ORIGINAL ARTICLE

Whole genome sequencing of a MAGIC population identified genomic loci and candidate genes for major fiber quality traits in upland cotton (*Gossypium hirsutum* L.)

Gregory N. Thyssen[1,2] · Johnie N. Jenkins[3] · Jack C. McCarty[3] · Linghe Zeng[4] · B. Todd Campbell[5] · Christopher D. Delhom[6] · Md. Sariful Islam[7] · Ping Li[1] · Don C. Jones[8] · Brian D. Condon[2] · David D. Fang[1]

Received: 29 August 2018 / Accepted: 27 November 2018 / Published online: 1 December 2018

Abstract

Key message **Significant associations between candidate genes and six major cotton fiber quality traits were identified in a MAGIC population using GWAS and whole genome sequencing.**

Abstract Upland cotton (*Gossypium hirsutum* L.) is the world's major renewable source of fibers for textiles. To identify causative genetic variants that influence the major agronomic measures of cotton fiber quality, which are used to set discount or premium prices on each bale of cotton in the USA, we measured six fiber phenotypes from twelve environments, across three locations and 7 years. Our 550 recombinant inbred lines were derived from a multi-parent advanced generation intercross population and were whole-genome-sequenced at 3× coverage, along with the eleven parental cultivars at 20× coverage. The segregation of 473,517 single nucleotide polymorphisms (SNPs) in this population, including 7506 non-synonymous mutations, was combined with phenotypic data to identify seven highly significant fiber quality loci. At these loci, we found fourteen genes with non-synonymous SNPs. Among these loci, some had simple additive effects, while others were only important in a subset of the population. We observed additive effects for elongation and micronaire, when the three most significant loci for each trait were examined. In an informative subset where the major multi-trait locus on chromosome A07:72-Mb was fixed, we unmasked the identity of another significant fiber strength locus in gene *Gh_D13G1792* on chromosome D13. The micronaire phenotype only revealed one highly significant genetic locus at one environmental location, demonstrating a significant genetic by environment component. These loci and candidate causative variant alleles will be useful to cotton breeders for marker-assisted selection with minimal linkage drag and potential biotechnological applications.

Introduction

Genome-wide association studies (GWAS) have proven valuable at identifying genomic locations that affect traits in humans, livestock, and crops (Bouwman et al. 2018; Huang and Han 2014; MacArthur et al. 2016; Welter et al. 2013). Most frequently, these depend on collecting genotypic and phenotypic information from natural populations and must overcome issues of population structure to control for shared regions that do not harbor a causative genetic polymorphism (McCarthy et al. 2008). However, the algorithms developed for GWAS can also be applied to carefully designed multi-parent populations, including nested association mapping (NAM) and multi-parent advanced generation intercross (MAGIC) populations (Buckler et al. 2009; Huang et al. 2015; Rakshit et al. 2012; Yu et al. 2008). While these populations may not capture the full complement of diversity that is present in a natural population, they avoid many of the issues of kinship and population structure (Huang et al. 2015). In species with a narrow genetic base, such as cultivated upland cotton (*Gossypium hirsutum* L.), the selection of sufficiently diverse parents may allow the identification of important haplotypes and causative genes or alleles (Fang et al. 2013; Islam et al. 2016; Percy et al. 2006). The recombinations between haplotypes observed in

Communicated by Yunbi Xu.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-018-3254-8) contains supplementary material, which is available to authorized users.

✉ David D. Fang
david.fang@ars.usda.gov

Extended author information available on the last page of the article

a MAGIC population are likely to be more numerous and more evenly distributed than in natural populations, resulting in significantly lower linkage disequilibrium (LD) and thus finer mapping of quantitative trait loci (QTLs) (Dell'Acqua et al. 2015; Huang et al. 2015, 2011). Ongoing advances in genotyping technology allow ever more precise delineation of haplotype blocks, and whole genome sequencing (WGS) of members of a population is now feasible (Cao et al. 2011; The 1000 Genomes Project Consortium 2010).

Cotton is the major renewable source of fibers for textiles and is an important cash crop worldwide (Paterson et al. 2012). In the USA, every bale of cotton is graded to establish a premium or discounted price. This analysis is carried out with a high volume instrument (HVI) which measures cotton fiber quality characteristics including elongation (ELO), micronaire (MIC), short fiber index (SFI), fiber strength (STR), upper half mean length (UHML), and uniformity index (UI), although ELO and SFI are not frequently used in setting the price. These characteristics are influenced by genotype, environment, management practices, and their interactions (Dabbert et al. 2017; Gore et al. 2014; Paterson et al. 2003; Said et al. 2015). Recently, researchers have used GWAS to identify important loci that affect cotton fiber and agronomic traits, using various DNA marker technologies, including SNP arrays, genotyping-by-sequencing (GBS) and WGS (Fang et al. 2017; Huang et al. 2017; Islam et al. 2016; Li et al. 2018; Ma et al. 2018; Su et al. 2016, 2018; Sun et al. 2017; Yuan et al. 2018).

Here, we present GWAS of a cotton MAGIC population that consists of 550 recombinant inbred lines (RILs) derived from crosses between ten diverse cotton cultivars and one improved, but not cultivated, cotton line (M240). We collected phenotypic data from three locations, spanning 7 years, for a total of twelve environments, or location–years. We identified significant QTLs for the major cotton fiber quality characteristics. We found masking of a secondary STR QTL by a major, multi-trait locus and additive effects for ELO and the highly environment-dependent MIC. Importantly, we were able to directly identify candidate gene variants at these loci using WGS of all the parents and RILs.

Materials and methods

Plant materials

A set of eleven diverse *G. hirsutum* cotton lines (Table S1) from major breeding programs across the USA were used as parents to develop a MAGIC population. The details of the MAGIC population development were previously described (Fang et al. 2014; Islam et al. 2016; Jenkins et al. 2008). Briefly, the eleven parents were crossed in a half-diallel to establish 55 families. These were randomly mated by a bulked pollen approach for five generations (C_5), followed by six generations of single seed descent (S_6) to establish the 550 C_5S_6 RILs in Starkville, MS, USA.

DNA isolation and whole genome sequencing

Five hundred fifty RILs along with their eleven parental lines were grown in a greenhouse in 2013 in New Orleans, LA, USA. Young leaves were collected from ten plants of each RIL or parent and stored at − 80 °C. The genomic DNA was extracted from frozen leaves following the protocol previously described with an additional RNAase A digestion step before binding DNA to the column (Islam et al. 2014). The quality and quantity of DNA were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) as well as on an 1.5% agarose gel. DNAs were sent to Novogene Corporation (Chula Vista, CA, USA) for library preparation, and whole genome sequencing using Illumina HiSeq 2500 with paired-end 150 bp reads. For the eleven parental lines, each was sequenced at 20× coverage (about 50 Gb), and for the 550 RILs, each was sequenced at 3× coverage (about 8 Gb).

Processing of sequence data and variant calling

Sequencing reads of the eleven parents were first aligned to the draft reference *G. hirsutum* cv. Texas Marker-1 genome (Zhang et al. 2015) with GSNAP software, using the "-Q -n 1" option which requires that mapped reads have a single, uniquely best match in the genome (Wu and Nacu 2010). Reads with multiple, equally good matches in the genome were excluded. Variants were identified with samtools mpileup (with "-Ego" flags) and bcftools ("call-vm") software (Danecek et al. 2011; Li et al. 2009). We then required that at least one of the eleven parents must have been scored as homozygous for the reference allele and at least one scored as homozygous for the alternate allele. A SNP was also discarded if three or more parental lines had missing data or were heterozygous.

For the RILs, the resulting variant call format (VCF) files were first filtered for variants that segregated in the RIL population using a custom Python script. The filtering criteria were: reference allele $\geq 2.5\%$, alternative allele $\geq 2.5\%$, missing score $\leq 20\%$, and heterozygous rate $\leq 30\%$. Then, the minor allele frequency (MAF) among the parents and RILs were checked separately for each of the filtered SNPs. The SNPs with significant MAF difference ($p > 0.05$ based on Chi-square test) between the parents and RILs were excluded from the final variant set. Remaining missing data were imputed with the *k*-nearest neighbors (*k*-NN) approach implemented in TASSEL 5.0 software with default parameters (Bradbury et al. 2007). Non-synonymous SNPs were identified as before (Thyssen et al. 2014).

Field experiments and phenotyping

The RILs and eleven parents were planted in Florence, SC, in 2014–2016, Starkville, MS, in 2009–2011, and 2014–2016, and in Stoneville, MS, in 2013–2015. The 550 RILs were divided in two subpopulations of 275, called Set A and Set B. Usually, only one set was grown at each location–year (Supplemental Table S2). Field plot arrangements were according to a randomized complete block design in Starkville in 2009 and according to an alpha lattice at all other location–years. Single row plots were 12 m long with approximately 120 plants per plot and two replicate plots per line at each location–year. Field practices were applied according to the prevailing conventions at each location–year. Twenty-five naturally opened bolls were manually harvested from each line and ginned using a 10-saw laboratory gin. The fiber quality attributes (ELO, MIC, SFI, STR, UHML, and UI) were measured using an HVI (USTER Technologies, Charlotte, NC, USA).

Phenotype variance analysis and normalization

Raw phenotypic data from 2015 across the three locations, and from Starkville, MS, USA across years 2010, 2011, 2015, 2016, were separately subjected to ANOVA using PROC MIXED in SAS software (SAS Institute, Cary, NC, USA). For GWAS of individual location–years, arithmetic means of phenotype values were computed between replicates. Raw phenotypic data were then also normalized across replicates, years, and locations using a best linear unbiased predictor (BLUP) implemented in R software using the lme4 package to fit the model: "model = lmer(phenotype ~ (1|line) + (1|location) + (1|year) + (1|(replicate: location):year) + (1|line: location) + (1|line:year))" (Bates et al. 2014). To investigate the location effect on micronaire, but control for year, we normalized data from each location separately by BLUP to fit the model: "model = lmer(phenotype ~ (1|line) + (1|year) + (1|rep:year) + (1|line:year))."

Association mapping analysis

The compressed mixed linear model (MLM) marker–trait association analysis was implemented with GAPIT software using the select sequencing variants, input parameter "PCA.total = 3," and phenotypic data, which was normalized and subsampled as described above (Lipka et al. 2012; Zhang et al. 2010). GAPIT calculated a kinship matrix according to the VanRaden method and performed GWAS using the default average clustering algorithm and default mean group kinship type (Lipka et al. 2012). We applied GAPIT's default correction for multiple sampling to establish the significance threshold of $-\log(p) > 7$.

RNA expression analysis

To present tissue-specific expression of annotated genes near the candidate loci, we retrieved publically available data from ccNET (http://structuralbiology.cau.edu.cn/gossypium) (You et al. 2016).

Results

Phenotypic variation

We collected phenotypes of six cotton fiber quality attributes: ELO, MIC, SFI, STR, UHML, and UI. We measured these phenotypes for cotton grown at three locations in the southeastern USA, Florence, SC, Starkville, MS, and Stoneville, MS, between 2009 and 2016, for a total of twelve environments (Fig. 1). The mean values of the RILs were similar to the parental lines for all traits; however, the standard deviations, maximum and minimum values were broader, indicating transgressive segregation (Table S3). We observed significant correlation between STR, and UI, SFI and UHML in both the parents and the RILs, while ELO and MIC varied independently of the other traits and each other (Table S4). ANOVA results revealed that variance of

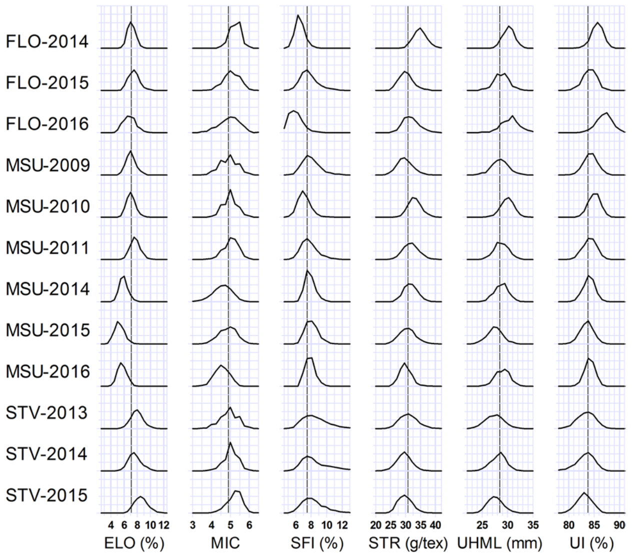


Fig. 1 Phenotypic distribution of six major cotton fiber quality phenotypes across twelve environments. Fiber elongation (ELO), micronaire (MIC), short fiber index (SFI), strength (STR), upper half mean length (UHML), and uniformity index (UI). Growing year is preceded by location at left. Florence, SC (FLO), Starkville, MS (MSU), and Stoneville, MS (STV)

genotypes and genotypes by environment effects are significant for all the tested traits. Among the traits, location had the most significant effect on the variance of MIC values by ANOVA, and this was more significant than the year component for MIC (Tables S5, S6).

Genotyping by whole genome sequencing

We generated 4.4 TB of sequences from 550 RILs, resulting in 3× coverage, each. We sequenced the eleven parents at 20× coverage. We selected 473,517 SNPs that did not exhibit significant segregation distortion in the MAGIC population and were distributed throughout the genome (Fig. 2). Of these, 7506 were non-synonymous mutations to annotated genes. We did not detect significant kinship among any of the lines (Fig. S1), which is congruent with our previous results based on the analysis of 1582 SSR markers (Fang et al. 2014) and of 6071 genotyping-by-sequencing (GBS)-SNP and 233 SSR markers (Islam et al. 2016).

GWAS at each location–year

Since we detected environmental contributions to phenotypic variance by ANOVA (Tables S5, S6), we first subjected data from each location–year to GWAS separately (Fig. S2). The congruence among the most highly significant loci identified at each location–year was very high for most traits, though noticeably less for MIC and UHML (Fig. S2). Next, we performed GWAS using the sequence variants and the full set of normalized phenotypes. We identified 460 SNPs at the $-\log(p) > 7$ significance threshold for at least one trait. The full list of 1546 SNPs with $-\log(p) > 4$ is presented in Table S7. The publically available tissue expression data for annotated genes at these loci are presented in Table S8.

Multi-trait fiber quality locus on chromosome (Chr.) A07 masks STR locus on Chr. D13

Three of the traits, STR, UI, and SFI, identified the same major effect locus on Chr. A07 near position 72-Mb (A07:72-Mb) in the whole data set (Fig. 3). This locus was identified before based on GBS-SNP marker analysis (Islam et al. 2016). At this location, we further identified non-synonymous mutations in two annotated genes (*Gh_A07G1744* and *Gh_A07G1769*) that had highly significant *p* values based on the GWAS (Table 1).

To identify additional loci that may be masked in the analysis by this large effect locus, we excluded the 15% of the RILs that contained the superior minor haplotype and analyzed the 465 RILs with the inferior A07:72-Mb haplotypes. This subset revealed a single SNP located on Chr. D13 that passed the significance threshold (Fig. 3). This SNP results in the loss of a stop codon from an in-frame, short

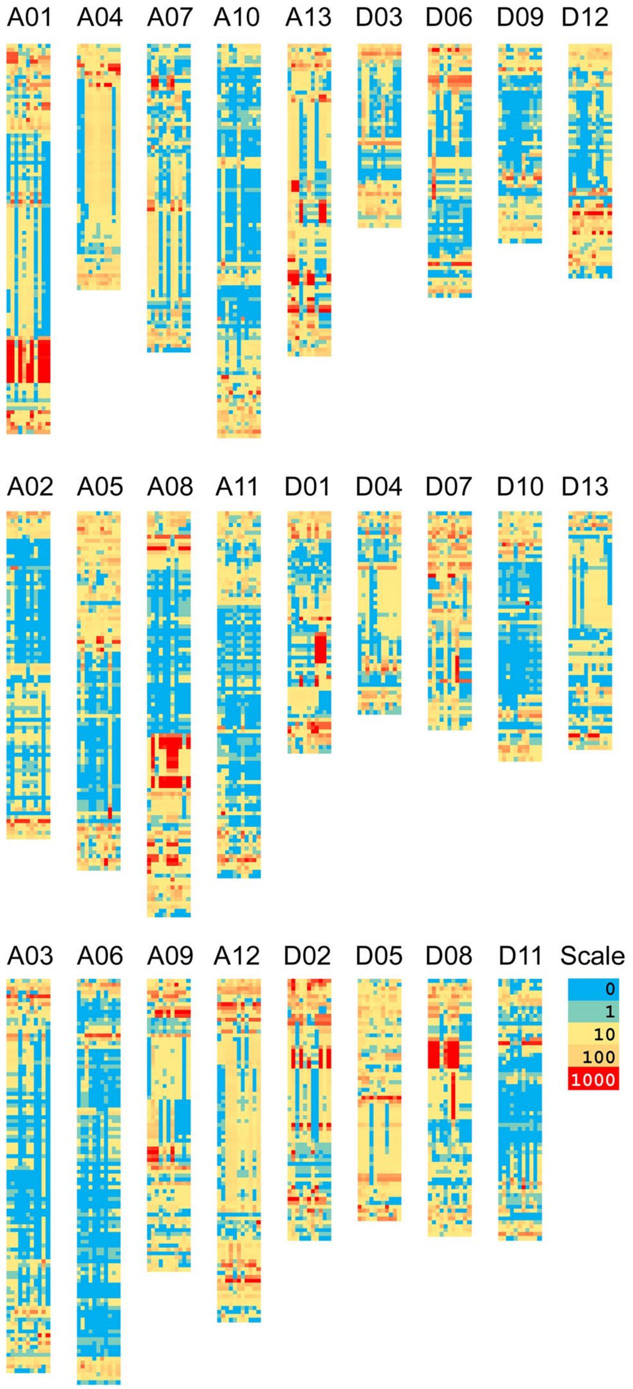


Fig. 2 Distribution of 473,517 SNPs in the eleven parental *G. hirsutum* varieties. Each chromosome is labeled and is composed of eleven vertical columns which represent the varieties in the same order as they are listed in Table S1. Each horizontal row represents a 1-Mb bin. The number of SNPs per Mb is color-coded according to the scale

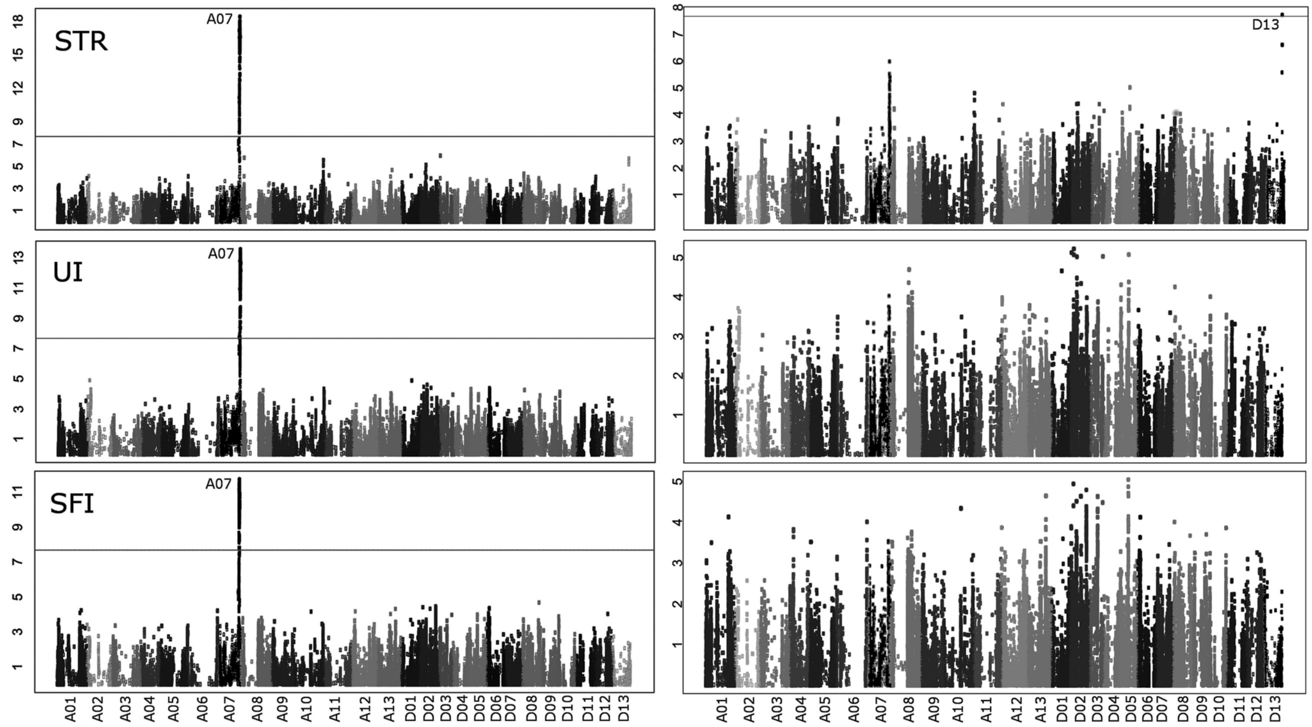


Fig. 3 Major multi-trait locus at chromosome A07:72-Mb masks a significant STR locus at Chr. D13:52,852,792. Three traits, STR, UI, and SFI, reveal a major locus, labeled A07 at left, when the full dataset is subjected to GWAS. When only the 465 RILs that contain the inferior haplotype at A07:72-Mb were analyzed, a significant SNP in the intron of *Gh_D13G1792* (D13) is evident in the STR data. Vertical axis is labeled with −log(*p*) values, and the significance threshold is indicated on each plot with a horizontal line

Table 1 Non-synonymous (NonSyn) mutations in annotated genes with significant GWAS *p* values

Trait	Chr	Position	*P* value	MAF	NonSyn	Gh_Gene	TAIR	TAIR description
STR	A07	71,156,748	1.58E−13	0.176	*V>M*	*Gh_A07G1744*	AT5G05260	Cytochrome p450 79a2
UI	A07	71,156,748	1.85E−11	0.176	*V>M*	*Gh_A07G1744*	AT5G05260	Cytochrome p450 79a2
SFI	A07	71,156,748	7.79E−09	0.176	*V>M*	*Gh_A07G1744*	AT5G05260	Cytochrome p450 79a2
STR	A07	72,203,768	2.85E−17	0.152	*K>E*	*Gh_A07G1769*	AT4G12780	Chaperone DnaJ-domain superfamily
UI	A07	72,203,768	5.73E−13	0.152	*K>E*	*Gh_A07G1769*	AT4G12780	Chaperone DnaJ-domain superfamily
SFI	A07	72,203,768	7.61E−12	0.152	*K>E*	*Gh_A07G1769*	AT4G12780	Chaperone DnaJ-domain superfamily
ELO	D01	57,996,968	1.34E−07	0.192	*F>L*	*Gh_D01G1959*	AT3G44150	Expp1 protein
ELO	D04	47,084,684	4.09E−07	0.154	*T>A*	*Gh_D04G1518*	AT5G64230	1,8-cineole synthase
ELO	D04	47,085,809	1.41E−08	0.188	*V>A*	*Gh_D04G1519*	AT5G64250	Aldolase-type TIM barrel family protein
ELO	D04	47,715,033	9.23E−14	0.204	*K>Q*	*Gh_D04G1559*	AT3G19670	Pre-mRNA-processing protein 40B
ELO	D04	47,720,280	6.28E−15	0.184	*Y>H*	*Gh_D04G1559*	AT3G19670	Pre-mRNA-processing protein 40B
ELO	D04	47,872,954	6.73E−12	0.215	*I>L*	*Gh_D04G1573*	AT3G19540	Protein of unknown function (DUF620)
MIC_MSU	D08	2,724,050	3.69E−08	0.458	*A>V*	*Gh_D08G0275*	AT4G30240	Syntaxin/t-SNARE family protein
MIC_MSU	D08	2,766,232	2.90E−09	0.447	*P>S*	*Gh_D08G0282*	AT5G45150	RNAse THREE-like protein 3
MIC_MSU	D08	2,959,509	3.95E−07	0.451	*R>H*	*Gh_D08G0305*	AT5G56240	Hapless protein
UHML	D11	24,009,361	2.87E−07	0.111	*R>C*	*Gh_D11G1928*	AT4G37040	Methionine aminopeptidase 1D
UHML	D11	24,030,087	9.86E−08	0.104	*E>D*	*Gh_D11G1929*	AT3G19150	KIP-related protein 6
UHML	D11	24,065,685	1.50E−07	0.110	*K>R*	*Gh_D11G1931*	AT3G19184	AP2/B3-like transcriptional factor family
STR_465	D13	52,852,792	1.90E−08	0.338	STOP>*R*	*Gh_D13G1792*	AT3G04470	Ankyrin repeat family protein

MIC_MSU is based on the subset of data from the Starkville, MS location. STR_465 is based on the 465 RILs that lack the superior multi-trait haplotype at A07:72-Mb

(63 bp), annotated intron in *Gh_D13G1792* (Table 1). This stop codon is present in the TM-1 reference sequence and four parents (Coker 315, STV825, PSC355, and STV474), but is replaced with an arginine residue in the other seven MAGIC parents.

Significant and additive QTLs for ELO

Our GWAS analysis of ELO, a measure of how far a cotton fiber sample can be stretched before breaking, revealed three highly significant loci, at Chr. D01:58-Mb, D04:47-Mb, and D05:43-Mb (Fig. 4). We identified six non-synonymous SNPs in five genes at two of these loci (Table 1). One of these, *Gh_D04G1519*, is very highly expressed in fiber cells and ovules (Table S8). We did not observe non-synonymous variants of genes at D05:43-Mb, which is an especially broad peak, with highly significant SNPs extending 8-Mb in both directions (Fig. 4 and Table S7). We found that these QTLs had additive effects, and the 136 RILs with high-ELO haplotypes at all three loci had

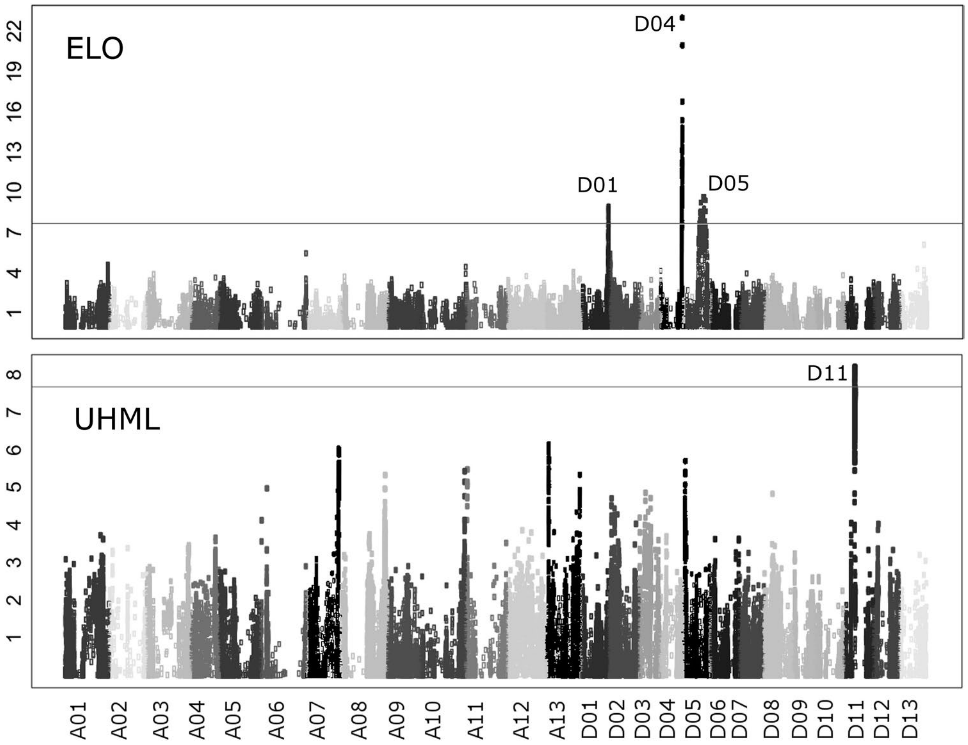


Fig. 4 GWAS of the full MAGIC population reveals significant loci that control ELO and UHML. The loci at D01:58-Mb, D04:47-Mb, D05:43-Mb, and D11:24-Mb are labeled with their respective chromosomes. Vertical axis is labeled with $-\log(p)$ values, and the significance threshold is indicated on each plot with a horizontal line

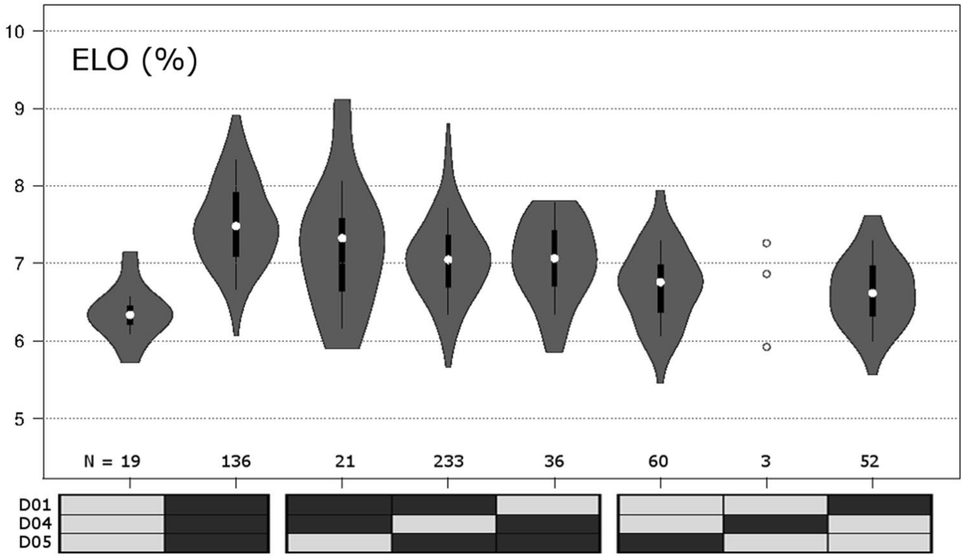


Fig. 5 Violin plot of ELO values for RILs based on genotypes at the three significant loci. Genotypes are presented along the horizontal axis, with the high-ELO haplotype indicated with a dark gray rectangle and the low-ELO haplotypes with a light gray rectangle. The number (*N*) of RILs in each group is indicated. See also Fig. 4. For pairwise *t* test *p* values see Table S9

much greater ELO than the 19 RILs with three low-ELO haplotypes (Fig. 5 and Table S9). None of the eleven parents had all three low-ELO haplotypes, while five parents (M240, DP90, SG747, PSC355, and STV474) had all three high-ELO haplotypes.

Significant QTL for UHML on Chr. D11

We identified a single highly significant locus for UHML at Chr. D11:24-Mb, which harbored three candidate genes (*Gh_D11G1928, Gh_D11G1929,* and *Gh_D11G1931*) with non-synonymous SNPs that are associated with fiber length variation (Fig. 4 and Table 1). These genes are expressed at similar levels in most tissues, including developing fiber cells (Table S8). One parent, HS26, contributed the low-UHML alternative haplotype, and the other parents shared the same, reference-type haplotype.

Location-dependent QTLs for MIC

Our analysis of MIC using the full dataset did not reveal any highly significant QTLs by GWAS. However, when we divided the data based on locations, we identified one significant locus at Chr. D08:3-Mb in the Starkville, MS location, but not in the other two locations (Fig. 6 and Fig. S2). This locus harbors three genes (*Gh_D08G0275, Gh_D08G0282,* and *Gh_D08G0305*) with non-synonymous SNPs relative to the reference sequence (Table 1). We observed that this peak was prominent in the overall MIC analysis, though below the significance threshold. The most significant peak from each of the other two locations was also visible in the overall analysis. The Florence, SC location, revealed a locus at A13:68-Mb and Stoneville, MS, at A08:63-Mb (Fig. 6 and Table S7). We found that these three loci exhibited an additive effect in the overall data and that the 23 RILs with low-MIC haplotypes at all three had significantly lower MIC than the 102 RILs with high-MIC haplotypes at each (Fig. S3 and Table S10). One parent, DP90, had all three low-MIC haplotypes, and two (HS26 and Pyramid) had all three high-MIC haplotypes.

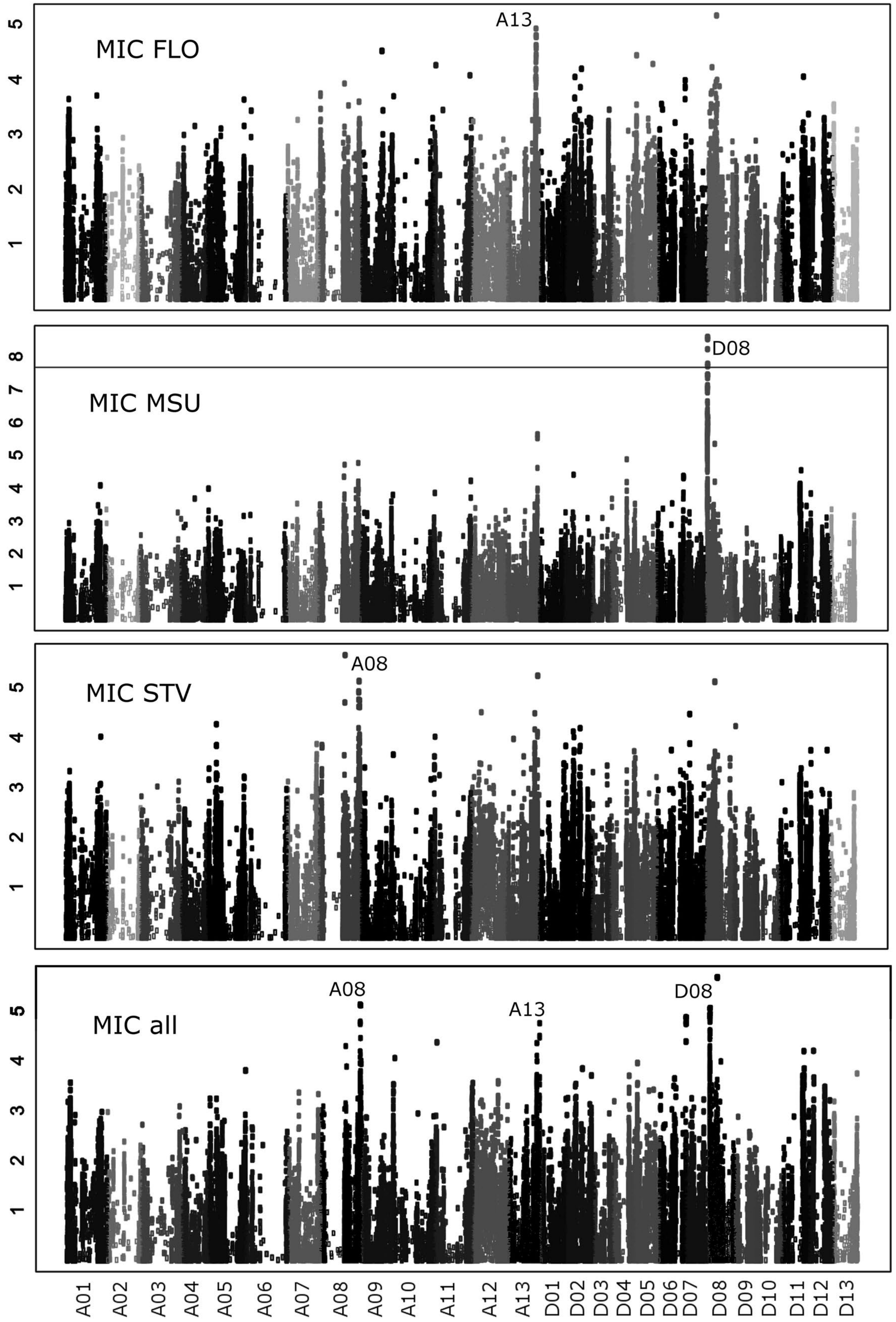
Discussion

Independent identification of candidate causative mutations for STR and UHML

Perhaps, the greatest advantage of genotyping by WGS for GWAS is the ability to observe variants of genes at the detected loci and directly observe these variants in each of the lines. The large number of recombinations that accumulated during the eleven generations of MAGIC population development further ensured a very high resolution to the GWAS mapping approach. We previously determined that 5000 markers would be sufficient for GWAS with this population, and here we report a WGS analysis with 100× that many markers (Islam et al. 2016). Recently, the WGS for GWAS approach was used with a collection of 419 cotton varieties (Ma et al. 2018). Interestingly, although their study used primarily Chinese varieties and relied on natural rather than MAGIC populations, Ma et al. did propose candidate causative mutations that were also identified in our analysis, particularly at the STR locus at Chr. A07:72-Mb and UHML locus at Chr. D11:24-Mb. At the Chr. A07:72-Mb locus, they proposed that the mutant allele of *Gh_A07G1769* is responsible for the superior fiber STR, and indeed this was our most significant non-synonymous mutation at the locus. However, we also detected another gene with non-synonymous SNPs, *Gh_A07G1744*, very close to the locus, and with a similarly significant *p* value (Fig. 3 and Table 1). However, since expression of *Gh_A07G1744* is low in all tissues, it may be a pseudogene (Table S8). Also, here we identified the non-synonymous SNP at Chr. D11:24-Mb in *Gh_D11G1929* that Ma et al. propose as a fiber length gene, GhFL2; however, we also identified non-synonymous SNPs in *Gh_D11G1928* and *Gh_D11G1931*, each less than 40-kb away from GhFL2 and expressed at similar levels (Fig. 4, Tables 1 and S8). The same group, in an earlier report based on the Cotton 63 K Illumina Infinium SNP array, proposed another candidate gene, which we can here independently corroborate and perhaps better explain (Hulse-Kemp et al. 2015; Sun et al. 2017). At the Chr. D13:53-Mb STR_465 locus, which we identified by only examining the 465 RILs that lack the high-STR A07:72-Mb haplotype, we found a SNP in an intron of *Gh_D13G1792* (Fig. 3 and Table 1). Sun et al. (2017) demonstrated that expression of this gene is very low in two low-quality cultivars and high in two normal-quality cultivars. Our sequence analysis suggests that the 63-bp annotated intron is usually retained in full-length transcripts from cotton lines that lack the stop codon and that the intron is likely only annotated as such due to the presence of the stop codon in the TM-1 cultivar that was used to establish the reference genome. A potential epistatic interaction between *Gh_D13G1792* and genes at the Chr. A07:72-Mb locus merits further study to explore the QTL masking we observed (Fig. 3).

Location-dependent QTLs for MIC

The molecular underpinnings of MIC are particularly elusive due to its complicated relationships with other fiber properties and developmental factors. MIC is a measure of the resistance to airflow of a sample of cotton fibers of a known weight that has been compressed to a known volume (Wakelyn and Chaudhry 2010). MIC is a complex trait that both the fineness and maturity of fibers and properties of the cell


MIC FLO
A13
MIC MSU
D08
MIC STV
A08
MIC all
A08
A13
D08
1
2
3
4
5
6
7
8
A01
A02
A03
A04
A05
A06
A07
A08
A09
A10
A11
A12
A13
D01
D02
D03
D04
D05
D06
D07
D08
D09
D10
D11
D12
D13

◂**Fig. 6** Micronaire (MIC) is highly influenced by growing location. Manhattan plots for each location and the full dataset (MIC_all) are shown, and the most prominent peaks at A08:63-Mb, A13:68-Mb, and D08:3-Mb are labeled with their chromosome. Vertical axis is labeled with −log(p) values, and the significance threshold is indicated on each plot with a horizontal line

wall can all affect the MIC value (Paudel et al. 2013). Growers have long reported a significant environmental influence on MIC values, with fibers harvested relatively early in the growing season exhibiting lower MIC values (Bradow and Davidonis 2000; Verhalen et al. 1975). It is interesting to find a difference in identified QTLs between locations, particularly between Starkville, MS, and Stoneville, MS, which are at similar latitudes (33.4°N) and only 200 km apart. Florence, SC, is only slightly north at 34.2°N, but is 840 km from Starkville, MS. Further research into the contributions of soil type and management practices to MIC values may be warranted. Perhaps, the candidate genes we present at Chr. D08:3-Mb will be useful for the development of niche cultivars that offer significant value to growers under specific environmental conditions or management practices (Fig. 6 and Table 1). Moreover, as MIC is a complex trait including maturity and fineness, identification of a major and stable MIC QTL may be difficult. Future research may require to accurately measure fiber maturity and fineness separately using special instrument such as CottonScope® (Rodgers et al. 2011) or cross section (Hequet et al. 2006) and then identify QTLs for fineness and/or maturity.

Transgressive segregation in MAGIC populations

Linkage drag is a major impediment to the efficient improvement in crops and livestock by traditional breeding, since beneficial alleles of genes may reside in chromosomal locations that are physically close to deleterious alleles of other genes. Our MAGIC population was originally developed as a breeding tool, to facilitate novel combinations of alleles from a number of diverse high-quality cotton lines (Jenkins et al. 2008). As we reported earlier, the five cycles of random mating and six generations of self-pollination via single seed descent efficiently shuffled the genomes of the eleven parents into a MAGIC population with no discernible structure or kinship and very low LD (Fig. S1) (Islam et al. 2016). This created opportunities for beneficial alleles that originate in different parents to exhibit novel additive or epistatic effects and to break the linkage between nearby genes. That this goal of transgressive segregation was achieved is readily apparent from the broader distribution of phenotypes observed in the RILs than in the parent lines, the masking of *Gh_D13G1792* and the additive effects of our ELO and MIC loci (Tables S3, S9, S10, Figs. 3, 5, and S3). Gene pyramiding is a daunting task for breeders, since several small-effect QTLs are laborious to identify and require multiple introgressions to be combined (Servin et al. 2004). Among our RILs are lines where the three ELO and three MIC loci have already been combined with the STR and UHML loci and are thus valuable breeding and research materials. The concordance of some of our candidate genes with independent studies is encouraging for the utility of our full list of candidates in breeding and biotechnological applications.

Author contribution statement D. D. F. conceived and coordinated the project. G. N. T. analyzed phenotypic and sequencing data and wrote the manuscript. J. N. J. and J.C. M. designed and developed the population. J. N. J., J. C. M., L. Z. and B. T. C. conducted field experiments. C. D. D. conducted fiber property measurements. M. S. I. conducted ANOVA analysis. M. S. I., P. L. and D. D. F. isolated DNA. D. C. J., B. D. C. and D. D. F. provided material support and edited the manuscript. All authors read, edited, and approved the final manuscript.

Acknowledgements This research was funded mainly by the USDA Agricultural Research Service CRIS Projects 6054-21000-017-00D (GNT, DDF), 6064-21000-016-00D (JNJ, JCM), 6066-21000-051-00D (LZ), and 6082-21000-008-00D (BTC). Additional funding was provided by Cotton Incorporated Projects 10-747 and 15-751 awarded to DDF, and Project 09-541 to JNJ. 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 USDA which is an equal opportunity provider and employer.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and materials All relevant data reported in this paper are within the paper and its online supplementary files.

References

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. ArXiv preprint arXiv:1406.5823

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS et al (2018) Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet 50:362–367

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

Bradow JM, Davidonis GH (2000) Quantitation of fiber quality and the cotton production-processing interface: a physiologist's perspective. J Cotton Sci 4:34–64

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43:956–963

Dabbert TA, Pauli D, Sheetz R, Gore MA (2017) Influences of the combination of high temperature and water deficit on the heritabilities and correlations of agronomic and fiber quality traits in upland cotton. Euphytica 213:6

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G et al (2015) Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. Genome Biol 16:167

Fang DD, Hinze LL, Percy RG, Li P, Deng D, Thyssen G (2013) 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 191:391–401

Fang DD, Jenkins JN, Deng DD, McCarty JC, Li P, Wu J (2014) Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). BMC Genom 15:397

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. Nat Genet 49:1089–1098

Gore MA, Fang DD, Poland JA, Zhang J, Percy RG, Cantrell RG et al (2014) Linkage map construction and quantitative trait locus analysis of agronomic and fiber quality traits in cotton. Plant Genome. https://doi.org/10.3835/plantgenome2013.07.0023

Hequet EF, Wyatt B, Abidi N, Thibodeaux DP (2006) Creation of a set of reference material for cotton fiber maturity measurements. Text Res J 76:576–586

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Ann Rev Plant Biol 65:531–551

Huang X, Paulo M-J, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. PNAS 108:4488–4493

Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P et al (2015) MAGIC populations in crops: current status and future prospects. Theor Appl Genet 128:999–1017

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 Biotechnol J 15:1374–1386

Hulse-Kemp AM, Lemm J, Plieske J, Ashrafi H, Buyyarapu R, Fang DD et al (2015) Development of a 63 K SNP array for cotton and high-density mapping of intra-and inter-specific populations of Gossypium spp. G3: Genes Genomes Genet. https://doi.org/10.1534/g3.115.018416

Islam MS, Zeng L, Delhom CD, Song X, Kim HJ, Li P, Fang DD (2014) Identification of cotton fiber quality quantitative trait loci using intraspecific crosses derived from two near-isogenic lines differing in fiber bundle strength. Mol Breed 34:373–384

Islam MS, Thyssen GN, Jenkins JN, Zeng L, Delhom CD, McCarty JC et al (2016) A MAGIC population-based genome-wide association study reveals functional association of GhRBB1_A07 gene with superior fiber quality in cotton. BMC Genom 17:903

Jenkins J, McCarty J, Gutierrez O, Hayes R, Bowman D, Watson C, Jones D (2008) Registration of RMUP-C5, a random mated population of upland cotton germplasm. J Plant Regist 2:239–242

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li C, Fu Y, Sun R, Wang Y, Wang Q (2018) Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in Upland cotton (*Gossypium hirsutum* L.). Front Plant Sci 9:1083

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G et al (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet 50:803–813

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E et al (2016) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45:D896–D901

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369

Paterson A, Saranga Y, Menz M, Jiang C-X, Wright R (2003) QTL analysis of genotype × environment interactions affecting cotton fiber quality. Theor Appl Genet 106:384–396

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D et al (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

Paudel DR, Hequet EF, Abidi N (2013) Evaluation of cotton fiber maturity measurements. Ind Crops Prod 45:435–441

Percy RG, Cantrell RG, Zhang J (2006) Genetic variation for agronomic and fiber properties in an introgressed recombinant inbred population of cotton. Crop Sci 46:1311–1317

Rakshit S, Rakshit A, Patil J (2012) Multiparent intercross populations in analysis of quantitative traits. J Genet 91:111–117

Rodgers J, Delhom C, Fortier C, Thibodeaux D (2011) Rapid measurement of cotton fiber maturity and fineness by image analysis microscopy using the Cottonscope®. Text Res J 82:259–271

Said JI, Knapka JA, Song M, Zhang J (2015) Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum* × *G. barbadense* populations. Mol Genet Genom 290:1615–1625

Servin B, Martin OC, Mézard M (2004) Toward a theory of marker-assisted gene pyramiding. Genetics 168:513–523

Su J, Fan S, Li L, Wei H, Wang C, Wang H, Song M, Zhang C, Gu L, Zhao S (2016) Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese upland cotton. Front Plant Sci 7:1576

Su J, Li L, Zhang C, Wang C, Gu L, Wang H, Wei H, Liu Q, Huang L, Yu S (2018) Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. Theor Appl Genet 131:1299–1314

Sun Z, Wang X, Liu Z, Gu Q, Zhang Y, Li Z et al (2017) Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. Plant Biotechnol J 15:982–996

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Thyssen GN, Fang DD, Turley RB, Florane C, Li P, Naoumkina M (2014) Next generation genetic mapping of the Ligon-lintless-2 (Li_2) locus in upland cotton (*Gossypium hirsutum* L.). Theor Appl Genet 127:2183–2192

Verhalen LM, Mamaghani R, Morrison WC, McNew RW (1975) Effect of blooming date on boll retention and fiber properties in cotton 1. Crop Sci 15:47–52

Wakelyn PJ, Chaudhry MR (2010) Cotton: technology for the 21st century. In: International cotton advisory committee, Washington, DC

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H et al (2013) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001–D1006

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881

You Q, Xu W, Zhang K, Zhang L, Yi X, Yao D, Wang C, Zhang X, Zhao X, Provart NJ (2016) ccNET: Database of co-expression networks with functional modules for diploid and polyploid *Gossypium*. Nucleic Acids Res 45:D1090–D1099

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Yuan Y, Wang X, Wang L, Xing H, Wang Q, Saeed M, Tao J, Feng W, Zhang G, Song X-L (2018) Genome-wide association study identifies candidate genes related to seed oil composition and protein content in *Gossypium hirsutum* L. Front Plant Sci 9:1359

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J et al (2015) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol 33:531–537

Affiliations

Gregory N. Thyssen[1,2] · Johnie N. Jenkins[3] · Jack C. McCarty[3] · Linghe Zeng[4] · B. Todd Campbell[5] · Christopher D. Delhom[6] · Md. Sariful Islam[7] · Ping Li[1] · Don C. Jones[8] · Brian D. Condon[2] · David D. Fang[1]

[1] Cotton Fiber Bioscience Research Unit, USDA-ARS-SRRC, New Orleans, LA 70124, USA

[2] Cotton Chemistry and Utilization Unit, USDA-ARS-SRRC, New Orleans, LA 70124, USA

[3] Genetics and Sustainable Agriculture Research Unit, USDA-ARS, Mississippi State, MS 39762, USA

[4] Crop Genetics Research Unit, USDA-ARS, Stoneville, MS 38776, USA

[5] Coastal Plain Soil, Water and Plant Conservation Research Unit, USDA-ARS, Florence, SC 29501, USA

[6] Cotton Structure and Quality Research Unit, USDA-ARS-SRRC, New Orleans, LA 70124, USA

[7] Sugarcane Production Research Unit, USDA-ARS, Canal Point, FL 33438, USA

[8] Cotton Incorporated, Cary, NC 27513, USA