Theor Appl Genet (2014) 127:2183–2192
DOI 10.1007/s00122-014-2372-1

ORIGINAL PAPER

Next generation genetic mapping of the Ligon-lintless-2 (Li_2) locus in upland cotton (*Gossypium hirsutum* L.)

Gregory N. Thyssen · David D. Fang · Rickie B. Turley · Christopher Florane · Ping Li · Marina Naoumkina

Received: 12 May 2014 / Accepted: 25 July 2014 / Published online: 15 August 2014

Abstract

Key message **Mapping-by-sequencing and novel subgenome-specific SNP markers were used to fine map the Ligon-lintless 2 (Li_2) short-fiber gene in tetraploid cotton. These methodologies will accelerate gene identification in polyploid species.**

Abstract Next generation sequencing offers new ways to identify the genetic mechanisms that underlie mutant phenotypes. The release of a reference diploid *Gossypium raimondii* (D_5) genome and bioinformatics tools to sort tetraploid reads into subgenomes has brought cotton genetic mapping into the genomics era. We used multiple high-throughput sequencing approaches to identify the relevant region of reference sequence and identify single nucleotide polymorphisms (SNPs) near the short-fiber mutant Ligon-lintless 2 (Li_2) gene locus. First, we performed RNAseq on 8-day post-anthesis (DPA) fiber cells from the Li_2 mutant and its wild type near isogenic line (NIL) *Gossypium hirsutum* cv. DP5690. We aligned sequence reads to the D_5 genome, sorted the reads into A and D subgenomes with PolyCat and called SNPs with InterSNP. We then identified SNPs that would result in non-synonymous substitutions to amino acid sequences of annotated genes. This step allowed us to identify a 1-Mb region with 24 non-synonymous SNPs, representing the introgressed region that differentiates Li_2 from its NIL. Next, we sequenced total DNA from pools of F_2 plants, using a super bulked segregant analysis sequencing (sBSAseq) approach. The sBSAseq predicted 82 non-synonymous SNPs among 3,494 SNPs in a 3-Mb region that includes the region identified by RNAseq. We designed subgenome-specific SNP markers and tested them in an F_2 population of 1,733 individuals to construct a genetic map. Our resulting genetic interval contains only one gene, an aquaporin, which is highly expressed in wild-type fibers and is significantly under-expressed in elongating Li_2 fiber cells.

Communicated by Christiane Gebhardt.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2372-1) contains supplementary material, which is available to authorized users.

G. N. Thyssen · D. D. Fang · C. Florane · P. Li · M. Naoumkina (✉)
Cotton Fiber Bioscience Research Unit, USDA-ARS-SRRC, 1100 Robert E. Lee Blvd, New Orleans, LA 70124, USA
e-mail: marina.naoumkina@ars.usda.gov

D. D. Fang · R. B. Turley
Crop Genetics Research Unit, USDA-ARS, 141 Experiment Station Road, Stoneville, MS 38776, USA

Introduction

Genetic mapping of a mutant phenotype generally progresses through three steps when a bi-parental mapping population is used. First, polymorphic genetic markers are identified between parents or between the DNA pools of mutant and wild-type individuals. Second, genome-wide analysis of linkage in a mapping population establishes a chromosome or chromosomal region for the mutant locus. Finally, fine mapping with a larger segregating population and additional genetic markers narrows down the interval until candidate genes are identified. Since the advent of next generation sequencing technologies, many researchers have attempted to streamline this process; however, most of these efforts have been in model species. The existence of a quality reference genome enabled mapping-by-sequencing approaches in *Arabidopsis thaliana*, *Saccharomyces cerevisae* and *Drosophila melanogaster* (Birkeland et al. 2010;

Blumenstiel et al. 2009; Schneeberger et al. 2009; Zuryn et al. 2010). More recently, the use of related genomes with local synteny has been shown to be useful for identification of a diverse region encompassing clusters of mutations that distinguish mutant from wild-type organisms and contain the mutant locus (Galvão et al. 2012; Wurtzel et al. 2010).

Upland cotton (*Gossypium hirsutum* L.) is a tetraploid crop plant that is the foundation of global natural textile production (Paterson et al. 2012). Allotetraploid cotton descends from an inter-specific hybridization event that occurred 1–2 million years ago between an A-genome diploid native to Africa and a Mexican D-genome diploid (Wendel and Cronn 2003). Recently, the genome sequence of a D_5 genome diploid species, *G. raimondii* Ulbr, was released (Paterson et al. 2012). Importantly, a bioinformatics tool, called PolyCat, enables the assignment of sequence reads from *G. hirsutum* to A and D subgenomes, after first aligning all reads to the *G. raimondii* genome (Page et al. 2013). Used together, PolyCat and the *G. raimondii* genome provide a workable reference for the chromosomes of the D-subgenome (chromosomes 14–26) of *G. hirsutum* and the gene content of the A subgenome (chromosomes 1–13). In addition to causing difficulties with sequence alignment, the presence of two similar subgenomes within *G. hirsutum* complicates the design of genetic markers such as PCR-based simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers, which may amplify products from both subgenomes, making it impossible in some cases to determine marker zygosity.

Our objective was to identify candidate genes responsible for the short-fiber mutant Ligon-lintless 2 (Li_2) phenotype by locating the Li_2 genetic locus on a physical map and quantifying expression of genes in this locus. The Li_2 mutation affects cotton fiber elongation and its identification will advance the understanding of plant cell elongation and enable the development of cotton varieties with higher fiber quality. Previous work has mapped the Li_2 locus to *G. hirsutum* chromosome 18 by phenotype association with aneuploid stocks and linkage analysis using restriction fragment length polymorphism and SSR markers (Hinchliffe et al. 2011; Kohel et al. 2002; Rong et al. 2005). In this study, we used RNA and DNA deep-sequencing to identify a region of dense polymorphism between Li_2 and its wide-type NIL *G. hirsutum* cv. DP5690. Filtering candidate SNPs for non-synonymous changes to annotated proteins enabled us to accurately identify the introgressed region between NILs using *G. raimondii* as a reference sequence. We were therefore able to streamline, in a non-model tetraploid crop species, the first two steps of mutant characterization, by direct RNA sequencing of NILs without a mapping population. We sequenced bulked F_2 segregants to confirm the region and to identify more candidate SNP markers. We designed novel subgenome-specific SNP markers, based on our sequencing data, which included both allele-specific and subgenome-specific primers for fine mapping in a population of 1,733 F_2 plants. The resulting genetic map defined an interval that contains only a single gene, an aquaporin. This gene, γTIP, does not contain any non-synonymous mutations, but is significantly underexpressed in elongating Li_2 fiber cells.

Materials and methods

Plant materials

The development of the two NILs of Upland cotton used in this study has been described previously (Hinchliffe et al. 2011). Briefly, the Li_2 gene was introgressed from a *G. hirsutum* cv. TM-1 background into *G. hirsutum* cv. DP5690 via five generations of backcrossing and single seed descent using DP5690 as the recurrent parent (Hinchliffe et al. 2011). Homozygous plants with the dominant (Li_2/Li_2) or recessive (li_2/li_2) phenotype were grown in the field in New Orleans, LA in 2013 for mRNA isolation. Parental NILs, Li_2 mutant and DP5690, were crossed and the resulting F_1 were self-pollinated to produce F_2 seeds. The F_2 population was grown in Stoneville, MS as described below.

RNA isolation, RNAseq and RT-qPCR

Cotton bolls were harvested at the following eight time-points during fiber development: 0, 3, 5, 8, 12, 16, 20, and 24 days post-anthesis (DPA). Cotton fibers were separated from ovules using a glass bead shearing technique (Taliercio and Boykin 2007). Total RNA was isolated from detached fibers as described previously (Naoumkina et al. 2014). RNA samples from three biological replicates of 8-DPA fiber were subjected to paired-end Illumina mRNA sequencing (RNAseq). Library preparation and sequencing were conducted by Data2Bio LLC (Ames, Iowa) following standard procedures. The libraries were sequenced using 101 cycles of chemistry and imaging, resulting in paired-end sequencing reads with length of 2 × 101 bp that flank 150-bp inserts. RNA from each of the above-mentioned time-points was used for reverse transcription quantitative polymerase chain reaction (RT-qPCR) analysis using primers listed in Table S1. A detailed description of reverse transcription, qPCR and expression analysis was previously reported (Naoumkina et al. 2014).

Super bulked segregant analysis sequencing (sBSAseq)

F_2 plants from a population of 536 individuals were selected for sequencing by a bulked segregant approach

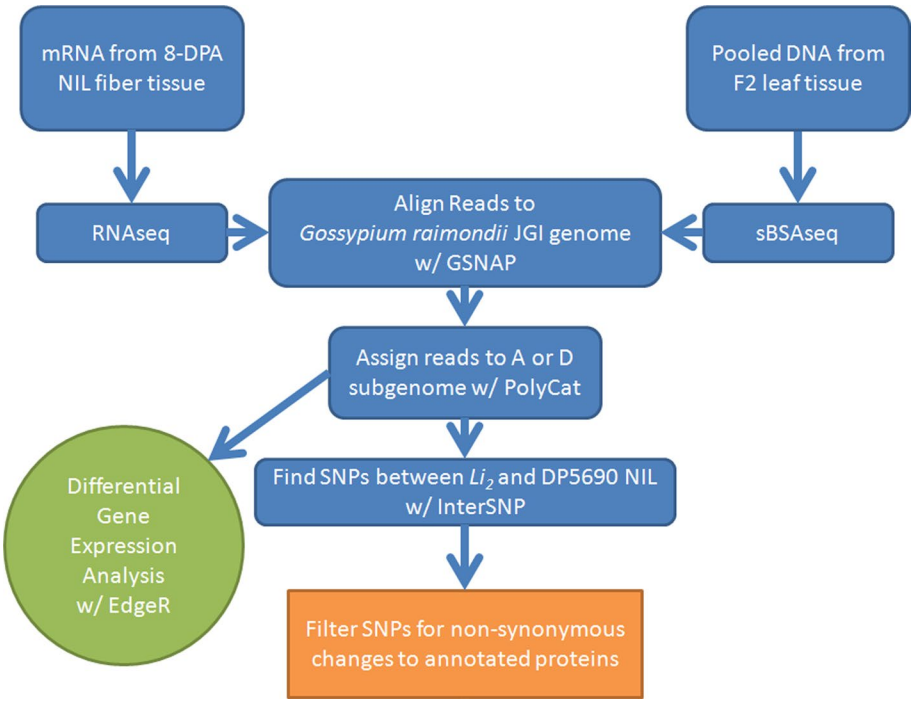

Fig. 1 Flowchart of data processing to identify the introgressed region by RNAseq and sBSAseq and to identify differentially expressed genes by RNAseq. Sequence reads from wild type or Li_2 mutant *Gossypium hirsutum* samples were aligned to the related *G. raimondii* genome with GSNAP, assigned to subgenomes with PolyCat, and compared for SNPs with InterSNP. SNPs were analyzed to find non-synonymous changes in annotated proteins. The RNAseq alignments were also used to compute differential gene expression with EdgeR

(Michelmore et al. 1991; Takagi et al. 2013). Two DNA pools were constructed: a pool of DNA from 100 short fiber (Li_2/Li_2, and Li_2/li_2) and a pool of DNA from 100 wild-type (li_2/li_2) plants. DNA was Illumina sequenced with paired 101-bp reads flanking 150-bp inserts by Data2Bio LLC (Ames, Iowa).

Identification of diverse genomic regions

The general outline of data processing is shown in Fig. 1. Sequence reads were aligned to the JGI *G. raimondii* genome sequence (v2.1) with the GSNAP software program (Paterson et al. 2012; Wu and Nacu 2010). These reads were assigned to the A or D subgenome using the PolyCat software program (Page et al. 2013). The binary alignment/map (BAM) files from different biological replicates of RNAseq data, but matching subgenome assignment, were combined using the SAM tools merge software (Li et al. 2009). InterSNP, part of the BamBam package that includes PolyCat, was used to identify candidate SNPs between Li_2 and its wild-type DP5690, for each subgenome separately and for RNAseq and sBSAseq data separately (http://sourceforge.net/projects/bambam/). InterSNP was run with default parameters, but we filtered the output file for SNPs with only one non-ambiguous variant per sample. These SNPs were then compared between Li_2 and DP5690 to identify non-synonymous SNPs in annotated exons. Histograms were generated by counting the instances of non-synonymous SNPs in each 1-Mb or 100-kb interval.

Differential gene expression

RNAseq expression analysis was conducted following the PolyCat pipeline (Page et al. 2013), with only two minor adjustments: (1) we only counted exonic reads; (2) we used the ratio of A-assigned to D-assigned reads to proportionally divide the total number of mapped reads for each gene. Differential gene expression was calculated by the negative binomial method of the EdgeR software using the tagwise estimation of dispersion (Robinson et al. 2010).

Subgenome-specific SNP primer design

Manual inspection of read alignments in RNAseq and sBSAseq data was used to identify true SNPs and nearby homeoSNPs. Primers were designed to end on the SNP nucleotide and to incorporate an additional mismatch, usually at the third base from the 3′ end, which has been shown to increase stringency (Drenkard et al. 2000). Forward primers were designed to be specific to the D-subgenome variant of the homeoSNP, while reverse primers were synthesized in two versions, for the mutant and wild-type variants of the true SNPs.

Mapping population

The 1,733 F_2 plants used in this study were grown over 2 years in Stoneville, MS. During 2012, 516 F_2 plants were grown, and 1,217 plants were grown in 2013. Standard conventional field practices were applied during the

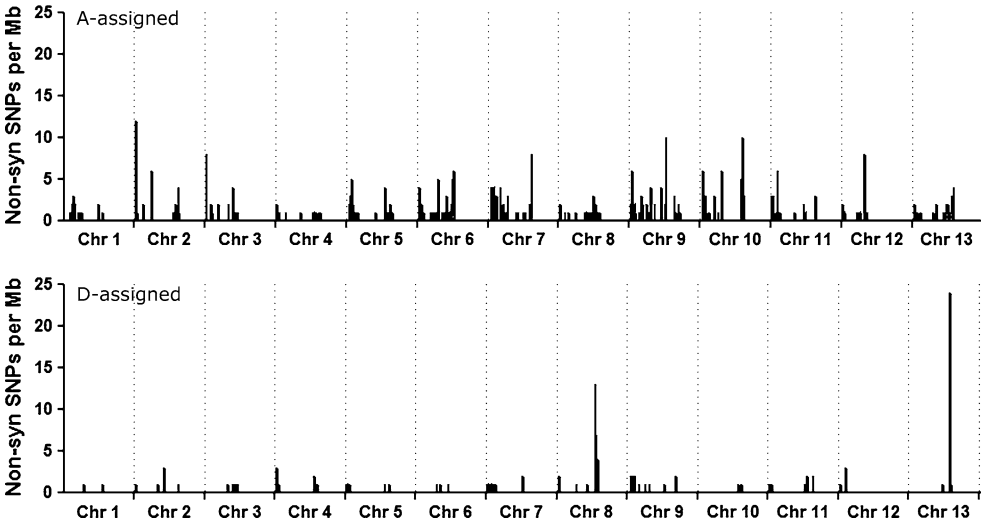

Fig. 2 Non-synonymous SNPs per Mb in RNAseq data. The 13 reference *G. raimondii* chromosomes are presented along the *x*-axis in each panel, with either A or D subgenome-assigned reads

plant growing season. Young leaves were collected for DNA isolation as described previously (Fang et al. 2010). SSR markers were scored as usual (Fang et al. 2010). The new SNP markers were first tested by running allele-specific qPCR reactions on parental NILs and F_1 plants. Unlabeled oligos, template DNA and a SYBR PCR mix were run on a C1000 thermal cycler with CFX96 Real Time System (Bio-Rad, Hercules, CA), which recorded fluorescent signal after each cycle. Samples were scored as positive or negative for each reaction by a C_t value threshold determined for each primer pair. After both allele-specific reactions were run for each individual F_2 plant, a SNP marker was scored as homozygous or heterozygous. The marker genotypes were analyzed by the JoinMap software using default parameters and LOD score of 10 to construct the genetic map (Van Ooijen 2006).

Results

Identification of the introgressed region containing the Li_2 locus based on RNAseq

We isolated and sequenced mRNA from 8-DPA fibers from field grown NILs that were either homozygous for the Li_2 short-fiber gene or the wild-type allele from the recurrent parent, DP5690. We analyzed the data according to the pipeline shown in Fig. 1. We first aligned reads to the reference *G. raimondii* genome using GSNAP, and then assigned the reads to A and D subgenomes using PolyCat (Page et al. 2013; Wu and Nacu 2010). We compared the Li_2 and wild-type alignments to identify SNPs in each subgenome using InterSNP. Next, we identified those SNPs that would alter the amino acid sequence of annotated proteins. We identified 391 non-synonymous SNPs between the Li_2 and DP5690 NILs in the A-assigned reads and 120 in the D-assigned reads. Therefore, the average non-synonymous SNP density across 749 Mb of the 13 analyzed chromosome assemblies of the D_5 genome for the A-assigned reads was 0.52/Mb and was 0.16/Mb for D-assigned reads. We observed a striking peak at 57 Mb of chromosome (Chr.) 13 in the histogram of D-assigned reads (Fig. 2). There are 24 putative non-synonymous SNPs in 17 genes located within the 57th Mb region, constituting the region of highest diversity between the NILs.

Super bulked segregant analysis sequencing (sBSAseq)

To confirm the region identified by RNAseq and to identify additional SNPs, including intergenic SNPs and those in genes that lacked expression in the 8-DPA fiber tissue of either or both parental NILs, we sequenced pooled DNA of 100 short fiber and 100 wild-type F_2 plants. Again, after aligning, assigning and filtering for non-synonymous SNPs, we observed a striking peak near the telomere of reference Chr. 13 (Fig. 3). This peak was centered at the 55.7 Mb region, which is 1.5 Mb away from the region identified by RNAseq at 57.2 Mb. The 3 Mb region between 55 Mb and the end of Chr. 13 contains 3,494 putative SNPs of which 82 are non-synonymous mutations in 48 genes.

Subgenome-specific SNP markers

We designed primer sets to interrogate SNPs that were found in the RNAseq and sBSAseq data. We found that the wild-type allele primer pair would amplify from both wild type (li_2/li_2) and mutant (Li_2/Li_2) parental NIL templates, even though the

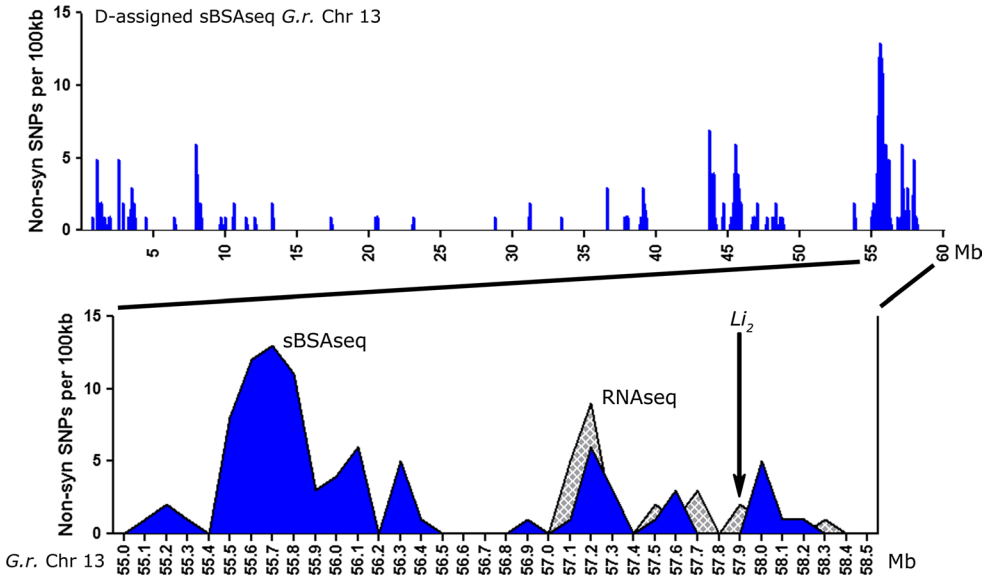

Fig. 3 Non-synonymous SNPs per 100 kb in sBSAseq data. Chromosome 13 of *G. raimondii*, which corresponds to *G. hirsutum* chromosome 18, is shown, and the 3-Mb peak is expanded for detail. *Solid blue* indicates non-synonymous SNPs in sBSAseq data; *hatched gray* indicates RNAseq data in the inset. Li_2 indicates the locus according to the genetic map (see Fig. 5)

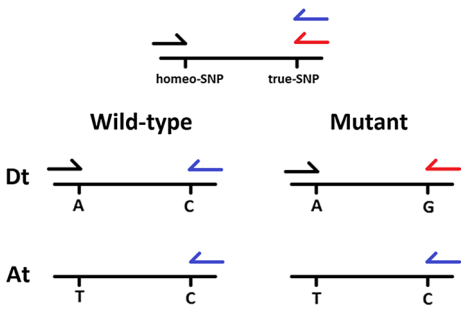

Fig. 4 Subgenome-specific SNP marker design. Subgenome-specific primer is *black*, mutant allele-specific primer is *red*, and wild-type allele-specific primer is *blue*. Only templates with two annealed primers will successfully amplify

mutant-specific primer pair would only amplify from a mutant parental NIL template. We deduced that the homeologous chromosome in our tetraploid *G. hirsutum* DNA was also acting as a template for our primers. Therefore, we designed subgenome-specific forward primers to pair with allele-specific reverse primers (Fig. 4; Table S2). We found homeoSNPs less than 600 bp from our true SNPs by manual inspection of alignment files and designed primers that would end on the homeoSNP. This resulted in a significant improvement in our conversion of putative SNPs to viable PCR markers.

Segregation of markers in 1,733 F_2 progeny

Since the RNAseq and sBSAseq implicated a 3 Mb introgressed region, we generated a large mapping population for fine genetic mapping of the Li_2 locus. We tested our SNP markers for linkage to Li_2 and used the alignment of reads to the reference chromosome as our guide to identify and develop markers that flanked Li_2 as closely as possible. We also tested two SSR markers that were previously identified as linked to Li_2 on *G. hirsutum* Chr. 18 by our group, NAU3391 and DPL0922, on the large segregating population (Hinchliffe et al. 2011). Ultimately, we constructed a genetic map (Fig. 5) which confines the Li_2 locus to the end of *G. hirsutum* Chr. 18, which is orthologous to *G. raimondii* Chr. 13 (Blenda et al. 2012; Wang et al. 2013). The Li_2 locus is flanked by two SNP markers, CFB5851 and CFB5852, which are 0.084 and 0.216 cM away, respectively. These two SNPs are located less than 6 kb apart on the reference sequence, flanking a single gene, Gorai.013G265400, a gamma tonoplast-intrinsic protein (γTIP) or aquaporin. However, there are no differences in the coding sequence of this gene between the Li_2 and DP5690 NILs.

Differential expression of genes near the Li_2 locus based on RNAseq

To investigate the expression of genes near the Li_2 locus, we subjected the subgenome-assigned RNAseq reads to statistical analysis with EdgeR (Robinson et al. 2010). We analyzed the region bound by our second closest flanking markers, CFB5850 and CFB5853. We found nine genes on the interval that were significantly different (adjusted p value <0.05) between Li_2 and wild-type 8-DPA fiber cells (Gorai.013G265900, Gorai.013G265400, Gorai.013G265200, Gorai.013G265100, Gorai.013G264600, Gorai.013G264300, Gorai.013G264100, Gorai.013G263800, Gorai.013G263400). Of these, only four (Gorai.013G265900, Gorai.013G265400, Gorai.013G265200, Gorai.013G263800) are at least twofold different (Table 1). γTIP is nearly fourfold down regulated in Li_2 fibers and is also the most highly expressed gene on the interval, even more highly expressed than the ribosomal protein, Gorai.013G265000. Just outside

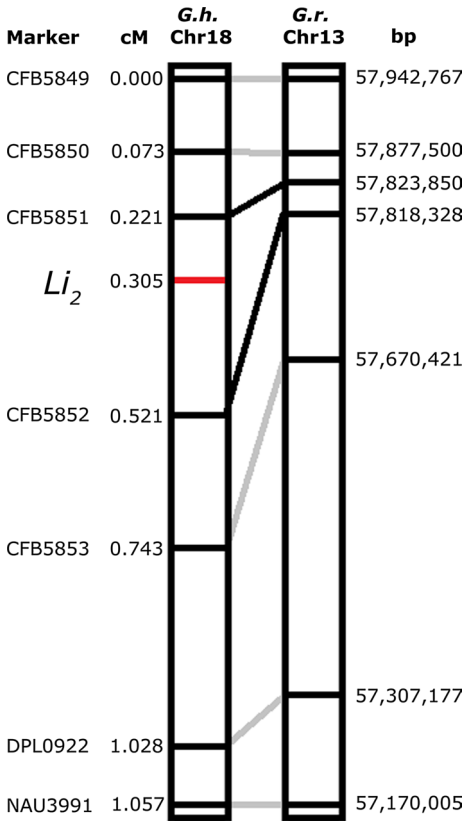

Fig. 5 Li_2 genetic and physical map based on 1,733 F_2 progeny. SNP and SSR markers associated with the Li_2 gene are shown on *G. hirsutum* chromosome 18 and on *G. raimondii* chromosome 13. Genetic map locations are shown in centiMorgans (cM) and physical locations are shown in base pairs (bp)

the CFB5851/CFB5852 interval is a C2H2-type zinc finger family protein (ZnF), Gorai.013G265200, which is more than 50-fold over-expressed in Li_2 fiber, although its average expression is 175-fold less than γTIP at 8-DPA.

RT-qPCR of the selected genes during fiber development

To confirm the results of the RNAseq expression analysis and to investigate the activation of genes near the Li_2 locus throughout the development of cotton fiber cells, we chose six genes from the CFB5850/CFB5853 interval for RT-qPCR across eight time-points (Fig. 6). In addition to γTIP and ZnF, we chose the two closest transcription factors, Gorai.013G265700 (NOT), a NOT family transcription factor and Gorai.013G265300 (MYB), an MYB family transcription factor. We also chose the most highly expressed of the significantly misregulated genes on each side of the flanking markers, Gorai.013G265900 (PK), a protein kinase, and Gorai.013G264100 (ACT), an actin (Fig. 6). Three of the genes, PK, NOT and MYB showed significant differences between wild type and Li_2 expression during the late stage of fiber development, at 20-DPA, with PK and MYB remaining over-expressed in Li_2 at 24-DPA. ACT expression is reduced in Li_2 fibers during part of early and late elongation phases at 5, 8, 16, 20 and 24-DPA, however, is indistinguishable from wild type at 0, 3, and 12-DPA. γTIP is not expressed at 0-DPA in either sample and is expressed equally in Li_2 and wild type at the beginning of elongation, 3-DPA. Thereafter, the wild-type γTIP gene increases dramatically during the peak of elongation, while expression in Li_2 is essentially flat and consistently below wild type. ZnF is expressed at similar levels at 0-DPA in both wild type and Li_2 fibers, before being essentially silenced in both at 3-DPA. At 5-DPA, ZnF rebounds to about one-tenth its expression at 0-DPA in Li_2 fibers only, while wild-type expression remains null. Thereafter, ZnF expression drops off slowly in Li_2 and only becomes detectable in wild-type fibers at 16-DPA.

Discussion

Mapping-by-sequencing mRNA and bulked DNA of NILs

Sequencing of bulked segregants has successfully identified causative mutations in other species (Schneeberger et al. 2009; Wenger et al. 2010). Since tetraploid *G. hirsutum* does not have a published genome sequence, we relied on the related diploid *G. raimondii* reference sequence and a bioinformatics tool to assign reads to A and D subgenomes (Page et al. 2013; Paterson et al. 2012). Large-scale synteny between *G. raimondii* and the D subgenome of *G. hirsutum* has been shown by the correspondence of genetic and physical maps, and our work supports the utility of local, megabase-scale, synteny for mapping-by-sequencing (Blenda et al. 2012; Wang et al. 2013).

The direct sequencing of RNA from fiber tissues of the Li_2 mutant and its wild-type NIL enabled us to map the Li_2 gene to a 1-Mb chromosomal region without the creation of a mapping population. We found putative SNPs between Li_2 and its wild-type NIL throughout the genome, but after filtering for non-synonymous SNPs we found a striking peak on *G. raimondii* Chr. 13 which corresponds to *G. hirsutum* Chr. 18 (Fig. 2) (Blenda et al. 2012; Wang et al. 2013). Confining our search for SNPs to coding sequences probably diminished the significance of the divergence between *G. hirsutum* and *G. raimondii*. After following the same SNP filtering pipeline we used for mapping by RNAseq (Fig. 1), we pooled and sequenced DNA from 100 wild-type and 100 short-fiber F_2 plants, which confirmed the signature of a diverse introgressed region. However, this combined approach did not itself produce a clear candidate gene, mainly because the putative SNPs in the introgressed region were so numerous, with 82 non-synonymous SNPs among 3,494 SNPs in the 3-Mb region (Fig. 3).

Table 1 RNAseq differential expression of annotated genes near the Li_2 locus in 8-DPA fiber cells position indicates base on Chr. 13 of *G. raimondii*

Gene/marker	Position	log2(Li_2/wt)	AvConc	adj. *p* val	Description
CFB5850	57,877,500				
Gorai.013G266000_D	57,876,948	0.79	2.29	1.07E–01	RNA-binding (RRM/RBD/RNP motifs) family protein
Gorai.013G265900_D	**57,860,416**	**1.64**	**2.25**	**1.09E–05***	**"PK" Protein kinase superfamily protein**
Gorai.013G265800_D	57,858,386	–0.08	1.89	1.00E+00	Histone superfamily protein
Gorai.013G265700_D	**57,850,915**	**0.65**	**8.17**	**6.92E−02**	**"NOT" NOT2/NOT3/NOT5 family**
Gorai.013G265600_D	57,847,645	–0.42	0.19	9.39E−01	Tetratricopeptide repeat (TPR)-like superfamily protein
Gorai.013G265500_D	57,842,506	–0.71	0.1	9.39E−01	Protein of unknown function (DUF1635)
CFB5851	57,823,850				
Gorai.013G265400_D	**57,822,093**	**–1.92**	**235.67**	**3.43E−12***	**"γTIP" gamma tonoplast-intrinsic protein/aquaporin**
CFB5852	57,818,328				
Gorai.013G265300_D	**57,816,908**	**–0.32**	**0.5**	**9.19E−01**	**"MYB" myb domain protein 48**
Gorai.013G265200_D	**57,804,350**	**5.74**	**1.35**	**1.90E−18***	**"ZnF" C2H2-type zinc finger family protein**
Gorai.013G265100_D	57,798,242	–0.59	15.16	1.49E−02*	Ras-related small GTP-binding family protein
Gorai.013G265000_D	57,793,431	0	127.6	1.00E+00	Ribosomal protein S15A
Gorai.013G264900_D	57,789,590	–0.43	2.53	4.78E−01	Histone superfamily protein
Gorai.013G264800_D	57,785,158	–0.44	24.7	1.53E−01	Actin depolymerizing factor 4
Gorai.013G264700_D	57,782,217	–0.41	5.14	3.39E−01	Histone superfamily protein
Gorai.013G264600_D	57,777,495	0.59	19.8	3.64E−02*	RNA-binding (RRM/RBD/RNP motifs) family protein
Gorai.013G264500_D	57,772,133	–0.32	2.06	6.29E−01	ACT domain-containing protein
Gorai.013G264400_D	57,767,399	–1.01	0.87	1.11E−01	Histone superfamily protein
Gorai.013G264300_D	57,762,909	–0.59	37.91	2.61E−02*	Ubiquitin carrier protein 7
Gorai.013G264200_D	57,755,340	2.58	0.05	9.39E−01	WUSCHEL related homeobox 2
Gorai.013G264100_D	**57,747,900**	**–0.64**	**48.07**	**3.65E−03***	**"ACT" actin 1**
Gorai.013G264000_D	57,742,532	0.89	1.53	1.72E−01	U-box domain-containing protein kinase family protein
Gorai.013G263900_D	57,739,683	–0.1	21.57	9.58E−01	Tim10/DDP family zinc finger protein
Gorai.013G263800_D	57,734,440	1.21	3.21	2.81E−04*	O-Glycosyl hydrolases family 17 protein
Gorai.013G263700_D	57,727,015	–0.28	2.54	6.91E−01	Methyl-CPG-binding domain protein 5
Gorai.013G263600_D	57,714,657	–0.31	0.27	1.00E+00	Protein kinase 1B
Gorai.013G263500_D	57,711,487	0.05	0.26	1.00E+00	Protein kinase superfamily protein
Gorai.013G263400_D	57,701,606	0.71	7.99	7.50E−03*	XH/XS domain-containing protein
Gorai.013G263300_D	57,692,812	0	0.04	1.00E+00	ROTUNDIFOLIA like 5
Gorai.013G263200_D	57,687,612	–0.76	0.05	1.00E+00	Laccase 2
Gorai.013G263100_D	57,680,207	–0.35	11.92	2.53E−01	RING/U-box superfamily protein
Gorai.013G263000_D	57,677,654	2.38	0.16	1.42E−01	Tetratricopeptide repeat (TPR)-like superfamily protein
Gorai.013G262900_D	57,671,655	0.09	32.38	9.39E−01	Translocon at the outer envelope membrane of chloroplasts 75-III
CFB5853	57,670,421				

The log base 2 ratio of mutant to wild-type expression, average transcript concentration and adjusted *p* values are presented in the columns labeled log2(Li_2/wt), AvConc, and adj. pVal, respectively. Significant *p* values <0.05 are marked with an asterisk. The closest flanking markers are CFB5851 and CFB5852, while the next closest markers to the Li_2 mutation are CFB5850 and CFB5853 (see Fig. 5). Bolded genes were tested by RT-qPCR (see Fig. 6)

Subgenome-specific SNP marker development

Our deep-sequencing data provided us with many candidate SNPs in the introgressed region, but due to amplification from the homeologous chromosome, we could not score the zygosity of many of our early SNP markers. Therefore, we developed subgenome-specific SNP markers, where a forward primer that terminates on a homeoSNP was designed to ensure that only the D-subgenome was interrogated by the allele-specific reverse primers (Fig. 4). Prior techniques for subgenome-specific markers relied either on two-step nested amplification (Blake et al. 2004; Chai et al.

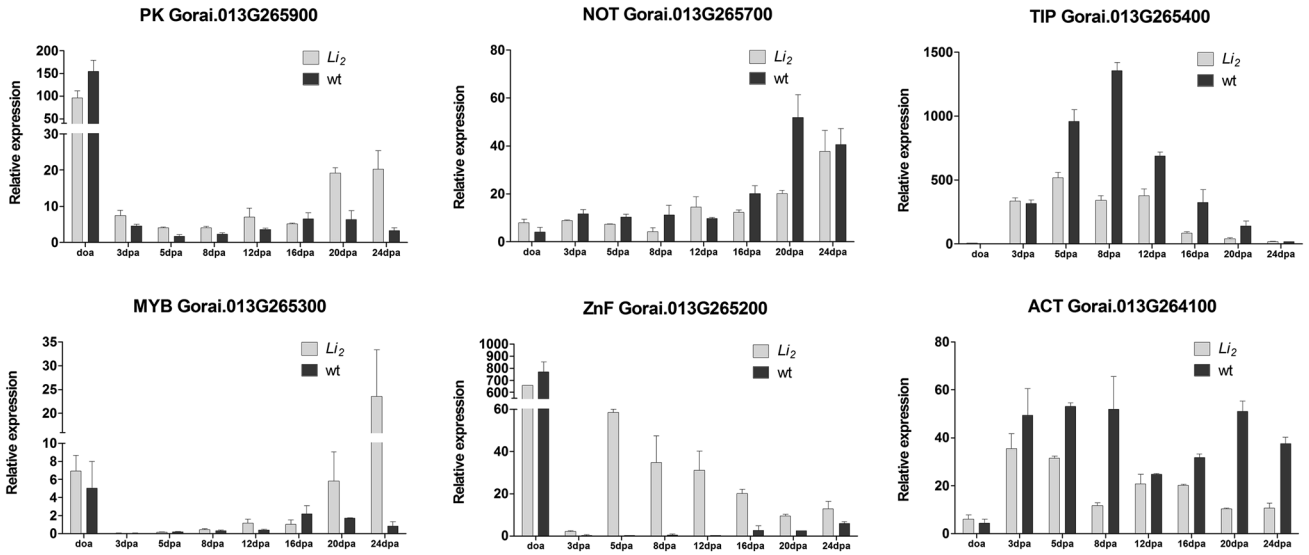

Fig. 6 RT-qPCR expression of select genes near the Li_2 locus during fiber development of Li_2 and wild-type DP5690 *G. hirsutum* fiber cells. *Error bars* indicate standard deviation from three biological replicates. Along the *x*-axis, dpa indicates the number of days post-anthesis, a measure of developmental time for cotton fiber cells

2010; Long et al. 2011) or incorporated the homeoSNP into the same primer as the true SNP (Byers et al. 2012). Since we only made a small number of SNP markers for the purpose of our fine mapping, we cannot compare the relative efficiencies of the different approaches. However, we believe that placing the homeoSNP at the 5′ position of the subgenome-specific primer should increase specificity and that our one-step technique will be useful to future work in cotton and other polyploid species.

Traditional fine genetic mapping

To narrow down the list of candidate genes from a 3-Mb to a single gene interval, we turned to traditional fine mapping. We tested our subgenome-specific SNP primers in a large segregating population of 1,733 F_2 plants. As our flanking markers closed in on the Li_2 gene, we returned to our sequencing data to identify more putative SNPs and developed more markers. Ultimately this led to a locus, 2.1 Mb away from the region identified by sBSAseq and 0.6-Mb away from the region identified by RNAseq, which was not predicted by either dataset to contain non-synonymous mutations (Fig. 3). The two closest flanking markers, CFB5851 and CFB5852, are less than 6 kb apart and flank a single gene γTIP, Gorai.013G26550, an aquaporin. The next two closest flanking markers define an interval that includes 32 annotated genes in *G. raimondii*. RNAseq and RT-qPCR confirmed that several of the genes on this interval show altered expression in Li_2 short-fiber cells during fiber development including γTIP (Table 1; Fig. 6).

Li_2 candidate genes γTIP and ZnF

Taken together, our data strongly implicate the aquaporin, γTIP (Gorai.013G265400), in the genetic mechanism of the Li_2 short-fiber trait. The closest *G. hirsutum* ortholog of the reference gene Gorai.013G265400 is GhγTIP (EF470294), which was previously identified as a highly expressed fiber transcript (Liu et al. 2008). The involvement of aquaporins in cell elongation has already been suggested both in cotton and *Arabidopsis* (Karlsson et al. 2000; Li et al. 2013; Liu et al. 2008; Ruan et al. 2001; Smart et al. 1998). By controlling the flow of water in and out of the vacuole, tonoplast-intrinsic aquaporins allow turgor pressure within the vacuole to drive cell expansion (Li et al. 2013; Liu et al. 2008). *G. hirsutum* has at least 71 aquaporins, only some of which are active in fiber cells (Park et al. 2010). Those aquaporins are highly expressed during the elongation phase of cotton fiber cells to facilitate the rapid influx of water into the central vacuole (Li et al. 2013; Liu et al. 2008). The peak expression of γTIP that we observe at 8-DPA corresponds well with the opening of fiber plasmodesmata, the import of solutes capable of increasing turgor pressure, and the expression of other aquaporins implicated in cotton fiber development (Li et al. 2013; Liu et al. 2008; Park et al. 2010; Ruan et al. 2001). The Li_2 short-fiber mutant fails to up-regulate γTIP during the peak of elongation. At 5 and 12-DPA expression of γTIP is about twofold less in Li_2 than DP5690 fiber cells and at 8-DPA the difference is even greater: threefold by RT-qPCR or nearly fourfold by RNAseq (Fig. 6; Table 1). This failure to up-regulate γTIP may prevent the accumulation of solutes and water in the

central vacuole of the mutant fiber and thereby impede the turgor-driven mechanism of cell elongation.

Very close to our mapped location is ZnF, which, although expressed at a much lower level than γTIP, is a transcription factor and has a greater difference in expression between Li_2 and wild-type fiber. Interestingly, the *Arabidopsis* homolog of ZnF, Zat12, is involved in stress responses including response to reactive oxygen species (ROS) (Davletova et al. 2005; Kiełbowicz-Matuk 2012). Our prior work that characterized global changes to Li_2 fiber transcript levels identified altered levels of genes and metabolites involved in ROS homeostasis and stress response (Hinchliffe et al. 2011; Naoumkina et al. 2013, 2014). ROS has been shown to be involved in cell elongation through loosening of cell walls in growing tissue and selective breeding for enhanced fiber growth seems to have influenced regulation of ROS related genes (Chaudhary et al. 2009; Liszkay et al. 2004). Furthermore, we recently found that Li_2 mutant fiber cells have altered gene expression in mitochondria, the site of much ROS production (Thyssen et al. 2014).

Our future work will attempt to modify the fiber phenotypes in cotton plants by modulating the activity of these candidate genes. We believe that both TIP and ZnF are plausible Li_2 candidates based on the current evidence. However, absent changes to coding sequences in these or the other genes in the vicinity of the Li_2 locus, the causative mutation is likely to affect a control sequence, which could potentially exert its influence on a distant gene (Clark et al. 2006; Guenther et al. 2014).

Author contributions GNT, DDF and MN conceived and designed the experiment. GNT analyzed the sequencing data, designed the SNP markers and wrote the paper. DDF analyzed the SSR data and oversaw the project. RBT developed the NILs and grew the F_2 populations. CF and PL conducted SNP and SSR marker analysis of the F_2 populations. MN performed the RT-qPCR experiments and analyzed the SNP marker data. All authors read and approved the manuscript.

Acknowledgments This project was financially supported by the USDA-ARS CRIS project # 6435-21000-0017D and Cotton Incorporated project # 12-210. We greatly appreciate the contributions of Dr. Xianliang Song in our group. We also appreciate Mrs. Sheron Simpson and Dr. Brian Scheffler at the Genomics and Bioinformatics Research Unit, USDA-ARS at Stoneville, MS, for their support in SSR marker analysis. 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 US Department of Agriculture that is an equal opportunity provider and employer.

Conflict of interest The authors declare that they have no conflict of interest.

References

Birkeland SR, Jin N, Ozdemir AC, Lyons RH, Weisman LS, Wilson TE (2010) Discovery of mutations in Saccharomyces cerevisiae by pooled linkage analysis and whole-genome sequencing. Genetics 186:1127–1137

Blake N, Sherman J, Dvořák J, Talbert L (2004) Genome-specific primer sets for starch biosynthesis genes in wheat. Theor Appl Genet 109:1295–1302

Blenda A, Fang DD, Rami J-F, Garsmeur O, Luo F, Lacape J-M (2012) A high density consensus genetic map of tetraploid cotton that integrates multiple component maps through molecular marker redundancy check. PLoS One 7:e45739

Blumenstiel JP, Noll AC, Griffiths JA, Perera AG, Walton KN, Gilliland WD, Hawley RS, Staehling-Hampton K (2009) Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. Genetics 182:25–32

Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet 124:1201–1214

Chai G, Bai Z, Wei F, King GJ, Wang C, Shi L, Dong C, Chen H, Liu S (2010) Brassica GLABRA2 genes: analysis of function related to seed oil content and development of functional markers. Theor Appl Genet 120:1597–1610

Chaudhary B, Hovav R, Flagel L, Mittler R, Wendel JF (2009) Parallel expression evolution of oxidative stress-related genes in fiber from wild and domesticated diploid and polyploid cotton (Gossypium). BMC Genom 10:378

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38:594–597

Davletova S, Schlauch K, Coutu J, Mittler R (2005) The zinc-finger protein Zat12 plays a central role in reactive oxygen and abiotic stress signaling in Arabidopsis. Plant Physiol 139:847–856

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiol 124:1483–1492

Fang DD, Xiao J, Canci PC, Cantrell RG (2010) A new SNP haplotype associated with blue disease resistance gene in cotton (Gossypium hirsutum L.). Theor Appl Genet 120:943–953

Galvão VC, Nordström KJ, Lanz C, Sulz P, Mathieu J, Posé D, Schmid M, Weigel D, Schneeberger K (2012) Synteny-based mapping-by-sequencing enabled by targeted enrichment. Plant J 71:517–526

Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM (2014) A molecular basis for classic blond hair color in Europeans. Nat Genet

Hinchliffe DJ, Turley RB, Naoumkina M, Kim HJ, Tang Y, Yeater KM, Li P, Fang DD (2011) A combined functional and structural genomics approach identified an EST-SSR marker with complete linkage to the Ligon lintless-2 genetic locus in cotton (Gossypium hirsutum L.). BMC Genom 12:445

Karlsson M, Johansson I, Bush M, McCann MC, Maurel C, Larsson C, Kjellbom P (2000) An abundant TIP expressed in mature highly vacuolated cells. Plant J 21:83–90

Kiełbowicz-Matuk A (2012) Involvement of plant C_2H_2 -type zinc finger transcription factors in stress responses. Plant Sci 185:78–85

Kohel R, Stelly D, Yu J (2002) Tests of six cotton (Gossypium hirsutum L.) mutants for association with aneuploids. J Hered 93:130–132

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S

(2009) The sequence alignment/map format and SAM tools. Bioinformatics 25:2078–2079

Li DD, Ruan XM, Zhang J, Wu YJ, Wang XL, Li XB (2013) Cotton plasma membrane intrinsic protein 2s (PIP2s) selectively interact to regulate their water channel activities and are required for fibre development. New Phytol 199:695–707

Liszkay A, van der Zalm E, Schopfer P (2004) Production of reactive oxygen intermediates (O_2^-, H_2O_2, and OH) by maize roots and their role in wall loosening and elongation growth. Plant Physiol 136:3114–3123

Liu D, Tu L, Wang L, Li Y, Zhu L, Zhang X (2008) Characterization and expression of plasma and tonoplast membrane aquaporins in elongating cotton fibers. Plant Cell Rep 27:1385–1394

Long Y, Wang Z, Sun Z, Fernando DW, McVetty PB, Li G (2011) Identification of two blackleg resistance genes and fine mapping of one of these two genes in a Brassica napus canola cultivar 'Surpass 400'. Theor Appl Genet 122:1223–1231

Michelmore RW, Paran I, Kesseli R (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci 88:9828–9832

Naoumkina M, Hinchliffe DJ, Turley RB, Bland JM, Fang DD (2013) Integrated metabolomics and genomics analysis provides new insights into the fiber elongation process in Ligon lintless-2 mutant cotton (Gossypium hirsutum L.). BMC Genom 14:155

Naoumkina M, Thyssen G, Fang DD, Hinchliffe DJ, Florane C, Yeater KM, Page JT, Udall JA (2014) The Li2 mutation results in reduced subgenome expression bias in elongating fibers of allotetraploid cotton (Gossypium hirsutum L.). PLoS One 9:e90830

Page JT, Gingle AR, Udall JA (2013) PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3: genes genomes. Genetics 3:517–525

Park W, Scheffler BE, Bauer PJ, Campbell BT (2010) Identification of the family of aquaporin genes and their expression in upland cotton (Gossypium hirsutum L.). BMC Plant Biol 10:142

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Rokhsar DS, Wang X, Schmutz J (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

Rong J, Pierce GJ, Waghmare VN, Rogers CJ, Desai A, Chee PW, May OL, Gannaway JR, Wendel JF, Wilkins TA (2005) Genetic mapping and comparative analysis of seven mutants related to seed fiber development in cotton. Theor Appl Genet 111:1137–1146

Ruan Y-L, Llewellyn DJ, Furbank RT (2001) The control of single-celled cotton fiber elongation by developmentally reversible gating of plasmodesmata and coordinated expression of sucrose and K+ transporters and expansin. Plant Cell Online 13:47–60

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen J-E, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6:550–551

Smart LB, Vojdani F, Maeshima M, Wilkins TA (1998) Genes involved in osmoregulation during turgor-driven cell expansion of developing cotton fibers are differentially regulated. Plant Physiol 116:1539–1549

Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J 74:174–183

Taliercio EW, Boykin D (2007) Analysis of gene expression in cotton fiber initials. BMC Plant Biol 7:22

Thyssen GN, Song X, Naoumkina M, Kim HJ, Fang DD (2014) Independent replication of mitochondrial genes supports the transcriptional program in developing fiber cells of cotton (Gossypium hirsutum L.). Gene 544:41–48

Van Ooijen J (2006) JoinMap 4 software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands

Wang Z, Zhang D, Wang X, Tan X, Guo H, Paterson AH (2013) A whole-genome DNA marker map for cotton based on the D-genome sequence of gossypium raimondii L. G3: genes genomes. Genetics 3:1759–1767

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Wenger JW, Schwartz K, Sherlock G (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. PLoS Genet 6:e1000942

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

Wurtzel O, Dori-Bachash M, Pietrokovski S, Jurkevitch E, Sorek R (2010) Mutation detection with next-generation resequencing through a mediator genome. PLoS One 5:e15628

Zuryn S, Le Gras S, Jamet K, Jarriault S (2010) A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics 186:427–430