Genomic variation mapping and detection of novel genes based on genome-wide survey of an elite upland cotton hybrid (*Gossypium hirsutum* L.)

Zhenyu Wang[1,†], Wei Li[1,†], Guanghui Xiao[2], Xiaojian Zhou[1], Xiaoyu Pei[1], Yangai Liu[1], Kehai Zhou[1], Kunlun He[1], Junfang Liu[1], Ying Li[1], Wensheng Zhang[1], Zhongying Ren[1], Qingqin Meng[1], Haifeng Wang[1], Xiongfeng Ma[1,*] and Daigang Yang[1,*]

[1]State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China
[2]College of Life Sciences, Shaanxi Normal University, Xi'an 710119, China

CCRI63, with the largest cultivated area among hybrids in China, is a successful promotion of elite upland cotton (*Gossypium hirsutum* L.) hybrid cultivar. We have constructed a detailed genomic variation map of CCRI63 by aligning whole-genome shotgun sequencing reads from CCRI63 to the TM-1 reference genome. Genomic single nucleotide polymorphism (SNP) and insertion-deletion (Indel) mutational hotspots were identified, most of which were located on chromosome D02, and associated with disease resistance and lipid glycosylation and modification. The density of heterozygous SNP sites showed 73 quantitative trait loci overlapped with peak intervals of high-density heterozygous SNPs, suggesting that the heterozygous sites in the peak are important for improvement of CCRI63 yield and fibre quality. To avoid loss of genetic components, unmapped reads were used for *de novo* assembly of the missing regions in the reference genome, and 153 novel functional genes were obtained. The large-scale genetic variation and novel functional genes identified in the CCRI63 genome can facilitate future gene-phenotype studies and provide an additional resource for the improvement of cotton.

Keywords: Cotton hybrid, genomic variation map, mutational hotspots, novel genes.

COTTON is an economically important crop and the largest natural fibre source in the world. It is the most widely used material in the textile industry due to many beneficial characteristics such as absorbency, strength, colour retention, heat-resistance and easy handling. The genus *Gossypium* includes approximately 45 diploid and five tetraploid species[1,2]. It is an excellent plant material for studying domestication and improvement[3]. Previous research on its physiology, biochemistry, metabolic mechanisms and genetic evolution has provided a theoretical foundation for molecular biology and genetic breeding[4–10]. Upland cotton (*Gossypium hirsutum* L., AADD), which has a high yield capacity and wide adaptability, is a major contributor to yield of cotton fibre, supplying approximately 95% of the total cotton fibre produced in the world. Correspondingly, it is cultivated on more than 90% of the agricultural land used to grow cotton worldwide[11]. As the most important cultivated cotton, upland cotton has long been the primary focus of researchers in cotton genetics and breeding for increasing yield and improving fibre quality. Heterosis is an effective way to achieve genetic improvement in cotton breeding[12,13]. As in the case of other crops, there is no clear explanation of the mechanism of heterosis in cotton. Population genetics research showed that dominance and over-dominance contributed to heterosis of cotton hybrid, and dominance played a more important role in cotton yield[14]. Over-dominance is a major factor in the heterosis of yield traits based on heterotic quantitative trait loci (QTL) mapping [15]. Epistasis and partial dominance also contribute to heterosis of cotton, but they do not exclude the contribution of dominance and over-dominance[16,17].

In some economically important crops such as rice, maize and sorghum, next-generation sequencing (NGS) has been used to characterize genomic variation[18–21]. Throughout a genome, mutation density is heterogeneous, and genes in regions with high single nucleotide polymorphism (SNP) density might contribute to phenotypic diversity[22]. At the same time, assembling unmapped reads provides more insight into the missing regions in the reference genome, e.g. the 101 novel genes identified in the sorghum genome[18]. Currently, completion of the draft genomes of allotetraploid cultivated upland cotton TM-1 (*G. hirsutum* L., AADD) has provided the foundation for analysis of cotton genomic variation and phenotypic diversity[7,23].

*For correspondence. (e-mail: yangdaigang@caas.cn)
†Equally contributed.

CCRI63, a successful hybrid cultivar *G. hirsutum*, is cultivated in the largest area in China. CCRI63 shows strong heterosis, superior to that of its parents in terms of boll weight, boll number, lint yield and disease resistance. However, the genome structure and mechanism of its heterosis are still unclear. In this study, we have sequenced the whole genome of CCRI63, producing 170.14 Gb of sequence (~67.9-fold coverage depth) using the Illumina HiSeq 4000 platform. We have identified a large number of genomic variations, including SNPs, insertion-deletion (Indels), structural variations (SVs) and copy number variations (CNVs) and shown that some of these mutations are located in hotspots. By constructing genome-wide heterozygous SNP distribution map, we have obtained the important genome segment that overlaps with 73 QTLs. This provides favourable clues for studying the mechanism of CCRI63 heterosis. Moreover, after assembling unmapped reads in the CCRI63 genome, we found 153 new functional genes through homology-based prediction. This study has identified many genomic variations and provides more comprehensive genomic information for use in future cotton genomics research.

Materials and methods

Plant materials and DNA extraction

The CCRI63 cultivar, a hybrid of 9053 (maternal) × P4 (an improved line of sGK9708, paternal), has been authorized for commercial production by the National Crop Variety Identification Committee in China with the following fibre characteristics: an upper half mean fibre length of 30.0 mm, breaking strength of 29.1 cN/tex, micronaire value of 4.8, elongation of 7.0%, and uniformity index of 84.2%. The National Medium-Term Gene Bank of Cotton in China provided seeds of *G. hirsutum* CCRI63 for this study. They were grown in a light incubator at 37°C and 75% humidity. When the cotyledon reached a fully expanded flat state, samples were collected and genomic DNA was immediately extracted using a modified CTAB method[24].

Library construction and sequencing

Before library construction, the quality and purity of DNA were ensured by agarose gel electrophoresis and OD 260/280 ratio, which ranged from 1.8 to 2.0. A total of 1.5 μg of genomic DNA per sample was used for library preparations. Acceptable DNA was randomly cut into 350 bp fragments using a hydrodynamic shearing system and sequentially processed to construct a paired-end sequencing library according to the manufacturer's instructions (Illumina). Sequencing was performed on the Illumina HiSeq 4000 platform, and 150 bp paired-end reads were generated at the Novogene Bioinformatics Institute, Beijing, China.

Filtering reads and read alignments

To avoid reads with artificial bias, adaptor and low quality reads were removed if they displayed the following characteristics: (a) the read contained more than 10% unidentified nucleotides (*N*); (b) the read had >10 nt aligned to the adapter, allowing ≤10% mismatches; (c) more than 50% of the bases had a Phred score <5; and (d) the read contained potential duplications generated by PCR amplification during the library construction process. The cleaned paired-end reads were mapped to the reference genome of *G. hirsutum* L. acc. TM-1 (http://mascotton.njau.edu.cn/html/Data/Genomefhsequence/2015/05/05/16ab0945-19e9-49f7-a09e-8e956ec866bf.html) using the BWA (Burrows–Wheeler Aligner) software with the command 'mem-t 10-k 32'[25]. The SAMtools software (settings: –bS –t) was used to convert and index the mapping results to the BAM files[26]. If multiple read pairs had identical external coordinates, only the pair with the highest mapping quality was retained; potential polymerase chain reaction (PCR) duplications were removed to improve the alignment results.

Variant detection and annotation

SNPs and Indels were identified by SAMtools mpileup (settings: mpileup -m 2 -F 0.002 -d 1000). We characterized the results with a coverage depth ≥4 and ≤1000 and RMS mapping quality ≥20 as potential variations. SVs were detected using BreakDancer[27] (settings –q 20), while CNVs were detected using CNVnator software[28]. To obtain further support for the identified candidate duplications and deletions, we used CNVnator to slide 100 bp along each chromosome and compare the observed RD (reads depth) with the GC-matched RD of the same chromosome. CNVs were retained if the $\log_2$ ratio of the counts of reads per sliding window was greater than 1.5 and less than 0.5, and the length of CNV was more than five consecutive windows. All of the variations were annotated using the ANNOVAR software[29].

De novo assembly and annotation of unmapped reads

We performed *de novo* assembly of unmapped reads using SOAPdenovo[30], which is a de Bruijn graph algorithm-based *de novo* genome assembler with the following parameters: the *k*-mer length is 63 bp and the *k*-mer frequency cut-off is 2. The protein sequences of *T. cacao* that shared a common ancestor with cotton (*T. cacao*: https://phytozome.jgi.doe.gov/pz/portal.html#!bulk?org=Org_Tcacao) and the annotated model plants *Arabidopsis*

thaliana and *Oryza sativa* were downloaded from Ensembl release 30 (*A. thaliana*: ftp://ftp.ensemblgenomes.org/pub/plants/release30/fasta/arabidopsis_thaliana/pep/; *O. sativa*: ftp://ftp.ensemblgenomes.org/pub/plants/release-30/fasta/oryza_sativa/pep/). These protein sequences were mapped onto the contigs from the unmapped reads using TBLASTN. Next, homologous genome sequences were aligned against the matching proteins using GeneWise (http://www.ebi.ac.uk/Tools/psa/genewise/) to define the gene models. Gene functions were assigned according to the best match of the alignment to the Swiss-Prot and TrEMBL databases (http://www.uniprot.org/help/publications), using BLASTP. We also mapped *G. hirsutum* genes to the KEGG pathway database to identify the best pathway for each gene.

Results

Genome resequencing and alignment

We sequenced the whole genome of CCRI63, producing 170.14 Gb of sequence (NCBI Sequence Read Archive database; accession: SRR3658873), using the Illumina HiSeq 4000 platform. In the resulting reads, 96.58% and 92.12% of the bases had quality scores ≥Q20 and ≥Q30 respectively. After strict filtering, a total of 1.23 Gb high-quality paired-end reads (169.36 Gb) were mapped to the TM-1 reference genome using the BWA software[7,25]. Finally, 99.58% of the sequence reads were mapped to 98% of the reference genome at least once with approximately 67.9-fold coverage depth (Table 1).

Detection of genomic variation

A large number of SNPs, Indels, SVs and CNVs were discovered by comparing the genome of CCRI63 with TM-1 (Figure 1). We identified 2,302,944 credible SNPs in the CCRI63 genome using strict criteria, of which 2,040,165 (88.59%) were in the intergenic regions, 67,527 in the exonic regions and 107,010 in the intronic regions (Supplementary Figure 1). Moreover, there were 30,759 synonymous and 36,065 non-synonymous SNPs in coding sequence (CDS) regions with a non-synonymous-to-synonymous (N/S) ratio of 1.17 (Supplementary Table 1). For crop plants, such as sorghum, rice and soybean, the N/S ratio is generally at or higher than 1.0 (soybean 1.37 (ref. 31); rice 1.2 (ref. 32); sorghum 1.0 (ref. 18)). We also found 922 large effective SNPs that impacted codon sequence translation in 818 genes by causing splicing or resulting in the gain or loss of a stop codon. A gene ontology term enrichment analysis of these genes indicated significant enrichment of the molecular function 'ADP binding' (GO: 0043531, 26 genes, *P*-value = 0.001) (Supplementary Figure 2). For further functional categorization, KEGG pathway analyses were also performed using KOBAS2.0 (http://kobas.cbi.pku.edu.cn/), and 13 genes were assigned to KEGG pathways. The pathways were associated with 'plant–pathogen interaction' (tcc04626, *P*-value = 0.10), which may be related to disease in CCRI63 (Supplementary Figure 3).

A total of 218,784 Indels were detected, including 102,832 insertions and 115,952 deletions with lengths from 1 to 50 bp. Most of the Indels (70.05%) were located in the intergenic regions, and 16.95% were distributed in the genic regions (Figure 2 *a*). We found that the length of most Indels (97.3%) ranged from 1 to 21 bp. Also, small fragments of Indels are more likely to occur

Table 1. Sequence summary and mapping statistics for CCRI63

Parameter	Value
Raw base (bp)	170,135,232,600
Clean base (bp)	169,358,837,100
Q20 (%)	96.58
Q30 (%)	92.12
GC content (%)	38.16
Mapping rate (%)	99.58
Average depth (X)	67.90
Coverage at least 1× (%)	98.00
Coverage at least 4× (%)	95.41

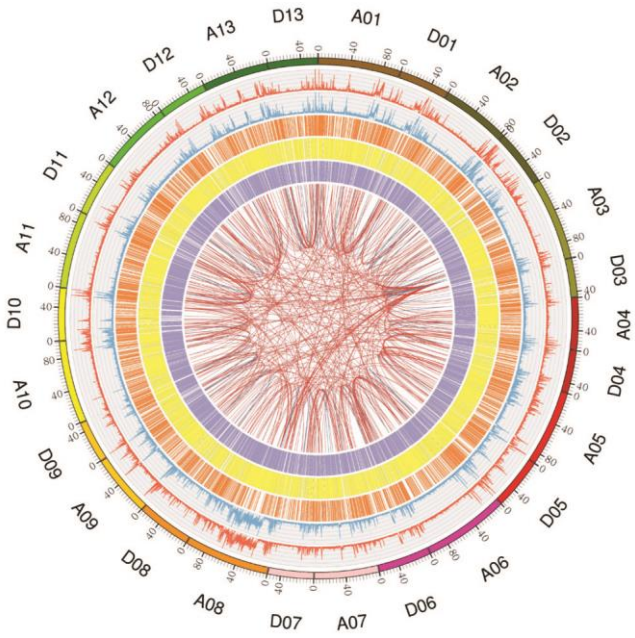

Figure 1. Distribution of genome-wide variation. SNP and Indel density per 100 kb, SV and CNV are counted and displayed using Circos 0.64. Data are arranged in rows by chromosome order. From outside to inside: SNP density, Indel density, CNV duplications, CNV deletions, SV (insertions, deletions and inversions), SV intra-chromosomal translocations (ITXs) and inter-chromosomal translocations (CTXs). SNP and Indel frequency per 100 kb are indicated by peaks. The orange and yellow lines represent CNV duplications and deletions in chromosome positions respectively. SV (insertions, deflections and inversions) is represented by the green line. ITXs and CTXs are represented by the red and blue curves respectively.

in the genome; especially, the deletion and insertion of a single base is the most common. However, the length of integer multiple of 3 bp produces more Indel variants in the CDS region (Supplementary Figure 4); these Indels of 3 bp did not result in frame shift mutations. This was strong evidence that the variation is likely conserved in the biological evolutionary process of cotton. We carried out a detailed classification of 37,092 Indels located in the genic regions, including 10,378 upstream, 1433 exonic and 16,396 intronic (Figure 2 *b*). Among the exonic regions, variants were further divided into 24 stop gain mutations, 10 stop loss mutations, 470 frameshift deletion mutations, 310 non-frameshift deletion mutations, 341 frameshift insertion mutations and 278 non-frameshift insertion mutations (Figure 2 *c*).

A total of 21,328 SVs were detected using BreakDancer. There were significant differences in the length distribution (Supplementary Figure 5 *a*) of SVs as follows: 50–200 bp accounted for 56.2%, 300–600 bp for 10.2%, and 600–1000 bp for 26.1%. Based on their position SVs in the reference genome, all SVs were classified as the following: 654 upstream, 461 exonic, 498 downstream, 950 intronic, 52 upstream/downstream, 18,708 intergenic and five at splicing junctions (Supplementary Figure 5 *b*). We found that most of the SVs were chromosomal translocations, with a proportion of 62.3% in the CCRI63 genome. Among them, intra-chromosomal translocation and inter-chromosomal translocation accounted for 9.4% and 52.9% respectively. A total of 40,130 CNVs were detected using CNVnator software, containing 3912 duplications and 36,218 deletions. According to the location of the annotated CNVs, we divided all CNVs into 3079 upstream, 4269 exonic, 1750 downstream 1142 intronic, 329 upstream/downstream and 29,560 intergenic categories (Supplementary Figure 6).

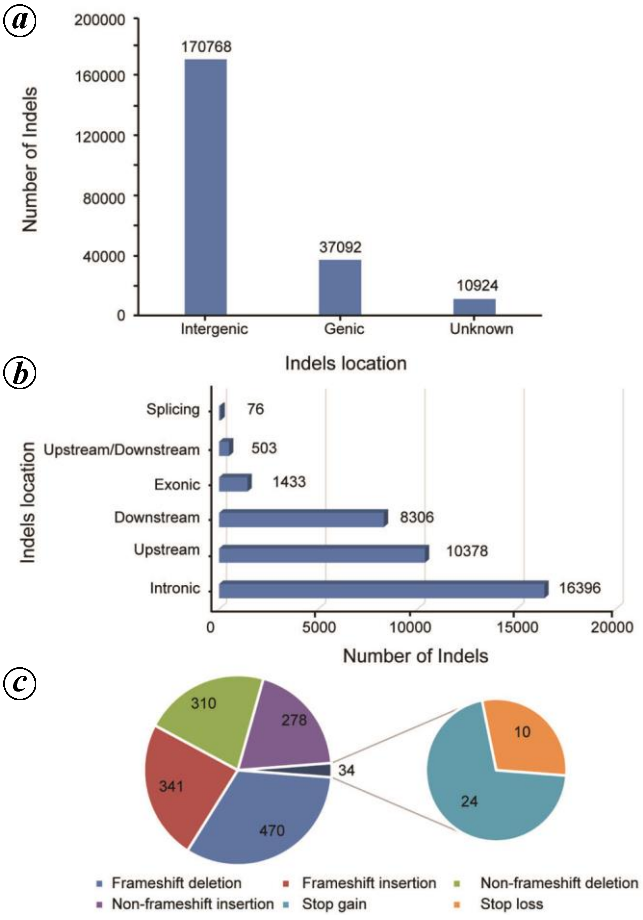

Figure 2. Characteristics of Indels in CCRI63. ***a***, Numbers of genic and intergenic Indels. ***b***, Distribution of genic Indels in CCRI63. ***c***, Distribution of exonic Indels in CCRI63. Upstream: 1 kb region upstream of a gene; Exonic: Indels located in exons; Intronic: Indels located in introns; Downstream: 1 kb downstream region of a gene; upstream/downstream: Indels located in the 1 kb region upstream of one gene and downstream of another gene; Intergenic: Indels located in intergenic regions. Stop gain is a mutation that introduces a stop codon. Stop loss is a mutation that eliminates a stop codon.

Detection and annotation of mutational hotspots

SNP and Indel densities were calculated with a window size of 100 kb and step size of 50 kb. The frequency of SNPs and Indels in the chromosomes was similar (Supplementary Figure 7). A map of the density of SNP and Indels on 26 chromosomes was constructed to show the variation distribution (Figure 3). The highest average mutation density appeared on chromosome A08, with SNP density of 0.0022974 and Indel density of 0.0024704. Windows with mutation density higher than 0.007 were extracted and merged to identify mutational hotspots. N/S ratio for genes in mutational hotspots (1.80) was marginally higher than the whole-genome average (1.17). In mutational hotspots, there were 347 genes containing 3746 SNPs, and the average number of SNPs was 10.8 per gene. The distribution of mutational hotspots in the chromosomes was uneven (Figure 4 *a*). A total of 1724 SNPs were located on chromosome D02, which was significantly more than the other chromosomes. A total of 75 genes containing 89 Indels was found in Indel hotspots and 31 Indels were located on chromosome D02 (Figure 4 *b*).

According to the reference genome annotation information, here were many functionally important variants on chromosome D02, such as DC1 domain-containing proteins, disease resistance proteins, leucine-rich receptor-like proteins, kinase family proteins and selenium-binding proteins. To further study the impact of these mutations on gene function, a gene ontology (GO) term enrichment analysis was performed for these genes. We detected five significant GO terms with *P*-values <0.05, containing 34 genes located on chromosome D02. Seven genes participated in lipid glycosylation processes (GO: 0030259, *P*-value = 0.0004); seven genes participated in lipid modification processes (GO: 0030258, *P*-value = 0.0177); four genes were associated with selenium binding (GO: 0008430, *P*-value = 0.0008); eight genes were

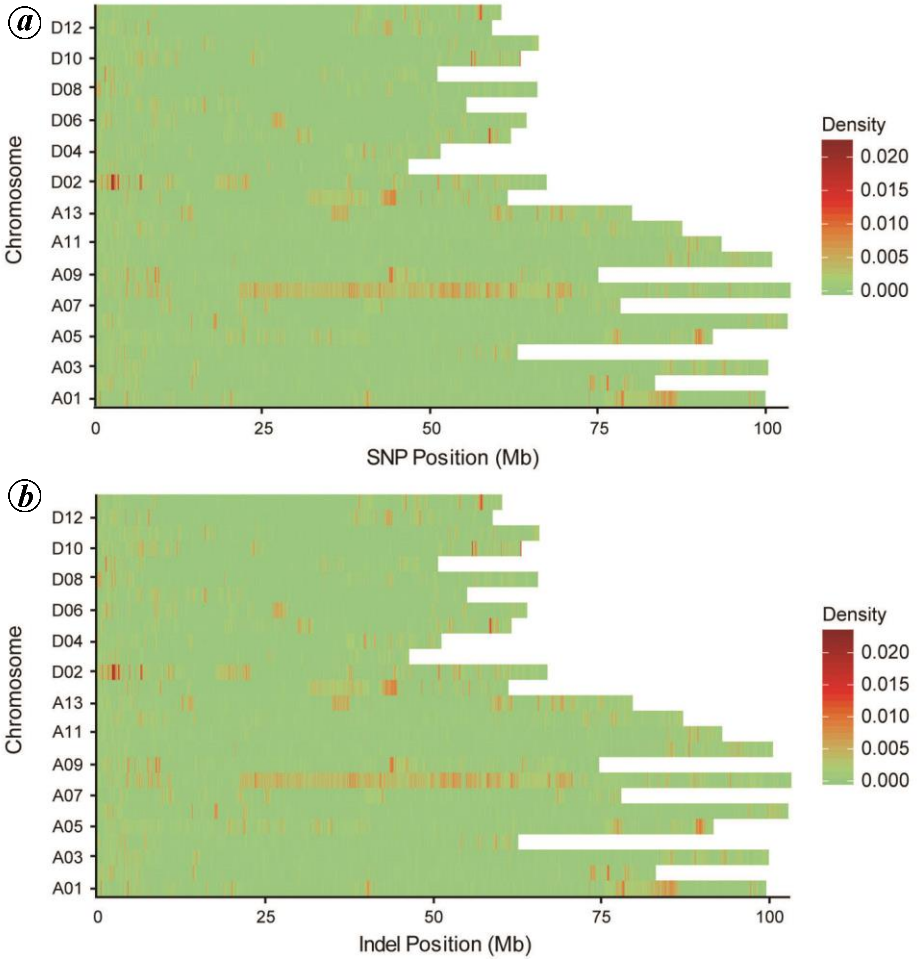

Figure 3. Density distributions of SNPs and Indels in chromosomes: ***a***, SNP density distribution; ***b***, Indel density distribution. SNP and Indel densities have been calculated with a window size of 100 kb and step size of 50 kb. Deeper red represents higher density.

associated with oxidoreductase activity, acting on sulphur group donors with NAD(P) as an acceptor (GO: 0016668, *P*-value = 0.0177); and eight genes were associated with protein-disulphide reductase activity (GO: 0016668, *P*-value = 0.0177, Supplementary Figure 8).

Distribution and potential effects on traits of heterozygous SNPs in CCRI63

Heterosis was positively related to the degree of dissimilarity between gametes and parents[33,34]. Heterozygous SNP loci are a direct manifestation of the differences among these gametes. A large number of heterozygous SNP sites (1,454,988) were distributed in CCRI63, accounting for 63.18% of all SNPs. To further study the distribution and potential function of these heterozygous SNP sites in CCRI63, densities of the heterozygous SNPs were calculated with a window size of 1 Mb and step size of 200 kb. The minimum and maximum density distribution within the window was 0.23e-04 and 0.77e-02 respectively, with an average of 0.50e-03. Based on the calculation, a map of the density of heterozygous sites on 26 chromosomes was constructed (Figure 5 and Supplementary Figure 9). The average density was a maximum of 0.20e-02 on the A08 chromosome and a minimum of 0.21e-03 on the D13 chromosome. The average density on 26 chromosomes was 0.48e-03.

In addition, we found that 73 QTLs overlapped with these high-density heterozygous mutation peaks (Figure 5 and Supplementary Figure 9). These QTLs contained the major agronomic traits, fibre quality traits and resistance traits in cotton, such as boll weight, boll number, seed index, lint percentage, fibre length, fibre strength and verticillium wilt resistance (Supplementary Table 2). Among these QTLs, 18 were located in the A subgenome and 55 in the D subgenome. This result shows that these high-density heterozygous intervals in chromosomes have important potential biological functions and may be involved in cotton traits. The results may provide favourable clues for further studies on the formation mechanism of cotton heterosis. Moreover, the differences in the

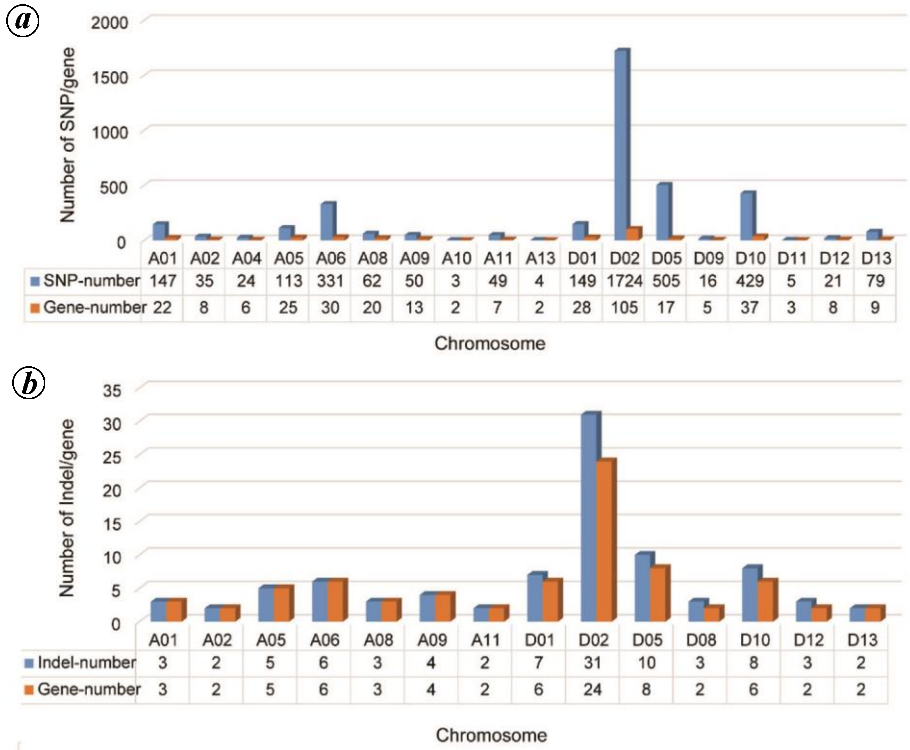

Figure 4. Distribution of (***a***) SNPs and genes, and (***b***) Indels and genes in mutational hotspots.

number of overlapped QTLs in the A and D subgenomes suggested that the D subgenome may make a more significant contribution to the improvement of CCRI63. It also implies that the D subgenome may be more important for the formation of CCRI63 heterosis.

Identification of novel genes

Although the majority of sequence reads obtained from CCRI63 were expected to align with the reference genome, the short reads derived from CCRI63 specific regions or regions missing in the reference genome cannot be aligned and may contain information important to the CCRI63 cultivar. There was considerable genetic information missing from the reference *G. hirsutum* genome. We retrieved 6.7 Mb of unmapped reads from short-insert paired-end libraries (Supplementary Table 3) and *de novo* assembled these reads. From this assembly, we identified 8121 contigs ≥150 bp (2.62 Mb in length) that were supported by at least two unmapped reads per base (Supplementary Table 4).

Characterization of the functions of genes in these missing sequences predicted 163 candidate novel genes; 87 (53.37%) of them were annotated functionally in publicly available databases (Supplementary Table 5). Aligning these novel genes to the reference genome yielded 47 hits, while only four of them were perfectly matched with 100% identity. In order to ensure the novelty of these genes, we discarded sequences with more than 95% identity, and 37 aligned genes were retained (Supplementary Table 6). Ultimately, 153 novel genes were retained, of which 116 (71.17%) could not be aligned to the reference completely and 37 genes that were aligned to the cotton genome with less than 95% identify (Supplementary Table 7).

Discussion

The rapid development of next-generation sequencing technologies and bioinformatic tools makes the *de novo* assembly and resequencing of many polyploidy species possible, which provides insight into the genetic variation and diversity occurring at the genome scale[35]. Since the sequencing of the allotetraploid cultivated upland cotton TM-1 (*G. hirsutum*, AADD) genome in 2015 (refs 7, 23), additional information on the genetic variation in different upland cotton cultivars will benefit breeding and crop improvement. By resequencing the elite upland cotton cultivar CCRI63, we have uncovered 2,302,944 SNPs, 218,784 Indels, 57,906 SVs and 45,764 CNVs. A majority of SNPs, accounting for 88.59% of the total, were located in the intergenic regions, which was close to soybean (86.5%) and sorghum (83%), and higher than that in rice (56.4%)[18–20,36]. A possible reason is that the cotton genome has a larger proportion of intergenic regions than rice[7,37]. Large effective SNPs that impact codon sequence translation by causing splicing or resulting in the gain or loss of a stop codon often affect important biological

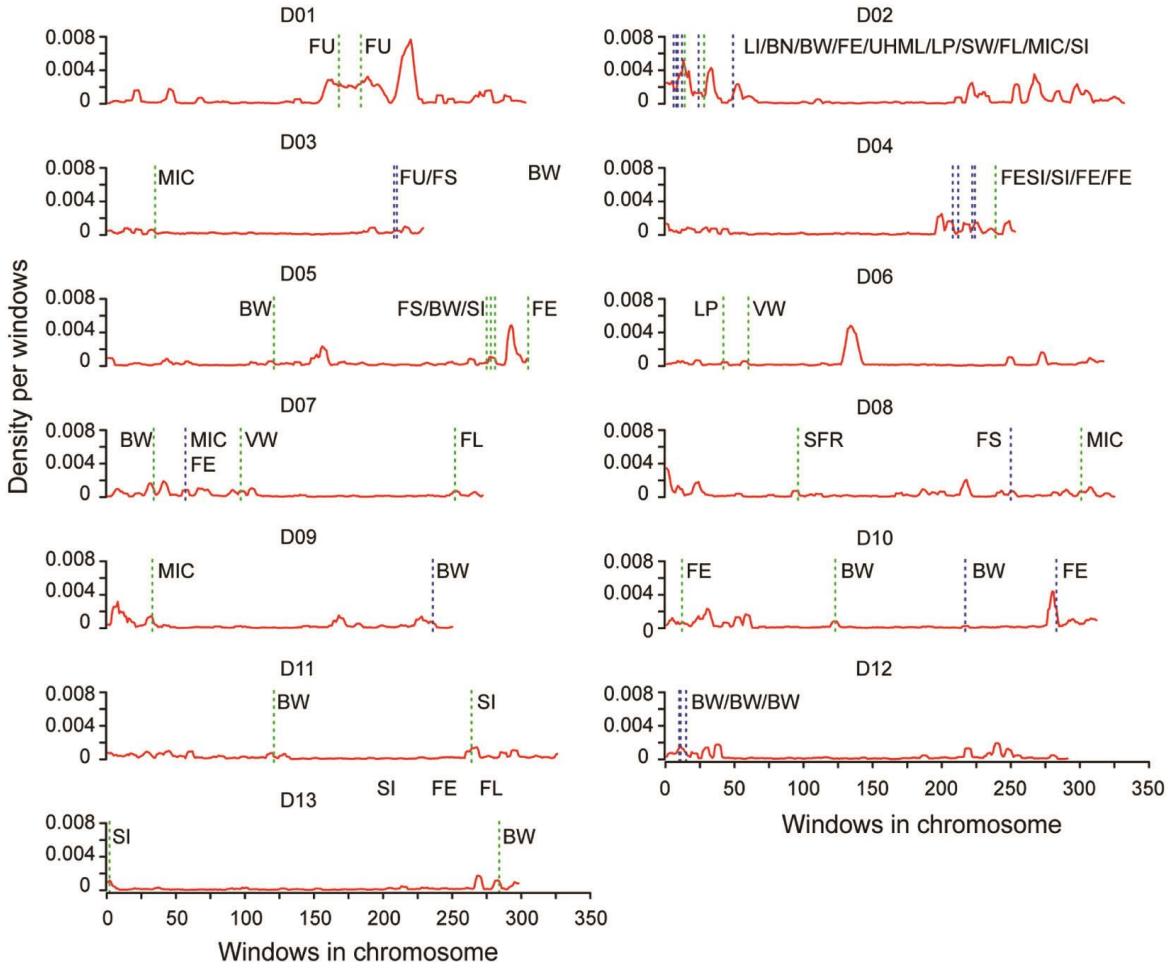

Figure 5. Map of the density of heterozygous sites on D subgenome. The 13 cotton chromosomes are indicated by D1–D13. Horizontal red curves indicate the heterozygous SNP loci density calculated with a window size of 1 Mb and step size of 200 kb. The axis of abscissa indicates sliding windows with a window size of 1 Mb and step size of 200 kb. Corresponding density per window is marked on the left vertical scale. The blue and green dotted lines represent QTL and GWAS loci respectively, and the related traits are marked near them (BN, Boll number; BW: Boll weight; FE, Fibre elongation; FL, Fibre length; MIC, Fibre micronaire; FS, Fibre strength; FU, Fibre uniformity; LI, Lint index; LP, Lint percentage; SI, Seed index; SFR, Short fibre rate; VW, Verticillium wilt).

metabolites or pathways[36,38]. In the present study, 922 large effective SNPs were uncovered in CCRI63, which affected 818 genes. Fourteen of these genes are involved in plant–pathogen interaction, which can provide clues for further analysis of CCRI63 resistance.

By scanning the distribution of genome-wide variation, we can explore high-frequency mutation regions and relatively conserved regions. The latter may involve manual selection in material, while the former may contain important genomic segments and genes that reflect individual characteristics[39,40]. Polymorphism of trait may also be related to these high-frequency mutant genomic segments[22]. In CCRI63, the highest average mutation density appeared on chromosome A08, with a SNP density of 0.0022974 and Indel density of 0.0024704. In mutational hotspots of CCRI63, the N/S ratio for genes (1.80) was larger than the whole-genome average (1.17). This increase in the ratio indicates that variation sites have undergone a positive manual selection[41]. So we infer that more favourable mutations in hotspots were retained with positive selection, which may be the result of artificial selection. Genes located in hotspots participated in many important metabolic pathways such as phosphatidylinositol signalling system, inositol phosphate metabolism and plant–pathogen interaction. These hotspot regions can provide a new insight into studying the genomic characteristics of CCRI63. Assembling unmapped reads provides further insights on the missing regions in the reference genome or unique regions in variety[18]. To prevent the loss of genetic information, we took full advantage of unmapped reads to reconstruct the 'missing sequences' for the CCRI63 genome. As a result, a total of 153 putative novel genes were discovered by assembling the unmapped reads. Among them, the functions of 87 genes

were predicted in different databases (Supplementary Table 5). According to the prediction results, these genes were found to be involved in some important biological functions such as photosynthesis metabolism. The genes may help explore the unique genetic characteristics of CCRI63.

Heterosis is a common genetic phenomenon in nature and the main way to improve crops. Heterosis is positively related to the degree of dissimilarity between gametes and parents[33,34]. The overdominance hypothesis attributes heterosis to heterozygous locus[42,43]. We found that 63.18% of all SNPs were heterozygous alleles in CCRI63, with a distribution density of one heterozygous SNP per 1.7 kb, which is much higher than that in conventional cotton varieties[44]. QTL mapping is not only a good strategy for analysing the function of cotton genomic fragments, but also an important method to study heterosis[14,15,17]. These peak intervals of high-density distribution of heterozygous SNP sites on each chromosome overlapped with 73 credible QTLs across the chromosomes (Figure 5 and Supplementary Figure 9). These QTLs were closely related to the main agronomic traits, fibre quality traits and resistance traits in cotton (Supplementary Table 2). Eighteen of these QTLs were located in the A subgenome, of which ten were related to yield traits and eight with fibre quality traits. There were three important QTL loci associated respectively, with boll weight, boll number and lint percentage on the A11 chromosome. Six QTLs associated with fibre quality were located on the A13 chromosome. These results show that heterozygous sites on A11 and A13 chromosomes respectively, are important for the improvement of CCRI63 yield and fibre quality. Compared with the A subgenome, more functional QTLs were distributed in the D subgenome. A total of 55 QTLs were located in the high-density heterozygous region of the D subgenome. Among them, 28, 25 and 2 were related to yield, fibre quality and disease resistance respectively. Judging from the number of QTLs, the high-density heterozygous region on the D subgenome may have a greater effect than the A subgenome on the CCRI63 trait. QTL clusters were formed in high-density heterozygous regions on D02, D04 and D05 chromosomes (Figure 5). There were 17 QTL loci associated with yield and fibre quality traits in the QTL cluster located on D02 chromosome, which highlights the importance of this heterozygous region (Supplementary Table 2). These heterozygous loci are of importance to explore the formation of heterosis in CCRI63.

1. Fryxell, P., A revised taxonomic interpretation of *Gossypium* L. (Malvaceae). *Rheedea*, 1992, **2**, 108–165.
2. Brubaker, C. L., Bourland, F. and Wendel, J. F., The origin and domestication of cotton. In *Cotton: Origin, History, Technology and Production*, John Wiley, New York, 1999, pp. 3–31.
3. Kim, H. J. and Triplett, B. A., Cotton fibre growth *in planta* and *in vitro*. Models for plant cell elongation and cell wall biogenesis. *Plant Physiol.*, 2001, **127**, 1361–1366.
4. Shangguan, X. X., Yang, C. Q., Zhang, X. F. and Wang, L. J., Functional characterization of a basic helix–loop–helix (bHLH) transcription factor GhDEL65 from cotton (*Gossypium hirsutum*). *Physiol. Plant.*, 2016, **158**(2), 200–212.
5. Tiwari, N., Sharma, P. K. and Malathi, V. G., Functional characterization of βC1 gene of cotton leaf curl multan betasatellite. *Virus Genes*, 2013, **46**, 111–119.
6. Fang, L., Tian, R., Li, X., Chen, J., Wang, S., Wang, P. and Zhang, T., Cotton fibre elongation network revealed by expression profiling of longer fibre lines introgressed with different *Gossypium barbadense* chromosome segments. *BMC Genomics*, 2014, **15**, 838.
7. Zhang, T. *et al.*, Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. Acc. Tm-1) provides a resource for fibre improvement. *Nature Biotechnol.*, 2015, **33**, 531–537.
8. Wang, K., Huang, G. and Zhu, Y., Transposable elements play an important role during cotton genome evolution and fibre cell development. *Sci. China Life Sci.*, 2016, **59**, 112–121.
9. Tang, M. *et al.*, Rapid evolutionary divergence of *Gossypium barbadense* and *G. hirsutum* mitochondrial genomes. *BMC Genomics*, 2015, **16**, 770.
10. Liu, Y. *et al.*, A *Gossypium* BAC clone contains key repeat components distinguishing sub-genome of allotetraploidy cottons. *Mol. Cytogenet.*, 2016, **9**, 27.
11. Campbell, B. T. *et al.*, Status of the global cotton germplasm resources. *Crop Sci.*, 2010, **50**, 1161.
12. Shang, L. *et al.*, Genetic analysis of upland cotton dynamic heterosis for boll number per plant at multiple developmental stages. *Sci. Rep.*, 2016, **6**, 35515.
13. Li, B. *et al.*, Genetic effects and heterosis of yield and yield component traits based on *Gossypium barbadense* chromosome segment substitution lines in two *Gossypium hirsutum* backgrounds. *PLoS One*, 2016, **11**, e0157978.
14. Liu, R. *et al.*, Quantitative trait loci mapping for yield and its components by using two immortalized populations of a heterotic hybrid in *Gossypium hirsutum* L. *Mol. Breed.*, 2012, **29**, 297–311.
15. Guo, X. *et al.*, Mapping heterotic loci for yield and agronomic traits using chromosome segment introgression lines in cotton. *J. Integr. Plant Biol.*, 2013, **55**, 759–774.
16. Liang, Q., Shang, L., Wang, Y. and Hua, J., Partial dominance, overdominance and epistasis as the genetic basis of heterosis in upland cotton (*Gossypium hirsutum* L.). *PLOS ONE*, 2015, **10**, e0143548.
17. Shang, L., Wang, Y., Cai, S., Wang, X., Li, Y., Abduweli, A. and Hua, J., Partial dominance, overdominance, epistasis and QTL by environment interactions contribute to heterosis in two upland cotton hybrids. *G3: Genes|Genomes|Genetics* (*Bethesda, Md*), 2015, **6**, 499–507.
18. Mace, E. S. *et al.*, Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Commun.*, 2013, **4**, 2320.
19. Lin, J. *et al.*, Genome re-sequencing and bioinformatics analysis of a nutraceutical rice. *Mol. Genet. Genomics: MGG*, 2015, **290**, 955–967.
20. Kim, M. Y. *et al.*, Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* sieb. and zucc.) genome. *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 22032–22037.
21. Chung, W. H. *et al.*, Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res.*, 2014, **21**, 153–167.
22. Srivastava, S. K., Wolinski, P. and Pereira, A., A strategy for genome-wide identification of gene based polymorphisms in rice reveals non-synonymous variation and functional genotypic markers. *PLoS One*, 2014, **9**, e105335.
23. Li, F. *et al.*, Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnol.*, 2015, **33**, 524–530.

24. Paterson, A. H., Brubaker, C. L. and Wendel, J. F., A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Rep*., 1993, **11**, 122–127.
25. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 2009, **25**, 1754–1760.
26. Li, H. *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, **25**, 2078–2079.
27. Chen, K. *et al.*, Break dancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Meth*., 2009, **6**, 677–681.
28. Abyzov, A., Urban, A. E., Snyder, M. and Gerstein, M., CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*., 2011, **21**, 974–984.
29. Wang, K., Li, M. and Hakonarson, H., ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res*., 2010, **38**, e164–e164.
30. Li, R. *et al.*, *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*., 2010, **20**, 265–272.
31. Lam, H. M. *et al.*, Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genet*., 2010, **42**, 1053–1059.
32. McNally, K. L. *et al.*, Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 12273–12278.
33. Shull, G. H., What is 'heterosis'? *Genetics*, 1948, **33**, 439–446.
34. Fu, D. *et al.*, What is crop heterosis: new insights into an old topic. *J. Appl. Genet*., 2015, **56**, 1–13.
35. Schuster, S. C., Next-generation sequencing transforms today's biology. *Nat. Meth*., 2008, **5**, 16–18.
36. Zhou, D. *et al.*, Pedigree-based analysis of derivation of genome segments of an elite rice reveals key regions during its breeding. *Plant Biotechnol. J*., 2016, **14**, 638–648.
37. Ouyang, S. *et al.*, The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res*., 2007, **35**, D883–D887.
38. Lai, J. *et al.*, Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genet*., 2010, **42**, 1027–1030.
39. Wang, M. *et al.*, Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genet*., 2017, **49**, 579–587.
40. Fang, L. *et al.*, Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. *Genome Biol*., 2017, **18**, 33.
41. Yang, Z. and Bielawski, J. P., Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol*., 2000, **15**, 496–503.
42. Crow, J. F., 90 years ago: the beginning of hybrid maize. *Genetics*, 1998, **148**, 923–928.
43. East, E. M., Inbreeding in corn. *Rep. Conn. Agric. Exp. Stn*, 1908, **1907**, 419–428.
44. Fang, L. *et al.*, Genomic analyses in cotton identify signatures of selection and loci associated with fibre quality and yield traits. *Nature Genet*., 2017, **49**, 1089–1098.

ACKNOWLEDGEMENTS. This work was supported by the Natural Science Foundation of Henan Province of China (No. 152300410010), the National Natural Science Foundation of China (No. 31401431) and National the Key Research and Development Program of China (No. 2016YFD0100300). We thank the staff of Novogene Bioinformatics Institute for technical support.

Received 11 December 2017; accepted 3 May 2018

doi: 10.18520/cs/v115/i4/701-709