Research

New Phytologist

Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton

Zhonghua Li[1]*, Pengcheng Wang[1]*, Chunyuan You[2], Jiwen Yu[3], Xiangnan Zhang[1], Feilin Yan[1], Zhengxiu Ye[1], Chao Shen[1], Baoqi Li[1], Kai Guo[1], Nian Liu[1], Gregory N. Thyssen[4], David D. Fang[4], Keith Lindsey[5], Xianlong Zhang[1], Maojun Wang[1] and Lili Tu[1]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, Hubei, China; [2]Cotton Research Institute, Shihezi Academy of Agriculture Science, Shihezi 832000, Xinjiang, China; [3]State Key Laboratory of Cotton Biology, Cotton Institute of the Chinese Academy of Agricultural Sciences, Anyang 455000, China; [4]Cotton Fibre Bioscience Research Unit, USDA-ARS, Southern Regional Research Center, New Orleans, LA 70124, USA; [5]Department of Biosciences, Durham University, Durham, DH1 3LE, UK

Authors for correspondence:
Maojun Wang
Tel: +86 02787283955
Email: mjwang@mail.hzau.edu.cn

Lili Tu
Tel: +86 02787283955
Email: lilitu@mail.hzau.edu.cn

Received: *2 September 2019*
Accepted: *28 January 2020*

New Phytologist (2020) **226:** 1738–1752
doi: 10.1111/nph.16468

Key words: cell elongation, cell wall synthesis, cotton fibre, eQTL, genetic regulation.

Summary

- The cotton fibre serves as a valuable experimental system to study cell wall synthesis in plants, but our understanding of the genetic regulation of this process during fibre development remains limited.
- We performed a genome-wide association study (GWAS) and identified 28 genetic loci associated with fibre quality in allotetraploid cotton. To investigate the regulatory roles of these loci, we sequenced fibre transcriptomes of 251 cotton accessions and identified 15 330 expression quantitative trait loci (eQTL).
- Analysis of local eQTL and GWAS data prioritised 13 likely causal genes for differential fibre quality in a transcriptome-wide association study (TWAS). Characterisation of distal eQTL revealed unequal genetic regulation patterns between two subgenomes, highlighted by an eQTL hotspot (Hot216) that established a genome-wide genetic network regulating the expression of 962 genes. The primary regulatory role of Hot216, and specifically the gene encoding a KIP-related protein, was found to be the transcriptional regulation of genes responsible for cell wall synthesis, which contributes to fibre length by modulating the developmental transition from rapid cell elongation to secondary cell wall synthesis.
- This study uncovered the genetic regulation of fibre-cell development and revealed the molecular basis of the temporal modulation of secondary cell wall synthesis during plant cell elongation.

Introduction

Phenotypic diversity in a species is usually determined by a variety of genetic variations that can be investigated by genetic approaches such as genome-wide association studies (GWAS). The molecular roles of causal genetic variations in controlling phenotypes can be explored by investigating how they modulate the differential expression of genes (Cookson *et al.*, 2009). At a population level, gene expression variation can be measured quantitatively and mapped to a genomic locus (Cheung & Spielman, 2009). With the application of high-throughput technologies for gene expression profiling, expression quantitative trait locus (eQTL) mapping has emerged as a potentially powerful new approach to investigate the genetic architecture of expression variation in a variety of organisms, and to provide molecular links between genetic variation and phenotypic diversity (Lappalainen *et al.*, 2013; Zhu *et al.*, 2016; Aguet *et al.*, 2017; Hormozdiari *et al.*, 2018).

Over the past decade, eQTL mapping has been used to investigate the genetic architecture of regulatory variation in gene expression in model plants and major crops (West *et al.*, 2007; Wang *et al.*, 2010; Zhang *et al.*, 2011; L. Zhang *et al.*, 2017; X. Wang *et al.*, 2018). A comprehensive eQTL analysis allows the construction of regulatory networks (Keurentjes *et al.*, 2007; Fu *et al.*, 2013; Wang *et al.*, 2014). Furthermore, characterisation of eQTL provides an approach to address additional biological questions. In Arabidopsis, eQTL mapping has been conducted to explore the relationship between genes responding to the environment and genes regulated by genomic variation affecting their expression (Lowry *et al.*, 2013). In maize, an eQTL analysis has shown a critical role for noncoding genomic sequences in regulating expression variation, and a multitissue eQTL analysis revealed

*These authors contributed equally to this work.

the contribution of rare genetic variants to expression extremes (Liu *et al.*, 2017; Kremling *et al.*, 2018). Recently, an integrated analysis of eQTL and metabolic QTL revealed the metabolic breeding history of tomato (Zhu *et al.*, 2018).

Cotton (*Gossypium* spp.) is globally cultivated for the utilisation of natural renewable fibres in textiles, which dictates that the major goal of cotton breeding is to cultivate varieties producing fibres of superior quality. Genome-enabled breeding serves as an efficient approach to fulfil this goal, but requires an understanding of the genetic basis underpinning fibre quality-related traits and how genetic variations influence fibre development. The cotton fibre undergoes a staged cellular development to form a mature lint fibre after the ovule epidermal cell initiates an elongation process. This has been used as an excellent experimental system for studying polarised cell growth and cell wall biosynthesis in plants (Haigler *et al.*, 2012). Numerous genetic mapping efforts have identified many loci that contribute to fibre quality-related traits (Fang *et al.*, 2017; Huang *et al.*, 2017; Wang *et al.*, 2017; Liu *et al.*, 2018; Ma *et al.*, 2018; Thyssen *et al.*, 2019), and functional analysis has confirmed important roles for candidate genes (Tan *et al.*, 2013; Shan *et al.*, 2014). It has also been found that most characterised genes play a role in more than one biological process, such as redox homeostasis, sugar transport and cell wall metabolism (Han *et al.*, 2013; Guo *et al.*, 2016; Li *et al.*, 2016; Z. Zhang *et al.*, 2017). However, some intriguing questions remain poorly understood, such as how those genes work together to regulate the complex process of fibre development and how they determine the transition from rapid cell elongation to secondary cell wall biosynthesis.

Here, we present an analysis of fibre transcriptomes from a natural *G. hirsutum* population with 251 accessions. This allowed us to identify 15.330 eQTL associated with 9282 genes, and define the eQTL network for the elongating fibre. An eQTL hotspot has been characterised for regulating the expression of genes responsible for cell wall biosynthesis, which controls the trait of fibre length. This study opens the way for linking regulatory variants to gene transcription in cotton fibre development and provides a useful reference for accelerating the genetic improvement of fibre quality.

Materials and Methods

Plant materials

In a previous study, we constructed a genomic variation map by sequencing a natural population of upland cotton (*Gossypium hirsutum*) accessions (Wang *et al.*, 2017). To understand the regulatory mechanisms of fibre development, 251 upland cotton accessions were cultivated in the field at Shihezi (44°20′15″N, 86°3′28″E), Xinjiang, China in 2016. Shihezi has a suitable climate for cotton growth and is one of major locations for cotton cultivation in China. Cotton bolls were tagged on the day of flowering as 0 dpa (day post anthesis) samples. Three replicate populations of each accession were planted in three separate plots, with each replicate population arranged randomly in each plot. For each accession, fibres at 15 dpa from at least 10 cotton bolls of different plants were collected and bulked together, and were immediately frozen in liquid nitrogen. Fibre samples of individual replicate populations were used for RNA extraction.

RNA extraction and sequencing

For each accession, fibre samples were isolated for RNA extraction. Total RNA was extracted using a Spectrum Plant Total RNA Kit (Sigma, STRN250). These RNA samples were used to construct sequencing libraries using the Illumina TruSeq RNA Library Preparation Kit (Illumina, San Diego, CA, USA). Each library was sequenced on an Illumina HiSeq 4000 platform (paired-end 150 bp).

RNA-Seq data mapping and analysis

The raw RNA sequencing data were subject to adapter filtering and low-quality bases were removed using Trimmomatic (v.0.32) (Bolger *et al.*, 2014). The clean data were mapped to the reference genome sequence of *G. hirsutum* using HiSAT2 software (v.2.1.0) with default settings (Kim *et al.*, 2015; Wang *et al.*, 2019). The mapping reads were sorted to filter those reads representing PCR duplicates. Sequencing reads with mapping quality of less than 25 were filtered using Samtools (v.0.1.19) (Li *et al.*, 2009). Sequencing reads that were mapped to multiple genomic loci were filtered from the mapping files using customised Perl scripts. The remaining reads were used to calculate the expression levels of genes (fragments per kilobase of transcript per million mapped reads, FPKM) using Stringtie software (v.1.3.4) with default settings (Pertea *et al.*, 2015).

Identification of genomic SNPs

In this study, whole genome re-sequencing data of 251 accessions which were used for RNA-Seq analysis were mapped to the new reference genome sequence of *G. hirsutum* using BWA software (v.0.7.10-r789) using the mem method (-M -k 25) (Li & Durbin, 2009; Wang *et al.*, 2019). The unique mapping reads were parsed to identify single nucleotide polymorphisms (SNP) with Samtools and Gatk (v.3.1.1) as described previously (Li *et al.*, 2009; McKenna *et al.*, 2010; Wang *et al.*, 2017). Only those SNPs that were supported by both software tools were retained. The remaining SNPs were subject to a filtering process based on the mapping depth (>6×) and quality score (>100). VCFtools (v.0.1.15) was used to merge SNPs from each accession, of which those with a minor allele frequency (MAF) of less than 5% were filtered (Danecek *et al.*, 2011). Annovar software was used to annotate the remaining SNPs, including those representing genic variations (synonymous, nonsynonymous, intronic, un-translated regions, upstream 1 kb sequence of the transcription start site and downstream 1 kb sequence of transcription termination site) and intergenic variations (Hakonarson *et al.*, 2010).

GWAS on fibre quality-related traits

In a previous study, data on fibre quality-related traits for a natural population, including fibre length, fibre strength, elongation, uniformity and micronaire value, were collected over 3 yr (Wang *et al.*, 2017). In the current study, we collected data for these traits in 2016 and 2017, and conduct a GWAS analysis. Compared with our previous study (Wang *et al.*, 2017), the GWAS analysis here leveraged new phenotypic data measured in the same year when fibre samples were collected for RNA-Seq. GWAS was performed using genomic variation data (2372 767 SNPs) from those 251 accessions for eQTL analysis. Association analysis was carried out using TASSEL (v.5.0) and FASTLMM (v.2.02) programs (Kroon *et al.*, 2007; Lippert *et al.*, 2011). The population structure was calculated using the structure program and the kinship was derived from all SNPs (Falush *et al.*, 2003). The cutoff for determining significant associations was $P < 4.2 \times 10^{-7}$ ($1/n$, where n represents the total number of genomic SNPs). To help identify putative candidate genes for GWAS loci, RNA-Seq data for different fibre developmental stages were used to examine gene expression levels (Zhang *et al.*, 2015).

Identification of expression QTL (eQTL)

To identify eQTL for genes in fibre development, an analysis of gene expression levels was performed to discard those with expression levels less than 0.1 (FPKM < 0.1) in more than 95% of samples. This allowed us to filter 30 336 of 70 199 genes in the reference genome. The remaining genes were subject to a filtering process which requires an expression change of at least two fold between two samples representing the 5th percentile and the 95th percentile of sorted expression levels, respectively. In this step, 4098 genes were filtered. The expression levels of remaining genes were normalised using a normal quantile transformation. In total, we obtained 35 765 genes with transformed expression levels which were regarded as expression traits for GWAS analysis. The method of eQTL identification was similar to a previous study (Fu *et al.*, 2013), with minor modifications. In the current study, the merged genomic SNPs (MAF > 0.05) were used to perform GWAS for each gene using the FASTLMM program (Lippert *et al.*, 2011). The cutoff for determining significant associations was the same as that for GWAS. The significant SNPs for each trait (gene) were grouped into clusters with a maximum distance of 10 kb between two consecutive SNPs, and only those clusters with more than three significant SNPs were retained as putative eQTL. Each eQTL was represented by the most significant SNP of all (lead SNP). The putative eQTL which were identified in a LD region were further filtered as false-positive associations. In addition, to identify eQTL hotspots, the HOT_SCAN program was run for all eQTL in each chromosome (-m 5000 -s 0.05) (Holanda *et al.*, 2014).

TWAS

The idea of transcriptome-wide association studies (TWAS) was described in a previous study (Gusev *et al.*, 2016), which provided an approach for identifying significant *cis* genetic correlation between expression and phenotypic traits. This method used a reference population with available gene expression and genetic variation data to impute the *cis* genetic component of expression into another set of phenotyped individuals for which genetic variation has been identified. The fusion program was used to conduct TWAS in this study (http://gusevlab.org/projects/fusion/). As no measured population expression data in cotton have been generated previously, the normalised expression data of 3690 genes with local eQTL and SNP data from upstream 500 kb to downstream 500 kb of these genes were extracted to compute expression weights, which represent the premodelled relationship between SNPs and expression levels of genes (Gusev *et al.*, 2016). In this analysis, heritability was calculated using GCTA (v.1.26.0) and expression weights were computed with the option – models top1, blup, lasso, enet (Yang *et al.*, 2011). The genome-wide GWAS SNP data were used as the reference data. The GWAS statistics Z-score was computed based on GWAS *P*-values and effect size for each respective fibre quality-related trait. The TWAS was performed on each chromosome with significant GWAS associations for traits. The TWAS *P*-values were corrected with the Bonferroni method in R (v.3.2.3).

Network construction

In the eQTL hotspot data, we found that some genes were regulated by several eQTL. To systematically understand the regulation associations, we leveraged a genetic network involving genes (traits) and eQTL data to visualise the relationship between eGenes and eQTL. eQTL hotspots were highlighted in the genetic network. This network included the relationship between eQTL and their regulated genes, and between eQTL hotspots and regulated genes. In this analysis, those genes which included the lead SNPs of eQTL were regarded as causal regulators. The construction of the network was performed using CYTOSCAPE (v.3.6.1) with the edge-weighted force directed layout method (Shannon *et al.*, 2003).

Differential expression and gene set enrichment analyses

The differential expression analysis of genes in cotton accessions with extreme long fibre and extreme short fibre was performed using the DESEQ package in R with a significance of false-discovery rate (FDR ≤ 0.05) (Anders & Huber, 2010). The gene ontology (GO) enrichment analysis of genes in the eQTL network or exhibiting differential expression was performed using the AgriGO webserver (v.2.0) with genes showing expression levels in fibre development as a reference (Tian *et al.*, 2017). GO terms with a FDR threshold of 0.05 were considered as significant terms.

Data availability

The raw RNA-Seq data generated in this study are available in the BioProject database under accession no. PRJNA433615.

Results

QTL identification for fibre quality-related traits using GWAS

In a previous study, we generated genomic re-sequencing data for a natural population, which provided a resource to identify favourable variants associated with fibre quality-related traits (Wang *et al.*, 2017). In the present study, another 15 accessions were sequenced and integrated with previous data to generate an individual set of 251 accessions. We aligned sequence data of the 251 accessions against our recently published reference genome sequence of *G. hirsutum* acc. Texas Marker-1 (TM-1) (Wang *et al.*, 2019), and identified 2372 767 SNPs with a MAF of > 0.05. We then used the multiple phenotypic data for fibre quality, including those measured under the same environmental conditions used for sample collection for transcriptome sequencing in 2016 (Supporting Information Table S1), to perform a GWAS analysis of fibre quality-related traits, including fibre length (FL), strength (FS), elongation (FE), uniformity (FU) and micronaire value (MV). In total, 28 associations were identified for these traits, including 10, six, eight and four loci for length, strength, elongation and uniformity, respectively (Fig. 1; Tables S2, S3). No significant candidate loci were identified for MV. Among these associations, 16 were previously uncharacterised (Table 1).

Transcriptome sequencing and eQTL mapping

GWAS analysis identified candidate loci associated with fibre quality, but how these loci contribute to different fibre properties between accessions remains largely unknown. We sought to find an answer to this question by exploring expression variation of genes that may be influenced by those loci across a population. We sequenced fibre transcriptomes for 251 *G. hirsutum* accessions, with fibre samples at 15 d post anthesis (dpa), which represents a late elongation period before the developmental transition to secondary cell wall synthesis. The accessions for RNA-Seq analysis are the same as those for genome re-sequencing. In total, we generated 10 billion paired-end reads with an average of 40 million for each accession (Table S4). These RNA-Seq data were mapped to the TM-1 reference genome to quantify the expression levels of genes. In this analysis, we found 39 863 genes with expression levels at this stage of fibre development. In total, 35 765 genes that exhibited expression variation, representing 50.9% of the annotated genes (70 199) in TM-1 genome, were used for subsequent analysis.

We performed eQTL mapping using the same genomic SNP dataset as that for GWAS analysis of fibre quality-related traits. In total, we identified 15 330 eQTL associated with the expression of 9282 genes (eGenes regulated by eQTL). These eQTL and eGenes were from all 26 chromosomes (Fig. 2a). We found that the associations of eQTL and eGenes located on the same chromosome had a higher significance than those located on different chromosomes (two-sided Wilcoxon rank sum test, *P*-value $< 2.2 \times 10^{-16}$; Fig. S1). For the interchromosomal associations, those that occurred in homoeologous chromosomes between A- and D-subgenomes (691 of 15 330) were prevalent (Fisher's exact test, *P*-value $< 2.2 \times 10^{-16}$; Fig. 2a shown by two red arrows). One intriguing observation is that the associations between chromosome D11 and other chromosomes (962 of 15 330) were enriched (Fisher's exact test, *P*-value $< 2.2 \times 10^{-16}$; Fig. 2a shown by an orange arrow).

Based on the distance between eQTL and eGenes, we categorised all eQTL into 5370 local eQTL (<1 Mb) and 9960 distal eQTL (>1 Mb or in different chromosomes). We found that local eQTL had a larger effect on expression variation than did distal eQTL (two-sided Wilcoxon rank sum test, *P*-value $< 2.2 \times 10^{-16}$; Fig. 2b), which agrees with similar findings in other organisms (Wang *et al.*, 2010; Zhang *et al.*, 2011; L. Zhang *et al.*, 2017; X. Wang *et al.*, 2018). It was also found that the distance between local eQTL and eGenes shows a preferential distribution at *c.* 5 kb (Fig. 2c). Of the total eQTL, local eQTL account for 35%, while distal eQTL account for 65%, of which 15% occurred on the same chromosome (distal_intraChr) and the other 50% were found on different chromosomes (distal_interChr). The distal_interChr eQTL were further divided into three groups, that is within the A-subgenome (At–At; 9%), within the D-subgenome (Dt–Dt; 15%), and between the A- and D-subgenomes (At–Dt; 26%) (Fig. 2d). Of the eGenes, we found that 5027 were regulated by local eQTL and 6220 by distal eQTL, and the majority of eGenes (6049, 65.1%) were regulated by only one eQTL (Fig. 2e).

Global roles of eQTL in the genetics of fibre quality-related traits

Recent studies have demonstrated that TWAS represents a powerful approach for prioritising causal genes for GWAS loci, using the information of *cis*-eQTL (Gusev *et al.*, 2016). TWAS establishes a connection between gene expression, one kind of molecular phenotype, and other physical phenotypes in organisms (Gusev *et al.*,). Here, on the basis of characterising local eQTL, we attempted to prioritise likely causal genes for fibre quality-related GWAS loci using TWAS.

We modelled 3690 expression weights using SNPs from genomic region 500 kb both upstream and downstream of each eGene that was transcriptionally regulated by local eQTL. These expression weights were leveraged to perform a TWAS on each chromosome with GWAS loci. The TWAS identified 10 transcriptome-wide-significant associations, including four for FL, three for fibre elongation and three for fibre uniformity (Bonferroni corrected *P*-value < 0.05; Table 2; Fig. S2). In these associations, we found that a MYB transcription factor gene (*Ghir_D05G027990*) was prioritised to be a candidate gene for a pleiotropic locus associated with FL and fibre uniformity. We also noted that the TWAS prioritised two genes (*Ghir_D04G017090* and *Ghir_D04G016300*) for a fibre elongation-related GWAS association on chromosome D04, and two genes (*Ghir_D12G015450* and *Ghir_D12G015670*) for a fibre uniformity-related GWAS association on chromosome D12. Even though TWAS failed to identify gene-trait associations for

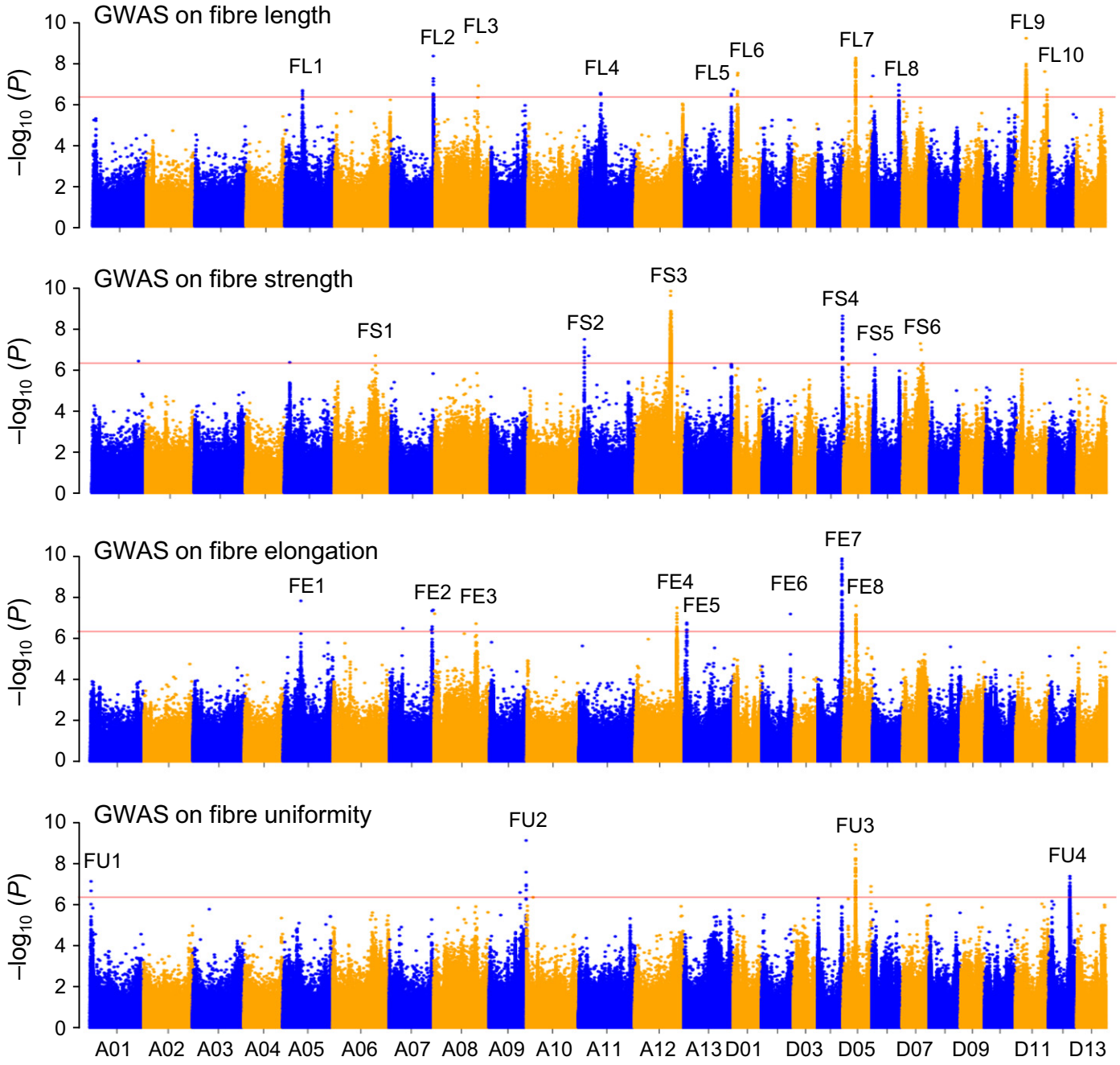

Fig. 1 Genome-wide association study (GWAS) on cotton fibre quality-related traits. These traits include fibre length (FL), fibre strength (FS), fibre elongation (FE) and fibre uniformity (FU). This analysis was performed using 2372 767 single nucleotide polymorphisms (SNPs) in 251 cotton accessions. The horizontal red lines show the significance threshold of GWAS ($1/n$; 6.4). The *x*-axes show the 26 chromosomes (A01–A13 and D01–D13) in *Gossypium hirsutum*. Each chromosome is scaled by the physical chromosome length. The significant GWAS associations are also shown in Table 1.

all GWAS loci, this approach represents an effective way to prioritise likely causal genes in the genetics of fibre quality-related traits.

Unequal subgenome transcription regulation as revealed by eQTL analysis

The hybridisation event between two diploid species that occurred 1–2 Ma resulted in the appearance of allotetraploid *G. hirsutum*, which significantly increased regulatory complexity of gene transcription with an additional feature of intersubgenomic transcription regulation relative to diploid ancestors (Adams *et al.*, 2003). However, few studies have been carried out to estimate the effect of intersubgenomic regulation on gene transcription in cotton (M. Wang *et al.*, 2018). The mapping of distal eQTL provides an avenue to investigate intersubgenomic genetic regulation for gene transcription in fibres. In this study, we found that 57.1% of eQTL in the A-subgenome, including

Table 1 Summary of genome-wide association study (GWAS) results for fibre quality-related traits in *Gossypium hirsutum*.

QTL	Chr.	SNP no.[1]	Lead SNP position	Major allele	Minor allele	Minor allele (%)	*P*-value	References[3]
FL1	A05	4	38429540	T[2]	A	12.7	2.0×10^{-7}	
FL2	A07	8	93511292	T[2]	C	10.8	4.1×10^{-9}	Ma *et al.* (2018)
FL3	A08	2	92006893	T[2]	A	5.2	8.8×10^{-10}	Liu *et al.* (2018)
FL4	A11	2	45735171	A[2]	C	6.8	2.7×10^{-7}	
FL5	A13	2	103375434	T[2]	C	15.9	2.9×10^{-7}	Fang *et al.* (2017); Liu *et al.* (2018)
FL6	D01	5	9210491	G[2]	A	8.0	2.8×10^{-8}	
FL7	D05	135	27233958	T[2]	C	5.6	5.2×10^{-9}	Ma *et al.* (2018)
FL8	D06	6	59462878	C	T[2]	21.5	1.1×10^{-7}	
FL9	D11	90	24616418	G[2]	A	17.1	6.6×10^{-10}	Ma *et al.* (2018); Thyssen *et al.* (2019)
FL10	D11	3	65633365	T[2]	C	5.2	2.4×10^{-8}	
FS1	A06	1	90655795	C[2]	T	6.8	1.8×10^{-7}	
FS2	A11	7	10086019	C	T[2]	5.2	2.8×10^{-8}	Liu *et al.* (2018)
FS3	A12	567	77056039	G	A[2]	9.6	8.4×10^{-11}	
FS4	D04	23	53480934	C	T[2]	6.8	2.0×10^{-9}	
FS5	D06	1	5570632	T	C[2]	17.1	1.6×10^{-7}	Ma *et al.* (2018)
FS6	D07	2	39082751	T	G[2]	7.6	4.6×10^{-8}	Huang *et al.* (2017)
FE1	A05	1	38430737	T[2]	C	6.8	1.3×10^{-8}	
FE2	A07	4	93524441	C[2]	T	8.4	4.1×10^{-8}	Ma *et al.* (2018)
FE3	A08	1	92006893	T[2]	A	5.2	6.3×10^{-11}	
FE4	A12	16	91652556	A	G[2]	29.1	2.9×10^{-8}	
FE5	A13	6	5588715	C	T[2]	24.7	1.6×10^{-7}	
FE6	D02	1	61894966	G	A[2]	8.0	5.9×10^{-8}	
FE7	D04	64	52570413	C	T[2]	48.2	8.9×10^{-11}	Huang *et al.* (2017); Wang *et al.* (2017); Thyssen *et al.* (2019)
FE8	D05	71	27233958	T[2]	C	5.6	2.3×10^{-8}	Thyssen *et al.* (2019)
FU1	A01	2	2247417	G[2]	A	13.1	7.1×10^{-8}	
FU2	A09	5	81113800	G[2]	T	8.0	7.2×10^{-10}	
FU3	D05	28	27260152	G[2]	A	5.6	1.2×10^{-9}	Ma *et al.* (2018)
FU4	D12	21	46644449	G[2]	A	33.1	3.9×10^{-8}	

[1]Number of significant single nucleotide polymorphisms (SNP).
[2]Favourable SNP alleles.
[3]Overlapping GWAS loci with previous studies.

52.6% for intersubgenomic regulation and 4.5% for both inter- and intrasubgenome regulations, are associated with the transcription of genes in the D-subgenome. An analysis of eQTL in the D-subgenome showed 58.2% (46.5% and 11.7%) of eQTL are responsible for transcriptional regulation of genes in the A-subgenome (Fig. 3a).

In terms of the number of genes, we found that 44.3% of the eGenes identified in the A-subgenome are regulated by eQTL in the D-subgenome, including 29.9% regulated by intergenomic eQTL and 14.4% regulated by both inter- and intrasubgenomic eQTL. However, an analysis of eGenes in the D-subgenome shows that only 23.4% of eGenes (9.9% and 13.5%) have eQTL regulation in the A-subgenome (Fig. 3b). This difference indicates that a larger number of genes in the A-subgenome are transcriptionally regulated by the D-subgenome at this fibre developmental stage (Pearson's chi-squared test, *P*-value $< 2.2 \times 10^{-16}$), highlighting unequal transcriptional regulation patterns between the two subgenomes. A GO enrichment analysis shows that eGenes in the A-subgenome with intersubgenomic eQTL regulation are involved in biological processes linked to fibre development, such as REDOX homeostasis, plant cell wall organisation or biogenesis, and cell tip growth (Fig. 3c; Table S5). In comparison, eGenes in the D-subgenome that are regulated by the A-subgenome are enriched in fundamental processes such as regulation of DNA metabolism and neutral amino acid transport.

To specifically explore the intersubgenomic regulation of fibre quality, we investigated the colocalisation of distal eQTL and GWAS loci for fibre quality-related traits. We found that eight GWAS associations had significant SNPs that overlapped with intersubgenomic eQTL, including two for FL on the D05 and D11 chromosomes, two for fibre strength on the A12 chromosome, three for fibre elongation on the A13, D04 and D05 chromosomes, and one for fibre uniformity on the D12 chromosome (Table S6). These colocalised eQTL regulated 1042 eGenes, of which 488 of 507 (96.2%) in the A-subgenome were regulated by eQTL in the D-subgenome and 30 of 535 (5.6%) in the D-subgenome were regulated by eQTL in the A-subgenome, which suggests a primary role for the D-subgenome in regulating genetic loci associated with fibre quality in terms of intersubgenomic regulation. Of note is the observation that a fibre length-associated GWAS locus on chromosome D11 was colocalised with eQTL regulating 479 eGenes in the A-subgenome. These data suggest that further characterisation of the functional roles

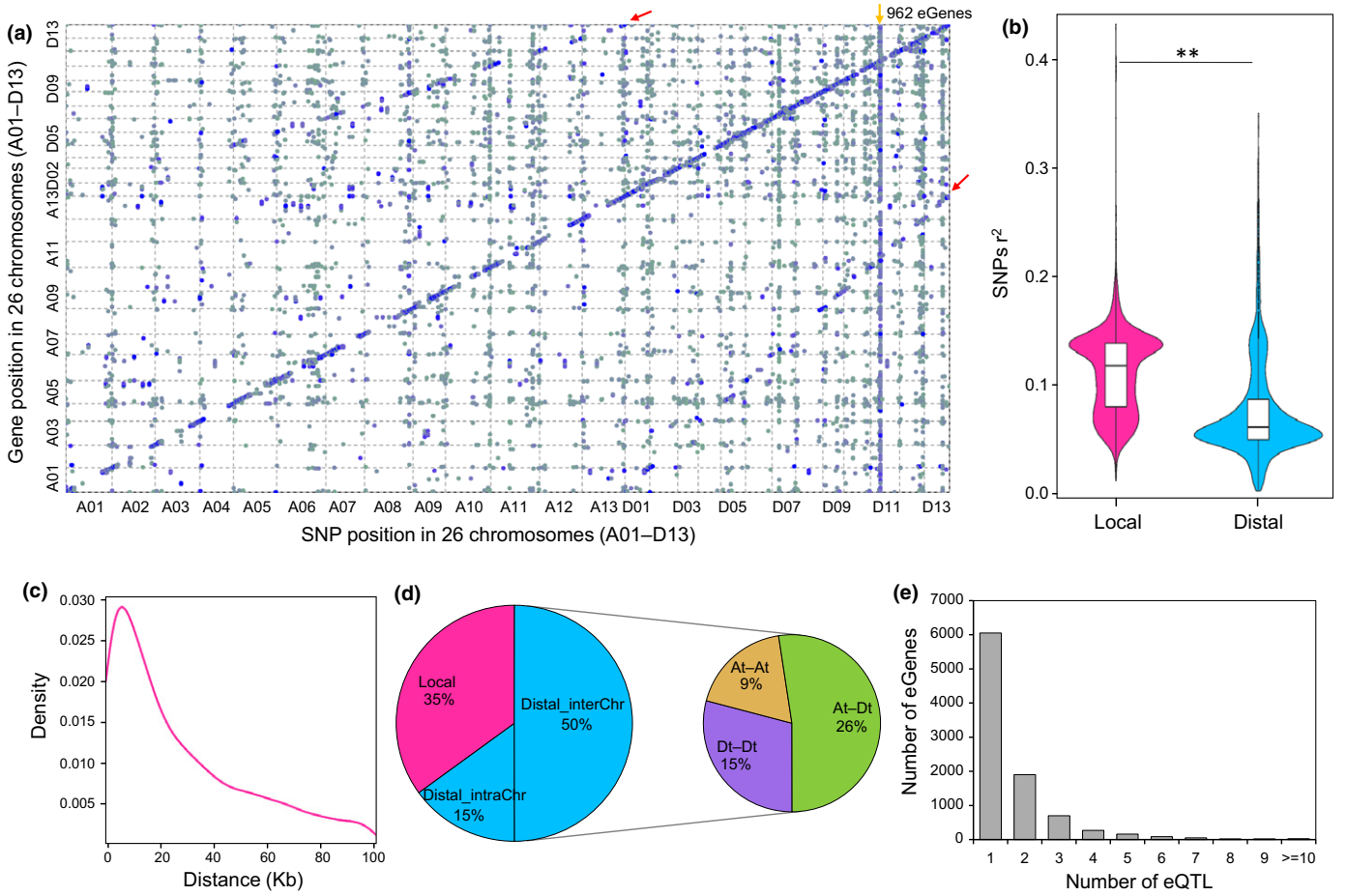

Fig. 2 Identification of eQTL using RNA-Seq data in cotton fibre development. (a) Dot plot showing eQTL and their regulated genes in 26 chromosomes. *x*-axis shows the single nucleotide polymorphism (SNP) position (bp) in each chromosome and *y*-axis shows gene position (bp) in each chromosome, with a chromosome order of A01 to D13 from left to right (*x*-axis) or from lower to upper (*y*-axis). The colour of each dot represents the significance (*P*-value) of each eQTL-gene association, with low significance in green and high significance in blue. Each chromosome is scaled by the physical chromosome length. Dots in the diagonal line show the intrachromosomal associations. The three red arrows show the enrichment of intersubgenomic associations in homoeologous chromosomes and interchromosomal associations between the D11 and other chromosomes. (b) Difference of explanation rate (r^2) of SNPs for expression variation between local eQTL (< 1 Mb) and distal eQTL (> 1 Mb). The violin plots show the distribution density and box plots show the distribution quantiles. Two-sided Wilcoxon rank sum test, $^{**}P$-value $< 2.2 \times 10^{-16}$. (c) Distribution of distance (< 100 kb) between eQTL and regulated genes. (d) Proportions of local eQTL and distal eQTL. Distal eQTL were divided into three groups, including associations between the A-subgenome (At) and D-subgenome (Dt; At–Dt), between the At and At (At–At) and between the Dt and Dt (Dt–Dt). (e) Distribution of the number of eQTL for genes which were regulated by eQTL.

Table 2 Identification of significant gene-trait associations in *Gossypium hirsutum* using transcriptome-wide association study (TWAS).

QTL	GWAS lead SNP position	GWAS *P*-value	Significant TWAS gene	TWAS FDR	Homologue	Description[1]
FL6	D01:9210491	2.8×10^{-8}	Ghir_D01G006540	9.2×10^{-3}	AT4G27270	Quinone reductase family protein
FL7	D05:27233958	5.2×10^{-9}	Ghir_D05G027990	2.3×10^{-4}	AT1G22640	MYB domain protein 3
FL9	D11:24616418	6.6×10^{-10}	Ghir_D11G020340	2.9×10^{-2}	AT3G19150	KIP-related protein 6
			Ghir_D11G020430	2.4×10^{-4}	AT1G09760	U2 small nuclear ribonucleoprotein A
FE5	A13:5588715	1.6×10^{-7}	Ghir_A13G004410	5.5×10^{-4}	AT2G14820	Phototropic-responsive NPH3 family protein
FE7	D04:52570413	8.9×10^{-11}	Ghir_D04G017090	1.3×10^{-6}	AT4G36720	HVA22-like protein K
			Ghir_D04G016300	5.9×10^{-3}	AT1G21070	Nucleotide-sugar transporter family protein
FU3	D05:27260152	1.2×10^{-9}	Ghir_D05G027990	7.4×10^{-4}	AT1G22640	MYB domain protein 3
FU4	D12:46644449	3.9×10^{-8}	Ghir_D12G015450	6.0×10^{-4}	AT5G50850	Transketolase family protein
			Ghir_D12G015670	6.3×10^{-3}	AT3G48000	Aldehyde dehydrogenase 2B4

[1]Characterised function of homologous genes in Arabidopsis.

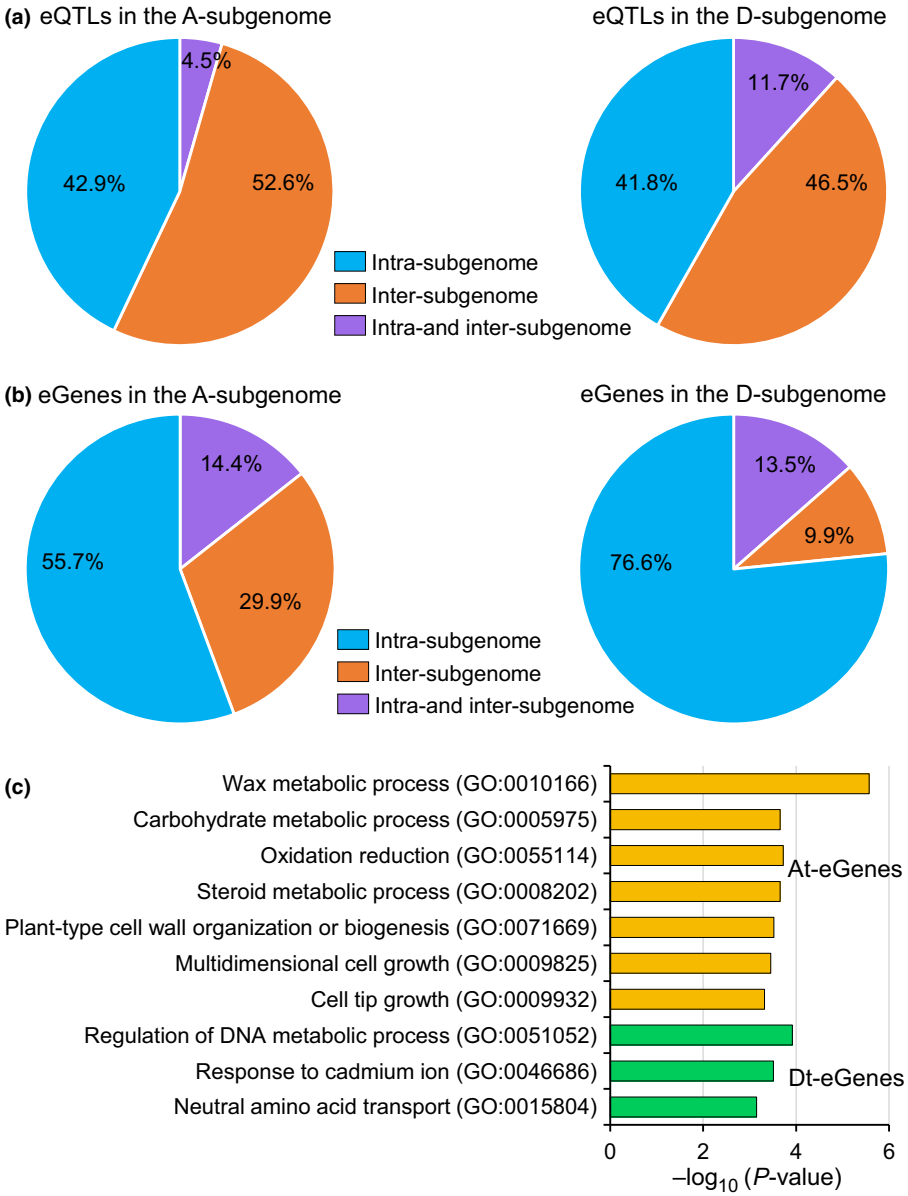

Fig. 3 Analysis of eQTL and their regulated genes at the subgenome level in *Gossypium hirsutum*. (a) Summary of eQTL in the A- and D-subgenomes involved in intrasubgenomic or intersubgenomic regulation. (b) Summary of eGenes in the A- and D-subgenomes that were regulated by intrasubgenomic or intersubgenomic eQTL. For (a) and (b), the percentages show the proportions of intra-/intersubgenomic eQTL in all distal eQTL or eGenes regulated by intra-/intersubgenomic eQTL. (c) GO enrichment of genes in the A-subgenome (At-eGenes), which were regulated by eQTL in the D-subgenome (yellow bar charts) and genes in the D-subgenome (Dt-eGenes) which were regulated by eQTL in the A-subgenome (green bar charts). For each subgenome, all expressed genes at this fibre developmental stage were used as a reference for GO enrichment analysis.

of these GWAS loci should be integrated with analysis of their intersubgenomic regulatory roles in gene transcription.

Characterisation of eQTL hotspots and a regulatory network for fibre length

As shown in previous studies, some eQTL are located in a genomic region which affects the expression of many genes, and such a region is known as an eQTL hotspot (Liu *et al.*, 2017; X. Wang *et al.*, 2018). We investigated whether there was a similar phenomenon in cotton fibre development, and 243 hotspots were identified (Table S7), regulating the expression of 3820 genes (41.1% of all identified eGenes). 125 of these eQTL hotspots were identified in the A-subgenome and 118 in the D-subgenome. For each hotspot, the number of eGenes varies from 3 to 962 (Table S7). We predicted possible key regulators for these hotspots affecting the expression of downstream eGenes by integrating genetic position information for both eQTL and gene annotations.

The identification of eQTL hotspots allows the characterisation of the complex regulatory relationship between eQTL and eGenes. Here, we constructed an eQTL network consisting of 44

eQTL hotspots, 220 eQTL and 1896 eGenes (Fig. 4a). In this network, the eQTL hotspot 216 (Hot216) on chromosome D11 (24.43–24.62 Mb) was highlighted, and was found to regulate the expression of 962 genes, including 479 in the A-subgenome and 483 in the D-subgenome (Table S8). This explains why the genome-wide analysis showed that expression variation of a large number of genes is associated with a genomic region in chromosome D11 (Fig. 2a). A detailed analysis of those genes showed that the expression of 293 genes is also regulated by other eQTL or hotspots, and the remaining 669 genes are only regulated by Hot216 (Fig. 4b).

To further characterise the biological role of Hot216, we performed a GO enrichment analysis of the 962 genes. It was observed that these genes showed an enrichment in microtubule cytoskeleton for the category of cellular component (CC); protein kinase and cellulase activity for the category of molecular function (MF); and cell wall organisation or biogenesis for the category of biological process (BP; Fig. 4c; Table S9). Given these relevant predicted functions, the large number of eGenes regulated by the Hot216 indicates that it represents a powerful eQTL hotspot for the transcriptional regulation of fibre development.

Effect of genomic and transcriptional variability on fibre length

To characterise the details of the 962 genes regulated by the Hot216, we performed an expression analysis across the 251 accessions. It was found that the expression patterns of these genes cluster into two groups (group-I and group-II; Fig. 5a). Group-I includes 287 genes which were highly expressed in 91 accessions and group-II includes the other 675 genes which were highly expressed in the other 160 accessions. GO analysis showed that genes in group-I are enriched in BP of plant-type cell wall organisation or biogenesis (GO:0071669) and cell wall biogenesis (GO:0042546), while genes in group-II are enriched in processes including activation of mitogen-activated protein kinase kinase (MAPKK) activity (GO:0000186), developmental growth (GO:0048589) and cell tip growth (GO:0009932). We hypothesised that the significant GO terms for group-II represent processes which may promote fibre-cell elongation and development.

We then compared the fibre quality-related traits of accessions between cluster-I and cluster-II. It was found that FL in cluster-I is significantly shorter than that in cluster-II (two-sided Wilcoxon rank sum test, P-value $< 7.4 \times 10^{-5}$), and the other three traits show no significant difference between the two clusters (Fig. S3). This suggests that the analysis of gene expression patterns categorised the accessions into two clusters with significant differences of FL. Accessions in cluster-I exhibit a short fibre phenotype with high expression of genes responsible for cell wall synthesis, while accessions in the cluster-II produce longer fibre accompanied by high expression of genes associated with cell growth. This result is supported by previous functional characterisation of genes amongst these 962 eGenes (Fig. 5b). For example this includes three genes encoding two fasciclin-like arabinogalactan proteins (FLA7: Ghir_A08G005490 and FLA11: Ghir_D11G035910) and a trichome birefringence-like protein (TBL3: Ghir_D13G010200; Ghir_A04G010010), with established roles in secondary wall cellulose synthesis (Bischoff *et al.*, 2010; MacMillan *et al.*, 2010). Two genes encoding an irregular xylem-related protein (IRX9: Ghir_A09G016060) and a xyloglucan endotransglucosylase/hydrolase (XTH30: Ghir_A08G016210) have been shown to be involved in xylan and xyloglucan metabolic processes respectively (Bourquin *et al.*, 2002; Lee *et al.*, 2007). We also highlighted three MYB transcriptional factors (MYB46: Ghir_A13G022890, MYB61: Ghir_A07G014020 and MYB103: Ghir_A08G012250; Ghir_D08G012890) which are involved in the positive regulation of secondary cell wall biogenesis (Kim *et al.*, 2013; Taylor-Teeples *et al.*, 2015).

To understand whether gene expression patterns of the Hot216-guided network have a genetic association with FL, we overlapped this hotspot region with GWAS signals. It is found that the Hot216 has the same genomic location as a significant GWAS locus for FL on chromosome D11 (24.44–24.62 Mb; Figs 1, 5c). Integration of eQTL and GWAS data led us to identify a candidate gene encoding KIP-related protein 6 (KRP6) which contains a nonsynonymous mutation in the first exon (G to T transition). Interestingly, the expression of *KRP6* was also regulated by the same genetic region which might act in *cis*-eQTL regulation (Fig. 5d). Genetic modification of KIP-related protein in *Arabidopsis* has been found to affect the expression of genes involved in plant cell wall organisation and heterochromatin modification, and regulated cell elongation (Jégu *et al.*, 2013). Here, we found that two different genotypes of *KRP6* correspond to differential FL (Fig. 5e). Moreover, *KRP6* exhibits differential expression levels between accessions with two different genotypes (Fig. 5f). In addition, many eGenes in Hot216, known to be involved in secondary cell wall organisation or biogenesis, such as *FLA7/11* and *MYB46/103*, exhibit differential expression levels between accessions with different genotypes of *KRP6* (Fig. 5f). These results indicate that *KRP6* is a candidate gene for differential fibre-cell length via transcriptional regulation of a large number of genes in fibre development.

To further support the view that differential expression of genes responsible for cell wall synthesis may contribute to varied FL, we made a comparative transcriptome analysis of 60 accessions with extreme FL. 30 accessions with short fibres were found to exhibit upregulated expression of 1163 genes and down-regulated expression of 507 genes compared with 30 accessions with long fibres (Fig. S4a). A GO analysis showed that these upregulated genes were enriched in plant-type cell wall organisation or biogenesis, xylan biosynthetic process and microtubule depolymerisation, while the down-regulated genes were enriched in categories designated response to stimulus and inorganic anion transport (Fig. S4b). We found that genes encoding cellulose synthase A catalytic subunit (CesA) 4, 7 and 8, which are required for secondary cell wall biosynthesis (Taylor-Teeples *et al.*, 2015), are highly expressed in accessions with short fibres (Fig. 6). Even though these *CesAs* were not found to be directly regulated by the Hot216, they act downstream of secondary wall cellulose synthesis and may be transcriptionally regulated by other genes with direct regulation from the Hot216 such as *MYB46* (Kim *et al.*, 2013; Taylor-Teeples *et al.*, 2015).

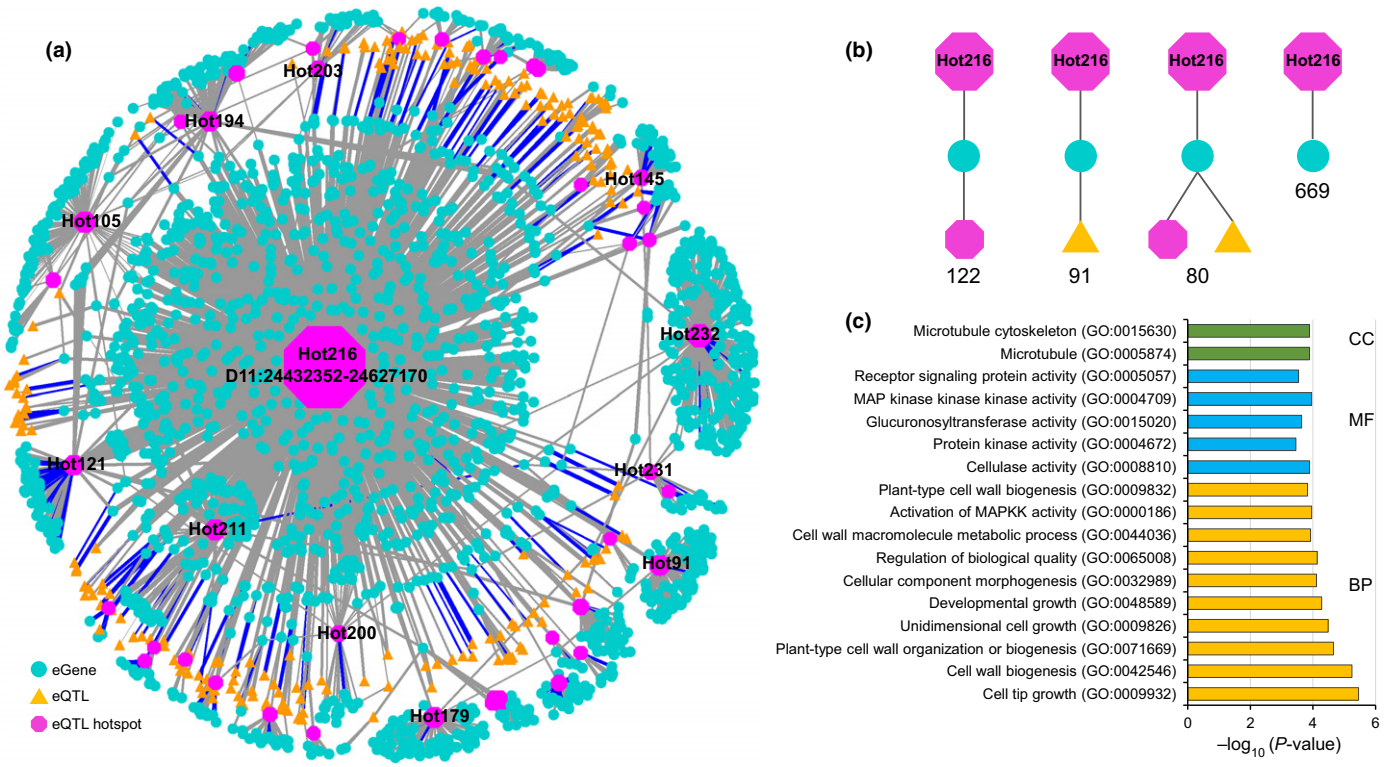

Fig. 4 Construction of eQTL regulatory network involving of eQTL hotspots. (a) Genetic network between eQTL and genes. The green circle nodes represent genes that were regulated by eQTL, the yellow triangle nodes represent eQTL, and the octagon nodes represent eQTL hotspots. The eQTL hotspot 216 (Hot216) on chromosome D11 is enlarged with a chromosome location from 24 432 352 bp to 24 627 170 bp. The blue network edges represent local eQTL associations and grey edges represent distal eQTL associations. For Hot216, only distal eQTL associations are shown. (b) Summary of the number of Hot216-involved associations in different categories. The three different symbols have the same meanings as those in panel (a). (c) Gene ontology (GO) enrichment of genes which are regulated by Hot216. These GO terms include those representing cellular component (CC), molecular function (MF) and biological processes (BP).

We conclude from these data that genetic variations at the GWAS locus on chromosome D11, specifically the candidate gene *KRP6*, can induce the expression of genes responsible for cell wall synthesis, and contributes to the early biosynthesis of secondary cell wall that leads to the formation of a shorter cotton fibre cell.

Discussion

eQTL analysis links regulatory variants to gene transcription

With the recent advances in sequencing the cotton genome, a very large number of genetic variants have been identified in different accessions, but only a few loci with variants have been found to be associated with agronomic traits by GWAS (Fang *et al.*, 2017; Huang *et al.*, 2017; Wang *et al.*, 2017; Liu *et al.*, 2018; Ma *et al.*, 2018; Thyssen *et al.*, 2019). It is believed that many other genetic variants may have a regulatory role in gene expression, but their regulatory targets remain undetermined. In a previous study, we integrated DNase I digestion followed by sequencing (DNase-Seq) and high-throughput chromosome conformation capture (Hi-C) data to annotate regulatory variants in *cis*-regulatory elements of promoters and distal enhancers of genes (Wang *et al.*, 2017). Here, we demonstrate that eQTL mapping represents another high-throughput approach to link regulatory variants to gene expression in fibre development. Our mapping of 15 330 eQTL represents the discovery of regulatory variants only at the transition from cell elongation to secondary cell wall development. Nevertheless, it would be expected that more regulatory variants would be identified if RNA-Seq data from different fibre developmental stages were generated. Nevertheless we demonstrate the practical application of TWAS for prioritising likely causal genes for GWAS loci. Compared with functional analysis of lead SNPs or homology-based prediction of causal genes for GWAS loci, TWAS establishes a direct connection between gene expression and phenotype using eQTL data. This suggests that an investigation of the intermediate omics–transcriptome between variome and phenome can facilitate the understanding of the regulatory roles of genetic variants in shaping phenotypic differences.

Intersubgenomic regulation increases regulatory complexity of gene transcription

It is known that many agronomic traits in allopolyploids, such as fibre quality in allotetraploid cotton, are regulated by the coordination of different subgenomes (Yuan *et al.*, 2015; Fang *et al.*, 2017; Wang *et al.*, 2017; Ma *et al.*, 2018; Thyssen *et al.*, 2019).

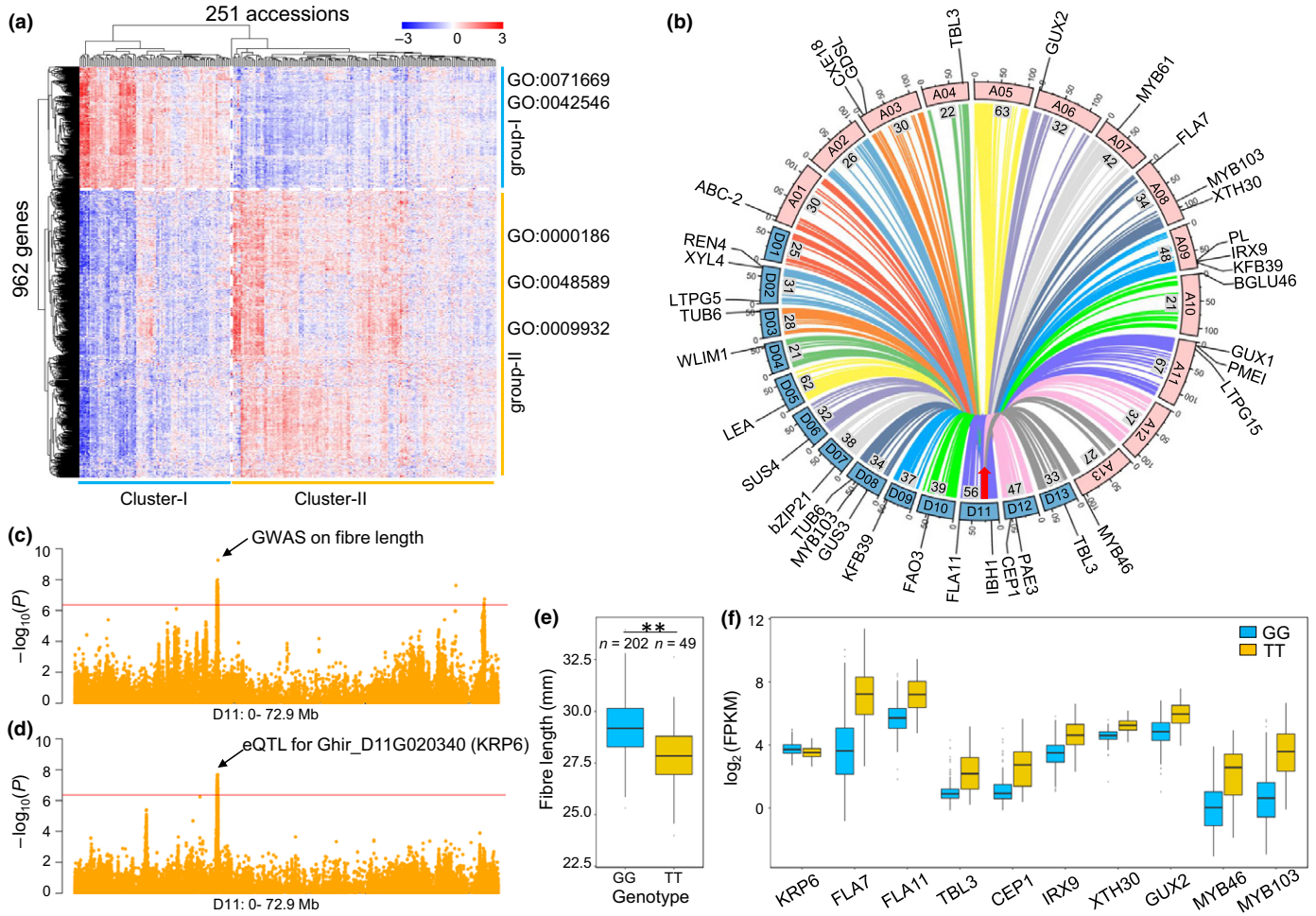

Fig. 5 Characterisation of eQTL hotspot 216 (Hot216) on chromosome D11. (a) Clustering analysis of the expression levels of eGenes which are regulated by Hot216. The 962 genes were categorised into two groups (group-I with 287 genes and group-II with 675 genes) and the 251 cotton accessions are clustered into two clusters (cluster-I with 91 accessions and cluster-II with 160 accessions). The significant GO terms in each gene group are shown. (b) Genome-wide distribution of eGenes regulated by Hot216. Some representative genes are shown with the full description of gene names in Supporting Information Table S8. (c) Manhattan plot of GWAS signal on chromosome D11 which is associated with fibre length. (d) Manhattan plot of the eQTL signal of *Ghir_D11G020340*. For (c) and (d), the horizontal red lines show the significance threshold ($1/n$; 6.4). (e) Distribution of fibre length with two different genotypes of *Ghir_D11G020340*. Two-sided Wilcoxon rank sum test, $^{**}P$-value $< 2.2 \times 10^{-16}$. The horizontal lines in box plots show median values, and ranges show the first and third quartiles. (f) Normalised expression (FPKM) of representative genes in two cotton groups categorised by the SNP site of *Ghir_D11G020340*.

Decoding the role of each subgenome in regulating desirable traits will enhance our understanding of the effects of polyploidisation on plant development. In cotton, previous studies only compared the expression patterns of homoeologous genes between A- and D-subgenomes in fibre development, which uncovered subgenome expression dominance or gene expression bias (Hovav *et al.*, 2008). In this study, we used eQTL data to identify widespread intersubgenomic regulation, and showed that the D-subgenome had a relatively large regulatory effect on the A-subgenome in terms of the number of eGenes, indicating the important role of the D-subgenome in fibre development. Even though we do not yet have similar evidence for the entire fibre development process, one of conclusions that can be made is that the D-subgenome has a large regulatory effect on the development of spinnable fibre in cultivated tetraploid cotton, because more selection signals during domestication were identified in the D-subgenome, which drove the phenotypic change from short and pigmented to long and white fibres (Wang *et al.*, 2017).

eQTL data represent genetic evidence for intersubgenomic regulation of gene expression. Previously, we used Hi-C data to identify a number of subgenomic chromatin interactions which were proposed to play a role in the coordination of expression of homoeologous genes in tetraploid cotton (M. Wang *et al.*, 2018). It remains to be determined whether the intersubgenomic eQTL regulation of gene transcription relies on spatio-chromatin contact between chromosomes. Interestingly, we observed a pattern of subgenome contacts using Hi-C data in leaf (Fig. S5), which was similar to the pattern of eQTL regulation (Fig. 2a). This open question should be investigated further using high-throughput chromatin contact matrix and eQTL data from the same tissue. From an evolutionary view, both subgenomes were divergent

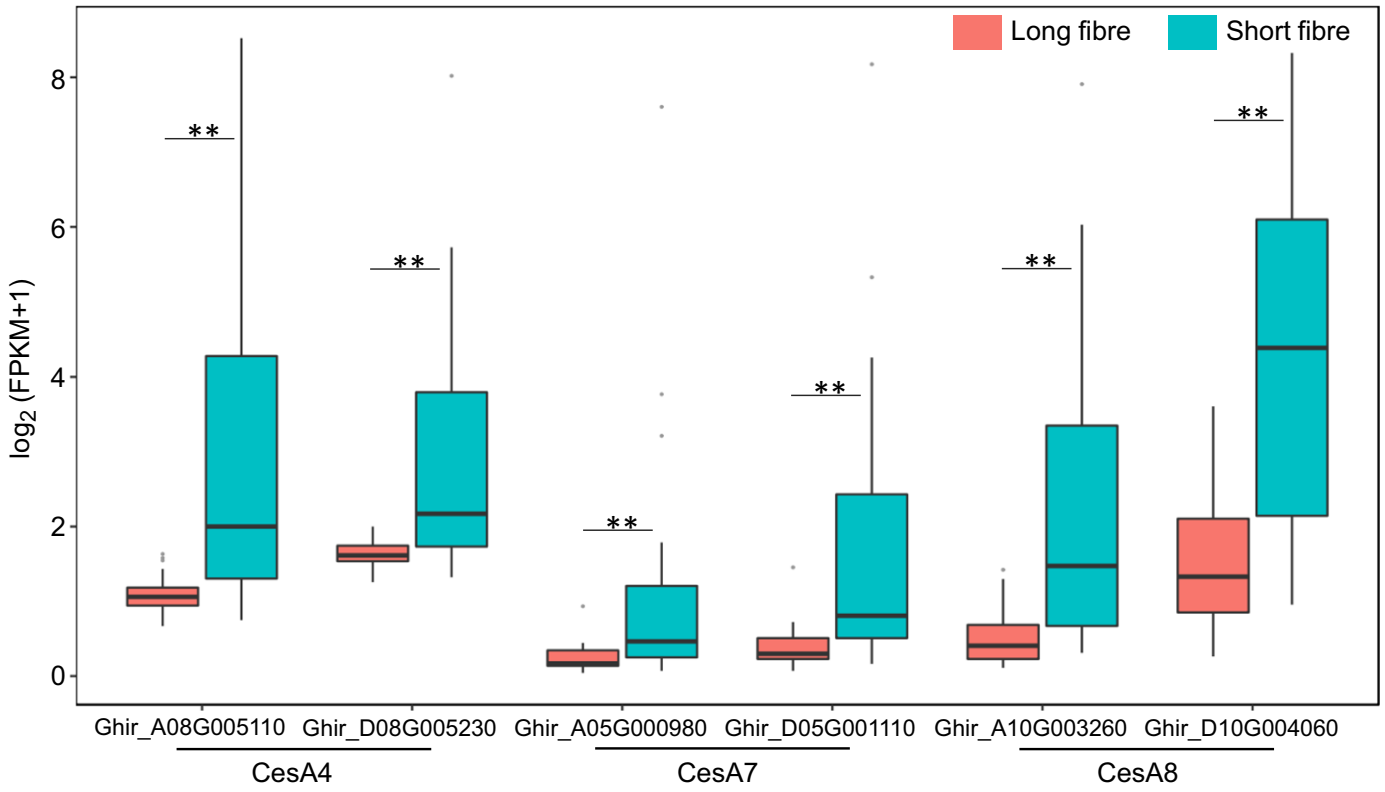

Fig. 6 Expression analysis of *CesA* genes in cotton accessions with extreme long or short fibres. 30 cotton accessions with extreme long fibres were compared with 30 accessions with extreme short fibres to identify differentially expressed genes. The long-fibre and short-fibre accessions are the same as those in Supporting Information Fig. S4. The horizontal lines in box plots show median values, and ranges show the first and third quartiles. All these *CesA* genes show differential expression levels between the two groups of accessions (**FDR ≤ 0.05).

from a common ancestor and had a high sequence similarity, so future studies should explore whether homoeologous sequences in one subgenome could regulate the expression of genes in the other subgenome. The analysis of intersubgenomic regulation may provide a strategy for exploring the regulatory mechanism of expression novelty in polyploids and inform our understanding of phenotypic advantage.

Hot216-guided genetic regulatory network orchestrates the initiation of secondary cell wall development in plants

After initiation, cotton fibre cells undergo a rapid elongation stage followed by a secondary cell wall synthesis stage in which cellulose is synthesised. The duration of elongation plays a vital role in determining mature FL, and the biosynthesis of the secondary cell wall restricts fibre elongation, as supported by metabolomic and transcriptomic data (Haigler *et al.*, 2012; Tuttle *et al.*, 2015). In this study, we integrate GWAS, TWAS, eQTL networks and transcriptome analysis to address the question of the genetic regulation of FL, and identify an important role for a genetic locus on chromosome D11. Even though this locus has been identified in previous studies (Ma *et al.*, 2018; Thyssen *et al.*, 2019), the regulatory mechanism underlying FL has remained largely unknown. We demonstrate that the likely causal gene *KRP6* acts as an eQTL hotspot (Hot216) to regulate the expression of 962 genes that are involved in a regulatory network. This large number of genes regulated by Hot216 provides a similar picture to results of experiments using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) of KRP5 in Arabidopsis, which showed an enrichment of genes involved in cell wall organisation (Jégu *et al.*, 2013). This suggests a conserved regulatory role of KRPs in promoting cell elongation between cotton and Arabidopsis.

In plants, KRPs have been found to act as inhibitors of cyclin-dependent kinases and have a role in endoreduplication (Barrôco *et al.*, 2006; Jégu *et al.*, 2013; García-Ramírez *et al.*, 2017; Zhao *et al.*, 2017). In cotton, a previous study discovered that developing fibre cells undergo endoreduplication but not cell division (Van't Hof, 1999). The single-cell characteristic of the cotton fibre prompts us to propose that a key function of *KRP6* might be to ensure a maintained cell cycle interphase that contributes to fibre-cell elongation, whereas an abnormal function of *KRP6* or down-regulated expression would lead to an earlier transition to secondary cell wall synthesis. This possibility is supported by the identification, in this genetic regulatory network, of several upregulated secondary cell wall synthesis genes in cotton accessions with relatively short fibres. Therefore, we propose the view that the interplay between secondary wall synthesis and endoreduplication may play a vital role in determining the rate or timing of fibre elongation and final length (Sablowski & Carnier Dornelas, 2013). Characterisation of this genetic locus and the related regulatory network suggests that future improvement of

fibre quality, especially for FL, should focus on the temporal manipulation of secondary cell wall synthesis by genetic selection of eQTL hotspots such as Hot216.

In conclusion, we provide an example of the power of integrating genotypic, gene expression and phenotypic analysis to gain insights into the genetic regulation of fibre quality-related traits, which informs our understanding of the role of secondary cell wall in modulating cell elongation in plants.

Acknowledgements

This project was financially supported by the National Natural Science Foundation of China (31830062; 31471540) to XZ and LT, and was also supported by National Key Research and Development Programme of China (2016YFD0100505-05) and Fundamental Research Funds for the Central Universities (2662017JC030) to LT. Funding was also provided by the National Postdoctoral Programme for Innovative Talents (BX201700094) to MW. The authors declare no competing interests.

Author contributions

LT, MW and Xianlong Zhang conceived and designed the project. CY and JY managed cotton cultivation in the field. PW, ZL, Xiangnan Zhang, FY, ZY, KG and NL collected fibre samples and performed experiments. LT, ZL and PW performed the RNA sequencing. MW, CS, PW, BL and ZL analysed the data. GNT, DDF and KL contributed to project discussion. MW wrote the manuscript draft. XZ, LT, KL and DDF revised the manuscript. ZL and PW contributed equally to this work.

ORCID

David D. Fang https://orcid.org/0000-0003-0036-5459
Keith Lindsey https://orcid.org/0000-0002-7992-6804
Xianlong Zhang https://orcid.org/0000-0002-7703-524X
Maojun Wang https://orcid.org/0000-0002-4791-3742
Lili Tu https://orcid.org/0000-0003-4294-1928

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences, USA* **100**: 4649–4654.

Aguet F, Brown AA, Castel SE, Davis JR, Yuan H, Jo B, Mohammadi P, Park YS, Parsana P, Segrè AV, *et al.* 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**: R106.

Barrôco RM, Peres A, Droual A-M, De Veylder L, Nguyen LSL, De Wolf J, Mironov V, Peerbolte R, Beemster GTS, Inzé D *et al.* 2006. The cyclin-dependent kinase inhibitor orysa; KRP1 plays an important role in seed development of rice. *Plant Physiology* **142**: 1053–1064.

Bischoff V, Nita S, Neumetzler L, Schindelasch D, Urbain A, Eshed R, Persson S, Delmer D, Scheible WR. 2010. *TRICHOME BIREFRINGENCE* and its homolog *AT5G01360* encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiology* **153**: 590–602.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bourquin V, Nishikubo N, Abe H, Brumer H, Denman S, Eklund M, Christiernin M, Teeri TT, Sundberg B, Mellerowicz EJ. 2002. Xyloglucan Endotransglycosylases have a function during the formation of secondary cell walls of vascular tissues. *Plant Cell* **14**: 3073–3088.

Cheung VG, Spielman RS. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews Genetics* **10**: 595–604.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**: 184–194.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Fang L, Wang Q, Hu Y, Jia Y, Chen J, Liu B, Zhang Z, Guan X, Chen S, Zhou B *et al.* 2017. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nature Genetics* **49**: 1089–1098.

Fu JJ, Cheng YB, Linghu JJ, Yang XH, Kang L, Zhang ZX, Zhang J, He C, Du XM, Peng ZY *et al.* 2013. RNA sequencing reveals the complex regulatory network in the maize kernel. *Nature Communications* **4**: 2832.

García-Ramírez E, Rosas-Bringas OG, Godínez-Palma SK, Vázquez-Ramos JM, Rosas-Bringas FR, Zamora-Zaragoza J. 2017. Two maize Kip-related proteins differentially interact with, inhibit and are phosphorylated by cyclin D–cyclin-dependent kinase complexes. *Journal of Experimental Botany* **68**: 1585–1597.

Guo K, Du X, Tu L, Tang W, Wang P, Wang M, Liu Z, Zhang X. 2016. Fibre elongation requires normal redox homeostasis modulated by cytosolic ascorbate peroxidase in cotton (*Gossypium hirsutum*). *Journal of Experimental Botany* **67**: 3289–3301.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA *et al.* 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**: 245–252.

Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, Song L, Safi A, McCarroll S, Neale BM, Ophoff RA *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics* **50**: 538–548.

Haigler C, Betancur L, Stiff M, Tuttle J. 2012. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. *Frontiers in Plant Science* **3**: 104.

Hakonarson H, Li M, Wang K. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**: e164.

Han LB, Li YB, Wang HY, Wu XM, Li CL, Luo M, Wu SJ, Kong ZS, Pei Y, Jiao GL *et al.* 2013. The dual functions of WLIM1a in cell elongation and secondary wall formation in developing cotton fibers. *Plant Cell* **25**: 4421–4438.

Holanda AJ, Silva IT, Nussenzweig MC, Jankovic M, Rosales RA. 2014. Identification of chromosomal translocation hotspots via scan statistics. *Bioinformatics* **30**: 2551–2558.

Hormozdiari F, Gazal S, van de Geijn B, Finucane HK, Ju CJT, Loh P-R, Schoech A, Reshef Y, Liu X, O'Connor L *et al.* 2018. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature Genetics* **50**: 1041–1047.

Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendel JF. 2008. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proceedings of the National Academy of Sciences, USA* **105**: 6191–6195.

Huang C, Nie X, Shen C, You C, Li W, Zhao W, Zhang X, Lin Z. 2017. Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnology Journal* **15**: 1374–1386.

Jégu T, Latrasse D, Delarue M, Mazubert C, Bourge M, Hudik E, Blanchet S, Soler M-N, Charon C, De Veylder L *et al.* 2013. Multiple functions of Kip-related protein5 connect endoreduplication and cell elongation. *Plant Physiology* **161**: 1694–1705.

Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJM, Vreugdenhil D, Koornneef M, Jansen RC. 2007. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences, USA* **104**: 1708–1713.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.

Kim W-C, Ko J-H, Kim J-Y, Kim J, Bae H-J, Han K-H. 2013. MYB46 directly regulates the gene expression of secondary wall-associated cellulose synthases in *Arabidopsis*. *The Plant Journal* **73**: 26–36.

Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES. 2018. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**: 520–523.

Kroon DE, Buckler ES, Bradbury PJ, Casstevens TM, Ramdoss Y, Zhang Z. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al.* 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.

Lee C, O'Neill MA, Tsumuraya Y, Darvill AG, Ye Z-H. 2007. The irregular xylem9 mutant is deficient in Xylan Xylosyltransferase activity. *Plant and Cell Physiology* **48**: 1624–1634.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li Y, Tu L, Pettolino FA, Ji S, Hao J, Yuan D, Deng F, Tan J, Hu H, Wang Q *et al.* 2016. GbEXPATR, a species-specific expansin, enhances cotton fibre elongation through cell wall restructuring. *Plant Biotechnology Journal* **14**: 951–963.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**: 833–835.

Liu H, Luo X, Niu L, Xiao Y, Chen L, Liu J, Wang X, Jin M, Li W, Zhang Q *et al.* 2017. Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. *Molecular Plant* **10**: 414–426.

Liu R, Gong J, Xiao X, Zhang Z, Li J, Liu A, Lu Q, Shang H, Shi Y, Ge Q *et al.* 2018. GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. *Frontiers in Plant Science* **9**: 1067–1067.

Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, DeRose-Wilson LJ, McKay JK, Sen S, Juenger TE. 2013. Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in *Arabidopsis*. *Plant Cell* **25**: 3266–3279.

Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G, Wu L, Li Z, Liu Z, Sun G *et al.* 2018. Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nature Genetics* **50**: 803–813.

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in *Arabidopsis* and *Eucalyptus*. *The Plant Journal* **62**: 689–703.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The genome analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–295.

Sablowski R, Carnier Dornelas M. 2013. Interplay between cell growth and cell cycle in plants. *Journal of Experimental Botany* **65**: 2703–2714.

Shan C-M, Shangguan X-X, Zhao B, Zhang X-F, Chao L-M, Yang C-Q, Wang L-J, Zhu H-Y, Zeng Y-D, Guo W-Z *et al.* 2014. Control of cotton fibre elongation by a homeodomain transcription factor GhHOX3. *Nature Communications* **5**: 5519.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**: 2498–2504.

Tan J, Tu L, Deng F, Hu H, Nie Y, Zhang X. 2013. A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. *Plant Physiology* **162**: 86–95.

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R *et al.* 2015. An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* **517**: 571–575.

Thyssen GN, Jenkins JN, McCarty JC, Zeng L, Campbell BT, Delhom CD, Islam MS, Li P, Jones DC, Condon BD *et al.* 2019. Whole genome sequencing of a MAGIC population identified genomic loci and candidate genes for major fiber quality traits in upland cotton (*Gossypium hirsutum* L.). *Theoretical and Applied Genetics.* **132**: 989–999.

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. 2017. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research* **45**: W122–W129.

Tuttle JR, Nah G, Duke MV, Alexander DC, Guan X, Song Q, Chen ZJ, Scheffler BE, Haigler CH. 2015. Metabolomic and transcriptomic insights into how cotton fiber transitions to secondary wall synthesis, represses lignification, and prolongs elongation. *BMC Genomics* **16**: 477.

Van't Hof J. 1999. Increased nuclear DNA content in developing cotton fiber cells. *American Journal of Botany* **86**: 776–779.

Wang J, Yu H, Weng X, Xie W, Xu C, Li X, Xiao J, Zhang Q. 2014. An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of Experimental Botany* **65**: 1069–1079.

Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, Li X, Xiao J, Zhang Q. 2010. A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *The Plant Journal* **63**: 1063–1074.

Wang M, Tu L, Lin M, Lin Z, Wang P, Yang Q, Ye Z, Shen C, Li J, Zhang L *et al.* 2017. Asymmetric subgenome selection and *cis*-regulatory divergence during cotton domestication. *Nature Genetics* **49**: 579–587.

Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G *et al.* 2019. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature Genetics* **51**: 224–229.

Wang M, Wang P, Lin M, Ye Z, Li G, Tu L, Shen C, Li J, Yang Q, Zhang X. 2018. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nature Plants* **4**: 90–97.

Wang X, Chen Q, Wu Y, Lemmon ZH, Xu G, Huang C, Liang Y, Xu D, Li D, Doebley JF *et al.* 2018. Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population. *Molecular Plant* **11**: 443–459.

West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175**: 1441–1450.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**: 76–82.

Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L *et al.* 2015. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. *Scientific Reports* **5**: 17662.

Zhang L, Su W, Tao R, Zhang W, Chen J, Wu P, Yan C, Jia Y, Larkin RM, Lavelle D *et al.* 2017. RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications* **8**: 2264.

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE *et al.* 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc.

TM-1) provides a resource for fiber improvement. *Nature Biotechnology* **33**: 531–537.

Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Research* **21**: 725–733.

Zhang Z, Ruan Y-L, Zhou N, Wang F, Guan X, Fang L, Shang X, Guo W, Zhu S, Zhang T. 2017. Suppressing a putative sterol carrier gene reduces plasmodesmal permeability and activates sucrose transporter genes during cotton fiber elongation. *Plant Cell* **29**: 2027–2046.

Zhao XA, Bramsiepe J, Van Durme M, Komaki S, Prusicki MA, Maruyama D, Forner J, Medzihradszky A, Wijnker E, Harashima H *et al.* 2017. RETINOBLASTOMA RELATED1 mediates germline entry in *Arabidopsis*. *Science* **356**: eaaf6532.

Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, Lin T, Qin M, Peng M, Yang C *et al.* 2018. Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**: 249–261.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM *et al.* 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**: 481–487.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Comparison of significance (*P*-values) between intrachromosomal eQTL (SameChr_eQTL) and interchromosomal eQTL (DiffChr_eQTL).

Fig. S2 Visualisation of GO hierarchical graphic design for three likely causal genes.

Fig. S3 Comparison of fibre quality-related traits between cotton accessions in cluster1 and cluster2.

Fig. S4 Differential expression analysis of genes in accessions with extreme fibre length.

Fig. S5 Whole genome Hi-C matrix in *G. hirsutum*.

Table S1 Descriptive statistics for phenotypic variations and broad-sense heritability for five fibre quality-related traits.

Table S2 Summary of significant SNPs in GWAS loci.

Table S3 Summary of genes in GWAS loci with putative function in fibre development.

Table S4 Summary of RNA-Seq data in this study.

Table S5 GO enrichment of genes in one subgenome regulated by eQTL in the other subgenome.

Table S6 Colocalisation of GWAS associations and subgenomic eQTL regulation.

Table S7 Summary of eQTL hotspot in this study.

Table S8 Information for 962 eGenes regulated by the Hot216. Columns 2–6 show the expression levels (FPKM) of eGenes in fibre development.

Table S9 GO enrichment of genes which were regulated by the eQTL hotspot in chromosome D11.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.