunying, W.; Fang, L.; Kunbo, W. Genome Wide SSR High Density Genetic Map Construction from an Interspecific Cross of *Gossypium hirsutum* X *Gossypium tomentosum*. *Front. Plant Sci.* **2016**, *7*, 436. [[CrossRef](#)] [[PubMed](#)]
28. Ramakrishnan, M.; Ceasar, S.A.; Duraipandiyar, V.; Dhabi, N.A.A.; Ignacimuthu, S. Assessment of genetic diversity, population structure and relationships in Indian and non-Indian genotypes of finger millet (*Eleusine coracana* (L.) Gaertn) using genomic SSR markers. *SpringerPlus* **2016**, *5*, 120. [[CrossRef](#)] [[PubMed](#)]

29. Van Becelaere, G.; Lubbers, E.L.; Paterson, A.H.; Chee, P.W. Pedigree vs. DNA marker-based genetic similarity estimates in cotton. *Crop Sci.* **2005**, *45*, 2281–2287. [[CrossRef](#)]
30. Iqbal, M.J.; Aziz, N.; Saeed, N.A.; Zafar, Y. Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. *Theor. Appl. Genet.* **1997**, *94*, 139–144. [[CrossRef](#)] [[PubMed](#)]
31. Bakht, J.; Iqbal, M.; Shafi, M. Genetic diversity and phylogenetic relationship in different genotypes of cotton for future breeding. *Int. Quart. J. Biol. Sci.* **2017**, *5*, 25–29.
32. Rahman, M.; Yasmin, T.; Tabbasam, N.; Ullah, I.; Asif, M.; Zafar, Y. Studying the extent of genetic diversity among *Gossypium arboreum* L. genotypes/cultivars using DNA fingerprinting. *Genet. Resour. Crop. Evol.* **2008**, *55*, 331–339. [[CrossRef](#)]
33. Shaheen, N.; Pearce, S.R.; Khan, M.A.; Mahmood, T.; Yasmin, G.; Hayat, M.Q. AFLP mediated genetic diversity of *Malvaceae* species. *J. Med. Plant Res.* **2010**, *4*, 148–154.
34. Liu, S.; Cantrell, R.G.; McCarty, J.C.J.; Stewart, J.M. Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. *Crop Sci.* **2000**, *40*, 1459–1469. [[CrossRef](#)]
35. Liu, D.; Guo, X.; Lin, Z.; Nie, Y.; Zhang, X. Genetic diversity of Asian cotton (*Gossypium arboreum* L.) in china evaluated by microsatellite analysis. *Genet. Res. Crop Evol.* **2006**, *53*, 1145–1152. [[CrossRef](#)]
36. Hinze, L.L.; Gazave, E.; Gore, M.A.; Fang, D.D.; Scheffler, B.E.; Yu, J.Z.; Jones, D.C.; Frelichowski, J.; Percy, R.G. Genetic Diversity of the Two Commercial Tetraploid Cotton Species in the *Gossypium* Diversity Reference Set. *J. Hered.* **2016**, *107*, 274–286. [[CrossRef](#)] [[PubMed](#)]
37. Zhang, Y.; Wang, X.F.; Li, Z.K.; Zhang, G.Y.; Ma, Z.Y. Assessing genetic diversity of cotton cultivars using genomic and newly developed expressed sequence tag-derived microsatellite markers. *Genet. Mol. Res.* **2011**, *10*, 1462–1470. [[CrossRef](#)] [[PubMed](#)]
38. Dongre, A.; Bhandarkar, M.; Banerjee, S. Genetic diversity in tetraploid and diploid cotton (*Gossypium* spp.) using ISSR and microsatellite DNA markers. *Indian J. Biotechnol.* **2007**, *6*, 349–353.
39. Liu, B.; Wendel, J.F. Intersimple sequence repeat (ISSR) polymorphisms as a genetic marker system in cotton. *Mol. Ecol. Notes* **2001**, *1*, 205–208. [[CrossRef](#)]
40. Meng, K.; Wei, S.J.; Wang, Y.Q.; Zhou, D.U.; Ma, L.; Fang, D.; Yang, W.H.; Ma, Z.Y. Development of a core set of SNP markers for the identification of upland cotton cultivars in China. *J. Integr. Agric.* **2016**, *15*, 954–962. [[CrossRef](#)]
41. Park, Y.J.; Lee, J.K.; Kim, N.S. Simple Sequence Repeat Polymorphisms (SSRPs) for Evaluation of Molecular Diversity and Germplasm Classification of Minor Crops. *Molecules* **2009**, *14*, 4546–4569. [[CrossRef](#)] [[PubMed](#)]
42. Campbell, B.T.; Williams, V.E.; Park, W. Using molecular markers and field performance data to characterize the pee dee cotton germplasm resources. *Euphytica* **2009**, *169*, 285–301. [[CrossRef](#)]
43. Kirungu, J.N.K.; Deng, Y.; Cai, X.; Mangwanga, R.O.; Zhou, Z.; Wang, X.; Wang, Y.; Zhang, Z.; Wang, K.; Liu, F. Simple Sequence Repeat (SSR) Genetic Linkage Map of D Genome Diploid Cotton Derived from an Interspecific Cross between *Gossypium davidsonii* and *Gossypium klotzschianum*. *Int. J. Mol. Sci.* **2018**, *19*, 204. [[CrossRef](#)] [[PubMed](#)]
44. Goodacre, N.F.; Gerloff, D.L.; Uetz, P. Protein Domains of Unknown Function Are Essential in Bacteria. *mBio* **2013**, *5*, e00744-13. [[CrossRef](#)] [[PubMed](#)]
45. Mulder, N.J.; Kersey, P.; Pruess, M.; Apweiler, R. In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol.* **2008**, *38*, 165–177. [[CrossRef](#)] [[PubMed](#)]
46. Bateman, A.; Coghill, P.; Finn, R.D. DUFs: Families in search of function. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2010**, *66*, 1148–1152. [[CrossRef](#)] [[PubMed](#)]
47. Littler, E. Combinatorial domain hunting: Solving problems in protein expression. *Drug Discov. Today* **2010**, *15*, 461–467. [[CrossRef](#)] [[PubMed](#)]
48. Hauser, R.; Pech, M.; Kijek, J.; Yamamoto, H.; Titz, B.; Naeve, F.; Tovchigrechko, A.; Yamamoto, K.; Szaflarski, W.; Takeuchi, N.; et al. RsfA (YbeB) proteins are conserved ribosomal silencing factors. *PLoS Genet.* **2012**, *8*, e1002815. [[CrossRef](#)] [[PubMed](#)]
49. Gruber, H.; Heijde, M.; Heller, W.; Albert, A.; Seidlitz, H.K.; Ulm, R. Negative feedback regulation of UV-B-induced photomorphogenesis and stress acclimation in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 20132–20137. [[CrossRef](#)] [[PubMed](#)]
50. Li, Q.; Zhao, P.; Li, J.; Zhang, C.; Wang, L.; Ren, Z. Genome-wide analysis of the WD-repeat protein family in Cucumber and Arabidopsis. *Mol. Genet. Genom.* **2013**, *289*, 103–124. [[CrossRef](#)] [[PubMed](#)]

51. Yu, J.; Fang, D.; Kohel, R.; Ulloa, M.; Hinze, L.; Percy, R.; Zhang, J.; Chee, P.; Schefer, B.; Jones, D. Development of a core set of SSR markers for the characterization of *Gossypium* germplasm. *Euphytica* **2012**, *187*, 203–213. [[CrossRef](#)]
52. Nei, M.; Tajima, F.; Tateno, Y. Accuracy of estimated phylogenetic trees from molecular data. *J. Mol. Evol.* **1983**, *19*, 153–170. [[CrossRef](#)] [[PubMed](#)]
53. Smith, C.W.; Cothren, J.T. *Cotton: Origin, History, Technology, and Production*; John Wiley & Sons: Hoboken, NJ, USA, 1999; p. 43.
54. Shu, K.; Yang, W. E3 Ubiquitin Ligases: Ubiquitous Actors in Plant Development and Abiotic Stress Responses. *Plant Cell Physiol.* **2017**, *58*, 1461–1476. [[CrossRef](#)] [[PubMed](#)]
55. Matteo, P.; Edward, M.M.; Michael, J.T.; David, E.; Todd, O.Y. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 4285–4288.
56. Tian, T.; Liu, Y.; Yan, H.; You, Q.; Yi, X.; Du, Z.; Xu, W.; Su, Z. agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **2017**, *45*, W12–W129. [[CrossRef](#)] [[PubMed](#)]
57. Dahab, A.A.; Saeed, M.; Mohamed, B.B.; Ashraf, M.A.; Puspito, A.N.; Bajwa, K.S.; Shahid, A.A.; Husnain, T. Genetic diversity assessment of cotton (*Gossypium hirsutum* L.) genotypes from Pakistan using simple sequence repeat markers. *Aust. J. Crop Sci.* **2013**, *7*, 261–267, ISSN 1835-2707.
58. Bardak, A.; Bolek, Y. Genetic Diversity of Diploid and Tetraploid Cottons Determined By Ssr and Issr Markers. *Turk. J. Field Crops* **2012**, *2*, 139–144.
59. Bertini, C.H.C.D.; Schuster, I.; Sediya, T.; Barros, E.G.; Moreira, M.A. Characterization and genetic diversity analysis of cotton cultivars using microsatellites. *Genet. Mol. Biol.* **2006**, *29*, 321–329. [[CrossRef](#)]
60. Qin, H.; Chen, M.; Yi, X.; Bie, S.; Zhang, C.; Zhang, Y.; Lan, J.; Meng, Y.; Yuan, Y.; Jiao, C. Identification of Associated SSR Markers for Yield Component and Fiber Quality Traits Based on Frame Map and Upland Cotton Collections. *PLoS ONE* **2015**, *10*, e0118073. [[CrossRef](#)] [[PubMed](#)]
61. Baytar, A.A.; Erdogan, O.; Frary, A.; Frary, A.; Doganlar, S. Molecular diversity and identification of alleles for Verticillium wilt resistance in elite cotton (*Gossypium hirsutum* L.) germplasm. *Euphytica* **2017**, *213*, 31. [[CrossRef](#)]
62. Iqbal, M.J.; Reddy, O.U.K.; El-Zik, K.M.; Pepper, A.E. A genetic bottleneck in the evolution under domestication of upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting. *Theor. Appl. Genet.* **2001**, *103*, 547–554. [[CrossRef](#)]
63. Rungis, D.; Llewellyn, D.; Dennis, E.S.; Lyon, B.R. Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton (*Gossypium hirsutum* L.) cultivars. *Aust. J. Agric. Res.* **2005**, *56*, 301–307. [[CrossRef](#)]
64. Lacape, J.M.; Dessauw, D.; Rajab, M.; Noyer, J.L.; Hau, B. Microsatellite diversity in tetraploid *Gossypium* germplasm: Assembling a highly informative genotyping set of cotton SSRs. *Mol. Breed.* **2007**, *19*, 45–58. [[CrossRef](#)]
65. Abdurakhmonov, I.Y.; Kohel, R.J.; Yu, J.Z.; Pepper, A.E.; Abdullaev, A.A.; Kushanov, F.N.; Salakhutdinov, L.B.; Buriev, Z.T.; Saha, S.; Scheffler, B.E.; et al. Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* **2008**, *92*, 478–487. [[CrossRef](#)] [[PubMed](#)]
66. Grover, C.E.; Gallagher, J.P.; Jareczek, J.J.; Page, J.T.; Udal, J.A.; Gore, M.A.; Wendel, J.F. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Mol. Phylogenet. Evol.* **2015**, *92*, 45–52. [[CrossRef](#)] [[PubMed](#)]
67. Noormohammadi, Z.; Farahani, Y.H.A.; Sheidai, M.; Baraki, S.G.; Alishah, O. Genetic diversity analysis in Opal cotton hybrids based on SSR, ISSR, and RAPD markers. *Genet. Mol. Res.* **2013**, *12*, 256–269. [[CrossRef](#)] [[PubMed](#)]
68. Noormohammadi, Z.; Sheidai, M.; Foroutan, M.; Alishah, O. Networking and Bayesian analyses of genetic affinity in cotton germplasm. *Nucleus* **2015**, *58*, 33–45. [[CrossRef](#)]
69. Wu, J.F.; Wang, Y.; Wu, S.H. Two New Clock Proteins, LWD1 and LWD2, Regulate Arabidopsis Photoperiodic Flowering. *Plant Physiol.* **2008**, *148*, 948–959. [[CrossRef](#)] [[PubMed](#)]
70. Biedermann, S.; Hellmann, H. The DDB1a interacting proteins ATCSA-1 and DDB2 are critical factors for UV-B tolerance and genomic integrity in Arabidopsis thaliana. *Plant J.* **2010**, *62*, 404–415. [[CrossRef](#)] [[PubMed](#)]
71. Strohmaier, H.; Spruck, C.H.; Kaiser, P.; Won, K.A.; Sangfelt, O.; Reed, S.I. Human F-box protein hDcd4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line. *Nature* **2001**, *413*, 316–322. [[CrossRef](#)] [[PubMed](#)]
72. Brown, W.L. Genetic diversity and genetic vulnerability: An appraisal. *Econ. Bot.* **1983**, *37*, 4–12. [[CrossRef](#)]

73. Flint-Garcia, S.A.; Anne-Ce line, T.; Yu, J.; Pressoir, G.; Romero, S.M.; Mitchell, S.E.; Doebley, J.; Kresovich, S.; Goodman, M.M.; Buckler, E.S. Maize association population: A high resolution platform for quantitative trait locus dissection. *Plant J.* **2005**, *44*, 1054–1064. [[CrossRef](#)] [[PubMed](#)]
74. Zhang, J.; Stewart, J.M. Economical and rapid method for extraction cotton genomic DNA. *J. Cotton Sci.* **2000**, *4*, 193–201.
75. Zhang, J.; Guo, W.; Zhang, T. Molecular linkage map of allotetraploid cotton (*Gossypium hirsutum* L. X *Gossypium barbadense* L.) with a haploid population. *Theor. Appl. Genet.* **2002**, *105*, 1166–1174. [[CrossRef](#)] [[PubMed](#)]
76. Liu, K.; Muse, S.V. PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* **2005**, *21*, 2128–2129. [[CrossRef](#)] [[PubMed](#)]
77. Anderson, J.A.; Churchill, G.A.; Autrique, J.E.; Tanksley, S.D.; Sorrellis, M.E. Optimizing parental selection for genetic linkage maps. *Genome* **1993**, *36*, 181–186. [[CrossRef](#)] [[PubMed](#)]
78. XLSTAT. *Data Analysis and Statistical Solutions for Microsoft Excell*; Addinsoft: Paris, France, 2014.
79. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945–959. [[PubMed](#)]
80. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **2003**, *164*, 1567–1587. [[PubMed](#)]
81. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [[CrossRef](#)] [[PubMed](#)]
82. Dent, A.E.; Bridgett, M.V. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **2012**, *4*, 359–361. [[CrossRef](#)]
83. Liu, K.J.; Goodman, M.; Muse, S.; Smith, J.S.; Buckler, E.; Doebley, J. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **2003**, *165*, 2117–2128. [[PubMed](#)]
84. Wang, K.; Wang, Z.; Li, F.; Ye, W.; Wang, J.; Song, G.; Yue, Z.; Cong, L.; Shang, H.; Zhu, S.; et al. The draft genome of diploid *Gossypium raimondii*. *Nat. Genet.* **2012**, *44*, 1098–1103. [[CrossRef](#)] [[PubMed](#)]
85. Magwanga, R.O.; Lu, P.; Nyangasi Kirungu, J.; Diouf, L.; Dong, Q.; Hu, Y.; Cai, X.; Xu, Y.; Hou, Y.; Zhou, Z.; et al. GBS Mapping and Analysis of Genes Conserved between *Gossypium tomentosum* and *Gossypium hirsutum* Cotton Cultivars that Respond to Drought Stress at the Seedling Stage of the BC 2 F 2 Generation. *Int. J. Mol. Sci.* **2018**, *19*, 1614. [[CrossRef](#)] [[PubMed](#)]
86. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger database. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [[CrossRef](#)] [[PubMed](#)]
87. Priyam, A.; Woodcroft, B.J.; Rai, V.; Munagala, A.; Moghul, I.; Ter, F.; Gibbins, M.A.; Moon, H.K.; Leonard, G.; Rumpf, W.; et al. Sequenceserver: A modern graphical user interface for custom BLAST databases. *bioRxiv* **2015**. [[CrossRef](#)]
88. Saitou, N.; Nei, M. The neighbor joining method: A new method for reconstructing phylogenetic trees. *Mol. Bio. Evol.* **1987**, *4*, 406–425.
89. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **1985**, *39*, 783–791. [[CrossRef](#)] [[PubMed](#)]
90. Zuckerkandl, E.; Pauling, L. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*; Bryson, V., Vogel, H.J., Eds.; Academic Press: New York, NY, USA, 1965; pp. 97–166.

