Research Article

Genomic divergence in cotton germplasm related to maturity and heterosis

Running title: cotton maturity and heterosis

Shoupu He¹,²,†, Gaofei Sun¹,³,†, Longyu Huang²,†, Daigang Yang², Panhong Dai², Dayun Zhou², Yuzhen Wu², Xiongfeng Ma², Xiongming Du¹,², Shoujun Wei²,*, Jun Peng²,* and Meng Kuang¹,²,*

1. Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang, China
2. Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China
3. Department of Computer Science and Information Engineering/Data Mining Institute, Anyang Institute of Technology, Anyang, China

†These authors contributed equally to the work.

*Correspondances: Meng Kuang (kuangmeng007@163.com, Dr. Kuang is fully responsible for the distribution of all materials associated with this article); Jun Peng (jun_peng@126.com); Shoujun Wei (weisj@cricaas.com.cn)

Edited by: Jianbing Yan, Huazhong Agricultural University, China
Abstract

Commercial varieties of upland cotton (*Gossypium hirsutum*) have undergone extensive breeding for agronomic traits, such as fiber quality, disease resistance, and yield. Cotton breeding programs have widely used Chinese upland cotton source germplasm (CUCSG) with excellent agronomic traits. A better understanding of the genetic diversity and genomic characteristics of these accessions could accelerate the identification of desirable alleles. Here, we analyzed 10,522 high-quality single-nucleotide polymorphisms (SNP) with the CottonSNP63K microarray in 137 cotton accessions (including 12 hybrids of upland cotton). These data were used to investigate the genetic diversity, population structure, and genomic characteristics of each population and the contribution of these loci to heterosis. Three subgroups were identified, in agreement with their known pedigrees, geographical distributions, and times since introduction. For each group, we identified lineage-specific genomic divergence regions, which potentially harbor key alleles that determine the characteristics of each group, such as early maturity-related loci. Investigation of the distribution of heterozygous loci, among 12 commercial cotton hybrids, revealed a potential role for these regions in heterosis. Our study provides insight into the population structure of upland cotton germplasm. Furthermore, the overlap between lineage-specific regions and heterozygous loci, in the high-yield hybrids, suggests a role for these regions in cotton heterosis.

INTRODUCTION

Cotton (*Gossypium* spp.) is the world’s most important renewable natural textile fiber crop, and is also a significant source of oilseed (Hulse-Kemp et al. 2015). The cotton genus, *Gossypium*, consists of approximately 46 diploid and seven allotetraploid species (composed by At and Dt subgenomes) (Wendel and Cronn 2003; Grover et al. 2014; Gallagher et al. 2017). The tetraploid *Gossypium hirsutum*, also known as upland cotton, contains two sets of homoeologous chromosomes, the At and Dt subgenomes, and is now the most widely cultivated species, contributing over 95% of the total cotton fiber produced, worldwide (Fang et al. 2014).
For cultivated cotton, traditional and molecular breeding methods have aimed to identify beneficial alleles that allow breeders to develop cultivars with key agronomic traits, such as adaptation to local environments, high yield, disease resistance, and superior fiber quality (Campbell et al. 2010; Hinze et al. 2017). For example, the early maturity trait allows cultivation in environments with shorter growing seasons and may have been introduced from the source germplasm accession King.

Genetic diversity of germplasm plays an important role in crop breeding. Typically, to meet the goals of higher yield and better quality, most breeding programs have used only a limited number of elite lines; this has greatly reduced the genetic diversity of the resultant gene pools. A deeper understanding of the genetic properties of this source germplasm would assist in the identification of useful alleles for future cotton breeding programs.

Modern upland cotton was likely domesticated from one or several geographic landraces in Central America. The primitive gene pool of the upland cotton ‘Petit Gulf Series’ resulted from hybridization of the *G. hirsutum* and *G. barbadense* lineages (Wendel et al. 2010). Subsequently, breeders have developed four major cultivars (Acala, Plains, Delta and Eastern type) from this gene pool to generate cultivars that are adapted to various local environments, forming the basis of upland cotton breeding programs, worldwide (Meredith 1991).

In China, small-scale cultivation of upland cotton began in 1865 and then expanded with the rise of mechanized spinning. After 1920, multiple American varieties, with desirable agronomic traits, were introduced and widely cultivated in the Yellow River and Yangtze River regions, and some early maturity varieties were also introduced from the former Soviet Union and North Korea and cultivated in northern China.

Among these introduced varieties, we define “source germplasm” as those varieties that have stable and excellent genetic characteristics, adaptability, better general combining ability, and that were frequently used as parents in various breeding programs (Chen and Du 2006). For instance, one source germplasm accession, King, was suggested to have been the donor of the early maturity trait in
modern Chinese breeding programs (Chen and Du 2006). The Deltapine-15 variety was developed into many different varieties in China, through selective breeding and hybridization breeding, because of its excellent adaptability to different environments.

These cultivars were distributed among multiple Chinese ecological regions and evolved into different ecotypes with distinct phenotypic characteristics that were adapted to local environments or breeding purposes. For instance, cultivars developed in northern regions acquired early maturity, short height, and clustered fruiting traits (Dai et al. 2016; Chen et al. 2006). The varieties developed in the Yangtze River region developed large plant, late-maturity, and higher fiber yield traits, whereas most phenotypes of the cultivars from the Yellow River region were less extreme than the other varieties. These characteristics could be easily distinguished among different varieties, but their genetic basis remained unknown. Understanding the genetic basis for these differences would be beneficial for identifying more loci/genes associated with agronomically important traits.

Previously, by comprehensively considering the introduction periods, phenotypic characteristics, and historical breeding effects, we selected 47 accessions that represent the source germplasm of Chinese upland cotton (Du et al. 2004) and analyzed their genetic diversity using simple sequence repeat (SSR) markers (Chen and Du 2006). However, the number of markers and genome coverage was limited, leading to uncertainty about the population structure and, thus, hampering the identification of favorable alleles. Recent single-nucleotide polymorphism (SNP)-based studies were also unable to clearly illuminate the population structure of upland cotton (Sun et al. 2017; Wang et al. 2017; Fang et al. 2017).

Van Deynze et al. (2009) reported the first findings based on large-scale SNP results for cotton, identifying approximately 1000 SNPs and 300 insertion/deletion mutants (InDels) by re-sequencing expressed sequence tagged sites (ESTs). With the advent of next-generation sequencing technologies, cotton SNP discovery has accelerated and many additional SNP markers have since been reported (Byers et al. 2012; Lacape et al. 2012; Li et al. 2014). Due to their bi-allelic nature, many SNPs can be genotyped simultaneously in a high-throughput, cost-effective manner (Van
The first commercial high-density array for cotton (CottonSNP63K array) was recently developed and contains some 63,058 SNPs that can be genotyped by Infinium II assays (Hulse-Kemp et al. 2015). This array enables researchers to efficiently analyze very large numbers of SNP markers and obtain highly repeatable results. The genotype data obtained from these high-throughput assay can then be used to evaluate genetic diversity, construct genetic linkage maps, and dissect the genetic architecture of important traits, as well as other applications (Ganal et al. 2012; Truco et al. 2013; Huang et al. 2017; Sun et al. 2017).

In the present study, we used the CottonSNP63K array to examine Chinese upland cotton source germplasm (CUCSG). By aligning our results with the upland cotton TM-1 genome sequence (Zhang et al. 2015), we identified approximately 25,000 SNPs with unique physical locations in the genome. Our results reveal the population structure and lineage-specific haplotypes of CUCSG, as well as the potential roles of specific genomic regions in heterosis.

RESULTS

SNP marker screening, distribution, and genotyping
To examine the population structure of upland cotton germplasm, we first gathered a collection of 125 cotton accessions (Table S1), including 4 diploid progenitor cottons (2 G. arboreum cv. Shixiya-1 and 2 G. raimondii), ten G. barbadense, and 111 G. hirsutum inbred accessions. We next used the CottonSNP63K array to genotype the SNPs in these 125 accessions and to identify the genomic locations of the SNPs by BLAST searches of probe sequences against the G. hirsutum TM-1 genome. By comparing the BLAST results with our genotyping data, we identified 24,856 single-locus SNPs in the G. hirsutum genome (Figure 1, Table S2).

By following a step-by-step screening strategy (see Materials and Methods), we detected a subset of 10,552 high-quality SNPs. Of these high-quality markers, 4,813 and 5,739 were assigned to the At and Dt subgenomes, respectively. Chromosome A08 had the highest number of SNPs (2,528), followed by its homoeologous
chromosome D08 (1,581) (Table S2). Hence, this pair of homoeologous chromosomes showed more diversity than the other chromosomes in the G. hirsutum genome.

All 125 Gossypium inbred lines, including 111 G. hirsutum (upland cotton), 10 G. barbadense, 2 G. arboreum, and 2 G. raimondii, and the 12 hybrids were used to calculate their allele frequencies with 10,552 high-quality polymorphic SNPs (Table S1). The frequency ranges of each accession for the genotypic categories (inbred lines) were 36.7%–67.2% for the homozygote major alleles, 22.6%–48.0% for homozygote minor alleles, 1.5%–21.6% for heterozygotes, and 0.0%–25.7% for the missing alleles.

In comparisons among the CUCSG, the accession Trice, reportedly the first upland cotton to be introduced to China from America in 1918 (Du et al. 2004), showed the highest rate of heterozygosity. Prior analyses of genetic maps revealed that the diploid A and D genomes have high synten with the At- and Dt-subgenomes in tetraploid cotton (Liu et al. 2001; Rong et al. 2004; Desai et al. 2006). In our present study, the number of missing alleles from two diploid cotton accessions were much greater than that of tetraploid cotton accessions, indicating that our single-locus SNP identification method was effective (Table S1).

Linkage disequilibrium, genetic diversity, and population structure
In agricultural crops, linkage disequilibrium (LD) is a sensitive indicator to reveal the past recombination events of breeding history. After critically distinguishing the SNPs on the subgenomes, we constructed a whole-genome LD map for upland cotton (Figure S1). This revealed that LD blocks of various sizes were unevenly distributed among various chromosomes. Several chromosomes were occupied by large LD blocks, such as A08 and A01 in the At subgenome, and D01, D08, D02, and D06 in the Dt subgenome. The positioning of the large LD blocks suggests that they may have arisen by repression of recombination at the pericentromeric regions.

All 10,552 high-quality polymorphic SNPs were used to construct a phylogenetic tree (Figure 2A). The ancestral diploid cotton, G. raimondii, was used as the root.
Another diploid cotton, *G. arboreum*, was the first to diverge, followed by *G. barbadense*. In upland cottons, all accessions could be easily separated by their year of release, therefore, we defined them as "early" (blue branches) and "recent" (green branches) accessions, respectively (Figure 2A, Table S1). Moreover, according to their genotypic characteristics (Figure 2D), three groups of accessions were screened to represent typical genotypes for further analysis. The "early" accession contained two subgroups. Based on the pedigree information (Table S1), the varieties mainly derived from a primitive cultivar "King" (developed ca. 1890) were named as Group-1 (Figure 2A, orange symbols), this group was cultivated in northern China. The typical traits of these varieties exhibited small plant size and early maturity.

Further along the phylogenetic tree, another "early" group, named as Group-2, contained most of the accessions derived from Stoneville 2B (developed ca. 1938) and Deltapine 15 (developed ca.1947), which were mainly cultivated in the Yangtze River region (Figure 2A, blue symbols). A majority of the "recent" accessions are represented by a branch of new varieties developed in recent years (after ca. 2000) and were cultivated in either the Yellow River region or in a Yangtze River region that has undergone a dramatic reduction in cultivated area (Figure 2A, green symbols, named as Group-3).

The remaining *G. hirsutum* accessions had complex genetic backgrounds and comprised two other groups: Mixture-1 and Mixture-2. Mixture-1 mainly contained the "early" accessions along with mixed King and Deltapine 15 lineages, most of which were cultivated in the Yangtze River region and were introduced from exotic countries. However, Mixture-2 was mainly composed of "recent" accessions that were cultivated in the Yellow River region (Figure 2A, D).

When we compared the allelic diversity and LD decay distance, the early accessions showed significantly higher allelic diversity (Figure 2B) and a shorter LD decay distance (Figure 2C) than the recent accessions. Moreover, the population structure analysis showed that the early accessions harbored a more mixed ancestry than the recent accessions (Figure 2D, when *K* = 5). Given the above results, we concluded that CUCSG represented a typical population structure in accordance with
its pedigree, and that its genetic diversity significantly decreased as the breeding process progressed.

**Lineage-specific haplotypes of three subgroups**

According to the clustering results and pedigree records, we selected 14, 17, and 13 accessions to represent the King (Group-1), Deltapine 15 (Group-2), and new varieties (Group-3) in CUCSG (**Figure 2; Table S1**). To identify specific regions of genomic divergence, among these groups, we calculated the $F$-statistic ($F_{ST}$) of each SNP between each set of groups (see Materials and Methods). Only the top-5% of the $F_{ST}$ values for each pairwise comparison were deemed to be significantly divergent SNPs and were, therefore, defined as lineage-specific SNPs. We established that the significantly divergent SNPs showed regional continuity and enrichment in the chromosomes (**Figure 3A; Tables S3–S5**), which suggested that large distinct haplotype-associated groups of SNPs exist on different chromosomes among the different lineage-derived groups. For Group-1 (King lineage), the lineage-specific SNPs were mainly concentrated on D08 (155), A06 (36), and A01 (33); in Group-2 (Deltapine-15 lineage) they were primarily located on A08 (334), A05 (34), and D07 (29); and in Group-3 (new varieties lineage) they were mostly located on D08 (52), D12 (46), and A09 (38) (**Figure 3B**).

An in-depth analysis of the genotypes, within these three groups, showed that each group had specific SNPs in their lineage-specific genomic regions. For example, Group-1 and Group-2 had an extensive range of continuous minor alleles on Chromosomes A06 (from ~64.27 to 93.49 Mb) and A08 (from ~21.40 to 57.11 Mb), respectively (**Figure 3C**). Combined with their ecological distribution and pedigrees, this suggested that these distinct regions are genomic signatures for CUCSG, derived from different lineages and ecological origins. These regions might harbor important alleles that confer environmental adaptation and elite traits.

**The lineage-specific haplotype on Chromosome A06 might contribute to early maturity in upland cotton**
In our study, Group-1 represented the primitive donor germplasm for the early maturity trait in CUCSG. Both chromosome-wide LD block identification (Figure S1) and comparisons of \( F_{ST} \) value among the groups (Figure 3A) indicated that strong regions of genetic divergence are enriched on several chromosomes, such as D08 and A06. According to the threshold value (top 5% \( F_{ST} \) value), we determined that most of the strongest \( F_{ST} \) signals from Group-1 were located on chromosome A06 (Figure 3A; Table S3). A cluster of SNPs ranged from ~89.5 Mb to ~93.5 Mb and composed a strong LD block in this region (Figure 4A). Two SSR, NAU5433 and NAU0874, tested in 419 core collections of upland cotton, were strongly associated with flowering time (NAU5433 with \(-\log P = 4.40\), NAU0874 with \(-\log P = 4.62\) (unpublished data) and were located in this block. Because most of the markers in this microarray were developed from EST data, we found that 7 of the 16 markers in this block were located in the exons of six genes, and two of them carried nonsynonymous mutations in two genes (A06G1245 and A06G1253) (Figure 4A, B).

The gene expression profiles further confirmed that A06G1245 was specifically expressed in floral organs (Figure 4B), and its homologue has been shown to regulate flowering time in Arabidopsis (Endo et al. 2013). Interestingly, we determined that the SNPs specific to the Group-1 in this block were missing in Group-2 and Group-3 (Figure 4C), which might explain why the major early maturity loci from the King lineage (Group-1) was missing in most of the modern cultivars.

**Genetic architecture of hybrid upland cotton cultivars**

In China, the upland cotton hybrids were mainly developed and cultivated in the Yangtze River and Yellow River regions. By analyzing the genetic diversity in the parents of 12 representative hybrid cultivars, we established that the Chinese cotton hybrids had different genetic backgrounds. The Yangtze River region hybrids were developed by crossing early accessions (belonging to Group-1 or Mixture-1) and recent accessions (belonging to Mixture-2), whereas nearly all of the Yellow River region hybrids were developed by crossing two recent accessions, except in the case of CCRI-51 and CCRI-72 (Figure 5A, B).
In hybridization breeding, the number of heterozygous loci in hybrids is determined by the genetic diversity between the two parents. In our study, we established that both the number of predicted (the number of polymorphic loci between two parents that could theoretically produce heterozygous loci) and observed (the number of heterozygous loci detected with microarray) heterozygous loci, in the hybrids, were significantly correlated with the genetic distance between their paired-parents (Figure 5C). Among all of the hybrids, CCRI-63 and CCRI-94 harbored the most and least heterozygous loci, respectively.

To reveal the genetic differences among hybrids, a total of five representative hybrids were selected according to their combination type and number of theoretical heterozygous loci (Figure 5D). By comparison, we identified only a small amount (1.86%, 122/6,550) of heterozygous loci shared among all five hybrids. However, different combination types harbored different numbers of unique heterozygous loci, such that CCRI-63 (13.80%, 905/6,550) and CCRI-70 (9.53%, 624/6,550) harbored the most unique heterozygous loci (Figure 5D).

Once we selected for the unique heterozygous loci for each representative hybrid, we further determined that the unique loci were concentrated in several chromosomes. For CCRI-63, most of the heterozygous loci were distributed among chromosomes A08 (341), D07 (206), and D02 (80). For CCRI-70, they were mainly distributed among chromosomes A13 (126), A10 (94), and D13 (58) (Figures 5E, S2, S3). Meanwhile, we established that most (85.1%) of the highly divergent loci (top 5%) in Group-2 (Figure 3A, B) overlapped with the heterosis loci in CCRI-63 (Figure 5F). These heterozygous loci-enrichment regions might well be related with heterosis in upland cotton.

**DISCUSSION**

**Accuracy of genotyping tetraploid cotton based on the CottonSNP63K Array**

The first commercial high-density CottonSNP63K Array that was used in this study provides a new resource for studying the genetic architecture of upland cotton germplasm. However, the explicit physical locations of the 63,058 SNPs markers on
this array were not clear, since the assay was developed before the upland cotton TM-1 genome sequence was released (Zhang et al. 2015). Tetraploid cotton contains two sets of homoeologous chromosomes, the At and Dt subgenomes, hence, many markers on the array were mapped to multiple possible positions, while the majority were located on both homoeologous chromosomes (according to the BLAST results). Determining which homoeologous chromosome each SNP actually belongs on is quite difficult and poses a challenge for future research, such as linkage group construction and association studies.

To solve this problem, we re-sequenced 46 representative CUCSGs (Table S1) at more than 10×depth by aligning the results with the TM-1 reference genome. According to the BLAST results of SNP positions on the microarray, the re-sequencing data could precisely distinguish their accurate positions in the subgenomes. Finally, we identified 24,856 single-locus SNP markers that were evenly distributed throughout the whole genome and accounted for 39.4% of the microarray. This set of single-locus SNPs not only facilitated the further analyses that are reported in this study, but also provides important supplementary information for the effective utilization of the CottonSNP63K array in the future.

The GenTrain score is a synthetic indicator reflecting the accuracy of a genotype, with a possible range from 0.00–1.00. Genotypes with lower scores are located further from the cluster center and, thus, have lower reliability. From the 24,856 single-locus SNPs, a subset of high-quality markers was selected, according to several filtering criteria. The GenTrain scores for this set of high-quality markers ranged from 0.3469–0.9592, which showed a two-type clustering pattern of polymorphic markers having GenTrain scores > 0.6 and those of 0.30–0.59, according to the classification of the CottonSNP63K array, as described by array developer (Hulse-Kemp et al. 2015). Both of types of markers showed three clearly definable clusters, and did not require any notable manual adjustment of their cluster positions to obtain accurate genotype results, thus providing an important foundation for further genetic analysis.
Genomic characteristics of CUCSG subgroups

Most prior genetic diversity studies have suggested that no obvious genetic structure exists in Chinese upland cotton populations (Fang et al. 2017; Sun et al. 2017; Wang et al. 2017). In our study, we present strong evidence that CUCSG has highly divergent genomic signatures for the different lineages. We identified three major lineages, called King (Group-1), Deltapine-15 (Group-2), and new varieties (Group-3), and this categorization strongly agreed with their times since introduction and geographic distributions.

King was a very primitive variety (ca. 1890) directly selected from the upland cotton ancient gene pool, Green Seed (Bowman 2006). This variety was introduced into China ca. 1919 and was mainly cultivated in northern China. The characterized varieties in this region include early-maturity, and King, which were also suggested to have been the major donor of the early maturity trait for subsequent breeding. ‘Deltapine-15’ was the variety most broadly used in Chinese upland cotton breeding programs, likely because of its extensive geographical adaptability. More than 400 varieties were developed from the Deltapine-15 cross in the Yangtze River and Yellow River regions in China until 2000 (Chen and Du 2006). These two major lineages represent the early accessions of CUCSG (Figure 2). Along with the migration of the major cultivation region for Chinese upland cotton, the recent accessions were mainly cultivated in the Yellow River region.

The third lineage, new varieties lineage (Group-3), is composed of varieties developed only since the start of the 21st century. Interestingly, together with the principal components analysis (PCA) results (Figure 5A), we showed that Group-3 may not have derived from the early Chinese upland cotton gene pool (Group-1 and Group-2). Therefore, it might have developed from other unknown sources through interspecific hybridization.

Identification of the lineage-specific haplotypes from different groups could accelerate our understanding of the formation of ecological phenotypes in the upland cotton germplasm. We observed strong genomic divergence between groups that were cultivated in distinct environments. These genomic regions were likely selected
by nature or humans to hasten local environmental adaptation. For example, the divergent regions in Group-1 (chromosomes A06 and D08) may be associated with the early-maturity trait that is characteristic of King lineage accessions, whereas the divergent regions in Group-2 (chromosome A08) may be associated with the extensive adaption and high yield traits that are characteristic of the Deltapine-15 lineage.

By integrating this information with RNA sequencing data, we identified a nonsynonymous mutation in a gene (A06G1245) that regulates flowering time in *Arabidopsis*, which might contribute to the early maturity observed in most King lineage (Group-1) cultivars. We also showed that the associated haplotype was missing in subsequent cultivars. Furthermore, we noticed that most of the lineage-specific haplotypes were continuous regions that spanned a large range of the chromosomes. Furthermore, these haplotypes were also located within the identified LD blocks (Figure S1). For the King specific-lineage on A06 (Figure 4A), this haplotype length extended ~4 Mb. The early maturity loci in this region might also be linked with unfavorable loci (such as weak disease resistance or poor fiber quality), which may have resulted in the elimination of the whole haplotype during the breeding process.

A possible explanation for the large LD blocks was that they formed through chromosomal re-arrangement, such as inversion and translocation (Hoffmann et al. 2011). Chromosome inversions are considered to be the major evolutionary force driving population divergence when adapting to local environments or in artificial selection (selective breeding), since they maintain the beneficial variations in a population by repressing genetic recombination (Rieseberg 2001).

**Potential role for lineage-specific haplotypes in upland cotton heterosis breeding**

The utilization of hybrid vigor is a powerful way to improve the yield and quality of a crop. Several classic models have been established in some crops. Complementation and the over-dominant theory suggests that heterozygous loci (F<sub>1</sub>) perform better than
homozygous loci (parents) in tomato (Krieger et al. 2010). In rice, the heterosis in various hybridization systems is caused by different heterozygous loci (Huang et al. 2016).

Our study also provides evidence that cotton heterosis was driven, in part, by different loci in different combination types. For instance, both CCRI-63 and CCRI-66 are considered to be the highest yield hybrid cultivars that are suitable for cultivation in the Yangtze River region (Yang et al. 2009). In our study, CCRI-63 and CCRI-66 harbored the most theoretical and unique heterozygous loci, and the number of heterozygous loci increased with increasing genetic distance from their parents (Figure 5C). The CCRI-63 heterozygous loci were mainly distributed in large genomic regions on chromosomes A08, D02, and D07, which overlap substantially with the top-5% of SNPs in Group-2 (Figure 5F), and its female parent was derived from the Deltapine-15 lineage (Figure 4A). This suggests that the heterozygous status of these regions might be responsible for this cultivar's high yield performance (Figure 4E). As these regions spanned a large range of chromosomes, traditional QTL analysis and genome-wide association mapping should be carried out to identify the loci or genes that are likely driving the yield performance of hybrid cultivars. In light of these comprehensive results, our study has revealed a plausible genetic basis for heterosis in cotton and provided an empirical basis for parent selection in future hybridization breeding.

MATERIALS AND METHODS

Plant materials and genotyping

A total of 137 accessions was used in this study (Table S1), including 4 diploid progenitor cottons (2 G. arboreum cv. Shixiya-1 and 2 G. raimondii). Ten G. barbadense and 114 G. hirsutum inbred accessions were selected according to the considerations described by Du et al. (2004) and screening from major breeding programs across three cotton cultivation regions (Yangtze River, Yellow River, and Xinjiang regions) and different breeding periods, representing the genetic diversity of the present Chinese inbred G. hirsutum as much as possible. The remaining 12
representative commercial hybrid varieties were used to study heterosis.

Genomic DNA was extracted from a single cotton seed for each accession following the protocol described by Kuang et al. (2010). The quality and quantity of DNA were estimated by measuring the optical density at 260/280 nm with an ultraviolet spectrophotometer, and was adjusted to 50 ng/μL prior to genotyping. Normalized DNA was processed according to Illumina protocols and hybridized to the CottonSNP63K array. Single-base extension was performed and the chips were scanned by the Illumina iScan system. The SNP genotypes were generated for each panel member, according to the previously reported procedure (Hulse-Kemp et al. 2015).

Identification of the single-locus SNPs

All SNPs and their flank sequences were downloaded from the CottonSNP63K array original article (Hulse-Kemp et al. 2015). To obtain the explicit physical localization of the 63,058 SNPs markers on the array, the probe sequences for each SNP were used as queries in local BLAST (Altschul et al. 1990) searches against the G. hirsutum TM-1 reference genome (Zhang et al. 2015). As a result, 61,343 SNP markers were assigned to one or multiple positions on the TM-1 reference genome, and most markers mapped to multiple genomic positions. For determining the precise physical positions for the markers with multiple positions (At or Dt subgenomes), we re-sequenced 46 upland cotton accessions (selected from the accessions used in this study) to a > 10× depth, and analyzed the genomic variation based on the TM-1 reference genome. Comparisons revealed that the physical locations of the 25,101 polymorphic SNP markers with no more than two positions on the chromosome were consistent with the BLAST results: 24,856 were single-locus SNP markers mapped to a unique position on the reference genome, and the remaining 245 were in fact double-locus markers. And the genotypes coincidence rate of selected single-locus from CottonSNP63K array and whole genome resequencing reached ~95.1% (Table S6). Therefore, the single-locus SNP markers, as detailed in Table S2, were used in the further analyses.
Screening of high-quality SNPs based on array genotyping

The raw data obtained from the iScan system were imported into GenomeStudio (Illumina) for the genotype analysis. The high-quality SNP markers were selected from the above-mentioned 24,856 single-locus markers with a stepwise screening strategy that used three filtering criteria: (1) ‘AA Freq > 0 and BB Freq > 0’ was set as a new indicator for the polymorphism evaluation instead of the usual indicator (‘The minor allele frequency [MAF] value’). (2) SNPs with a missing data rate exceeding 10% of the sites were discarded altogether. (3) ‘AA R mean ≥ 0.4 and BB R mean ≥ 0.4’ was set as an indicator for signal intensity, and those SNPs with a ‘Norm R value < 0.4’ were also discarded. In this way, a final set of 10,552 high-quality SNP markers was obtained. The GenTrain scores reflected the accuracy of the data, and ranged from 0.3469 to 0.9592, with an average value of 0.6519.

Genetic diversity and population genetic analysis

The 10,552 high-quality SNPs were used to construct a phylogenetic tree (UPGMA model) in PowerMarker (v.3.25, http://statgen.ncsu.edu/powermarker/index.html) and in the percentage calculations of the major, minor, missing, and heterozygous alleles. For the population structure analysis, the microarray formatted genotype was first transformed into corresponding .ped and .map formats – which are compatible with Plink software (v.1.9, http://www.cog-genomics.org/plink2/) – using homemade python scripts. These were further transformed into the bed format in Plink. Then, the population structure was calculated with the ADMIXTURE software (v.1.23) (Alexander et al. 2009). For the $F_{ST}$ value calculation (i.e., $F$-statistics), the groups were categorized; for example, in the top panel of Figure 3a, all accessions except Group-1 composed the Group-1-remaining group. Then the $F_{ST}$ was calculated for Group-1 and Group-1-remaining. The $F_{ST}$ value for each SNP was calculated by VCFtools software (v.0.1.14) (Danecek et al. 2011), and the top 5% of the $F_{ST}$ values (significant difference at 0.05 level) were separately defined as the thresholds in each comparison.
ACKNOWLEDGEMENTS

We thank National Cotton Germplasm Middle-term Genebank (Anyang) for providing the inbred seeds, and all the breeders: Gentu Yan, Youlu Yuan, Chaozhu Xing, and Hongmei Wang at the Institute of Cotton Research of the Chinese Academy of Agricultural Sciences for kindly providing the hybrids seeds and valuable suggestions for our work. We are very grateful to the Beijing Compass Biotechnology Company for their excellent technical assistance with the microarray experiments. This work was supported by grants from the National Key Research and Development Program of China (2017YFD0102000 and 2016YFD0100306) and the National Natural Science Foundation of China (31301365).

AUTHOR CONTRIBUTIONS

S. H. conducted the experiments. S. H., G.i S. and M. K. analyzed the data and wrote the manuscript; D. Z., Y. W., and P. D. performed the SSR experiments. X. M., D. Y., and X. D. provided the materials; and L. H., S. W., J. P. revised the manuscript. All authors read and approved the paper.

REFERENCES


Figure legends:

Figure 1. Genomic distribution of all 24,856 SNPs
Each red vertical band indicates an SNP.

Figure 2. Genetic diversity and population structure of CUCSG
(A) Phylogenetic tree of the accessions used in this study. *Gossypium raimondii* was used as the root. The branches of the diploid cottons and *G. barbadense* (red), and those of the "early" (blue) and "recent" (green) *G. hirsutum* accessions are indicated with the specified colors. Representative varieties are labeled and the representative groups are highlighted with red-shaded circles. The colored branch nodes represented accessions developed from exotic (black), northern (orange), Yangtze River (blue), and Yellow River regions (green). (B) Allelic diversity between the early and recent accessions groups. A Mann Whitney U test was used to determine the significant difference between the means of the two groups (shown as boxplots). (C) Linkage disequilibrium (LD) decay plot for the "early" and "recent" accessions. (D) Model-based clustering analysis of all accessions (when $K = 5$). The x-axis indicates the accessions, with an order matching the phylogenetic tree. The y-axis indicates the cluster membership, in different colors. The groups (corresponding with Figure 2A) were indicated.

Figure 3. Identification of lineage-specific divergent SNPs in three represented groups
(A) The $F_{ST}$ values for all SNPs between the specified groups and remaining accessions. Different panels indicate the various comparisons made: Group-1 vs. remaining accessions (top), Group-2 vs. remaining accessions (middle), and Group-3 vs. remaining accessions (bottom). Thresholds (i.e., top 5% $F_{ST}$ values) of the three comparisons are indicated by the horizontal dotted lines. The top 5% $F_{ST}$ values in top, middle, and bottom panels were 0.51, 0.38, and 0.52, respectively. (B) Number of highly divergent SNPs in the three groups across all chromosomes. (C) Schematic of the genotypes in the local and highly divergent regions of Group-1 on the Chromosome A06 (top) and of Group-2 on the Chromosome A08 (bottom). The y-axis shows the SNP genotypes ordered by their physical position, and the x-axis shows...
the accessions with the same order as in Figure 2D.

**Figure 4. Early maturity-related haplotype on chromosome A06**

**(A)** All SNP locations and linkage disequilibrium states in a region of chromosome A06. Exonic SNPs are indicated in red. SSR-1 and SSR-2 indicated SSR markers NAU5433 (\(-\log P = 4.40\)) and NAU0874 (\(-\log P = 4.62\)), respectively. **(B)** Expression profiles of genes containing exonic SNPs in various tissues and developmental stages. FPKM, fragments per kilobase of transcript per million fragments mapped. NS and S in the brackets indicate non-synonymous and synonymous SNPs at exonic regions, respectively. NS and S in the brackets indicate non-synonymous and synonymous SNPs at exonic regions, respectively. **(C)** Allele distribution of the SNP with the highest F\(_{ST}\) value (SNP_A06_89808670, i11180Gh) among the three groups.

**Figure 5. Lineage-specific divergent SNPs and heterosis**

**(A)** PCA plot of all *G. hirsutum* accessions. Different categorized groups (described in Figure 2D) are identified by circles of different colors. The represented varieties are named and the parent pair symbols are zoomed and connected by red dotted lines. All combinations are indicated with capital letters. **(B)** List of the combination types for all hybrids. **(C)** The relationship between the genetic distance of the parents (x-axis) and the number of heterozygous loci in hybrids (y-axis, theoretically predicted and detected by observation). Correlation coefficients and significant differences were calculated and tested, respectively, by Pearson’s r. **(D)** Venn diagram of all the shared and unique theoretical heterozygous loci among the five represented hybrids in the genome. **(E)** Schematic of theoretical heterozygous loci distributions in each hybrid. Red lines indicate the heterozygous loci (theoretical). Blue and black boxes indicate the three regions with the highest concentration of unique heterozygous loci (theoretical) in CCRI-63 and CCRI-70, respectively. **(F)** Venn diagram of the heterozygous loci (theoretical) in CCRI-63 and the significantly divergent SNPs in Group-2.

**SUPPORTING INFORMATION**

**Figure S1.** Whole-genome linkage disequilibrium (LD) map of upland cotton based on all the inbred *G. hirsutum* lines and 10,552 single-locus SNP markers
**Figure S2.** Venn diagrams of all shared and unique theoretical heterozygous loci among the five hybrids represented in the At subgenome.

**Figure S3.** Venn diagrams of all shared and unique theoretical heterozygous loci among the five hybrids represented in the Dt subgenome.

**Table S1.** Information and genotyping statistics of all cotton accessions used in this study

**Table S2.** Complete list of the 24,856 single-locus markers

**Table S3.** List of significantly divergent (top 5%) SNPs between Group-1 vs. remaining (Fst > 0.51)

**Table S4.** List of significantly divergent (top 5%) SNPs between Group-2 vs. remaining (Fst > 0.38)

**Table S5.** List of significantly divergent (top 5%) SNPs between Group-2 vs. remaining (Fst > 0.52)

**Table S6.** The genotypes coincidence rate of CottonSNP63K array and whole genome resequencing

---

**Figure 1**
Figure 2

Figure 3
Figure 4
Figure 5