Genome-wide association study identifies the loci and genes related to days to flowering, fiber length and strength in upland cotton

Zhiying Ma
Hebei Agricultural University
Baoding, China
2018.5.30

Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield

Zhiying Ma1,2, Shoupu He2,3, Xingfen Wang1,2, Junling Sun2,3, Yan Zhang1,4, Guiyin Zhang1,4, Liqiang Wu1,4, Zhihong Li2,3, Zhihao Liu1,4, Gaofei Sun5, Yuanyuan Yan1,5, Yinhua Jia2,5, Jun Yang3,5, Zhong Pan6, Qiwen Gu7, Xueyuan Li8, Zhengwen Sun9, Panhong Dai1,2, Zhengwen Liu1,3, Wenfang Gong1, Jinhua Wu1, Mi Wang1, Hengwei Liu1, Keyun Feng1, Hufeng Ke1, Junduo Wang1, Hongyu Lang1, Guoning Wang1, Jun Peng2, Nan Wang1, Liru Wang1, Baoxian Pang2, Zhen Peng2, Ruiqiang Li2, Shilin Tian1,2 and Xiongming Du1,2

Upland cotton is the most important natural-fiber crop. The genomic variation of diverse germplasms and alleles underpinning fiber quality and yield should be extensively explored. Here, we resequenced a core collection comprising 419 accessions with 6.55-fold coverage depth and identified approximately 3.66 million SNPs for evaluating the genomic variation. We performed phenotyping across 12 environments and conducted genome-wide association study of 13 fiber-related traits. 7,393 unique SNPs were significantly associated with these traits and were located within or near 4,820 genes; more associated loci were detected for fiber quality than fiber yield, and more fiber genes were detected in the D than the A subgenome. Several previously undescribed causal genes for days to flowering, fiber length, and fiber strength were identified. Phenotypic selection for these traits increased the frequency of elite alleles during domestication and breeding. These results provide targets for molecular selection and genetic manipulation in cotton improvement.
1. Phenotyping of 419 cotton accessions

- **A core collection**: representing ~5.7% of 7,362 upland cottons
- **Phenotyping**: six agro-ecologically diverse locations in 2014 and 2015
- **13 phenotypic traits**: important for fiber yield and quality

Fig. 1 The geographic distribution of 419 cottons

Fig. 2 The correlation coefficients among the traits
2. Resequencing of 419 cotton accessions

- generated **6.35 Tb** high-quality sequences
- **99.56%** of the reads covered **92.20%** of reference genome
- average of **6.5-fold depth**
- identified **~3.6 million SNPs**

### Table 1 Summary of categorized SNPs

<table>
<thead>
<tr>
<th>SNP category</th>
<th>SNP number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream</td>
<td>34,088</td>
</tr>
<tr>
<td>Downstream</td>
<td>34,689</td>
</tr>
<tr>
<td>Upstream / Downstream</td>
<td>2,182</td>
</tr>
<tr>
<td>Stop gain</td>
<td>1,050</td>
</tr>
<tr>
<td>Stop loss</td>
<td>210</td>
</tr>
<tr>
<td>Synonymous</td>
<td>29,546</td>
</tr>
<tr>
<td><strong>Non-synonymous</strong></td>
<td><strong>47,995</strong></td>
</tr>
<tr>
<td>Intronic</td>
<td>145,036</td>
</tr>
<tr>
<td>Splicing</td>
<td>364</td>
</tr>
<tr>
<td>Intergenic</td>
<td>3,369,870</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,665,030</strong></td>
</tr>
</tbody>
</table>
This population was not highly structured, with moderate LD, and suitable for the GWAS.

Fig. 3 Phylogenetic tree, PCA and Genetic differentiation and LD decay of the 419 accessions
4. GWAS for 13 fiber-related traits

Table 2  The associated SNP number

<table>
<thead>
<tr>
<th>Trait</th>
<th>Associated SNP number</th>
<th>Trait</th>
<th>Associated SNP number</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>1661</td>
<td>BW</td>
<td>119</td>
</tr>
<tr>
<td>FS</td>
<td>735</td>
<td>LP</td>
<td>1049</td>
</tr>
<tr>
<td>M</td>
<td>533</td>
<td>SI</td>
<td>119</td>
</tr>
<tr>
<td>E</td>
<td>297</td>
<td>LI</td>
<td>674</td>
</tr>
<tr>
<td>LU</td>
<td>688</td>
<td>FWPB</td>
<td>731</td>
</tr>
<tr>
<td>MAT</td>
<td>1614</td>
<td>FD</td>
<td>1268</td>
</tr>
<tr>
<td>SCI</td>
<td>1538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>11026</td>
</tr>
</tbody>
</table>

- 3,665,030 SNPs for GWAS
- log10(\(P\)) < 10^{-6}
- 11,026 SNP signals are significantly associated with 13 traits (7,383 excluding the repeated)
- 3,806 SNPs identified at least three times
- 7,398 genes were detected across the 13 traits
- 4,820 genes (excluding the repeated)
- 3,089 (64.1%) genes high expressed across fiber developmental stages
5. Characteristics

① a core collection (85% different from previous)
② a large sample size
③ phenotyping across 12 environments
④ deeper resequencing depth
⑤ more phenotypic traits
⑥ more unique associated SNPs identified
⑦ GWAS & RNAseq
⑧ genes function validation

Fig.5 A comprehensive diagram for the relationships among chromosomes, associated SNPs and genes, traits, fiber developmental stages and transcriptome analysis
6. Identification of flowering and fiber-initiation genes

- **FD** was positively correlated with following traits (FL, FS, E, LU, SCI, LP, LI, FWPB)

- 94.6% of the associated SNPs were located on Dt03

6.1 Identification of the FD causal gene **GhCIP1**

Fig. 6 The FD causal gene **GhCIP1** on chro. Dt03
6.2 Identification of FD causal gene GhUCE

- encoding ubiquitin-conjugating enzyme, ortholog AtUCE and function unknown
- containing three significantly associated intronic SNPs

Fig. 7 The FD causal gene GhUCE on chro. Dt03
7. Identification of fiber-length-related genes

- **FL-associated SNP**: 1,661
- **646 (38.9%)、755 (45.5%) SNP** located in At10 and Dt11

### 7.1 Identification of the FL causal gene *GhFL1*

**Fig. 8** The FL causal gene *GhFL1* on chro. Dt11
7.2 Identification of the FL causal gene *GhFL2*

- encoding KIP-related protein 6 (KRP6), *AtKRP6* function unknown
- *AtKRP5* is required for cell elongation in Arabidopsis

**Fig. 8** The FL causal gene *GhFL2* on chro. Dt11
8. Identification of fiber-strength-related genes

- FS: 735 significantly associated SNP
- 391 (53.1%) and 239 (32.5%) SNP located in At07 and Dt11
- At07: 72.17~72.23 Mb, contained 68 SNPs and three genes

**Identification of the Fs causal gene Gh_A07G1769**

**Fig.9** The causal FS gene for the peak on chro. At07
Acknowledgements

Supporting Funds
- China Agriculture Research System
- Science & Technology Support Program of Hebei
- National Major Science & Technology Program
- National Key Research & Development Program

Materials
- National Mid-term Gene Bank for Cotton

Affiliation
- Hebei Agricultural University; CRI, CAAS; Novogene Bioinformatics Institute