Islam et al. BMC Genomics (2016) 17:903
DOI 10.1186/s12864-016-3249-2

BMC Genomics

RESEARCH ARTICLE Open Access

A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton

Md Sariful Islam[1], Gregory N. Thyssen[2], Johnie N. Jenkins[3], Linghe Zeng[4], Christopher D. Delhom[5], Jack C. McCarty[3], Dewayne D. Deng[3], Doug J. Hinchliffe[2], Don C. Jones[6] and David D. Fang[1*]

Abstract

Background: Cotton supplies a great majority of natural fiber for the global textile industry. The negative correlation between yield and fiber quality has hindered breeders' ability to improve these traits simultaneously. A multi-parent advanced generation inter-cross (MAGIC) population developed through random-mating of multiple diverse parents has the ability to break this negative correlation. Genotyping-by-sequencing (GBS) is a method that can rapidly identify and genotype a large number of single nucleotide polymorphisms (SNP). Genotyping a MAGIC population using GBS technologies will enable us to identify marker-trait associations with high resolution.

Results: An Upland cotton MAGIC population was developed through random-mating of 11 diverse cultivars for five generations. In this study, fiber quality data obtained from four environments and 6071 SNP markers generated via GBS and 223 microsatellite markers of 547 recombinant inbred lines (RILs) of the MAGIC population were used to conduct a genome wide association study (GWAS). By employing a mixed linear model, GWAS enabled us to identify markers significantly associated with fiber quantitative trait loci (QTL). We identified and validated one QTL cluster associated with four fiber quality traits [short fiber content (SFC), strength (STR), length (UHM) and uniformity (UI)] on chromosome A07. We further identified candidate genes related to fiber quality attributes in this region. Gene expression and amino acid substitution analysis suggested that a regeneration of bulb biogenesis 1 (*GhRBB1_A07*) gene is a candidate for superior fiber quality in Upland cotton. The DNA marker *CFBid0004* designed from an 18 bp deletion in the coding sequence of *GhRBB1_A07* in Acala Ultima is associated with the improved fiber quality in the MAGIC RILs and 105 additional commercial Upland cotton cultivars.

Conclusion: Using GBS and a MAGIC population enabled more precise fiber QTL mapping in Upland cotton. The fiber QTL and associated markers identified in this study can be used to improve fiber quality through marker assisted selection or genomic selection in a cotton breeding program. Target manipulation of the *GhRBB1_A07* gene through biotechnology or gene editing may potentially improve cotton fiber quality.

Keywords: Cotton, Fiber quality, Genome wide association study, Genotyping-by-sequencing, Multi parent advanced generation inter-cross

* Correspondence: david.fang@ars.usda.gov
[1]Cotton Fiber Bioscience Research Unit, USDA-ARS, Southern Regional Research Center, New Orleans, LA 70124, USA
Full list of author information is available at the end of the article

Background

Cotton is one of the most important natural fibers for the textile industry and is a significant food source for livestock and human consumption. The industries associated with cotton fiber production and processing have a significant impact on the world economy [1, 2]. Although there are more than 50 species in the G*ossypium* genus, only four (*G. barbadense* L., *G. hirsutum* L., *G. arboreum* L. and *G. herbaceum* L.) are domesticated for cultivation [3]. Among these four cultivated species, Upland cotton (*G. hirsutum*) constitutes about 95 % of the world cotton production [2, 4]. Hence, most breeding efforts focus on the improvement of Upland cotton with the primary goal to improve yield and fiber quality. However, a simultaneous improvement of fiber quality and yield is challenging due to the presence of a negative genetic correlation between yield and quality [5]. Cotton breeders need to break this negative association in order to successfully breed improved cultivars.

In traditional QTL mapping, a bi-parental population is usually utilized to identify the genomic location and magnitude of effect of a locus that affects a phenotypic trait [5–7]. QTL mapping using such bi-parental populations is usually low in resolution since only two alleles per locus are analyzed and genetic recombination is limited [8]. A multi-parent advanced generation inter-cross (MAGIC) strategy has been anticipated to have higher genetic diversity, smaller haplotype blocks, higher recombination and better mapping resolution [8]. Thus, a MAGIC population in cotton has a better chance to break the negative linkage between yield and fiber quality [9]. Very recently, use of a MAGIC population to identify QTL has become a new approach, thanks to the advancement of next generation sequencing (NGS) techniques and novel statistical and bioinformatics tools to analyze large data sets.

Genome-wide association study (GWAS) falls into two categories: broad-based and narrow-based. Usually, a broad-based GWAS uses germplasm, landraces, cultivars, and natural populations to identify marker-trait associations. A narrow-based GWAS uses a man-made population involving more than two parents such as a MAGIC population to identify marker-trait associations. GWAS has several advantages over traditional QTL analysis, such as higher diversity, better utility of the identified QTL across diverse germplasm, and reduced time required for population development [10–12]. GWAS has been successfully employed in many plants, such as Arabidopsis [13], maize [10, 14], barley [15, 16], wheat [17, 18], rice [19], oat [20], sorghum [21] and soybean [11, 22]. In cotton, a few genetic studies using GWAS have been reported [2, 4, 23, 24]. Very recently another two GWAS papers in cotton have been published [25, 26]. All of those studies were based on using simple sequence repeat (SSR) markers.

With the advent of NGS technologies, GBS has already been proven to rapidly identify and genotype large numbers of single nucleotide polymorphism (SNP) markers in many crops [14] as well as in Upland cotton [27]. Although scientists have successfully identified marker-trait associations using GBS-based SNPs in many plant species [14, 21, 28, 29], so far no GWAS study has been reported in Upland cotton using SNPs. There are a few GWAS studies that were reported using SSR markers in Upland cotton [2, 4, 23, 24, 26, 30] as mentioned earlier. Among those reports, our laboratory previously identified 131 fiber QTL and 37 QTL clusters using 1582 SSR markers and 275 RILs of the same MAGIC population [4]. Recently, some candidate genes for fiber strength have been detected through gene expression and mapping-by-sequencing analysis [31–36]. So far, no report has been published to identify candidate genes in conjunction with GWAS analysis in cotton. In this study, a large number of GBS-based SNPs and a set of 223 SSRs were used to identify associations between fiber quality traits and DNA markers. We also used phenotypic data obtained from four environments for all RILs of the MAGIC population. In order to identify fiber quality candidate genes, we also investigated the most significant genomic location in more depth including whole genome sequencing (WGS) of the parental lines and gene expression analysis. We identified a positive correlation between the candidate gene regeneration of bulb biogenesis 1 (*GhRBB1_A07*) and fiber qualities in both the MAGIC RILs and a panel of 105 commercial cultivars. The marker-trait associations and candidate genes identified in this study may be useful to improve fiber quality in a cotton breeding program via marker-assisted selection (MAS) or genomic selection.

Results

Upland cotton MAGIC population structure assessment

Ten cultivars and one non-commercial variety (Additional file 1) that represent broad Upland cotton diversity in the US were crossed in a half-diallel design to produce the MAGIC RIL population (C_5S_6) (Fig. 1 and Additional file 2) [9]. The effectiveness of random-mating and its consequence on structure of this population was evaluated using 6039 SNP and 223 SSR markers. The relatedness between the 547 RILs is presented as a heat map (Additional file 3). The relatedness matrix value of RIL pairs ranged from 0.253 to 0.637 which indicated that no specific population structure exists among the RILs. Kinship analysis also showed no specific clustering pattern of the RILs (Additional file 4). These results confirmed our previous conclusion that no obvious population structure was observed in this population [4].

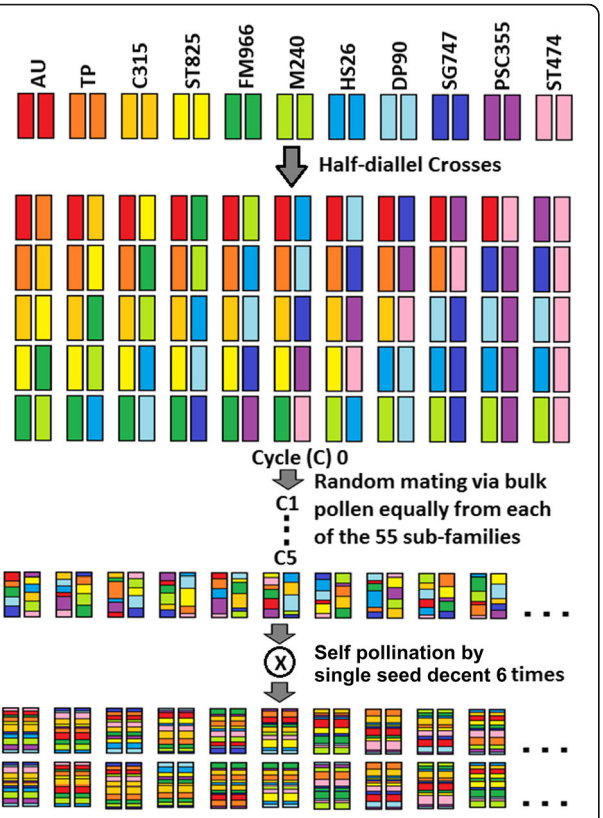

Fig. 1 Breeding scheme of Upland cotton MAGIC population. Recombinant inbred lines (RILs) were developed according to the crossing scheme as below and 55 sub-families (10 shown here) were established. Individual colors represents the 11 parents (AU - Acala Ultima, TP - Tamcot Pyramid, C315 - Coker 315, ST825 - Stoneville 825, FM966 – Fibermax 966, M240 - M-240RNR, HS26 - Paymaster HS-26, DP90 - Deltapine Acala 90, SG747 – Suregrow 747, PSC355 – Phytogen 355, ST474 – Stoneville 474)

Fiber quality analysis

Fiber quality data of 11 parents and RILs were collected from 3 years (2009, 2010 and 2011) in Starkville, MS and 1 year (2013) in Stoneville, MS. In this study, six fiber quality traits were measured and results are shown in Table 1. Parents' trait mean and RILs' mean were similar. However, ranges of all the traits of RILs were broader possibly due to transgressive segregation. The correlation coefficient analysis between all the traits of parents and RILs were conducted separately. Results revealed positive strong correlation between strength (STR) and length (UHM), STR and uniformity (UI), and UHM and UI, in both parents and RILs fiber data (Table 2). We previously analyzed the heritability (H^2) of fiber quality traits of this MAGIC population. Their heritability was moderate to high [4]. Analyses of variance for data collected in all environments indicated highly significant genotypic effects for all tested fiber quality traits (Additional file 5). The environment (location year) effects were significant for combined data for each trait under field conditions. RIL environment interactions under field conditions were also significant suggesting that there were differences in the response of some genotypes to environment over the location years of these experiments.

Alignment and distribution of SNP markers in allotetraploid Upland cotton TM-1 draft genome

Three RILs had very poor sequencing data, and were discarded from the study. Thus, the sequence data of 547 MAGIC RILs were used to obtain SNPs. A total of 128,212 SNP contigs were called. Of them, 6071 polymorphic SNPs were identified after filtering using the criteria: missing rate ≤ 20 %, minor allele frequency (MAF) ≥ 5 %, number of genotypes ≥ 2 and MAF difference between average of parents and RILs ≤ 20 %. In the case of SSRs, a total of 223 were selected based on our previous study result, and 255 SSR loci were scored. A summary of SNPs in both parents and RILs is shown in Table 3. Of the 11 parents, DP90 had the highest rate of SNP heterozygosity. All 6071 GBS-based SNP contigs and 223 SSR clone sequences were aligned to the allotetraploid Upland cotton TM-1 draft genome [37]. A total of 6039 SNPs and all 223 SSRs produced a hit, and were able to be aligned to the TM-1 genome (Additional file 6). The 32 non-aligned SNPs are

Table 1 Fiber quality measurements of parents and RILs collected in four environments

Trait	Parents					RILs				
	Mean	SE	SD	Min	Max	Mean	SE	SD	Min	Max
ELO (%)	6.10	0.131	1.23	3.75	8.90	6.03	0.029	1.56	2.80	9.93
MIC	4.63	0.047	0.44	3.17	5.55	4.67	0.011	0.46	3.29	6.25
SFC (%)	7.44	0.074	0.70	5.78	9.20	7.38	0.020	0.81	5.01	10.86
STR (g/tex)	30.68	0.250	2.34	25.80	36.8	30.59	0.056	2.29	24.13	40.97
UHM (mm)	28.19	0.152	1.27	24.64	30.99	28.19	0.051	1.78	22.61	34.04
UI (%)	83.68	0.112	1.06	81.25	85.78	83.76	0.030	1.22	79.20	88.13

ELO percent elongation of fibers before breaking, *MIC* a measurement of fiber fineness or maturity by an airflow instrument that measures the air permeability of a constant mass of cotton fibers compressed to a fixed volume, *SFC* short fiber content, calculated as the content (%) of fiber shorter than 12.7 mm, *STR* force required to break a bundle of fibers one tex unit in size, *UHM* upper half mean fiber length, the average length of the longer one-half of the fibers sampled, *UI* uniformity index, calculated as the (mean length/UHM) × 100, *SE* Standard error mean, *SD* standard deviation

Table 2 Pearson (r) correlations among fiber quality traits in RILs of MAGIC population

Trait	MIC	SFC	STR	UHM	UI
Parents					
ELO	0.34**	−0.36**	−0.14	0.13	0.30**
MIC		−0.24*	−0.05	−0.23*	0.16
SFC			−0.50**	−0.48**	−0.84**
STR				0.41**	0.48**
UHM					0.58**
RILs					
ELO	0.15**	−0.47**	−0.01	0.27**	0.43**
MIC		−0.07**	−0.14**	−0.39**	−0.04
SFC			−0.56**	−0.48**	−0.89**
STR				0.52**	0.61**
UHM					0.65**

* Significant at the p value <0.05. **Significant at the p value < 0.01

probably located in contigs that have not been assembled into chromosomes in the TM-1 draft genome. The marker distribution along each chromosome is shown in Additional file 7 as histograms generated by counting markers in each Mb interval.

Linkage disequilibrium (LD)

The square of correlation coefficient (r^2) between markers located on each chromosome was measured to create a LD relationship between loci by using TASSEL 5.0 software [38]. We prefer to use r^2 values over normalized LD coefficient (D′) since r^2 between two loci have more reliable sampling properties and are influenced by both mutation and recombination events in the population. The LD decay plots for each chromosome and sub-genome were created by plotting r^2 value onto physical distance measured in base pairs (Fig. 2). As anticipated, the r^2 value negatively correlated with the physical distance between the loci. Results revealed that LD decay varied between chromosomes. Both sub-genomes had similar physical distances of reaching the LD threshold ($r^2 = 0.2$) with ~520 kb and ~480 kb for A_t-subgenome and D_t-subgenome, respectively. Among the chromosomes, the slowest and most rapid LD decay was observed in chromosome A07 and D11, respectively (Fig. 2). Chromosome-wide LD contour plots are shown in Additional file 8.

Genome-wide association analysis of fiber quality traits

First, we performed GWAS using the best linear unbiased predictor (BLUPs) of RILs' mean over four environments

Table 3 SNP summary of 11 parents and 547 RILs

Sample[a]	Genotype (%)				Reads (million)	bp (million)
	Homozygote (major allele)	Homozygote (minor allele)	Heterozygote	Missing		
AU	59.2	16.1	18.8	5.9	2.55	163.07
C315	61.4	12.8	20.5	5.3	2.21	141.30
DP90	63.4	11.7	20.9	4.0	2.48	158.58
FM966	60.6	16.9	17.7	4.8	2.63	168.27
HS26	60.5	17.3	18.2	4.1	2.64	169.22
M240	61.0	16.9	17.8	4.4	2.73	174.61
PSC355	62.6	13.2	19.4	4.8	2.40	153.40
PSC355	62.6	13.2	19.4	4.8	2.40	153.40
SG747	65.5	12.1	17.9	4.5	2.35	150.30
ST474	66.4	10.6	19.6	3.3	2.17	138.96
ST825	63.8	14.0	17.1	5.1	2.20	141.11
TP	58.9	17.9	17.8	5.4	2.51	160.70
RIL avg.	62.1	14.8	16.1	6.9	2.12	135.93

[a]*AU* Acala Ultima, *C315* Coker 315, *DP90* Deltapine Acala 90, *FM966* Fibermax 966, *HS26* Paymaster HS-26, *M240* M-240RNR, *PSC355* Phytogen 355, *SG747* Suregrow 747, *ST474* Stoneville 474, *ST825* Stoneville 825, *TP* Tamcot Pyramid

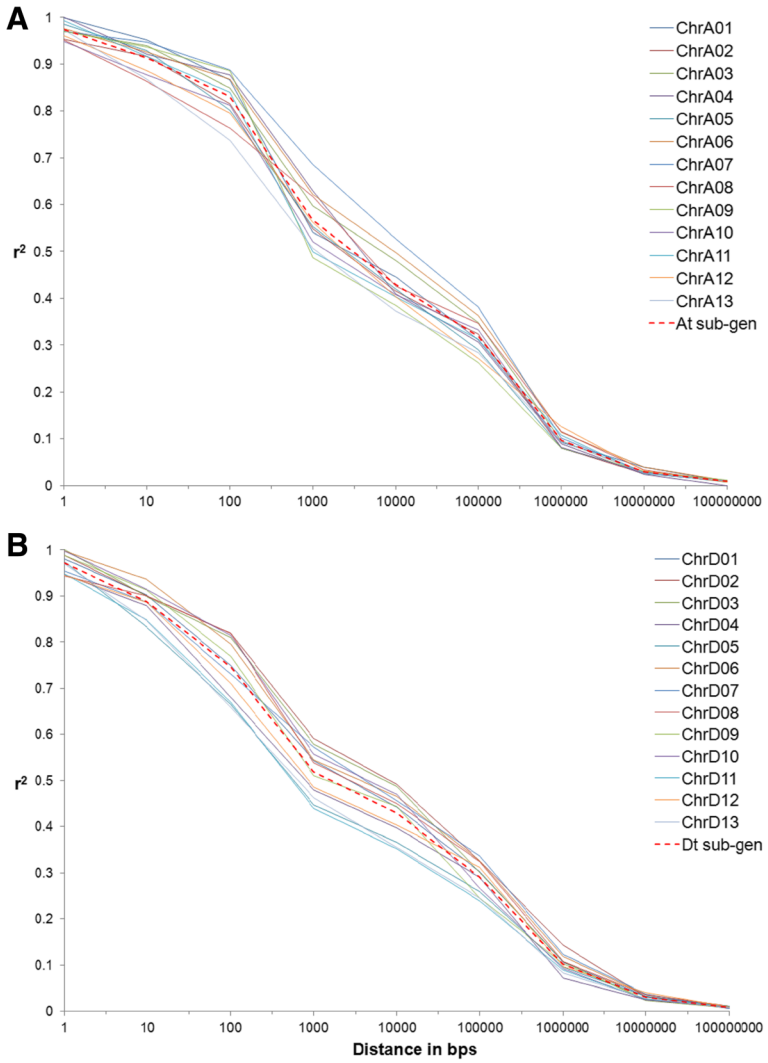

Fig. 2 Linkage disequilibrium (LD) decay across all chromosomes. **a** A_t sub-genome, **b** D_t sub-genome

in a general linear model (GLM) with incorporation of principle component analysis (PCA) as Q matrix (population structure) by employing TASSEL 5.0 software [38]. The quantile-quantile (Q-Q) plots representing expected and observed probability of obtaining association of markers with respective traits revealed the possibility of false positive associations, since many of the observed *p*-values deviated from the uniform distribution (Additional file 9). Hence, a mixed linear model (MLM) [39] was used for further analysis in order to control the genomic inflation effectively using both TASSEL 5.0 [38] and GAPIT [40] software. This time, both PCA as Q and relatedness between RILs as K matrices were incorporated in the association analysis. Updated Q-Q plots showed that most of the observed *p*-values follow a uniform distribution but the few that are in LD with a causal polymorphism have significant *p*-values in the tail (Additional file 10). The Manhattan plots for six tested fiber quality traits were generated from GAPIT and TASSEL software and are presented in Fig. 3 and Additional file 11, respectively. Results from both programs were more or less identical, so we decided to proceed further with the results generated from GAPIT. At p value ≤ 0.01, the MLM identified 357 unique markers associated with 86 fiber QTL on 24 chromosomes (Table 4).

The extremely significant (p value ≤ 0.0001) loci (*qELO-cD04*) associated with fiber elongation (ELO) comprises two SNPs (*CFB9477* and *CFB9479*) covering 152 Kb on chromosome D04 (Table 5, Additional file 12). The SNP (*CFB9477*) with strongest association with ELO (p value = 1.38E-12) is located at position 47,719,839 bp and explains 31 % of the phenotypic variation. Genotypes carrying the minor alleles for both the flanking SNPs (*GA*) at this QTL location had substantially higher ELO value than those carrying the flanking major alleles (*TC*) (Fig. 4).

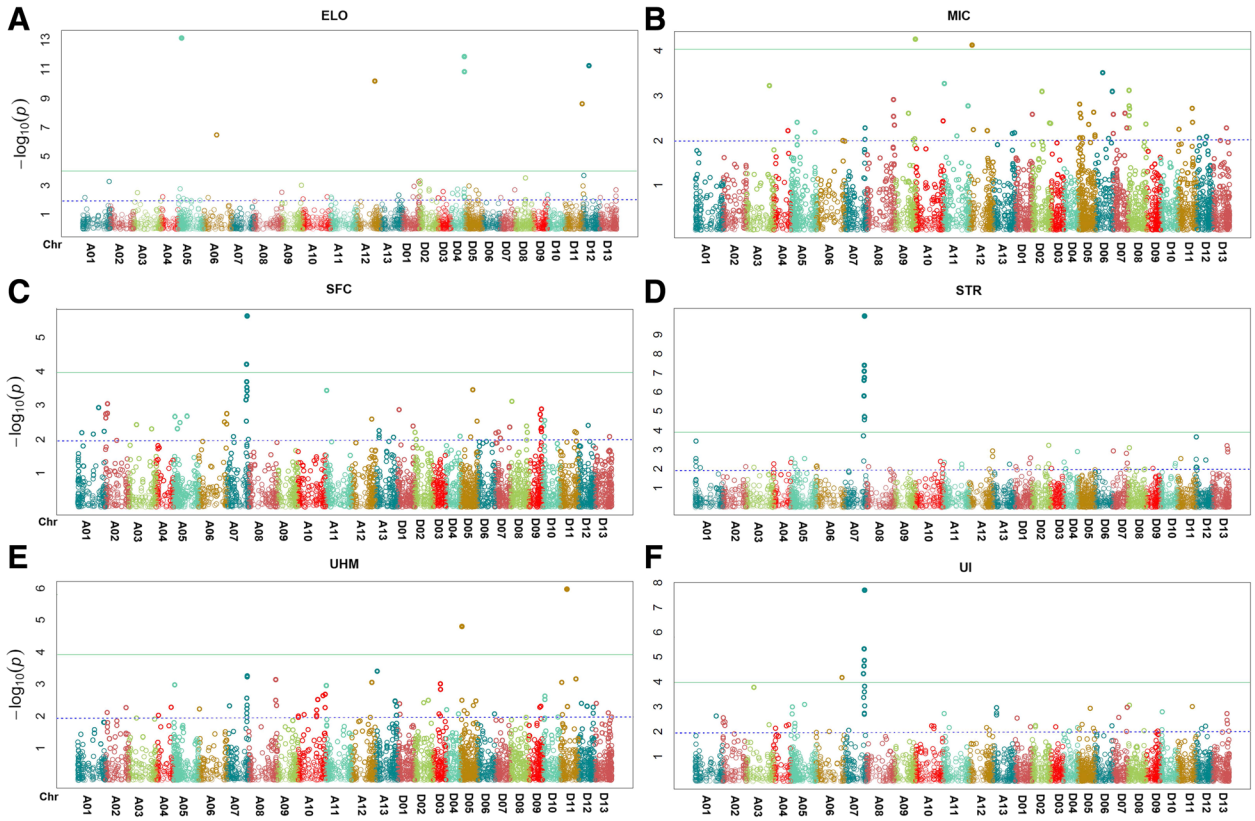

Fig. 3 Manhattan plots generated from GAPIT software for six fiber quality traits, **a** Elongation (ELO), **b** Micronaire (MIC), **c** Short fiber content (SFC), **d** Fiber strength (STR), **e** Upper half mean (UHM) fiber length, and **f** Uniformity (UI). The negative $\log_{10}$ transformed *p* values were plotted against the marker positions on the physical map of each of the 26 Upland cotton chromosome. The significant thresholds ($p = 0.01$ and 0.0001) are indicated by the purple and green horizontal dot line, respectively

On chromosome A07, one of the most significant (p value ≤ 0.0001) loci (*qSTR-cA07*) associated with fiber strength had nine markers spanning 2.5 Mb. The same genomic region also housed QTL for SFC (*qSFC-cA07*) and UI (*qUI-cA07*) at p value ≤ 0.0001; and UHM at p value ≤ 0.001. The most significant linked marker is SSR marker *C2-0114* located at position 72,331,925 bp which contributes 8, 18, 13, and 10 %, of phenotypic variation for SFC, STR, UHM and UI, respectively (Table 5). Since this genomic region between 70 to 76 Mb on chromosome A07 is significantly associated with four fiber quality traits, we further investigated the annotated genes located in this region. In order to first find the effective border of the QTL region, we analyzed all the available haplotype combinations affecting the STR phenotype. Results showed that RILs carrying the minor alleles (*TT*) of SNPs *CFB7267* and *CFB7300* located between

Table 4 Marker-trait associations and QTL identified at different significance level

Trait	Marker trait association				QTL[b]				
	$p \leq 0.01$	$p \leq 0.001$	$p \leq 0.0001$	Total	$p \leq 0.01$	$p \leq 0.001$	$p \leq 0.0001$	Total	# Chr.
ELO	54	8	7	69	13	1	1	15	10
MIC	58	6	2	66	16	0	0	16	14
SFC	54	9	2	65	10	0	1	11	10
STR	55	7	9	71	11	1	1	13	12
UHM	62	9	2	73	15	2	0	17	12
UI	58	8	6	72	13	0	1	14	13
Total	294[a]	41[a]	22[a]	357[a]	78	4	4	86	24[a]

[a]Some marker loci were associated with more than one trait
[b]Marker loci mapped within 5 Mb intervals were considered as a unique QTL

Table 5 The extremely significant ($p \leq 0.0001$) associations between markers and phenotypes

Trait	Marker	Chr.	Position (bp)	*p* value	MAF	R^2	Allelic effect
ELO	MGHES-021	A05	12471766	7.48E-14	0.49	0.32	0.306
	MUSS275	A06	47495244	3.28E-07	0.50	0.28	0.196
	DPL0644c	A12	66210831	7.02E-11	0.40	0.30	−0.279
	CFB9477	D04	47719839	1.38E-12	0.30	0.31	−0.313
	CFB9479	D04	47872771	1.58E-11	0.32	0.31	−0.305
	SHIN-0966	D11	61712633	2.30E-09	0.49	0.29	0.223
	MUSB0846	D12	20174365	5.74E-12	0.49	0.31	0.274
MIC	CFB7672	A09	73256219	5.53E-05	0.21	0.11	−0.089
	CFB8181	A12	9895997	7.48E-05	0.50	0.11	−0.138
SFC	CFB7265	A07	70370764	6.22E-05	0.12	0.07	−0.266
	C2-0114	A07	72331925	2.32E-06	0.14	0.08	−0.240
STR	CFB7265	A07	70370764	2.41E-07	0.12	0.15	0.897
	DPL0852	A07	71015749	8.34E-08	0.18	0.15	0.689
	CFB7266	A07	71506333	4.17E-08	0.23	0.16	−0.701
	CFB7267	A07	71510617	1.90E-05	0.18	0.14	0.572
	CFB7268	A07	71510635	2.77E-05	0.19	0.14	−0.559
	DPL0757	A07	71588662	1.73E-07	0.17	0.15	0.698
	CFB7269	A07	72135309	1.45E-06	0.19	0.14	−0.619
	C2-0114	A07	72331925	1.07E-10	0.14	0.18	0.868
	CFB7271	A07	72856692	1.98E-05	0.13	0.14	0.625
UHM	CFB9550	D05	5051755	1.57E-05	0.49	0.14	0.011
	CFB11211	D11	24009159	1.07E-06	0.10	0.15	0.018
UI	MUSS122	A06	93195954	6.92E-05	0.25	0.07	0.207
	SHIN-1138	A07	68260333	4.85E-05	0.19	0.07	−0.010
	CFB7265	A07	70370764	2.31E-05	0.12	0.08	0.384
	DPL0852	A07	71015749	4.66E-06	0.18	0.08	0.309
	DPL0757	A07	71588662	1.36E-05	0.17	0.08	0.304
	C2-0114	A07	72331925	2.01E-08	0.14	0.10	0.404

Chr. Chromosome, *MAF* Minor allele frequency (%)

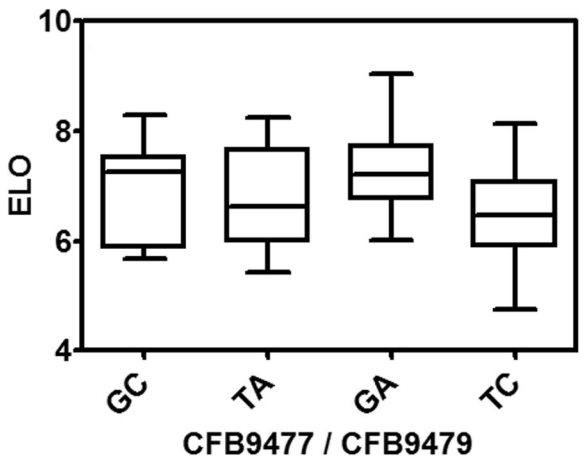

Fig. 4 The effect of two marker loci selection on fiber elongation of RILs. RILs were divided into four groups based on the allele combinations at two marker loci that flanked a significant ELO QTL

71,510,617 and 76,885,949 bp produced significantly improved fiber strength than RILs carrying major alleles (*GG*) of those two SNPs (Fig. 5b). RILs carrying both major and minor alleles of markers located up and down stream of this border did not differ significantly (Additional file 13). The same result was also true for SFC, UHM and UI (Fig. 5a, c, d).

Genomic architecture in the region flanked by SNP markers CFB7267 and CFB7300

The LD contour plot indicated that there are three large LD blocks present with a few small LD blocks in the genomic region between *CFB7267* and *CFB7300* (Additional file 14). Gene annotation of the Upland cotton TM-1 genome suggested 303 genes in this region. Based on the gene ontology, many of the genes are predicted to be transcription factors involved in cotton fiber development such as *CONSTANT-like 9* (*Gh_A07G1753*) and *oxydoreductase zinc-binding dehydogenase family protein* (*Gh_A07G1803*). Of the 303 genes, 142 had at least 1 read Per kilobase of transcript per million mapped reads (RPKM) expression in RNA-seq data from developing fiber cells of four Upland cotton cultivars and three cotton fiber mutant lines (Additional file 15).

Out of 142 genes, 18 genes, which were previously reported as candidates for fiber cell and/or cell wall development were selected for RT-qPCR gene expression analysis. In this experiment, we only selected two varieties Acala Ultima (AU, superior fiber quality) and Tamcot Pyramid (TP, inferior fiber quality) for comparative gene expression analysis through RT-qPCR. Of the 18 tested genes, the expression level of five genes (*Gh_A07G1753*, *Gh_A07G1758*, *Gh_A07G1784*, *Gh_A07G1795*, *Gh_A07G1803*) were significantly higher in AU developing fibers while one (*Gh_A07G1802*) showed down regulated expression in AU developing fibers compared with TP (Additional file 16).

Identification of non-synonymous mutations near the *CFB7267*/*CFB7300* interval

To identify mutations in the vicinity of the QTL cluster on chromosome A07 that were not represented in the GBS data, we analyzed whole genome sequence data from each of the 11 parents. By comparing AU with other parents, only 28 SNPs in nine genes produced mis-sense (27) and nonsense (1) substitutions of amino acid sequence (Table 6). The only nonsense amino acid mutation was observed in *Gh_A07G1913* (*Protein of unknown function, DUF593*). Since this gene was not well characterized and did not express in any of the developing fiber RNA-seq data (Additional file 15), we did not pursue further investigation. To investigate which mis-sense amino acid substitutions might have significant effect on fiber quality development, we compared the chemical nature of the substituted to the original amino acids and searched the Genebank for orthologous proteins with identical mutations. We found that 11 mis-sense mutations were conversions to chemically similar amino acids, while another 16 were synonymous to orthologous proteins (Table 6). However, one interesting mutation was detected in gene *Gh_A07G2049*

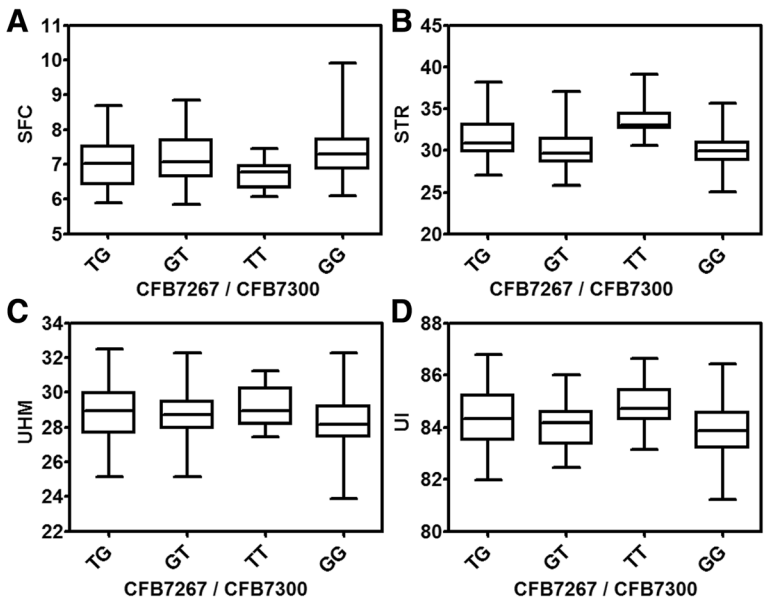

Fig. 5 The effect of two marker loci selection on four fiber quality traits of RILs. **a** Short fiber content (SFC), **b** Fiber bundle strength (STR), **c** Fiber length (UHM), and **d** Uniformity (UI). RILs were divided into four groups based on the allele combinations at two marker loci that flanked a significant QTL of the respective trait

Table 6 Amino acid change caused by the SNP between Acala Ultima (AU) and others based on Upland cotton TM-1draft genome

Gene	Position (bp)	Amino acid code		TAIR10 gene	Description
		AU	Others		
Gh_A07G1913	74778174	L	Stop	AT2G30690	Protein of unknown function, DUF593
Gh_A07G2012	76384997	V	L	AT3G01480	Cyclophilin 38
Gh_A07G2019	76483636	E	Q	AT3G16520	UDP-glucosyl transferase 88A1
	76483537	L	F		
	76483477	S	A		
	76483123	C	F		
Gh_A07G2025	76582935	H	R	AT3G01510	Like SEX4 1
Gh_A07G2041	76841398	G	S	AT5G14570	High affinity nitrate transporter 2.7
Gh_A07G2042	76843818	H	R	AT1G50170	Sirohydrochlorin ferrochelatase B
Gh_A07G2045	76868459	L	F	AT3G01570	Oleosin family protein
Gh_A07G2046	76871281	T	A	AT3G27640	Transducin/WD40 repeat-like superfamily protein
Gh_A07G2049	76921226	Q	R	AT5G40450	Regulation of bulb biogenesis, *GhRBB1_A07*
	76917593	K	E		
	76917404	G	E		
	76914909	C	Y		
	76914822	Q	H		
	76914298	I	T		
	76911876	K	N		
	76911659	InDel	-		
	76910857	T	I		
	76909194	L	P		
	76908681	D	V		
	76907407	Q	E		
	76906421	N	S		
	76904993	D	N		
	76904590	A	T		
	76901416	K	E		
	76900228	F	L		
	76899907	S	C		

(*Regeneration of bulb biogenesis 1, GhRBB1_A07*) at genomic location 76,911,659 bp which has an 18 bp deletion in the coding sequence in AU (Additional files 17 and 18). In order to find its possible effect on fiber development, we conducted RT-qPCR gene expression analysis using four [8, 14, 18 and 22 days post anthesis (DPA)] developmental stage fiber tissues of two parental lines AU and TP. The expression level of the gene was significantly higher in AU developing fibers than in TP (Fig. 6).

In order to explore the practical utility of this deletion, we designed an InDel marker (*CFBid0004*) (Forward primer, 5′ –TCTTTGATGACAACAACATTATAGA-3′ and reverse primer, 5′-AGAAACAGAAGAAACAGACA-TAAA-3′) and genotyped all 550 MAGIC RILs and 105 Upland cotton cultivars (Additional file 19) selected from the United States National Cotton Variety Test (NCVT)

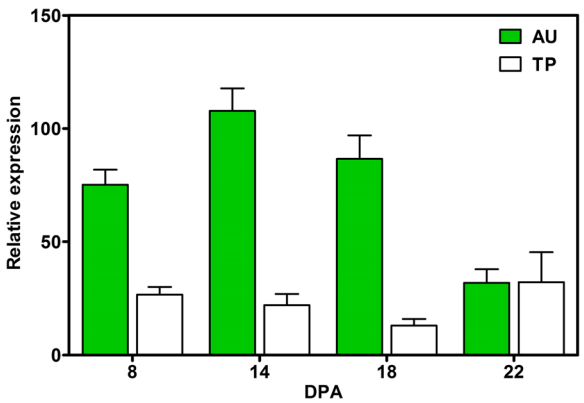

Fig. 6 RT-qPCR gene expression analyses of gene *Gh_A07G2049* (*GhRBB1_A07*) during fiber development between Acala Ultima (AU) and Tamcot Pyramid (TP)

program. The InDel marker *CFBid0004* locus has two alleles: 147 bp and 165 bp. Two alleles of the locus were grouped for One-way analyses. For comparison of the means of the grouping performance, Tukey's honest significant difference test was conducted for all pairs of levels. RILs carrying the 147 bp allele had a mean STR of 31.29 g/tex which was significantly higher than the mean STR (30.02 g/tex) of RILs carrying the reference allele 165 bp (Fig. 7b). Similarly, UHM and UI were also significantly increased while SFC was significantly decreased due to the presence of the deletion allele in the RILs genotype (Fig. 7a, c, d). The most remarkable results were found with historic phenotype data from NCVT trials. Upland cotton cultivars carrying the deletion allele (147 bp) had mean fiber strength and length of 29.46 cN/tex and 22.61 mm, which are significantly higher than mean fiber bundle strength and length (28.97 cN/tex and 21.24 mm), respectively, of cultivars carrying reference 165 bp allele (Fig. 7e, f).

Discussion

A MAGIC population is excellent for GWAS

Existence of various structures in a population can create spurious associations between markers and traits in a GWAS if the structure issue is not properly understood [41]. By combining high genetic diversity and low population structure, a MAGIC population creates positive characteristics for QTL study through association mapping [42]. By completing five random matings and six self-pollinations, we effectively reduced the population structure yet maintained relatively high allelic diversity within the MAGIC population. Both Kinship and

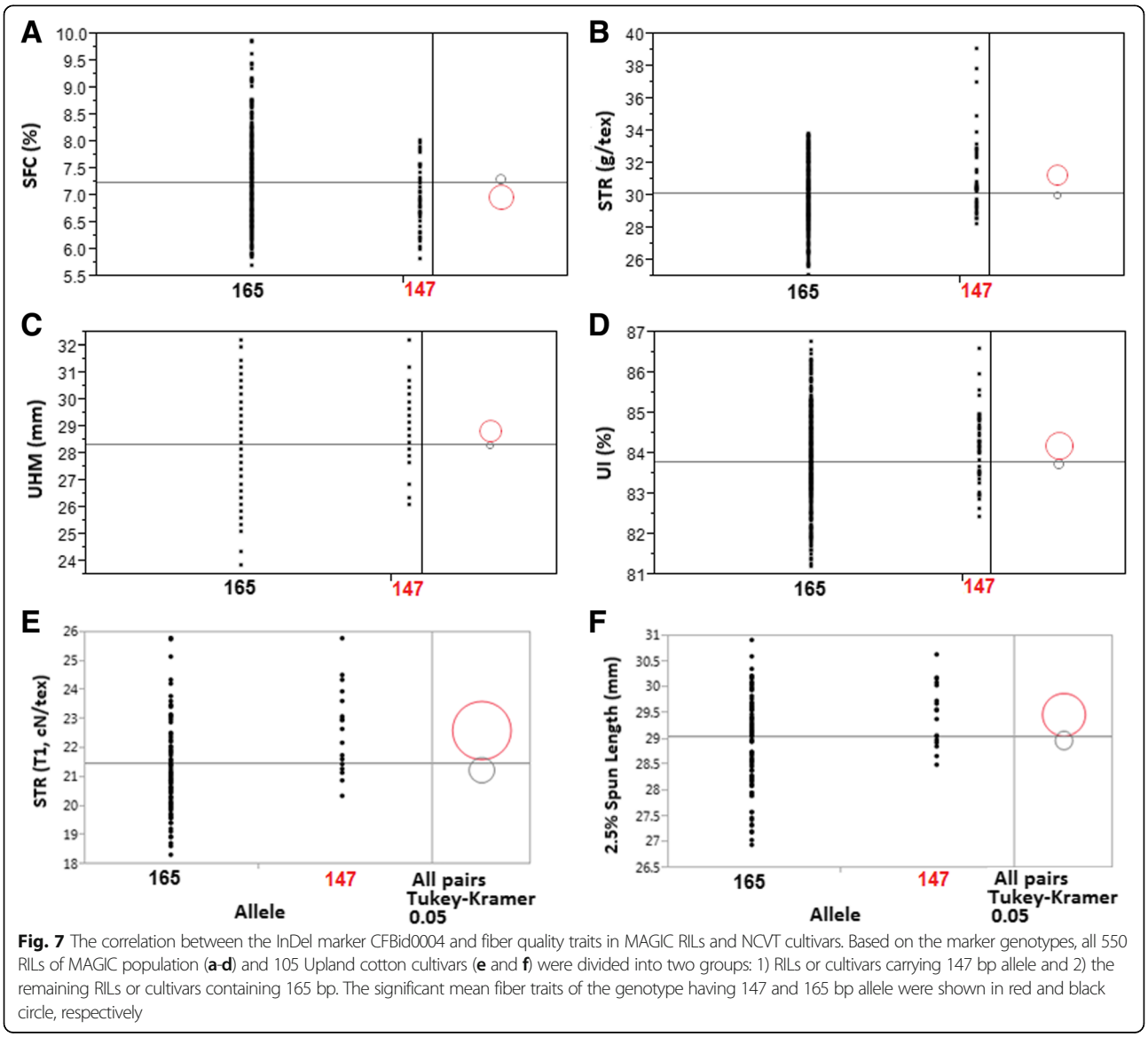

Fig. 7 The correlation between the InDel marker CFBid0004 and fiber quality traits in MAGIC RILs and NCVT cultivars. Based on the marker genotypes, all 550 RILs of MAGIC population (**a**-**d**) and 105 Upland cotton cultivars (**e** and **f**) were divided into two groups: 1) RILs or cultivars carrying 147 bp allele and 2) the remaining RILs or cultivars containing 165 bp. The significant mean fiber traits of the genotype having 147 and 165 bp allele were shown in red and black circle, respectively

relationship matrix revealed that there is essentially no particular population structure in the MAGIC population (Additional files 3 and 4). We previously reported similar result based on 1582 SSR markers [4]. Our results are also comparable with the genetic studies using MAGIC populations developed in other crops such as maize [43], rice [28] and wheat [44]. Thus, we believe that our Upland cotton MAGIC population is an excellent resource for GWAS as well as other genetic studies.

Linkage disequilibrium in the Upland cotton MAGIC population

One of the major factors for GWAS analysis is the LD between loci. The minimum number of markers that are required to conduct a successful association analysis depends on the extent of LD over physical distance in a given population. Many factors affect LD such as recombination and mutation rate, relatedness (kinship), gene conversion, selection (natural, artificial, and balancing), mating systems (self or cross pollination), genetic diversity, and population structure [29, 45]. In this study, the average LD decay was approximately 520 and 480 Kb in A_t and D_t sub-genome, respectively, and genome wide ~500 Kb (Fig. 2). The cotton genome is about 2.5 Gb. Thus, for this cotton MAGIC population, we expect that a minimum number of markers required for GWAS are about 5000, and we used 6326 SNP and SSR loci. This is the first study to use more than 6000 marker loci for a GWAS research in cotton to the best of our knowledge. There have been some reports on LD measurements in Upland cotton, however none of those studies were based on physical distance (bp) but instead relied on genetic distance (cM). Thus, it is hard to compare prior results with this study. If we convert genetic distance to physical distance at the rate of 1 cM = 681 Kb as reported by Hulse-Kemp et al. [46], then LD decay in this study is faster than the 5–6 cM [23], or 3 to 4 cM [30] previously reported. This faster LD decay might be due to the use of MAGIC population in this study. The five random-matings may have effectively broken possible linkages and consequently reduced LD. In this study, the LD blocks are more or less uniform across the genome except on a few chromosomes (Additional file 8). This uniformity may also be due to the lack of selection during population development, since human selection inflates LD [43].

GWAS of fiber quality traits

GWAS is a forward genetics approach which has been used to identify underlying causal genes, mutations and putative functional markers that affect complex quantitative traits [11, 14]. In this study, we identified 86 fiber QTL at $p \leq 0.01$. This number is lower than 131 we reported earlier using phenotypic data from two environments of 275 RILs analyzed by 1582 SSR markers [4]. In this study, we used 6326 marker loci with known physical locations, 547 RILs, and phenotypic data from 4 environments. Thus, the mapping accuracy is expected to be higher. After comparing the 86 QTL with the previously-identified 131 QTL, a great majority are congruent in genomic locations. The region encompassing a major fiber QTL and *GhRBB1_A07* gene was identified in both reports with very high confidence.

The 86 QTL detected in this study were distributed across 24 chromosomes (Table 4, Additional file 12). Chromosomes A02 and D05 did not have any QTL. With agreement with previous reports [47, 48], we identified 26.32 % more fiber QTL on D_t sub-genome than A_t sub-genome. Similarly, our previous study also reported 21 % more fiber QTL on D_t sub-genome when compared with A_t sub-genome using 275 RILs of the same MAGIC population [4]. Once again this emphasizes that the D_t sub-genome has greater impact in determining fiber quality in Upland cotton than the A_t sub-genome. QTL clusters are more interesting to cotton breeders since the markers flanking these regions can be used to select more than one trait through MAS. It is very common to identify QTL clusters in cotton [4, 47–49]. In the present study, we identified 16 QTL clusters ranging from two to five QTL in a cluster (Additional file 12).

The QTL cluster on chromosome A07 for STR, SFC, UHM and UI appeared to be particularly interesting and valuable for cotton breeding. From the allelic effect results, it is clear that this QTL cluster has a positive effect on fiber quality by increasing STR, UHM, and UI value while decreasing SFC value (Additional file 12). This QTL cluster region is co-localized with the QTL cluster identified on chromosome 07 in our previous report [4] and in other reports [7, 50, 51]. Cao et al. [7] suggested that the QTL for fiber length and strength might come from the introgression from *G. barbadense*. Because of the large effects of this QTL on multiple traits, it may serve as an excellent candidate for MAS to improve fiber quality in breeding. In order to prove the effectiveness of these loci in MAS, we examined the allelic effects of the flanking markers on the fiber phenotypes in the MAGIC RILs. The four traits (SFC, STR, UHM and UI) could be simultaneously improved by selecting the minor alleles of the two flanking SNPs *CFB7267* and *CFB7300* in the MAGIC population.

Identification of candidate gene for superior fiber quality

We followed a novel strategy to predict candidate genes for superior fiber quality in Upland cotton by integrating several approaches along with GWAS results. At first, the list of genes in the genomic region on chromosome A07 (71 to 77 Mb) encompassing a major fiber QTL and their annotation information were extracted from the allotetraploid cotton TM-1 draft genome. The ability to identify candidate genes in MAGIC population QTL

mapping was further improved by incorporating WGS and gene expression data of the founder lines. Moreover, our previous RNA-seq data from seven Upland cotton germplasm were used to identify the expressed fiber-related genes. Finally, RT-qPCR gene expression results confirmed expression of the candidate genes in developing fibers. Out of 18 possible candidate genes, six were differentially expressed in the superior fiber quality parent line (AU) as compared to an inferior parent line (TP) fiber tissue (Additional file 16).

By utilizing WGS data from the 11 founder lines, we were able to detect non-sense and mis-sense amino acid mutations as well as other kind of mutations among the founder lines. Of all the detected mutations, the 18 bp deletion at genomic location 76,911,659 bp on chromosome A07 in Acala Ultima was particularly interesting. This deletion is in the exon of the gene *Gh_A07G2049* (*Regeneration of bulb biogenesis 1, GhRBB1_A07*). Interestingly, *RBB1* corresponds to a very large protein of unknown function that is specific to plants, is present in the cytosol, and may associate with cellular membranes. This gene is involved in the regulation of vacuole morphology and may be involved in the establishment or stability of trans-vacuolar strands (TVS) and bulbs. TVS form predominantly along the root hair tip growing axis and are thought to deliver cytosolic components to the growing tip [52]. In this research, RT-qPCR results revealed that transcripts of *GhRBB1_A07* gene were highly abundant in developing fibers (8, 14 and 18 DPA) of the superior founder line AU as compared to the inferior founder line TP. Interestingly, only AU has the deletion, thus we hypothesize that the *GhRBB1_A07* allele in AU may transport higher amount of cytosolic liquid to developing fiber cells by increasing TVS in vacuoles which may lead to superior fiber quality. We also took advantage of the deletion in *Gh_A07G2049* by designing an InDel marker (*CFBid0004*) to validate its association with fiber quality by genotyping MAGIC RILs and NCVT cultivars. Results showed that RILs having the 147 bp allele had significantly more favorable SFC, STR, UHM, and UI than RILs with the 165 bp allele (Fig. 7). Furthermore, NCVT historic data independently confirmed the finding that the deletion allele has an impact on fiber quality in diverse Upland cotton germplasm. We believe that the *GhRBB1_A07* gene is involved in cotton fiber development and the InDel marker *CFBid0004* can be used in MAS to improve fiber quality in a cotton breeding program.

Conclusions

Use of a MAGIC population with little structure coupled with a high density marker coverage obtained through GBS enabled us to more precisely describe LD over the whole genome and to conduct GWAS at high resolution in Upland cotton. By employing an appropriate statistical model, GWAS identified markers significantly associated with fiber QTL. We further confirmed one major QTL cluster associated with four fiber quality traits (SFC, STR, UHM and UI) on chromosome A07. The availability of the reference Upland cotton genome and gene annotation also facilitated the identification of candidate genes, leading to the development of a functional marker for MAS. We were able to identify candidate genes by integrating WGS, and gene expression data of the founder lines. Finally, RT-qPCR gene expression results confirmed fiber expression of candidate genes for superior fiber quality in Upland cotton. Gene expression and amino acid mutation analysis suggested the *GhRBB1_A07* gene is likely one of the candidates for superior fiber quality in Upland cotton. The InDel marker *CFBid0004* has been proven to be associated with improved fiber quality in the MAGIC RILs and NCVT cultivars. The identified QTL can potentially be used to breed cotton cultivars with higher fiber quality through MAS or genomic selection. Further work will investigate the specific role of *GhRBB1_A07* gene in cotton fiber development.

Methods

Upland cotton MAGIC population

A set of 10 cultivars and one breeding line (Additional file 1) from major breeding programs across the United States were used as parents of the MAGIC population. The breeding scheme of the MAGIC population used in this study is shown in Fig. 1. The details of the population development were previously described by [4, 9, 27]. In brief, a half-diallel crossing scheme between 11 parents were followed to produce 55 F_1 in 2002. The 55 F_1 were considered as 55 half-sib families and designated as Cycle 0 (C_0). Five cycles (C_1 to C_5) of random mating were made by bulking the equal amount of pollen from each of the 55 families. After that, self-pollination was followed for six generations using single seed descent method to produce RILs. Ten RILs were randomly selected from each of the 55 families and used in this study (Additional file 2).

Field experiments, fiber quality measurement and phenotypic data analysis

Five hundred fifty MAGIC RILs and 11 parents were planted in Starkville (2009, 2010, and 2011) and Stoneville (2013), Mississippi, USA. Entries were laid out in single row plots 12 m long with about 120 plants per plot with two replications per RIL at each location-year. Standard field practices were applied over the plant growing seasons across years and locations. Twenty-five naturally opened bolls were harvested manually from the central part of a plant from each RIL and parent in all the location-years. Cotton bolls were ginned by using a 10-saw laboratory gin. Fiber quality attributes were measured using a High Volume Instrument (HVI, USTER technologies Inc., Charlotte, NC) for the following traits: ELO (%), MIC, SFC (%), STR (g/tex),

UHM (mm), and UI (%). Variance components were estimated using the following statistical model.

$$y_{ijk} = \mu + G_i + E_j + GE_{ij} + R_{j(k)} + \varepsilon_{ijk} \quad (1)$$

where y_{ijk} is an observed value, μ the overall mean, G_i the effect of the inbred line i, E_j the effect of location year j, GE_{ij} the interaction between inbred line i with the location year j, $R_{j(k)}$ the effect of replication k within environment j, and ε_{ijk} the residual. In this study, all effects except μ were considered as random to estimate variance components. Lines adjusted means were obtained from BLUP considering overall mean. Computations were performed by using PROC MIXED in SAS (SAS Institute, Cary, NC, USA). Simple Pearson correlation coefficients (r) were calculated among all traits based on the adjusted means of the 550 RILs and parents separately. Analyses of variance of fiber quality traits was conducted using SAS package as well.

DNA extraction and genotyping

Seeds from the RILs and parents were sown in small pots in a greenhouse in 2013 in New Orleans, Louisiana, USA. Young leaves were collected from ten plants of each of the entries and bulked. Leaves were stored at −80 °C. Total DNA was extracted from the frozen leaves following a protocol described earlier [27]. DNA quantity and quality was measured using a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA) as well as on a 1.5 % agarose gel.

Genotyping was conducted using both SSR and GBS-based SNP markers. A total of 223 SSR markers were selected based on our previous report [4] and used for genotyping all the RILs along with parents. SSR genotyping method was described earlier [4]. All the SSR marker sequences were BLASTed against Upland cotton TM-1 draft genome [37] in order to obtain their physical locations.

For GBS, DNA of RILs along with parents were sent to Institute for Genomic Diversity (IGD, Cornell University, Ithaca, NY) for library preparation, sequencing, and subsequent bioinformatics. The detail protocol for library preparation and sequencing was described earlier [53]. The raw reads alignment and SNP calling was conducted using reference genome sequence of *G. arboreum* [54] for sub-genome A and *G. raimondii* [55] for sub-genome D. The SNP sequences were filtered according to the criteria: missing rate ≤ 20 %, allele number = 2, MAF ≥ 5 %, number of genotypes ≥ 2. Polymorphisms and MAF among the parents and RILs were checked separately for each of the filtered SNPs. The SNPs with > 20 % MAF difference between the means of parents and RILs were discarded. Finally all filtered SNP sequences were aligned against allotetraploid Upland cotton TM-1 draft genome [37]. The SNP nomenclature was created in our laboratory starting with CFB (cotton fiber bioscience), followed by a serial number.

Linkage disequilibrium (LD) determination

The genome-wide LD between pairs of loci was performed by using the software TASSEL 5.0 [38]. The estimates of the LD were measured using the squared allele-frequency correlations (r^2) for pairs of loci. The distance in base pairs that loci could be expected to be in LD or LD decay was computed by plotting r^2 onto physical distance using the threshold $r^2 = 0.2$ as a cutoff. All markers with less than 20 % missing data and a minor allele frequency ≥ 5 % were used to measure LD decay. After getting r^2 values, data were summarized using R statistical software for each of the chromosomes individually as well as combining all chromosomes to test a sub-genome wide LD decay. LD contour plot were generated from JMP genomics 6.0.

Genome-wide association study

All marker-trait associations were performed using TASSEL 5.0 [38] and GAPIT (Genome Association and Prediction Integrated Tool-R package) [40]. All polymorphic markers that met the filtering criteria were used for GWAS analysis. A PCA was conducted to assess population structure, and a kinship (K) matrix was calculated using the VanRaden method and the EMMA method to determine the familial relatedness between lines. At first, the GLM including the PCA were tested for analyzing GWAS using TASSEL 5.0 [38]. An MLM [39] was also used for performing GWAS by incorporating K matrix along with PCA employing the program GAPIT and TASSEL 5.0. Q-Q plots were created to evaluate how much a significant result was produced by the analysis than expected by chance. A significant marker-trait association was declared when p value was equal to or smaller than 0.01. We also categorized the significant level as follows: significant $p \leq 0.01$, highly significant $p \leq 0.001$ and extremely significant $p \leq 0.0001$. When multiple loci were associated with a trait and were within a 5 Mb interval, they were considered as a single QTL. The QTL nomenclature was according to [56].

RNA extraction and gene expression analysis by reverse transcription quantitative PCR (RT-qPCR)

The parental lines were grown in Stoneville, MS in 2015 for RNA extraction. Total RNA was extracted from the developing cotton fibers (8, 14, 18 and 22 DPA) using the Sigma Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, St. Louis, MO) with DNaseI digestion according to the manufacturer's protocol. The quality and quantity of total RNA was determined using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA) and an Agilent Bioanalyzer 2100 (Agilent Technologies Inc., Santa Clara, CA, USA). The experimental procedures and

data analysis related to RT-qPCR were performed according to the minimum information for publication of quantitative real-time PCR experiments guidelines [57]. The detail descriptions of cDNA preparation, RT-qPCR and calculation were previously reported [31]. Three biological replications and two technical replications for each time-point were used for RT-qPCR. Primer sequences for tested genes are included in Additional file 20.

DNA and RNA sequencing and analysis

The 11 parental lines were sequenced at 20 × coverage with 101-bp paired end using Illumina HiSeq 2000. Sequence reads were aligned to the draft TM-1 genome using GSNAP software [37, 58]. DNA sequence polymorphisms in the vicinity of the major QTL cluster on Chr. A07 were identified with samtools and bcftools software and by manual inspection of alignment files with IGV software [59, 60]. Variants were analyzed for non-synonymous substitutions to annotated proteins as before [32, 61]. RNA-seq data previously published by our group was consolidated to identify fiber-expressed genes in the vicinity of the QTL cluster [31, 62, 63].

Analysis of the InDel marker in 105 cotton cultivars

The InDel marker (*CFBid0004*) was designed from the sequence of the gene *Gh_A07G2049* using NCBI Primer Blast tools (Additional file 18). Then this marker was used to genotype all the MAGIC RILs and 105 Upland cotton cultivars. These 105 cultivars (Additional file 19) were selected because they were tested in multiple locations and multiple years through the NCVT program, and fiber quality data are available at http://www.ars.usda.gov/main/docs.htm?docid=23813 [64]. The InDel marker genotyping method was according to [63]. Phenotypic data for two alleles of 165 and 147 bp of the locus are grouped for the one-way analyses. For comparison of the means of the grouping performance, Tukey's honest significant difference test was conducted for all pairs of levels.

Additional files

Additional file 1: Title: Eleven Upland cotton cultivars that were used for MAGIC population development. Description of data: This table contains names of 11 founder lines used to develop the MAGIC population. This table was taken from DD Fang, JN Jenkins, DD Deng, JC McCarty, P Li and J Wu [4]. (DOCX 13 kb)

Additional file 2: Title: MAGIC population RILs pedigree and their corresponding family information. Description of data: Five hundred fifty RIL names, their pedigree and family information were included in this table. (XLSX 78 kb)

Additional file 3: Title: A heat map showing the relationships between RILs. Marker data were used to measure the relationships among 547 RILs of the MAGIC population. The red diagonal represents perfect relationship of each RIL and the symmetric off-diagonal elements represent relationship measured [in this case identity by descent (IBD)] for pairs of lines. No cluster of related lines was found as no block of diagonal warmer color appeared. The dendrogram on the right shows the results of a cluster analysis on the IBD matrix. Description of data: The relationship between the RILs of MAGIC population derived from JMPGenomics 6.0 is included in this figure. (DOCX 111 kb)

Additional file 4: Title: Kinship matrix among the 547 RILs of Upland cotton MAGIC population using GBS based SNP and SSR marker. Description of data: This figure contains the information of Kinship relationship among the tested RILs of MAGIC population. (DOCX 400 kb)

Additional file 5: Title: Analyses of variance for fiber quality trait data from recombinant inbred lines (RILs) of the Upland cotton MAGIC population. Description of data: ANOVA analysis of fiber quality trait data in Starkville, MS. (DOCX 13 kb)

Additional file 6: Title: SNP and SSR markers and their physical location on TM-1 draft genome. Description of data: This figure contains 6039 GBS based SNP and 223 SSR markers name, associated allele and physical location on allotetraploid cotton TM-1 draft genome. (XLSX 343 kb)

Additional file 7: Title: Polymorphic SNP and SSR marker distribution across the TM-1 genome. The length of X axis for all chromosomes is based on the highest physical length chromosome for the respective sub-genome. Description of data: GBS based SNP and SSR markers distributions per 1 Mb across TM-1 draft genome are shown in this figure. (DOCX 49 kb)

Additional file 8: Title: Linkage disequilibrium (LD) contour plot by chromosome generated from JMP genomics 6.0. Description of data: The LD contour plots for all the 26 Upland cotton chromosomes are included in this figure. The LD contour plot were generated from the square of correlation coefficients (r^2) between markers located on each chromosome at different physical distances (bp) using JMP genomics 6.0 software. The X and Y axis have the physical distance of two markers. The square of correlation coefficients (r^2) presents as color code (blue to red – 1.000 to 0.000). (DOCX 1734 kb)

Additional file 9: Title: Quantile-quantile (Q-Q) plot of six fiber traits generated from GWAS analysis following general linear model (GLM) using TASSEL 5.0 software. A) Fiber elongation (ELO), B) Micronaire (MIC), C) Short fiber content (SFC), D) Fiber strength (STR), E) Upper half mean fiber length (UHM), and F) Uniformity index (UI). Description of data: Q-Q plots of six fiber traits generated from GWAS analysis following GLM are included in this figure. The X and Y axis have the expected and observed negative logarithm 10 of *p* value, respectively generated during GWAS analysis. (DOCX 538 kb)

Additional file 10: Title: Quantile-quantile (Q-Q) Plot of six fiber traits generated from GWAS analysis following mixed linear model (MLM) using GAPIT software. A) Fiber elongation (ELO), B) Micronaire (MIC), C) Short fiber content (SFC), D) Fiber strength (STR), E) Upper half mean fiber length (UHM), and F) Uniformity index (UI). Description of data: Q-Q plots of six fiber traits generated from GWAS analysis following MLM are included in this figure. The X and Y axis have the expected and observed negative logarithm 10 of *p* value, respectively generated during GWAS analysis. (DOCX 207 kb)

Additional file 11: Title: Manhattan plots generated from TASSEL 5.0 software for six fiber quality traits, A) Elongation (ELO), B) Micronaire (MIC), C) Short fiber content (SFC), D) Fiber strength (STR), E) Upper half mean (UHM) fiber length, and F) Uniformity (UI). The negative $\log_{10}$ transformed *p* values were plotted against the marker positions on the physical map of each of the 26 Upland cotton chromosome. The significant thresholds ($p = 0.01$ and 0.0001) are indicated by the purple and green horizontal dot line, respectively. Description of data: Manhattan plots generated from TASSEL 5.0 software for six fiber quality traits from GWAS analysis following MLM are included in this figure. The X and Y axis have chromosome name and observed negative logarithm 10 of *p* value, respectively. (DOCX 813 kb)

Additional file 12: Title: Significant associations between markers and fiber quality traits. Description of data: This table contains the information of significant QTL at $p \leq 0.01$ associated with six fiber quality traits detected from GWAS analysis following MLM. The first column have QTL name for the respective fiber traits (first letter 'q' for QTL then trait name and finally chromosome number). The second column contains marker name followed by chromosome number, physical location (bp), *p* vale, minor allele frequency (MAF), R^2 of model without SNP, R^2 of model with SNP, and estimated allelic effect. (XLSX 38 kb)

Additional file 13: Title: Effect of haplotype grouping near the major QTL on chromosome A07 on fiber bundle strength (g/tex). RILs were

divided into two groups (major and minor types) based on genotypes major and minor alleles of respective two SNPs. Description of data: This box plot shows the effect of major and minor allele combination of haplotype group near the QTL on chromosome A07 on fiber strength. The X axis has the fiber strength value (g/tex) and Y axis contain possible major and minor type haplotype groups and their physical locations. (DOCX 101 kb)

Additional file 14: Title: LD contour plot in the Upland cotton genomic region of 71 to 77 Mb on chromosome A07. LD contour was created from genotypic data of 547 RILs of Upland cotton MAGIC population using JMP genomics 6.0 software. X axis is physical distance in Mb and r^2 (CorrCoeff2) between marker pair is shown in different color block as per legend. Description of data: LD contour plot of genomic region of 71 to 77 Mb on chromosome A07 is included in this figure. Red color represents the higher r^2 value between two markers due to LD block. The square of correlation coefficients (r^2) presents as color code (red to blue – 1.000 to 0.000). (DOCX 69 kb)

Additional file 15: Title: Annotated genes within 71 M bp and 76 M bp region of chromosome A07 and their RNA-seq data. Description of data: This table contains gene name and associated annotation near the detected loci on chromosome A07. The RNA-seq data from our previous studies for the gene are also included in this table. First column have the gene name followed by physical location, Arabidopsis ortholog gene, gene description and RPKM value of RNA-seq data for the different RNA libraries. (XLSX 75 kb)

Additional file 16: Title: RT-qPCR gene expression analyses of selected genes related to cell wall activity during fiber development between Acala Ultima (AU) and Tamcot Pyramid (TP). A) CONSTANS-like 9 (*Gh_A07G1753*); B) RAB GTPase homolog A5E (*Gh_A07G1758*); C) Unknown protein family, DUF538 (*Gh_A07G1784*); D) Auxin-responsive protein (*Gh_A07G1795*); E) Oxidoreductase, zinc-binding dehydrogenase family protein (*Gh_A07G1803*); F) O-Glycosyl hydrolases family protein (*Gh_A07G1802*). Description of data: This file contains the results of RT-qPCR analysis data performed on six genes that were used to compare relative expression levels between superior line AU and inferior line TP. Four (8, 14, 28 and 22 DPA) developing fiber samples were used for RT-qPCR analysis. RT-qPCR values were corrected to 18S gene for each sample. For each treatment group two qPCR measurements were taken for each of three biological replicates and then averaged. (DOCX 161 kb)

Additional file 17: Title: Sashimi plot of parental lines showing 18 bp of deletion on parent Acala Ultima (AU) at genomic region 76,911,659 to 76,911,676 bp on chromosome A07. Description of data: Sashimi plot of parental lines generated from Integrated Genome Viewer (IGV) software is included in this file. The parent AU (row number 1) has an 18 bp deletion at genomic region 76,911,659 to 76,911,676 bp on exon 29 of gene *Gh_A07G2049* on chromosome A07 while other parental lines don't have that deletion. (DOCX 50 kb)

Additional file 18: Title: Alignment of Acala Ultima (AU) and TM-1 (Ref) alleles of *Gh_A07G2049*. Primer locations for CFBid0004 are highlighted yellow. Description of data: This file contains alignment of parental line AU (row number 1) and TM-1 (row number 2) alleles of gene *Gh_A07G2049*. The 18 bp deletion in the coding sequence in AU at 17,875 to 17,893 bp on gene has also been presented. The forward and reverse primer sequences for indel marker CFBid0004 has been highlighted as yellow at either side of the 18 bp deletion. (PDF 4439 kb)

Additional file 19: Title: The 105 Upland cotton cultivars used for finding the efficacy of deletion marker allele CFBid0004 in this study. Description of data: This table contains 105 NCVT cultivars name. The historic phenotypic data of those 105 cultivars were used to find out the practical efficacy of the InDel marker CFBid0004 in Upland cotton breeding. (XLSX 11 kb)

Additional file 20: Title: Sub-genome specific primer names and their sequences used in this study for RT-qPCR gene expression. Description of data: This file contains 19 RT-qPCR primer pair sequences that were used to analyze gene expression near the major QTL on chromosome A07. Corresponding gene name and descriptions are also included here. (XLSX 13 kb)

Abbreviations

AU: Acala Ultima; BLUP: Best linear unbiased predictor; DPA: Days post anthesis; ELO: Elongation; GBS: Genotyping-by-sequencing; GLM: General linear model; GWAS: Genome wide association study; HVI: High volume instrument; LD: Linkage disequilibrium; MAF: Minor allele frequency; MAGIC: Multi parent advanced generation inter-cross; MAS: Marker assisted selection; MIC: Micronaire; MLM: Mixed linear model; NCVT: National Cotton Variety Test; QTL: Quantitative trait locus (Loci); RBB1: Regeneration of bulb biogenesis 1; RIL: Recombinant inbreed line; RPKM: Reads per kilobase of transcript per million mapped reads; RT-qPCR: Reverse transcription quantitative polymerase chain reaction; SFC: Short fiber content; STR: Fiber strength; TP: Tamcot pyramid; TVS: Trans-vacuolar strands; UHM: Upper-half mean fiber length; UI: Uniformity index; WGS: Whole genome sequence

Acknowledgements

We thank Dr. David Stelly at Texas A&M University to have provided us the whole genome sequences of several parent lines. Our appreciation also goes to Dr. Russell Hayes for assisting with the field experiments. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U. S. Department of Agriculture which is an equal opportunity provider and employer.

Funding

This research was funded by United States Department of Agriculture-Agricultural Research Service CRIS project 6054-21000-017-00D and Cotton Incorporated projects 10–747 and 15–751.

Availability of data and materials

All relevant data reported in this paper are within the paper text and the 20 supplementary files included within this paper.

Authors' contributions

DDF conceived the research and revised the manuscript. MSI coordinated experiments; analyzed phenotypic and GBS marker data; performed GWAS and RT-qPCR, and wrote the manuscript. GNT analyzed WGS and RNAseq data, and helped identify mutations in genes. JNJ and JCM developed the MAGIC population, and conducted field trials in Starkville, MS. LZ conducted field trials in Stoneville, MS. CDD was responsible for fiber property measurement. DDD provided NCVT cultivar phenotypic data. DJH grew the parent lines and isolated RNAs. DCJ edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Cotton Fiber Bioscience Research Unit, USDA-ARS, Southern Regional Research Center, New Orleans, LA 70124, USA. [2]Cotton Chemistry and Utilization Research Unit, USDA-ARS, Southern Regional Research Center, New Orleans, LA 70124, USA. [3]Genetics & Sustainable Agriculture Research Unit, USDA-ARS, Mississippi State, MS 39762, USA. [4]Crop Genetics Research Unit, USDA-ARS, Stoneville, MS 38772, USA. [5]Cotton Structure and Quality Research Unit, USDA-ARS, Southern Regional Research Center, New Orleans, LA 70124, USA. [6]Cotton Incorporated, Cary, NC 27513, USA.

Received: 8 June 2016 Accepted: 2 November 2016
Published online: 09 November 2016

References

1. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD, et al. Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol. 2007;145:1303–10.
2. Zhang T, Qian N, Zhu X, Chen H, Wang S, Mei H, Zhang Y. Variations and transmission of QTL alleles for yield and fiber qualities in upland cotton cultivars developed in China. PLoS One. 2013;8:e57220.
3. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. In: Sparks DL, editor. Advances in Agronomy, vol. 78. San Diego: Academic; 2003. p. 139–85.

4. Fang DD, Jenkins JN, Deng DD, McCarty JC, Li P, Wu J. Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). BMC Genomics. 2014;15:397.
5. Yu J, Zhang K, Li S, Yu S, Zhai H, Wu M, Li X, Fan S, Song M, Yang D, et al. Mapping quantitative trait loci for lint yield and fiber quality across environments in a *Gossypium hirsutum* x *Gossypium barbadense* backcross inbred line population. Theor Appl Genet. 2013;126:275–87.
6. Islam MS, Zeng L, Delhom CD, Song X, Kim HJ, Li P, Fang DD. Identification of cotton fiber quality quantitative trait loci using intraspecific crosses derived from two near-isogenic lines differing in fiber bundle strength. Mol Breed. 2014;34:373–84.
7. Cao Z, Zhu X, Chen H, Zhang Z. Fine mapping of clustered quantitative trait loci for fiber quality on chromosome 7 using a *Gossypium barbadense* introgressed line. Mol Breed. 2015;35:215–28.
8. Cavanagh C, Morell M, Mackay I, Powell W. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol. 2008;11:215–21.
9. Jenkins JN, McCarty JC, Gutierrez OA, Hayes RW, Bowman DT, Watson CE, Jones DC. Registration of RMUP-C5, a random mated population of Upland cotton germplasm. J Plant Reg. 2008;2:239–42.
10. Yu J, Buckler ES. Genetic association mapping and genome organization of maize. Curr Opin Biotechnol. 2006;17:155–60.
11. Vuong TD, Sonah H, Meinhardt CG, Deshmukh R, Kadam S, Nelson RL, Shannon JG, Nguyen HT. Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC Genomics. 2015;16:593.
12. Abdurakhmonov IY, Abdukarimov A. Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics. 2008;2008:574927.
13. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature. 2010;465:627–31.
14. Pace J, Gardner C, Romay C, Ganapathysubramanian B, Lubberstedt T. Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genomics. 2015;16:47.
15. Matthies IE, Malosetti M, Röder MS, van Eeuwijk F. Genome-wide association mapping for kernel and malting quality traits using historical European barley records. PLoS One. 2014;9:e110046.
16. Visioni A, Tondelli A, Francia E, Pswarayi A, Malosetti M, Russell J, Thomas W, Waugh R, Pecchioni N, Romagosa I, Comadran J. Genome-wide association mapping of frost tolerance in barley (*Hordeum vulgare* L.). BMC Genomics. 2013;14:424.
17. Tadesse W, Ogbonnaya FC, Jighly A, Sanchez-Garcia M, Sohail Q, Rajaram S, Baum M. Genome-wide association mapping of yield and grain quality traits in winter wheat genotypes. PLoS One. 2015;10:e0141339.
18. Zegeye H, Rasheed A, Makdis F, Badebo A, Ogbonnaya FC. Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. PLoS One. 2014;9:e105593.
19. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42:961–7.
20. Newell MA, Cook D, Tinker NA, Jannink JL. Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theor Appl Genet. 2011;122:623–32.
21. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA. 2013;110:453–8.
22. Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, Jiang GL. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics. 2015;16:217.
23. Abdurakhmonov IY, Saha S, Jenkins JN, Buriev ZT, Shermatov SE, Scheffler BE, Pepper AE, Yu JZ, Kohel RJ, Abdukarimov A. Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. Genetica. 2009;136:401–17.
24. Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, et al. Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. Genomics. 2008;92:478–87.
25. Wang YY, Zhou ZL, Wang XX, Cai XY, Li XN, Wang CY, Wang YH, Fang L, Wang KB. Genome-wide association mapping of glyphosate-resistance in *Gossypium hirsutum* races. Euphytica. 2016;209:209–21.
26. Nie X, Huang C, You C, Li W, Zhao W, Shen C, Zhang B, Wang H, Yan Z, Dai B, et al. Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. BMC Genomics. 2016;17:352.
27. Islam MS, Thyssen GN, Jenkins JN, Fang DD. Detection, validation, and application of genotyping-by-sequencing based single nucleotide polymorphisms in Upland cotton. The Plant Genome. 2015;8:1–10.
28. Bandillo N, Raghavan C, Muyco PA, Sevilla MA, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R, et al. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice (NY). 2013;6:11.
29. Bastien M, Sonah H, Belzile F. Genome wide association mapping of resistance in soybean with a genotyping-by-sequencing approach. The Plant Genome. 2014;7:1–13.
30. Fang DD, Hinze LL, Percy RG, Li P, Deng D, Thyssen G. A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in Upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. Euphytica. 2013;191:391–401.
31. Islam MS, Fang DD, Thyssen GN, Delhom CD, Liu Y, Kim HJ. Comparative fiber property and transcriptome analyses reveal key genes potentially related to high fiber strength in cotton (*Gossypium hirsutum* L.) line MD52ne. BMC Plant Biol. 2016;16:36.
32. Islam MS, Zeng L, Thyssen GN, Delhom CD, Kim HJ, Li P, Fang DD. Mapping by sequencing in cotton (*Gossypium hirsutum*) line MD52ne identified candidate genes for fiber strength and its related quality attributes. Theor Appl Genet. 2016;129:1071–86.
33. Fang L, Tian R, Chen J, Wang S, Li X, Wang P, Zhang T. Transcriptomic analysis of fiber strength in upland cotton chromosome introgression lines carrying different *Gossypium barbadense* chromosomal segments. PLoS One. 2014;9:e94642.
34. Fang L, Tian R, Li X, Chen J, Wang S, Wang P, Zhang T. Cotton fiber elongation network revealed by expression profiling of longer fiber lines introgressed with different *Gossypium barbadense* chromosome segments. BMC Genomics. 2014;15:838.
35. Kim HJ, Hinchliffe DJ, Triplett BA, Chen ZJ, Stelly DM, Yeater KM, Moon HS, Gilbert MK, Thyssen GN, Turley RB, Fang DD. Phytohormonal networks promote differentiation of fiber initials on pre-anthesis cotton ovules grown in vitro and in planta. PLoS One. 2015;10:e0125046.
36. Kim HJ, Tang Y, Moon HS, Delhom CD, Fang DD. Functional analyses of cotton (*Gossypium hirsutum* L.) immature fiber (im) mutant infer that fiber cell wall development is associated with stress responses. BMC Genomics. 2013;14:889.
37. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33:531–7.
38. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
39. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42:355–60.
40. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28:2397–9.
41. Matthies IE, Hintum TV, Weise S, Röder MS. Population structure revealed by different marker types (SSR or DArT) has an impact on the results of genome-wide association mapping in European barley cultivars. Mol Breed. 2012;30:951–66.
42. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10:565–77.
43. Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Hlaing AL, Aung HH, Nelissen H, Baute J, et al. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. Genome Biol. 2015;16:167.
44. Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J, Gosman N, Greenland AJ, Horsnell R, Howells R, O'Sullivan DM, et al. An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3, Genes|Genomes|Genetics. 2014;4:1603–10.
45. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol. 2005;57:461–85.

46. Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD, Frelichowski J, Giband M, Hague S, Hinze LL, et al. Development of a 63K SNP array for cotton and high-density mapping of intra- and inter-specific populations of Gossypium spp. G3, Genes|Genomes|Genetics. 2015;5:1187–209.
47. Said JI, Lin Z, Zhang X, Song M, Zhang J. A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. BMC Genomics. 2013;14:776.
48. Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, et al. Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics. 2007;176:2577–88.
49. Lacape JM, Llewellyn D, Jacobs J, Arioli T, Becker D, Calhoun S, Al-Ghazi Y, Liu S, Palai O, Georges S, et al. Meta-analysis of cotton fiber quality QTLs across diverse environments in a *Gossypium hirsutum* x *G. barbadense* RIL population. BMC Plant Biol. 2010;10:132.
50. Lacape J, Nguyen T, Courtois B, Belot J, Giband M, Gourlot J, Gawryziak G, Roques S, Hau B. QTL analysis of cotton fiber quality using multiple *G. hirsutum* X *G. barbadense* backcross generations. Crop Sci. 2005;45:123–40.
51. Sun FD, Zhang JH, Wang SF, Gong WK, Shi UZ, Liu AY, Li JW, Gong JW HSH, Yuan YL. QTL mapping for fiber quality traits across multiple generations and environments in upland cotton. Mol Breed. 2012;30:569–82.
52. Han SW, Alonso JM, Rojas-Pierce M. Regulator of bulb biogenesis 1 (RBB1) is involved in vacuole bulb formation in Arabidopsis. PLoS One. 2015;10:e0125621.
53. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379.
54. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet. 2014;46:567–72.
55. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 2012;492:423–7.
56. McCouch S, Cho Y, Yano P, Blinstrub M, Morishima H, Kinoshita T. Report on QTL nomenclature. Rice Genet Newsl. 1997;14:11–3.
57. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem. 2009;55:611–22.
58. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.
59. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
60. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
61. Thyssen GN, Fang DD, Turley RB, Florane C, Li P, Naoumkina M. Next generation genetic mapping of the Ligon-lintless-2 (Li_2) locus in upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet. 2014;127:2183–92.
62. Naoumkina M, Thyssen GN, Fang DD. RNA-seq analysis of short fiber mutants Ligon-lintless-1 (Li_1) and - 2 (Li_2) revealed important role of aquaporins in cotton (*Gossypium hirsutum* L.) fiber elongation. BMC Plant Biol. 2015;15:65.
63. Thyssen GN, Fang DD, Zeng L, Song X, Delhom CD, Condon TL, Li P, Kim HJ. The Immature fiber mutant phenotype of cotton (*Gossypium hirsutum*) is linked to a 22- bp frame-shift deletion in a mitochondria targeted pentatricopeptide repeat gene. G3, Genes|Genomes|Genetics. 2016;6:1627–33.
64. Meredith JWR, Boykin DL, Bourland FM, Caldwell WD, Campbell BT, Gannaway JR, Glass K, Jones AP, May LM, Smith CW, Zhang J. Genotype x environment interactions over seven years for yield, yield components, fiber quality, and gossypol traits in the regional high quality tests. J Cotton Sci. 2012;16:160–9.