


Article

Genomic Insight into Differentiation and Selection Sweeps in the Improvement of Upland Cotton

Mian Faisal Nazir ¹, Yinhua Jia ¹, Haris Ahmed ¹, Shoupu He ^{1,2},
Muhammad Shahid Iqbal ^{1,3}, Zareen Sarfraz ¹, Mushtaque Ali ¹, Chenfan Feng ⁴,
Irum Raza ¹, Gaofei Sun ¹, Zhaoe Pan ¹ and Xiongming Du ^{1,5,*}

¹ Institute of Cotton Research, Chinese Academy of Agricultural Sciences, State Key Laboratory of Cotton Biology, Anyang 455000, China; mfn121@hotmail.com (M.F.N.); jiyayinhua@caas.cn (Y.J.); hafizahmed25@hotmail.com (H.A.); heshoupu@caas.cn (S.H.); shahidkooria@gmail.com (M.S.I.); zskpbg@hotmail.com (Z.S.); mushtaqjan6@gmail.com (M.A.); irumkhattak@gmail.com (I.R.); 20160380@ayit.edu.cn (G.S.); panzhaoe@caas.cn (Z.P.)

² Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450001, China

³ Ayub Agriculture Research Institute, Cotton Research Institute, Multan 66000, Pakistan

⁴ Department of Electronics and Information Engineering, Sichuan University, Chengdu 610000, China; Chenfanfeng1983@gmail.com

⁵ School of Agricultural Sciences, Zhengzhou University, Zhengzhou 450001, China

* Correspondence: dujffrey8848@hotmail.com

Received: 23 April 2020; Accepted: 29 May 2020; Published: 3 June 2020



Abstract: Upland cotton is the most economically important fibre crop. The human-mediated selection has resulted in modern upland cultivars with higher yield and better fibre quality. However, changes in genome structure resulted from human-mediated selection are poorly understood. Comparative population genomics offers us tools to dissect the genetic history of domestication and helps to understand the genome-wide effects of human-mediated selection. Hereby, we report a comprehensive assessment of *Gossypium hirsutum* landraces, obsolete cultivars and modern cultivars based on high throughput genome-wide sequencing of the core set of genotypes. As a result of the genome-wide scan, we identified 93 differential regions and 311 selection sweeps associated with domestication and improvement. Furthermore, we performed genome-wide association studies to identify traits associated with the differential regions and selection sweeps. Our study provides a genetic basis to understand the domestication process in Chinese cotton cultivars. It also provides a comprehensive insight into changes in genome structure due to selection and improvement during the last century. We also identified multiple genome-wide associations (GWAS associations) for fibre yield, quality and other morphological characteristics.

Keywords: upland cotton; phylogeny; domestication; selection sweeps

1. Introduction

Cotton (*Gossypium hirsutum*) is a major source of fibre for the textile industry, especially tetraploid cotton which covers 95% of the worldwide cotton production. Selection has been carried out in cotton to improve production and adaptation to the local environment, reduced growth period, and its defence against biotic and abiotic factors. Due to continuous selection pressure, the cotton crop is facing a narrowed genetic base in terms of diversity [1–4]. Therefore, insight into the genomic structure and changes occurring in genomic structure due to continuous selection and improvement can yield interesting information resulting in a better understanding of the process of domestication and improvement.

Cotton has been grown in China for centuries. However, the introduction of upland cotton cultivars throughout the world has changed the production scenario of the cotton crop worldwide since upland cotton is occupying most of the production area of cotton. Prior to the introduction of upland cotton in China, a diploid species *Gossypium arboreum* was mainly grown in China. The earliest reported evidence suggested that the first introduction of tetraploid cotton into China was during the French colonist era [5,6]. The Second introduction of tetraploid cotton into China was at the beginning of the 20th century when upland cotton was systematically introduced into China. During these years, cotton was mainly distributed in the Yangtze River region and the Yellow River region in China [7].

In China, cotton has been produced in three main regions; Yellow River Valley (YeRV): Hebei, Shandong and Henan, northwestern China: Xinjiang province, The Yangtze River Valley (YaRV): Hubei, Hunan, Jiangsu, Anhui [8]. During recent years, cotton production has been mainly shifted to Xinjiang province. Prior to the introduction and implementation of Seed Law (SL) and Plant Variety Protection Act (PVPA) in China [8], most of the southwest varieties (introduced in early 20th century from the United States of America in the Yangtze River region and Yellow River region in China [7]) were maintained by farmers over the years. From this diverse southern gene pool, which is not only from different ecosystems but also has lower human intervention, useful information can be excavated to study the diversity of cotton crop in China through the evaluation of current Chinese varieties and their genetic background with reference to southern varieties and landraces from Central America.

Improvements in genomic studies and resequencing technologies have established the tools to dissect the genetic basis of elite cultivars. Different techniques have been used to understand genetic diversity in upland cotton [2,9–14], i.e., pedigree breeding, morphological and biochemical markers, and molecular markers. In recent years, genome-wide association studies (GWAS) have proven to be a remarkable tool to dissect genetic diversity among cultivars to understand the genetic mechanism behind diseases and associations of putative candidate genes for morphological traits. GWAS have been widely implemented in maize, rice, Arabidopsis and legumes. In addition, single nucleotide polymorphism (SNP) genotyping techniques, third-generation sequencing has facilitated GWAS to provide better association results for morphological traits with their genetic background. In cotton, GWAS have been used to dissect genetic mechanisms underlying fibre quality traits [15–18], diseases [19] and verticillium wilt [20,21]. However, despite all the technological breakthroughs in genomics, contribution towards finding new genetic sources in crop plants has been limited.

Though multiple studies have been conducted to understand the genomic basis of domestication in different crops [22–24], there are very few studies addressing the domestication of the cotton crop with reference to high-density genomic data. Our study aims to provide a better insight into the changes in genomic structure due to human-mediated selection and improvement. It is worth mentioning that three distinct groups of genotypes viz. (i) Modern cultivars (mainly cultivated in YeRV and Xinjiang province), (ii) Obsolete cultivars previously grown in South China for more than 50 years without management, (iii) other identified *Gossypium hirsutum* landraces collected from North America, used in this study are the representation of the breeding history of cotton crop in China over the decades since the introduction of upland cotton in China. This study will provide insight into the differences between the genetic profile of current varieties, accessions collected from southwest China and geographical landraces of *G. hirsutum*, which could be further utilised in breeding to expand the narrowed genetic base of upland cotton.

2. Results

2.1. Population Classification and Structure Variations

We exploited the genetic relationship among all accessions using principle component analysis (PCA) and performed phylogenetic analysis to construct a phylogenetic tree using 4,329,838 single nucleotide polymorphisms (SNPs). The inference drawn from phylogenetic analysis, structure and PCA supported the classification into three groups (Figure 1). Group 1 comprised modern cultivars

(MCI), while group 2 and group 3 were biased towards obsolete cultivars from South China (OCI) and geographical landraces of *G. hirsutum* (GHL), respectively. Some accessions showed admixed ancestry suggesting the presence of introgression or gene flow during the breeding process (Supplementary Table S2). Phylogenetic tree complimented breeding history of *G. hirsutum* in China. In addition, linkage disequilibrium (LD) decay was measured as the physical distance on the chromosome (kb) when LD decreased to half of its maximum value. Linkage disequilibrium is critical in understanding and determining the location of causal loci through GWAS [25]. Furthermore, patterns of LD decay between different populations can present with the information regarding selective sweeps and selective pressure [26]. LD decay was observed at 357 kb (physical distance between SNPs) for MCI, while it was lower at 0.2 kb and 0.05 kb for OCI and GHL, respectively. These results indicated the linkage decay in the subpopulations of obsolete cultivars (OCI) and landraces of *G. hirsutum* (GHL) declined dramatically compared with that in modern cultivar (MCI) populations, which are in agreement with previously published statistics [27]. Furthermore, the extent of LD decay was higher in the cultivated group than in obsolete accessions and landraces, signifying the potential role of selection pressure, geneflow and nonrandom mating in shaping modern cultivars.

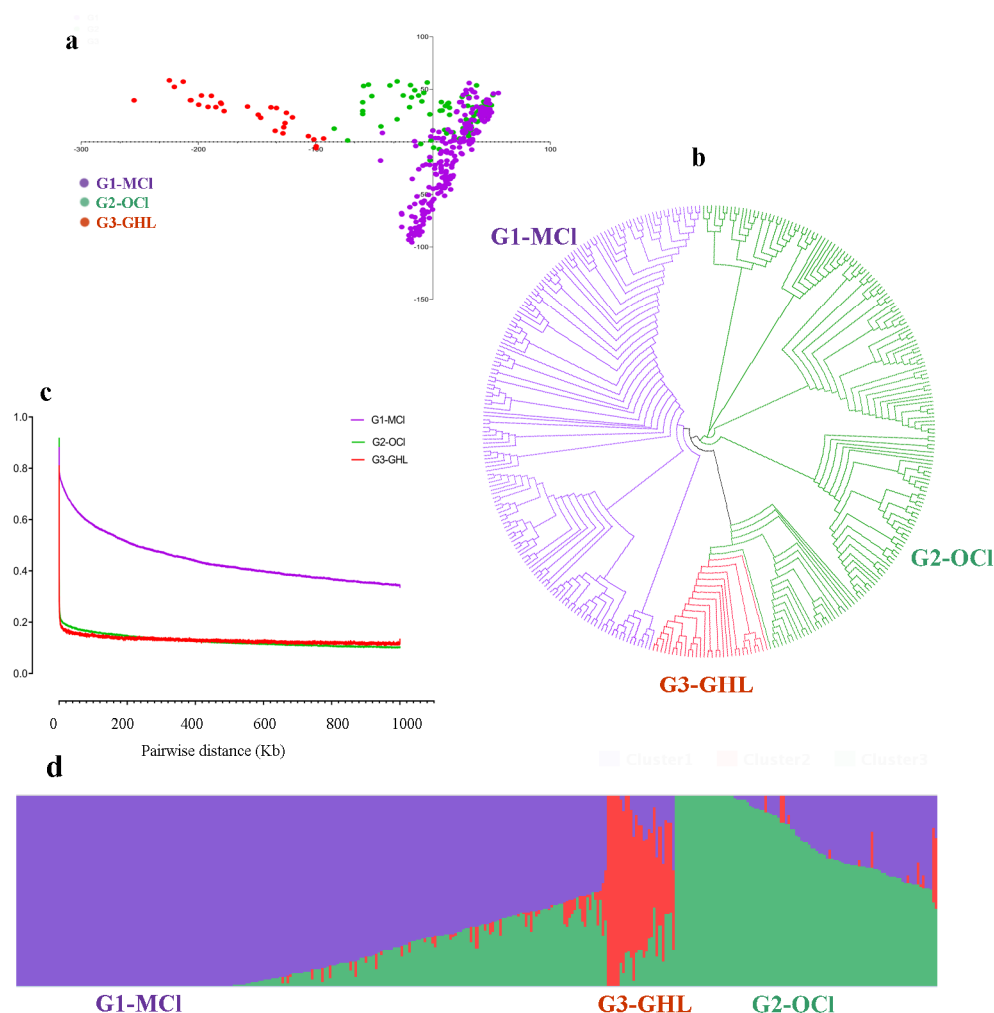


Figure 1. Population stratification. (a) Principal component analysis (PCA) plot of the first two PCAs, i.e., PCA1 (21.898%) and PCA2 (2.988%), for all accessions. Dot colour scheme is as G1-MCI = Modern cultivars G2-OCI = Obsolete Cultivars collected from south china, G3-GHL = Geographical landraces of *G. hirsutum*, (b) Phylogenetic tree constructed using whole-genome data, distributing genotypes into three clades as per original classification, (c) Pairwise linkage disequilibrium (LD) decay in each group, (d) Structure results for k-3.

2.2. Differentiation and Selection Signals between MCI and OCI Group

Modern upland cultivars have been developed from limited resources [18] and are spread worldwide in cotton-growing countries. Obsolete cultivars collected from Southwest China refer to the first systematic introduction of upland cotton in China. This distinct gene-pool, adapted to the local ecosystem, was mainly maintained by farmers without any organised breeding techniques. At the beginning of the 21st century, modern upland cultivars were systematically introduced worldwide, including China. As a result of rigorous selection, these cultivars referred to high yield and good quality. To understand this selection procedure and changes in genome structure due to continuous selection, we identified selection sweeps regions by comparing the genetic background of OCI, MCI and GHL. Our results pointed out differential selection patterns between different sub-populations, i.e., from landraces to OCI and from OCI to MCI.

First, the genetic differentiation between modern cultivars (MCI) and obsolete cultivars (OCI) was estimated (Figures 2a and 3a). Population fixation statistics (F_{st}) estimates resulted in understanding the differentiation between two groups of cultivars. Comparatively higher differentiation was associated with chromosomes A06, A08, A09 and A11 on sub-genome At (Figure 1a), while relatively high differentiation was estimated for D-genome on chromosomes D03, D04, D06, D07, D08, D10 and D11. Furthermore, we selected the top 5% threshold to select highly differentiated regions between the two groups of cultivars. With the threshold of $F_{st} > 0.2975$, we identified 193 highly differentiated regions. Among these regions, 103 reside on At sub-genome, and 90 reside on Dt sub-genome (Supplementary Table S3). Selection bottlenecks resulted in the loss of genetic diversity and depletion of elite alleles conferring favourable phenotypes in crop plants. To identify regions potentially associated with selection pressure for improvement, we scanned genomic regions with the highest reduction in diversity in modern cultivars and obsolete accessions. Furthermore, we selected a genome-wide top 1% threshold of reduction of diversity (ROD) $\pi_{OCI}/\pi_{MCI} > 10.425$ and categorise these regions as selection signals (Supplementary Table S4). A total of 311 improvement signals were identified, while 235 regions were located on At sub-genome, and the remaining 76 were located on the Dt sub-genome (Supplementary Table S4). The identified selection signals were compared with previously published reports. Some of the signals overlapped with previously reported QTLs for fibre yield and fibre quality. Contrary to previous reports, we found multiple hotspots for selection pressure on chromosome A02, A06, A11 on At sub-genome (Figure 2), D02, D10, D11, D12 and D13 on Dt sub-genome (Figure 3). Besides, we mapped the GWAS results of multiple traits with selection signals (Figures 2c and 3c).

2.3. Differentiation and Domestication between *G. hirsutum* Landraces and Cultivar Groups

To understand the differentiation and domestication between *G. hirsutum* landraces and cultivar groups, we compared the landraces of *G. hirsutum* with modern cultivars and obsolete accessions. Genome-wide population fixation statistics (F_{st}) suggested a wide range of genetic differentiation among these groups. *G. hirsutum* landraces showed significantly higher differentiation than modern cultivars and other obsolete accession collected from Southwest China, which is consistent with the breeding history of cotton (Figure 4). At sub-genome showed relatively higher differentiation as compare to Dt sub-genome (Figure 4a,d). Chromosome A03, A04, A12, A08, D05 and D08 showed less differentiation when compared with GHL, suggesting the conserved nature of these regions on the respective chromosomes. Furthermore, chromosomes A01, A02, A05, A06, A07, A10, D03, D04, D09, D10 and D11 depicted higher differentiation, suggesting the accumulation of changes due to adaptation, selection and improvement during the past few decades. We also investigated the diversity ratio to understand the genome-wide selection during domestication from landraces to modern cultivars. A large number of selection signals were identified genome-wide. Modern cultivars (MCI), as compared to Obsolete accessions from Southwest China (OCI) showed higher peaks representing selection sweeps, which is consistent with the breeding history of cotton in China and also emphasises the fact that modern cultivars are the result of rigorous selection over the period of time. Chromosome A06 showed significant selection signals, while the same region on chromosome

A06 showed a comparative differentiation among three groups; these results emphasised on change in genomic structure due to selection and improvement. A similar pattern of variation was observed in chromosome A13. Dt sub-genome also showed considerable selection sweep signals genome-wide. A similar pattern of selection sweeps was observed in At sub-genome, where modern cultivars showed higher selection peaks than obsolete cultivars.

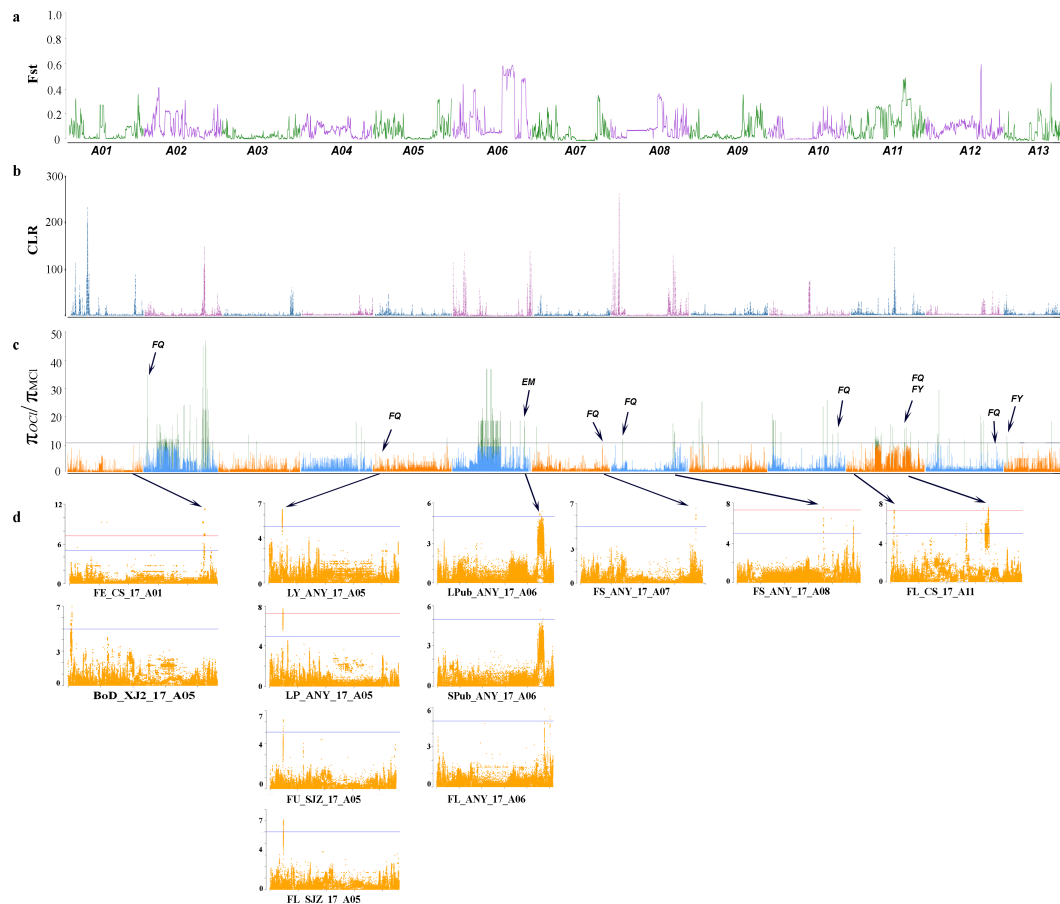


Figure 2. Genetic differentiation and identification of selection sweeps between OCI and MCI on At sub-genome (a) Population fixation statistics (F_{st}) for At sub-genome (Chr A01–A13). A threshold of top 1% is selected as highly differentiated regions, (b) CLR (Composite likelihood ratio) score as an estimate of genome-wide selection sweeps, (c) π_{OCI}/π_{MCI} values (genetic diversity in the cultivated group as compared to obsolete cultivars from Southwest China) for A genome (Chr A01–A13). π ratio was calculated using whole-genome data with a 100 kb sliding window. The horizontal dotted line represents the threshold of 1% values, whereas the threshold is represented with green columns. The annotations represent as FQ = Fibre quality [28–32], FY = Fibre yield [30] and EM = Early maturity [31,33,34] which donates to previously identified hotspots/quantitative trait loci (QTLs) on the corresponding location, (d) Genome-wide association studies' (GWAS) results as Manhattan plots of multiple traits where purple horizontal line represents suggestive significant threshold with $-\log_{10}(1 \times 10^{-5})$, and the red horizontal line represents genome-wide significant threshold with $-\log_{10}(5 \times 10^{-8})$. FE = Fibre elongation, BoD = Boll opening Days, LY = Lint yield, LP = Lint percentage, FU = Fibre length uniformity, FL = Fibre length, LPub = Leaf pubescence, SPub = Stem pubescence, FS = Fibre strength. ANY = Anyang, Henan Province, CS = Changsha, Hunan Province, SJZ = Shijiazhuang, Hebei Province.

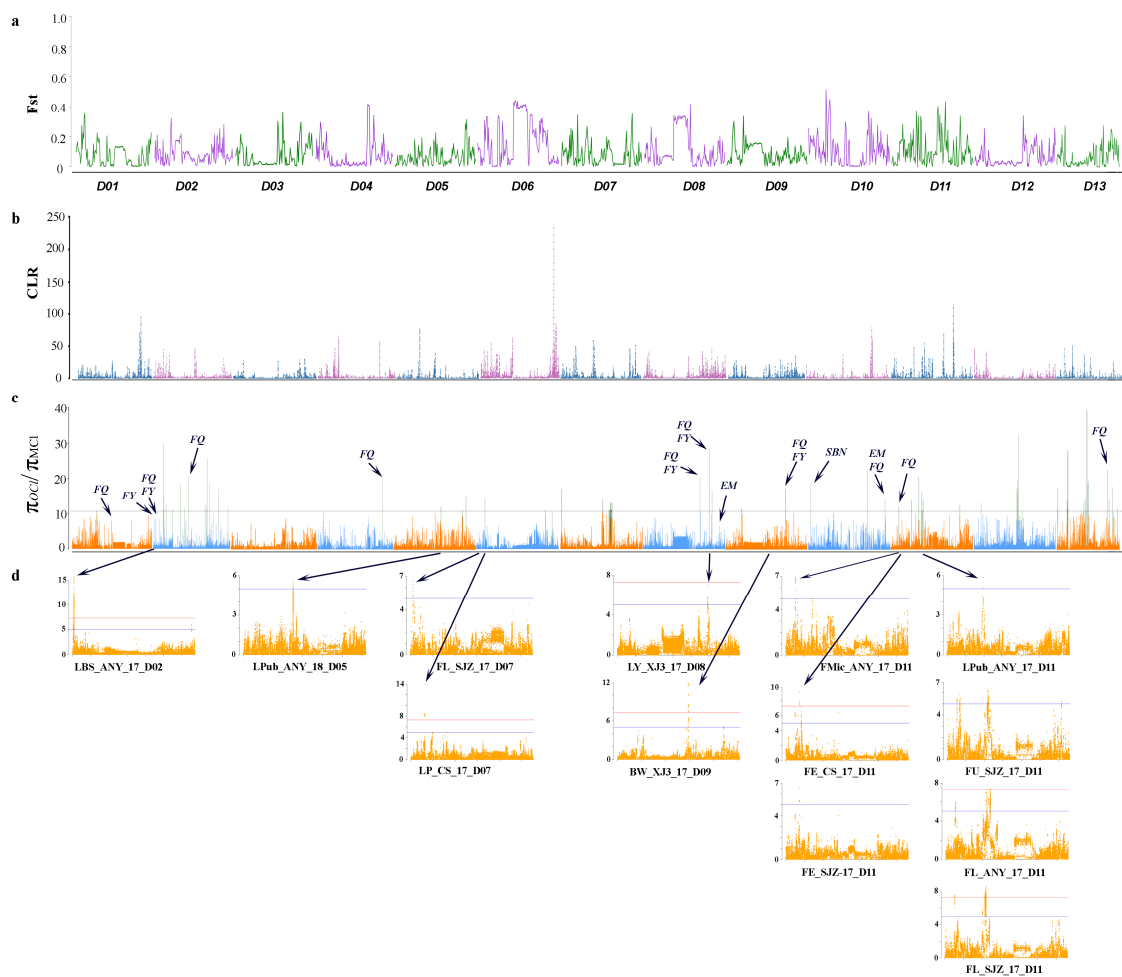


Figure 3. Genetic differentiation and identification of selection sweeps between OCI and MCI on Dt sub-genome (a) Population divergence (F_{st}) for Dt sub-genome (Chr D01–D13). A threshold of top 1% is selected as highly differentiated regions, (b) CLR (Composite likelihood ratio) score as an estimate of genome-wide selection sweeps, (c) π_{OCI}/π_{MCI} values (genetic diversity in the cultivated group as compare to obsolete cultivars from Southwest China) for D genome (Chr D01–D13). π ratio was calculated using whole-genome data with a 100 kb sliding window. The horizontal dotted line represents the threshold of top 1% values, whereas the threshold is represented with green columns. The annotations represent FQ = Fibre quality [28–32], FY = Fibre yield [30], SBN = Sympodial branch node and EM = Early maturity [31,33,34] which donates to previously identified hotspots/QTLs on the corresponding location, (d) GWAS results as Manhattan plots of multiple traits. LBS = Leaf base spot, BW = Boll weight, FMic = Fibre micronair, FE = Fibre elongation, LY = Lint yield, LP = Lint percentage, FU = Fibre length uniformity, FL = Fibre length, LPub = Leaf pubescence. ANY = Anyang, Henan Province, CS = Changsha, Hunan Province, SJZ = Shijiazhuang, Hebei Province, XJ3 = Shihezi, Xinjiang Province.



Figure 4. Genetic differentiation and selection signals among *G. hirsutum* landraces (GHL), modern cultivars (MCI) and Obsolete cultivars (OCI) (a) Population divergence (F_{st}) for At sub-genome (Chr A01–A13), (b) π_{GHL}/π_{MCI} (Purple columns) and π_{GHL}/π_{OCI} (Green columns) values (genetic diversity in *G. hirsutum* landraces (GHL) as compared to MCI and OCI for At sub-genome (Chr A01–A13), (c) Population divergence (F_{st}) for Dt sub-genome (Chr D01–D13), (d) π_{GHL}/π_{MCI} (Purple columns) and π_{GHL}/π_{OCI} (Green columns) values (genetic diversity in *G. hirsutum* landraces (GHL) as compare to MCI and OCI for Dt sub-genome (Chr D01–D13).

2.4. GWAS

Early maturity and improvement in fibre quality have been major objectives of breeding projects during the last century. To identify putative candidate genes for fibre yield, fibre quality and flowering time, we conducted a genome-wide association study (GWAS) using phenotypic data collected in 2017 and 2018. We selected 4,329,838 high-quality SNPs with minor allele frequency >0.05 . The high-density map was found to be superior to previous reports [19,27]. A total of 25 association signals were identified with $p < 4.9 \times 10^{-7}$ by using efficient mixed-model association expedited (EMMAX) (Supplementary Figure S1a–d). Very few of these associations have been previously characterised. We identified significant GWAS signals for fibre yield on chromosomes A05, D06, D08 and D09. These GWAS signals are also associated with significant improvement signals present on the respective chromosomes. We also identified 16 significant GWAS associations for fibre quality traits viz. fibre length (FL), fibre elongation (FE), fibre length uniformity (FU), fibre strength (FS) and fibre micronaire (FMic). These signals were present on chromosome A01 (FE), A05 (lint percentage (LP), FU, FL), A06 (FL), A07 (FS), A08 (FS), A09 (FL) on At sub-genome, while D06 (FL), D11 (FMic, FE, FU and FL) on Dt sub-genome. These identified signals, key SNPs, and their corresponding annotation have been presented in supplementary Table S5.

3. Discussion

The *Gossypium* genus includes 50 species distributed worldwide in tropical and subtropical regions. *Gossypium hirsutum* is the most important species among all due to its high yield and spinnable fibre

quality for industrial use [35]. Upland cotton, a type of *G. hirsutum*, dominates worldwide and has been the primary source of fibre production, as it has been growing on more than 90% of cotton-growing areas. Upland cotton originated in Mesoamerica, from where it spread worldwide through trade routes during the 18th century. In the early 20th-century, upland cotton was systematically introduced to cotton-growing countries worldwide, including China, India and Australia, which lead to the development of the modern cotton industry. Since upland cotton was developed from limited resources [36,37], it is considered that subsequent introduction and spread of upland cotton worldwide has reduced its genetic diversity. Reduction in diversity can negatively influence the development of superior crop varieties [38,39]. A comparison of two periods of the introduction of upland cotton in China can provide insight into the process of domestication and improvement in cultivars and changes in genomic structure. Thus, we evaluated three groups of genotypes, including modern cultivars, obsolete accessions collected from Southwest China and geographical landraces of *G. hirsutum*.

Phylogenetic, principal component and structure analyses indicated the divergent behaviour of modern cultivars with comparison to OCI and GHL, which were in agreement with genetic differentiation analysis. The divergent trends of landraces compared to modern cultivars are also in agreement with previous studies [18,27,38,40–42]. In support of previously published works, which have suggested narrowed genetic diversity among studied cultivars of upland cotton and also in other crops [9,27,43], our results also emphasised reduced differentiation on a genetic level corresponding to modern cultivars. Chen et al. [37] reported genetic diversity in source germplasm comprising of 43 upland cotton accession using simple sequence repeats (SSR) markers and concluded a decrease in genetic diversity in modern cultivars. Genetic bottlenecks in crop domestication may have resulted in the loss of genetic diversity and elite alleles in modern cultivars [44,45]. However, wild progenitors and landraces are excellent sources for developing desirable variations in current cultivars [46].

Furthermore, Obsolete accessions collected from southwest China are a rich source of genetic information for comparison of genetic variation in modern cultivars because of domestication and improvement. These genotypes comprised a distinct gene-pool, which is not only from a different ecosystem but also with less systematic selection. A comparison of these accessions with modern cultivars can provide insight into changes in genomic structure due to human-mediated selection. Therefore, we analysed these accessions and compared them with MCI and GHL. Our results suggested a marked differentiation between OCI, MCI and GHL. A comparison of GHL with OCI showed lower differentiation as compare to MCI. The divergent behaviour of geographical landraces of *G. hirsutum* was in accordance with the genetic differentiation analyses. This divergent trend of landraces compared to modern cultivars and obsolete accessions was also found in agreement with previous studies [18,27,38,40–42]. Furthermore, selection pressure as improvement/selection signals in obsolete cultivars (OCI) was lower than modern cultivars (MCI). These results are consistent with the breeding history of upland cotton in China [37].

Besides, we identified a large number of selection sweeps, suggesting the domestication bottlenecks. Some of the identified selection sweeps overlapped with highly differentiated regions on respective chromosomes, i.e., A06, A08, A09, A10, A11, D02, D04, D10, D12 and D13. This overlapping pattern suggests the occurrence of differentiation due to human-mediated selection. Further, to understand the genetic basis of domestication and improvement in fibre yield, fibre quality, maturity and other morphological traits, we compared the location of selection sweeps with the significant loci of GWAS analysis and narrowed down selection sweeps into corresponding small regions which will be helpful for future studies to determine and characterise new genes concerning domestication and selection in upland cotton. Some of the selection sweeps we identified have been previously reported for fibre yield [30], fibre quality [28–32] and maturity [31,33,34]. Fang et al. [18] performed a comprehensive experiment for identification of selection signature in 318 upland cotton accession and consequently identified 15 regions associated with improvement through comparison of whole-genome diversity between modern cultivars and landraces. However, with improved sequencing technology, our study

resulted in a better understanding of selection/improvement signals. Improvement signals ($\pi_{\text{OCI}}/\pi_{\text{MCI}}$) were lower than selection signals ($\pi_{\text{GHI}}/\pi_{\text{MCI}}$), suggesting a weaker selection pattern during modern genetic improvement than earlier selection [47,48]. These results can lead us to understand the changes at the genomic level caused by domestication, selection and improvement of upland cotton cultivars. Modern sequencing technology and GWAS has enabled us to better understand the genetic mechanisms behind the evolution of a specific trait, as previously described in different crops [40,49,50]. In this study, we identified multiple GWAS signals significantly associated with different traits, including fibre yield and quality. Moreover, loci associated with fibre yield (chromosome A05, D06 and D08), fibre quality (chromosome A01, A05, A07, A08, A11, D01, D07, D08 and D11) and other morphological traits (chromosome A01, A06, D02, D05 and D11) fall within the selection sweeps, and these loci have not been previously reported. Genotyping for these traits and identification of candidate genes and their functional analysis can reveal the potential impact of genes related to traits.

4. Materials and Methods

4.1. Plant Material

A total of 357 upland cotton accessions obtained from the gene bank of the Cotton Research Institute of the Chinese Academy of Agricultural Sciences (CRI-CAAS) with diverse genetic backgrounds were used for phenotyping. These accessions comprised three groups, i.e., group 1 belonged to modern cultivars (235) currently being cultivated, group two comprised 91 obsolete cultivars collected from southwest China and group 3 (31) comprised seven reported geographical landraces of *G. hirsutum*, i.e., Yucatanese, richmondi, morrilli, Marie-Galante, palmeri, punctatum and latifolium (Group 3 was not included in phenotyping as these landraces cannot flower in the test locations due to photoperiod sensitivity) (Supplementary Table S1). Two replications were planted in five agro-ecologically different environments viz. Shijiazhuang (SJZ) in Hebei Province, Changsha (CS) in Hunan province, Anyang (AY) in Henan Province (Yellow River region), Alaer and Shihezi in the Xinjiang (XJ) autonomous region (Northwest Inland), for two consecutive seasons 2017 and 2018. Two sets of genotypes were used for phenotyping Set 1 comprised 169 accessions whose phenotypic data was collected from SJZ, CS, AY, XJ2 and XJ3, Set 2 comprised 324 accessions whose phenotypic data was collected from AY. Some of the genotypes in both sets overlapped to give a proper representation of two groups viz. G1-MCI = Modern cultivars G2-OCI = Obsolete Cultivars. All standard field management practices were applied, including irrigation, pest management and fertilisation, following the usual local management practices in each test location. The cotton was sown in mid- to late-April and was harvested in mid- to late-October at all locations.

In all test locations, phenotypic traits were recorded following the same scoring standard. We characterised lint yield (LY), lint percentage (LP), fibre quality (fibre length (FL, mm), fibre length uniformity (FLU, %), fibre micronaire (FMic), fibre strength (FS), fibre elongation (FE, %)), flowering time (DF, days), boll opening days (BoD, days), leaf pubescence (LPub) and stem pubescence (SPub). Fibre quality was tested using twenty naturally opened balls from each accession. A High-volume instrument (HFT9000) was used for characterizing fibre quality parameters at the Cotton Quality Testing Center in Anyang, China. Flowering time was observed daily, and days to flowering (DF) were calculated from the sowing day to the day that the first flowers appeared in 50% of the plants. All samples were subjected to the High-volume instrument (HFT9000) for the estimation of quality parameters.

4.2. DNA Extraction, Sequencing, Alignment and SNP Detection

Total genomic DNA was extracted from the seedlings of five cultured seeds of each accession in a growth chamber. After three weeks of sowing, at the true leaf stage, young leaves were collected, and a Plant DNA Mini Kit (Aidlab Biotech, Beijing, China) was used to extract total genomic DNA. Three hundred and fifty base pair whole-genome libraries were constructed for each accession according to the manufacturer's specifications (Novogene Bioinformatics technology company, Beijing, China).

Subsequently, we used Illumina HiSeq X10 by a commercial service “Novogene” platform to generate 6.45-Tb raw sequences with 150-bp read length. Following alignment of high-quality reads with the genome of *G. hirsutum*, GATK (Genome Analysis Toolkit, version v3.1) was used for SNP calling. Sequencing data for G4; GH1 was obtained from published data [18].

4.3. Population Genetic Analyses

Population structure was studied using ADMIXTURE [51], which utilises a clustering method (mode-based) to draw population structure assuming different numbers of clusters (K). A total of 431,985 SNPs without missing genotypes were used. SNPhylo software was used to prune SNPs, which reduces SNP redundancy by linkage disequilibrium (LD). SNPs in the same LD block provide redundant lineage information. SNPhylo keeps only one informative SNP in a LD block, and subsequently, a relatively small number of SNPs (9.97%) were used for structure and phylogenetic analysis.

Principal component analysis was performed using the EIGENSOFT package with an embedded SMARTPCA program [52] using 4,329,838 SNPs without missing genotypes.

Phylogenetic analysis was performed to understand phylogenetic relationships among genotypes by constructing a phylogenetic tree using SNPs of all genotypes. SNPs were filtered with minor allele frequency, MAF = 0.05. Subsequently, a neighbour-joining tree was constructed using the maximum likelihood method with SNPhylo software [53]. To visualise the phylogenetic tree, we used Dendroscope.

4.4. Identification of Selection/Improvement Signals

The fixation index (F_{st}) is a measure of population differentiation as it provides insight into the genome-wide differentiation among different groups. Thus, we calculated population fixation statistics (F_{st}) using vcfTools with a sliding window of 100 kb and step size of 20 kb ($-fst-window-size$ 100,000 $-fst-window-step$ 20,000). The average F_{st} of all sliding windows was considered as the value at the whole genome level among different groups.

Highly diverged regions were selected by merging fragments with a distance of less than 50 kb after the initial selection of the top 1% π values. To identify the putative regions under selective pressure between landraces and cultivars, the nonsynonymous SNPs with the top 1% of F_{st} values were selected.

Nucleotide diversity (π) is an estimate of the degree of variability within population and species [54]. Nucleotide diversity was calculated using vcfTools with a 100 kb sliding window based on genotypes in different groups separately. Furthermore, genetic diversity ratios between different groups were calculated to estimate selection/improvement regions. π_{OCI}/π_{MCI} was used as an estimate of improvement signals, while π_{GHL}/π_{MCI} and π_{GHL}/π_{OCI} were used as an estimate of selection signals. The top 1% threshold was used to identify significant selection/improvement signals. Composite likelihood ratio (CLR) was calculated as an alternative estimate for selection/improvement signals, using SweepD. Diversity ratios and CLR scores were compared for better assessment of selection/improvement signals.

4.5. GWAS Analysis

For GWAS analysis, we categorised genotypes into two sets. Set 1 comprised 169 accessions whose phenotypic data were collected from SJZ, CS, AY, XJ2 and XJ3, Set 2 comprised 324 accessions whose phenotypic data were collected from AY. A total of 4,329,838 high-quality SNPs were subjected to filtering with MAF >0.05, missing rate <20% and 1,604,221, and 1,506,091 SNPs were kept in set 1 and set 2, respectively, and subsequently, GWAS was performed on both sets of genotypes separately. Accessions with missing SNPs data were excluded from analyses. We performed GWAS for multiple traits in efficient mixed-model association expedited (EMMAX) software [55,56]. Population stratification and hidden relatedness were modelled with a kinship (K) matrix in the emmax-kin-intel package

of EMMAX. The significant threshold for GWAS was kept constant with the suggestive significant threshold at $-\log_{10}(1 \times 10^{-5})$ and the genome-wide significant threshold at $-\log_{10}(5 \times 10^{-8})$.

5. Conclusions

Our study provides a genetic basis to understand the domestication process in upland cotton cultivars. It also provides a comprehensive insight into changes in genome structure due to selection and improvement during the last century. We also identified multiple GWAS associations for fibre yield, quality and other morphological characteristics. Further study is required to explore these novel loci associated with different traits to uncover causal genes related to these traits. Our study provides a comprehensive insight into the differentiation between modern cultivars, OCl and GH, which can be a useful tool for the cotton breeders to understand changes accumulated due to selection and improvement breeding strategies.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2223-7747/9/6/711/s1>, Figure S1 (a) Manhattan plots for GWAS results, (b) Manhattan plots for GWAS results, (c) Manhattan plots for GWAS results, (d) Manhattan plots for GWAS results; Table S1 list of accessions, cultivars and landraces used in this study; Table S2 Structure results and corresponding grouping for all accessions and landraces used in this study; Table S3 List of highly differentiated regions and genes located in the highly differentiated regions identified through pairwise *Fst* estimates (top 1% threshold); Table S4 Improvement signals, Top 1% threshold of π_{OCl}/π_{MCl} ; Table S5 Key SNPs identified for flowering time, fibre quality and anther colour, and their corresponding annotations.

Author Contributions: Contribution of authors is as follows Conceptualisation, X.D. and M.F.N.; methodology, M.F.N., Y.J., H.A.; software, S.H., G.S., C.F.; validation, M.F.N., S.H., H.A.; I.R. formal analysis and investigation, M.F.N., M.A., Z.S., I.R.; resources, Z.P.; data curation, M.F.N., G.S., M.S.I., S.H., C.F.; writing—original draft preparation, M.F.N.; writing—review and editing, X.D., M.F.N., S.H., M.S.I.; visualisation, M.F.N.; supervision, X.D.; project administration, X.D., Z.P.; funding acquisition, X.D., Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by funding from the National Key Technology R&D Program, the Ministry of Science and Technology (2016YFD0100203, 2016YFD0102105) and Crop Germplasm Conservation program of Ministry of Agriculture (2015NWB039).

Acknowledgments: Special thanks to Muhammad Kausar Nawaz (PMAS AAUR) for proofreading the manuscript.

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

References

- 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](#)]
- Tyagi, P.; Gore, M.A.; Bowman, D.T.; Campbell, B.T.; Udall, J.A.; Kuraparthy, V. Genetic diversity and population structure in the US Upland cotton (*Gossypium hirsutum* L.). *Theor. Appl. Genet.* **2014**, *127*, 283–295. [[CrossRef](#)] [[PubMed](#)]
- Zhang, J.; Percy, R.G.; McCarty, J.C. Introgression genetics and breeding between Upland and Pima cotton: A review. *Euphytica* **2014**, *198*, 1–12. [[CrossRef](#)]
- Chen, Y.H.; Gols, R.; Benrey, B. Crop domestication and its impact on naturally selected trophic interactions. *Annu. Rev. Entomol.* **2015**, *60*, 35–58. [[CrossRef](#)] [[PubMed](#)]
- Watt, S.G. *The Wild and Cultivated Cotton of the World*; Longmans, Green and Co. London: London, UK, 1907; p. 538.
- Watt, G. *Gossypium*. *Bull. Misc. Inf.* **1927**, *8*, 321–356. [[CrossRef](#)]
- He, S.; Sun, G.; Huang, L.; Yang, D.; Dai, P.; Zhou, D.; Wu, Y.; Ma, X.; Du, X.; Wei, S.; et al. Genomic divergence in cotton germplasm related to maturity and heterosis. *J. Integr. Plant. Biol.* **2018**. [[CrossRef](#)]
- Fok, M.; Xu, N. State and market interaction: Cotton variety and seed market development in China. In Proceedings of the ISSCRI International Conference “Rationales and evolutions of cotton policies”, Montpellier, France, 13–17 May 2008.
- 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](#)]

10. May, O.L.; Bowman, D.T.; Calhoun, D.S. Genetic Diversity of U.S. Upland Cotton Cultivars Released between 1980 and 1990. *Crop Sci.* **1995**, *35*, 1570–1574. [[CrossRef](#)]
11. Esbroeck, G.V.; Bowman, D.T. Cotton Germplasm Diversity and Its Importance to Cultivar Development. *J. Cotton Sci.* **1998**, *2*, 121–129.
12. Yu, J.Z.; Fang, D.D.; Kohel, R.J.; Ulloa, M.; Hinze, L.L.; Percy, R.G.; Zhang, J.; Chee, P.; Scheffler, B.E.; Jones, D.C. Development of a core set of SSR markers for the characterization of *Gossypium* germplasm. *Euphytica* **2012**, *187*, 203–213. [[CrossRef](#)]
13. Kuang, M.; Wei, S.-j.; Wang, Y.-q.; Zhou, D.-y.; 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](#)]
14. Su, J.; Li, L.; Pang, C.; Wei, H.; Wang, C.; Song, M.; Wang, H.; Zhao, S.; Zhang, C.; Mao, G.; et al. Two genomic regions associated with fiber quality traits in Chinese upland cotton under apparent breeding selection. *Sci. Rep.* **2016**, *6*, 38496. [[CrossRef](#)] [[PubMed](#)]
15. Cheng, F.; Sun, R.; Hou, X.; Zheng, H.; Zhang, F.; Zhang, Y.; Liu, B.; Liang, J.; Zhuang, M.; Liu, Y.; et al. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet.* **2016**, *48*, 1218–1224. [[CrossRef](#)] [[PubMed](#)]
16. Liu, R.; Gong, J.; Xiao, X.; Zhang, Z.; Li, J.; Liu, A.; Lu, Q.; Shang, H.; Shi, Y.; Ge, Q.; et al. GWAS Analysis and QTL Identification of Fiber Quality Traits and Yield Components in Upland Cotton Using Enriched High-Density SNP Markers. *Front. Plant Sci.* **2018**, *9*, 1067. [[CrossRef](#)]
17. Ma, Z.; He, S.; Wang, X.; Sun, J.; Zhang, Y.; Zhang, G.; Wu, L.; Li, Z.; Liu, Z.; Sun, G.; et al. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* **2018**, *50*, 803–813. [[CrossRef](#)]
18. Fang, L.; Wang, Q.; Hu, Y.; Jia, Y.; Chen, J.; Liu, B.; Zhang, Z.; Guan, X.; Chen, S.; Zhou, B. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **2017**, *49*, 1089. [[CrossRef](#)]
19. Huang, C.; Nie, X.H.; Shen, C. Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* **2017**, *15*. [[CrossRef](#)]
20. Du, X.; Huang, G.; He, S.; Yang, Z.; Sun, G.; Ma, X.; Li, N.; Zhang, X.; Sun, J.; Liu, M.; et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **2018**, *50*, 796–802. [[CrossRef](#)]
21. Gong, Q.; Yang, Z.; Chen, E.; Sun, G.; He, S.; Butt, H.I.; Zhang, C.; Zhang, X.; Yang, Z.; Du, X.; et al. A Phi-Class Glutathione S-Transferase Gene for Verticillium Wilt Resistance in *Gossypium arboreum* Identified in a Genome-Wide Association Study. *Plant Cell Physiol.* **2018**, *59*, 275–289. [[CrossRef](#)]
22. Shepherd, L.D.; de Lange, P.J.; Cox, S.; McLenachan, P.A.; Roskrug, N.R.; Lockhart, P.J. Evidence of a strong domestication bottleneck in the recently cultivated New Zealand endemic root crop, *Arthropodium cirratum* (Asparagaceae). *PLoS ONE* **2016**, *11*, e0204943. [[CrossRef](#)]
23. Wang, M.; Li, W.; Fang, C.; Xu, F.; Liu, Y.; Wang, Z.; Yang, R.; Zhang, M.; Liu, S.; Lu, S. Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* **2018**, *50*, 1435–1441. [[CrossRef](#)] [[PubMed](#)]
24. Lin, Z.; Li, X.; Shannon, L.M.; Yeh, C.-T.; Wang, M.L.; Bai, G.; Peng, Z.; Li, J.; Trick, H.N.; Clemente, T.E. Parallel domestication of the *Shattering1* genes in cereals. *Nat. Genet.* **2012**, *44*, 720. [[CrossRef](#)] [[PubMed](#)]
25. Gupta, P.K.; Rustgi, S.; Kulwal, P.L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol. Biol.* **2005**, *57*, 461–485. [[CrossRef](#)] [[PubMed](#)]
26. Kim, Y.; Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **2004**, *167*, 1513–1524. [[CrossRef](#)]
27. Wang, M.; Tu, L.; Lin, M.; Lin, Z.; Wang, P.; Yang, Q.; Ye, Z.; Shen, C.; Li, J.; Zhang, L.; et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **2017**, *49*, 579–587. [[CrossRef](#)] [[PubMed](#)]
28. Wang, B.; Guo, W.; Zhu, X. QTL mapping of fiber quality in an elite hybrid derived-RIL population of upland cotton. *Euphytica* **2006**, *152*. [[CrossRef](#)]

29. Wang, P.; Zhu, Y.; Song, X.; Cao, Z.; Ding, Y.; Liu, B.; Zhu, X.; Wang, S.; Guo, W.; Zhang, T. Inheritance of long staple fiber quality traits of *Gossypium barbadense* in *G. hirsutum* background using CSILs. *Theor. Appl. Genet.* **2012**, *124*. [[CrossRef](#)]
30. Liu, X.; Teng, Z.; Wang, J.; Wu, T.; Zhang, Z.; Deng, X.; Fang, X.; Tan, Z.; Ali, I.; Liu, D.; et al. Enriching an intraspecific genetic map and identifying QTL for fiber quality and yield component traits across multiple environments in Upland cotton (*Gossypium hirsutum* L.). *Mol. Genet. Genom.* **2017**, *292*. [[CrossRef](#)]
31. Liu, D.X.; Zhang, J.; Liu, X.Y.; Wang, W.W.; Liu, D.J.; Teng, Z.H.; Fang, X.M.; Tan, Z.Y.; Tang, S.Y.; Yang, J.H.; et al. Fine mapping and RNA-Seq unravels candidate genes for a major QTL controlling multiple fiber quality traits at the T1 region in upland cotton. *BMC Genom.* **2016**, *17*. [[CrossRef](#)]
32. Tang, S.; Teng, Z.; Zhai, T. Construction of genetic map and QTL analysis of fiber quality traits for upland cotton (*Gossypium hirsutum* L.). *Euphytica* **2015**, *201*. [[CrossRef](#)]
33. Ding, M.Q.; Ye, W.W.; Lin, L.F. The hairless stem phenotype of cotton (*Gossypium barbadense*) is linked to a copia-like retrotransposon insertion in a Homeodomain-Leucine Zipper Gene (HD1). *Genetics* **2015**, *201*. [[CrossRef](#)] [[PubMed](#)]
34. Niu, E.L.; Cai, C.P.; Bao, J.H.; Wu, S.; Zhao, L.; Guo, W.Z. Up-regulation of a homeodomain-leucine zipper gene HD-1 contributes to trichome initiation and development in cotton. *J. Integr. Agric.* **2018**, *17*. [[CrossRef](#)]
35. Ahmed, H.; Nazir, M.F.; Pan, Z.; Gong, W.; Iqbal, M.S.; He, S.; Du, X. Genotyping by Sequencing Revealed QTL Hotspots for Trichome-Based Plant Defense in *Gossypium hirsutum*. *Genes* **2020**, *11*, 368. [[CrossRef](#)] [[PubMed](#)]
36. Doebley, J.F.; Gaut, B.S.; Smith, B.D. The molecular genetics of crop domestication. *Cell* **2006**, *127*, 1309–1321. [[CrossRef](#)] [[PubMed](#)]
37. Chen, G.; Du, X.M. Genetic diversity of source germplasm of upland cotton in China as determined by SSR marker analysis. *Acta Genet. Sin.* **2006**, *33*. [[CrossRef](#)]
38. Zhou, Z.; Jiang, Y.; Wang, Z.; Gou, Z.; Lyu, J.; Li, W.; Yu, Y.; Shu, L.; Zhao, Y.; Ma, Y.; et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **2015**, *33*, 408–414. [[CrossRef](#)]
39. Tanksley, S.D.; McCouch, S.R. Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* **1997**, *277*, 1063–1066. [[CrossRef](#)]
40. Varshney, R.K.; Saxena, R.K.; Upadhyaya, H.D.; Khan, A.W.; Yu, Y.; Kim, C.; Rathore, A.; Kim, D.; Kim, J.; An, S.; et al. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* **2017**, *49*, 1082–1088. [[CrossRef](#)]
41. Qi, J.; Liu, X.; Shen, D.; Miao, H.; Xie, B.; Li, X.; Zeng, P.; Wang, S.; Shang, Y.; Gu, X.; et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **2013**, *45*, 1510–1515. [[CrossRef](#)]
42. Xu, X.; Liu, X.; Ge, S.; Jensen, J.D.; Hu, F.; Li, X.; Dong, Y.; Gutenkunst, R.N.; Fang, L.; Huang, L.; et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **2011**, *30*, 105–111. [[CrossRef](#)]
43. Rana, M.K.; Singh, V.P.; Bhat, K.V. Assessment of Genetic Diversity in Upland Cotton (*Gossypium hirsutum* L.) Breeding Lines by using Amplified Fragment Length Polymorphism (AFLP) Markers and Morphological Characteristics. *Genet. Res. Crop Evol.* **2005**, *52*, 989–997. [[CrossRef](#)]
44. Voss-Fels, K.P.; Stahl, A.; Hickey, L.T. Q&A: Modern crop breeding for future food security. *BMC Biol.* **2019**, *17*, 18. [[CrossRef](#)]
45. Fu, Y.-B. Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet.* **2015**, *128*, 2131–2142. [[CrossRef](#)] [[PubMed](#)]
46. Huang, X.; Han, B. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant. Biol.* **2014**, *65*, 531–551. [[CrossRef](#)] [[PubMed](#)]
47. Hufford, M.B.; Xu, X.; Van Heerwaarden, J.; Pyhäjärvi, T.; Chia, J.-M.; Cartwright, R.A.; Elshire, R.J.; Glaubitz, J.C.; Guill, K.E.; Kaeppeler, S.M. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **2012**, *44*, 808–811. [[CrossRef](#)] [[PubMed](#)]
48. Jia, G.; Huang, X.; Zhi, H.; Zhao, Y.; Zhao, Q.; Li, W.; Chai, Y.; Yang, L.; Liu, K.; Lu, H. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **2013**, *45*, 957–961. [[CrossRef](#)]

49. Zhang, C.; Zhao, X.; Qu, Y.; Teng, W.; Qiu, L.; Zheng, H.; Wang, Z.; Han, Y.; Li, W. Loci and candidate genes in soybean that confer resistance to *Fusarium graminearum*. *Theor. Appl. Genet.* **2019**, *132*, 431–441. [[CrossRef](#)]
50. Zhai, S.; Liu, J.; Xu, D.; Wen, W.; Yan, J.; Zhang, P.; Wan, Y.; Cao, S.; Hao, Y.; Xia, X.; et al. A Genome-Wide Association Study Reveals a Rich Genetic Architecture of Flour Color-Related Traits in Bread Wheat. *Front. Plant Sci.* **2018**, *9*, 1136. [[CrossRef](#)]
51. Alexander, D.H.; Novembre, J.; Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **2009**, *19*, 1655–1664. [[CrossRef](#)]
52. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. [[CrossRef](#)]
53. Lee1, T.-H.; Guo, H.; Wang, X.; Kim, C.; Paterson, A.H. SNPhylo, a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genom.* **2014**, *15*, 1471–2164. [[CrossRef](#)] [[PubMed](#)]
54. Tajima, F. Evolutionary relationship of DNA sequence in finite populations. *Genetics* **1983**, *105*, 437–460. [[PubMed](#)]
55. Kang, H.M. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **2010**, *42*. [[CrossRef](#)] [[PubMed](#)]
56. Li, M.X.; Yeung, J.M.Y.; Cherny, S.S.; Sham, P.C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **2012**, *131*. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).