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

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0119-7

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

Zhiying Ma[1,9*], Shoupu He[2,9], Xingfen Wang[1,9*], Junling Sun[2,9], Yan Zhang[1,9], Guiyin Zhang[1,9], Liqiang Wu[1,9], Zhikun Li[1,9], Zhihao Liu[3,9], Gaofei Sun[4], Yuanyuan Yan[1], Yinhua Jia[2], Jun Yang[1], Zhaoe Pan[2], Qishen Gu[1], Xueyuan Li[5], Zhengwen Sun[1], Panhong Dai[2,6], Zhengwen Liu[1], Wenfang Gong[2], Jinhua Wu[1], Mi Wang[6], Hengwei Liu[7], Keyun Feng[8], Huifeng Ke[1], Junduo Wang[5], Hongyu Lan[8], Guoning Wang[1], Jun Peng[2], Nan Wang[1], Liru Wang[2], Baoyin Pang[2], Zhen Peng[2], Ruiqiang Li[3], Shilin Tian[3*] and Xiongming Du[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. 7,383 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

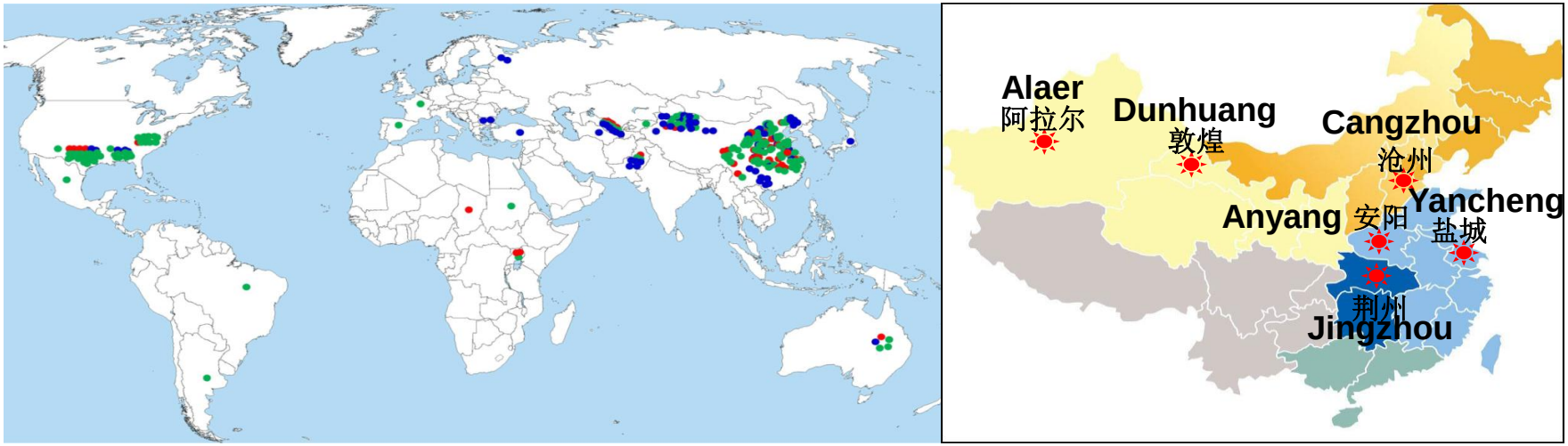


Fig.1 The geographic distribution of 419 cottons

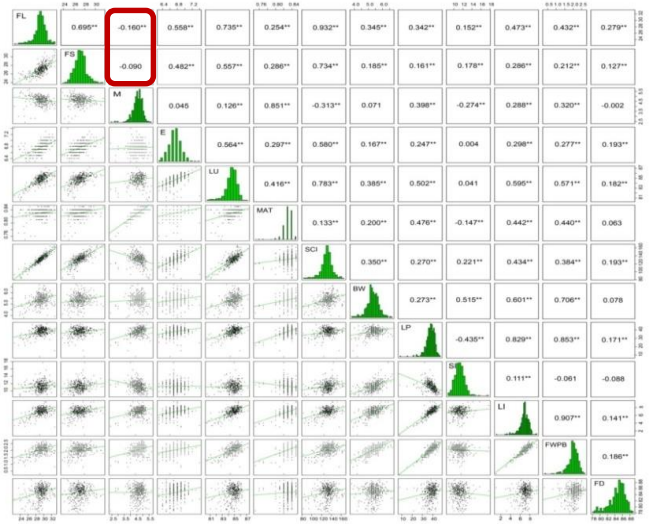


Fig. 2 The correlation coefficients among the traits

- **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 yeld and quality

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

SNP category	SNP number
Upstream	34,088
Downstream	34,689
Upstream / Downstream	2,182
Stop gain	1,050
Stop loss	210
Synonymous	29,546
Non-synonymous	47,995
Intronic	145,036
Splicing	364
Intergenic	3,369,870
Total	3,665,030

3. Genomic variation and population structure

This population was not highly structured, with moderate LD, and suitable for the GWAS.

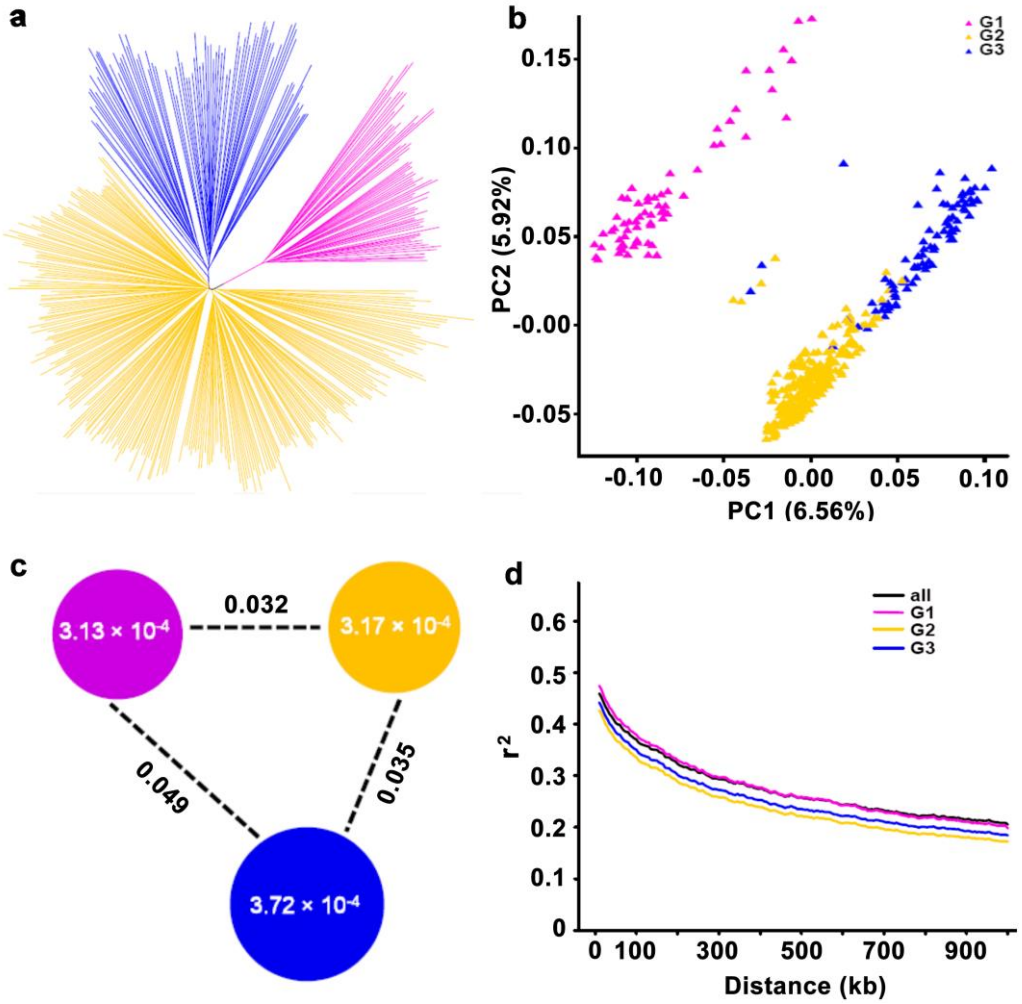


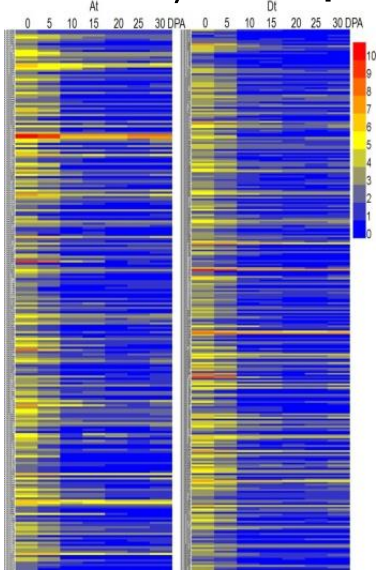
Fig. 3 Phylogenetic tree, PCA and Genetic differentiation and LD decay of the 419 accessions

4. GWAS for 13 fiber-related traits

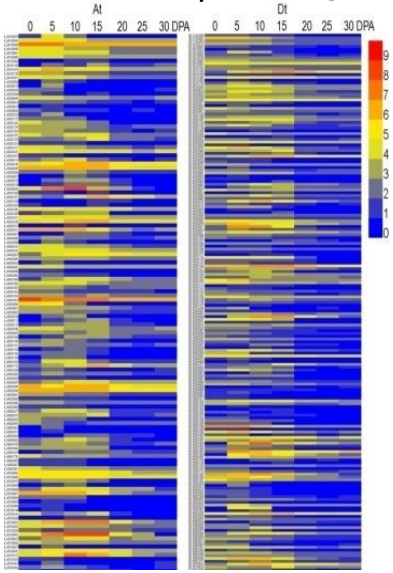
Table 2 The associated SNP number

Trait	Associated SNP number	Trait	Associated SNP number
FL	1661	BW	119
FS	735	LP	1049
M	533	SI	119
E	297	LI	674
LU	688	FWPB	731
MAT	1614	FD	1268
SCI	1538		
Total		11026	

Pattern I, 1185个

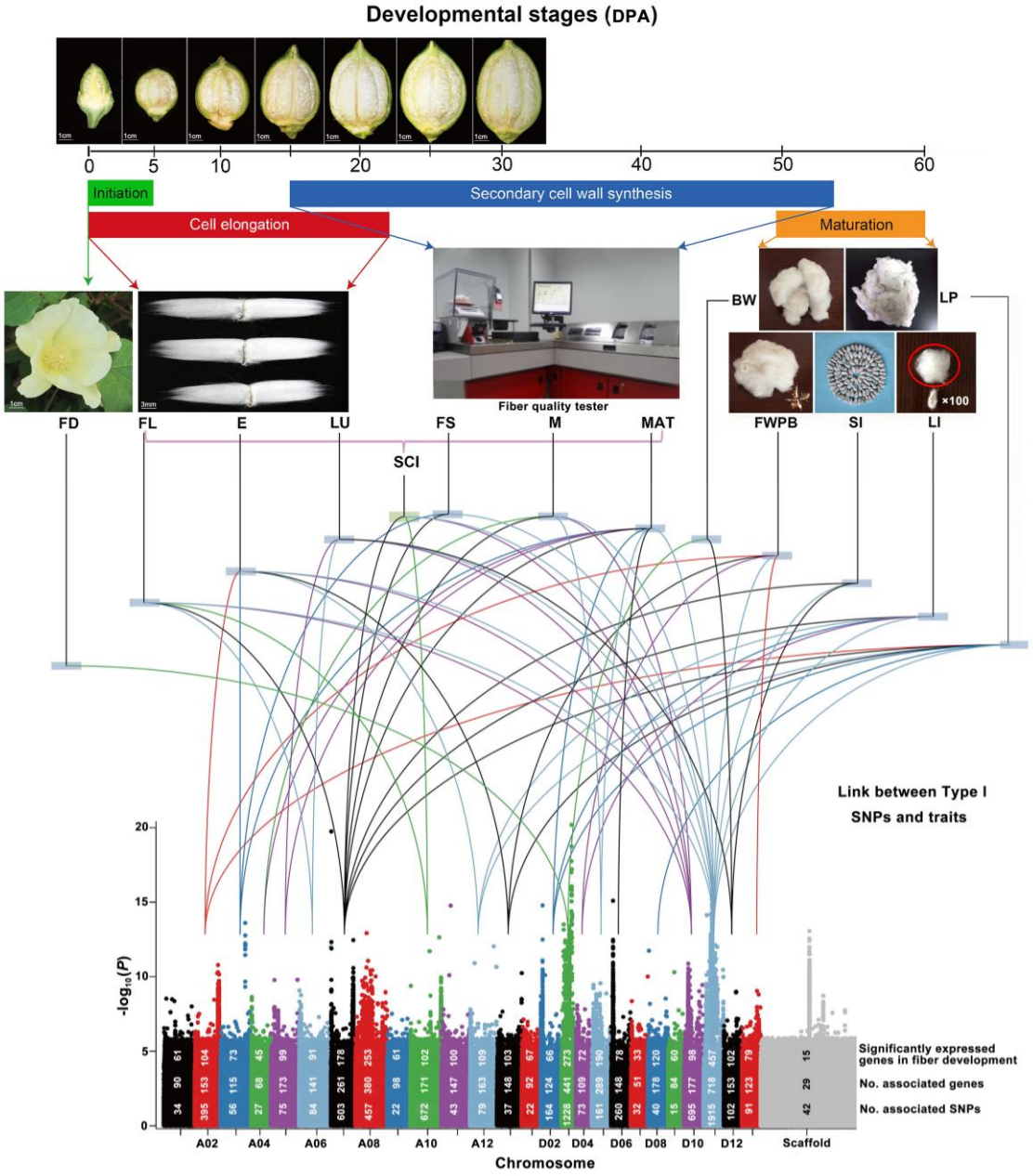


Pattern II, 708个



- **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)**
- **3,806SNPs 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*

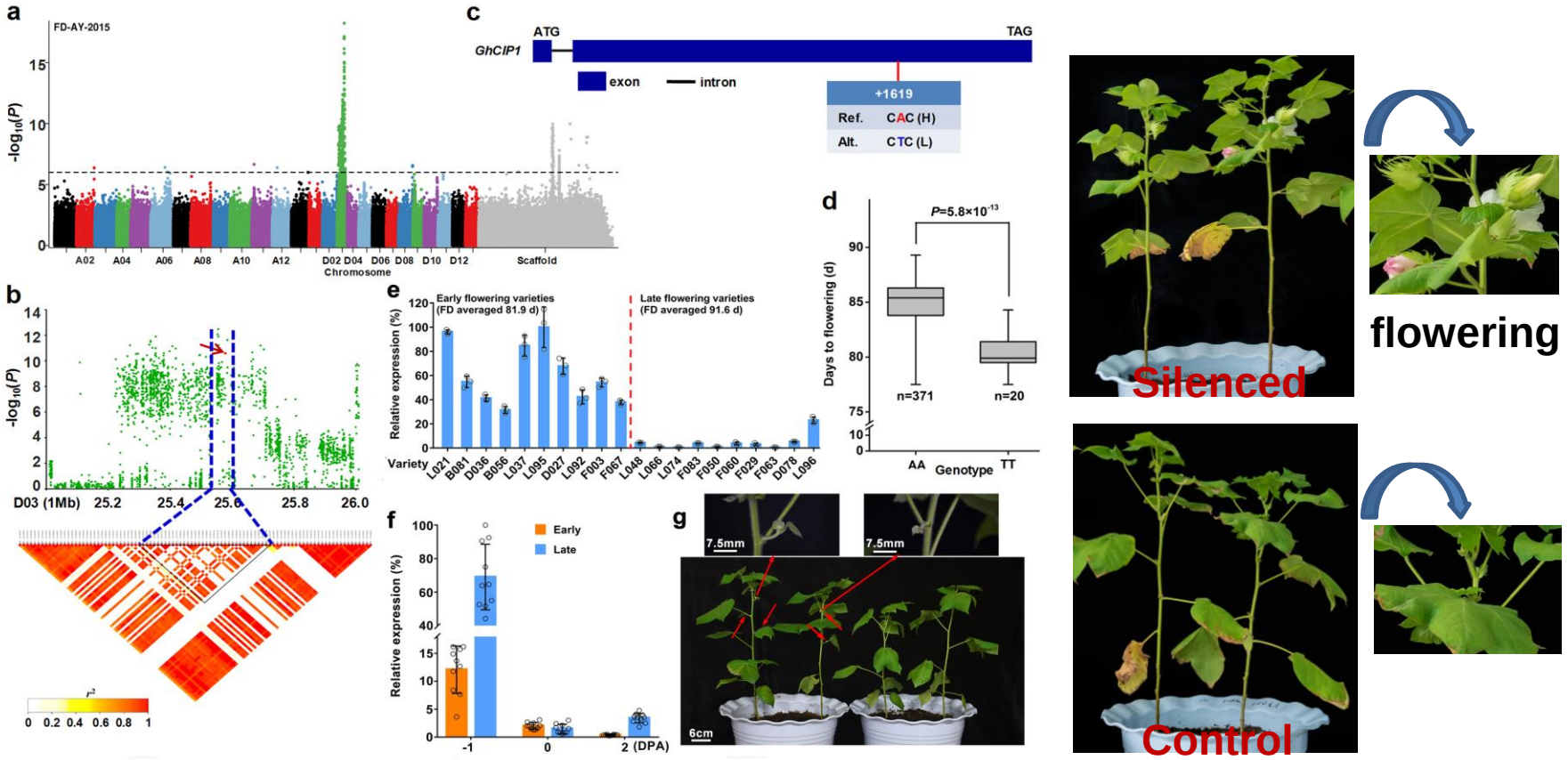


Fig. 6 The FD causal gene *GhCIP1* on chro. Dt03

6.2 Identification of FD causal gene *GhUCE*

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

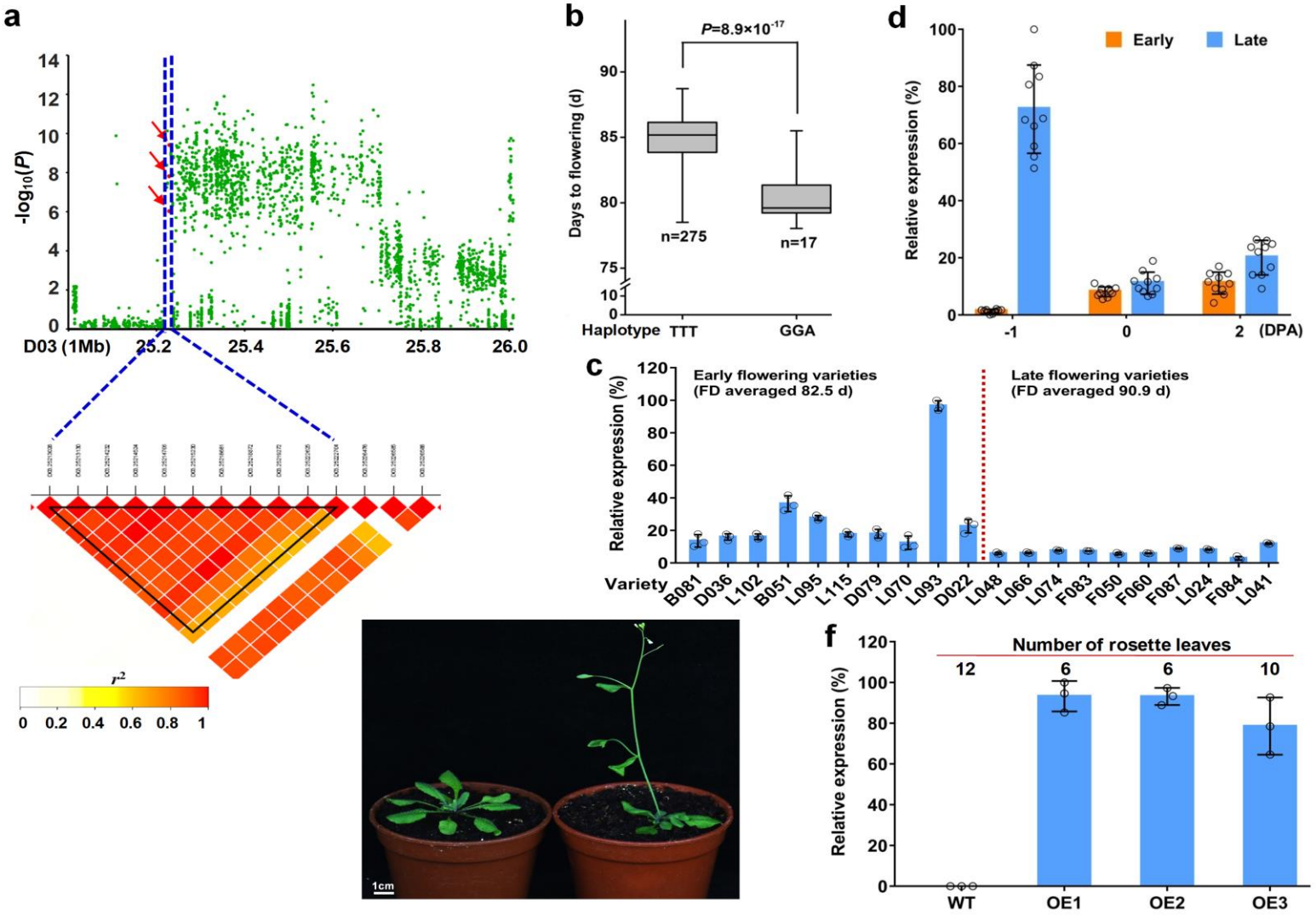
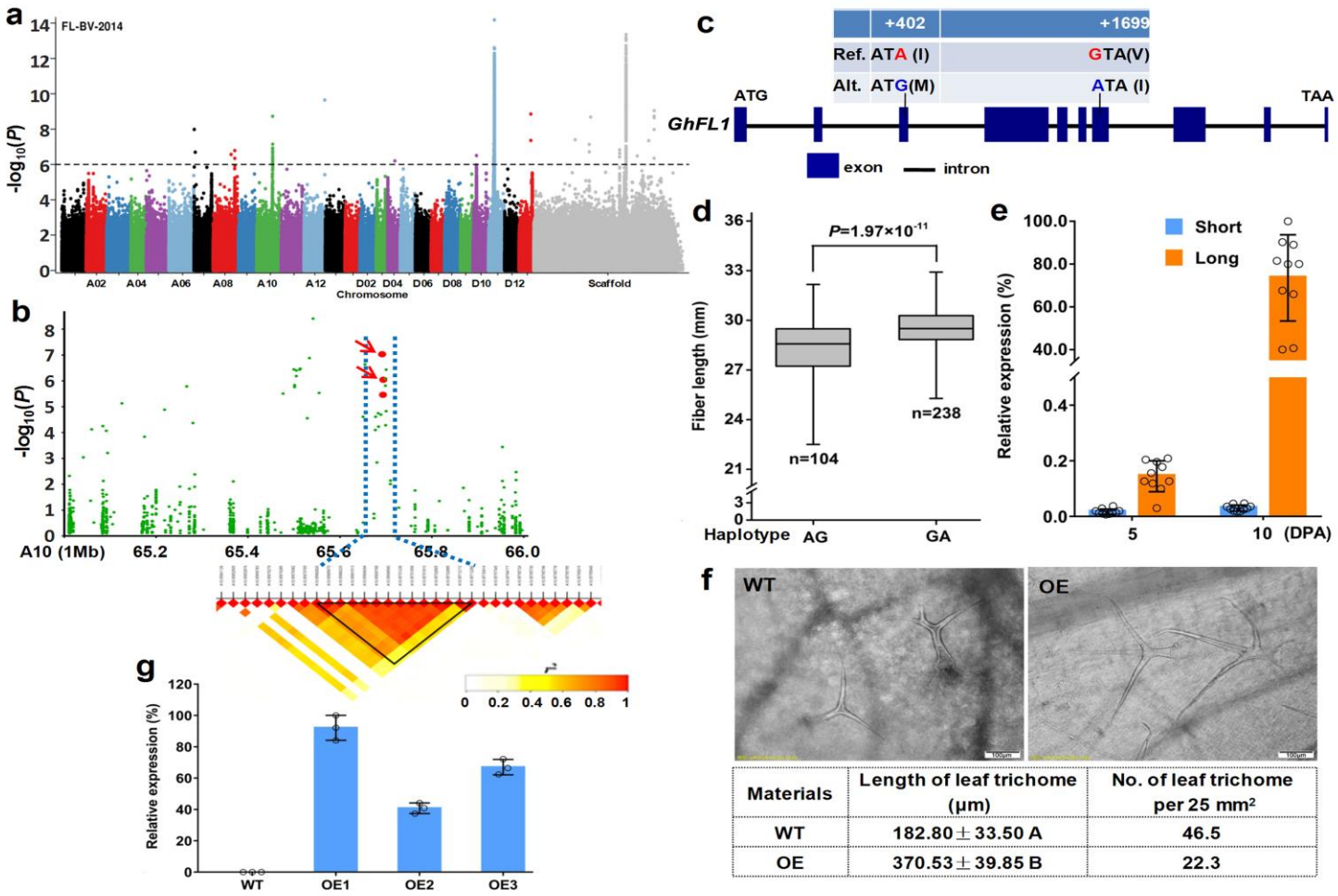


Fig. 7 The FD causal gene *GhUCE* on chro. Dt03

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*

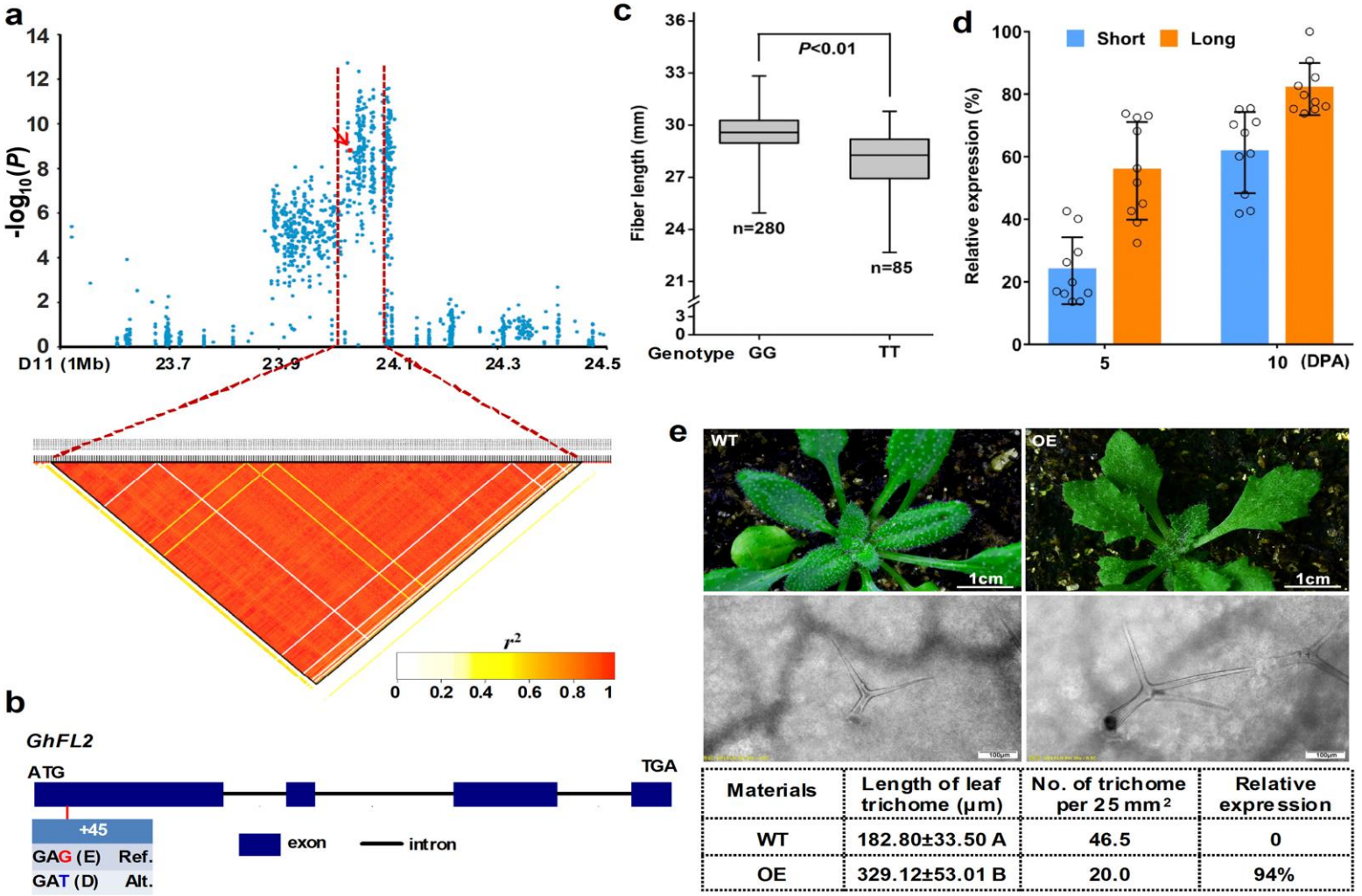


Materials	Length of leaf trichome (μm)	No. of leaf trichome per 25 mm²
WT	182.80±33.50 A	46.5
OE	370.53±39.85 B	22.3

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



Materials	Length of leaf trichome (μm)	No. of trichome per 25 mm²	Relative expression
WT	182.80±33.50 A	46.5	0
OE	329.12±53.01 B	20.0	94%

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*

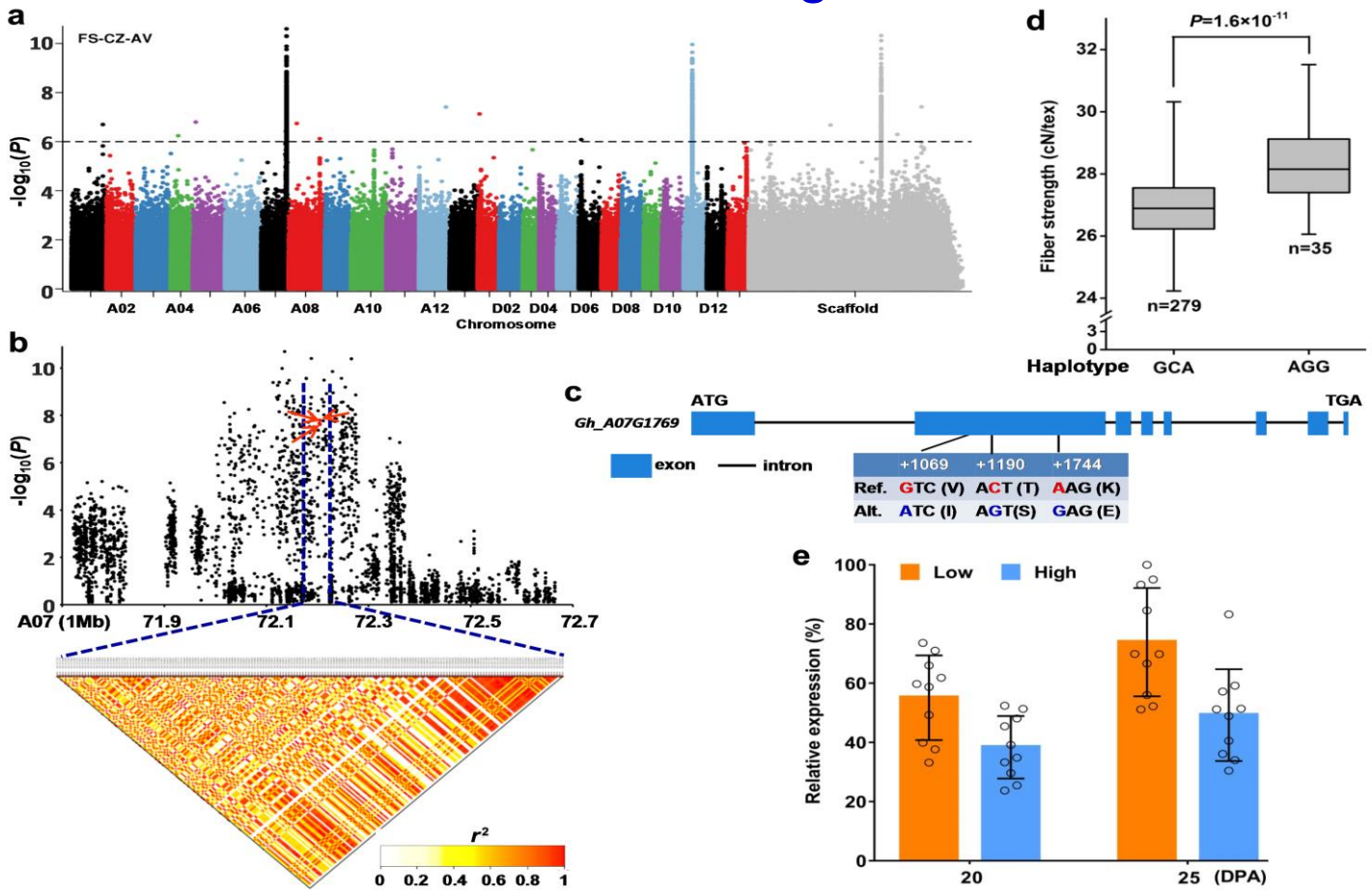


Fig.9 The causal FS gene for the peak on chro. At07

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