Genomics 96 (2010) 369–376

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Transcriptome profiling of early developing cotton fiber by deep-sequencing reveals significantly differential expression of genes in a fuzzless/lintless mutant

Qin Qin Wang [a], Fei Liu [a], Xu Sheng Chen [b], Xiao Jie Ma [b], Hou Qing Zeng [a], Zhi Min Yang [a,*]

[a] Department of Biochemistry and Molecular Biology, College of Life Science, Nanjing Agricultural University, Nanjing, China
[b] Laboratory of Cotton Breeding, Industrial Crop Institute, Jiangsu Academy of Agricultural Sciences, Nanjing, China

ARTICLE INFO

Article history:
Received 30 July 2010
Accepted 31 August 2010
Available online 7 September 2010

Keywords:
Gossypium hirsutum
Fiber
Transcriptome
Deep-sequencing
MicroRNA target gene

ABSTRACT

Cotton fiber as a single-celled trichome is a biological model system for studying cell differentiation and elongation. However, the complexity of its gene expression and regulatory mechanism allows only marginal progress. Here, we report the high-throughput tag-sequencing (Tag-seq) analysis using Solexa Genome Analyzer platform on transcriptome of −2 to 1 (fiber initiation, stage I) and 2–8 (fiber elongation, stage II) days post anthesis (DPA) cotton (*Gossypium hirsutum*) ovules (wild type: WT; Xuzhou 142 and its mutant: fuzzless/lintless or *fl* M, in the same background). To this end, we sequenced 3.5–3.8 million tags representing 0.7–1.0 million unique transcripts for each library (WT1, WT2, M1, and M2). After removal of low quality tags, we obtained a total of 2,973,104, 3,139,306, 2,943,654, and 3,392,103 clean sequences that corresponded to 357,852, 280,787, 372,952, and 382,503 distinct tags for WT1, WT2, M1, and M2, respectively. All clean tags were aligned to the publicly available cotton transcript database (TIGR, http://www.tigr.org). About 15% of the distinct tags were uniquely mapped to the reference genes, and 31.4% of existing genes were matched by tags. The tag mapping to the database sequences generated 23,854, 24,442, 23,497, and 19,957 annotated genes for WT1, WT2, M1, and M2 libraries, respectively. Analyses of differentially expressed genes revealed the substantial changes in gene type and abundance between the wild type and mutant libraries. Among the 20 most differentially expressed genes in WT1/M1 and WT2/M2 libraries were cellulose synthase, phosphatase, and dehydrogenase, all of which are involved in the fiber cell development. Overall, the deep-sequencing analyses demonstrate the high degree of transcriptional complexity in early developing fibers and represent a major improvement over the microarrays for analyzing transcriptional changes on a large scale.

1. Introduction

Cotton fiber is a single-celled epidermal trichome. Development of fiber cells undergoes several distinct but successive morphological processes including fiber initiation, elongation, secondary wall solidification, and maturation. At molecular levels, the cellular differentiation is accompanied by multiple changes in transcription, post-transcription, and translation [1–3]. Significant progress has been made to understand genes controlling early fiber initiation and late development [4]. Several mutants with impaired leaf trichome and cotton fiber development have been characterized [5–11]. The cotton *TTG1* was identified as a trichome-initiated responsive gene because the *TTG1* mutant from *Arabidopsis* can be rescued by the cotton *TTG1*, suggesting that equivalent *TTG1* functionality mediates both *Arabidopsis* leaf trichome and cotton fiber development [12]. *Arabidopsis GLABROUS 1* (*GL 1*) coding MYB transcription factor has been identified as an essential gene for leaf trichome initiation [13], and the *GL1* mutant shows a trichome-less phenotype [14]. Overexpression of a putative cotton ortholog *GL1* encoding a similar MYB protein in tobacco induced supernumerary epidermal trichomes on cotyledons and other organs [15]. Other mutants that bear loss of function genes for trichome elongation also have been identified recently [7,16].

Recently, genome-wild analyses of gene expression provide insights into the mechanisms for fiber initiation and development [4,17–21]. Several groups used microarrays to profile a large number of gene expression at a particular stage of fiber development [2,3,10,22–24]. An early study using differentially expressed cDNA libraries revealed that 60 genes were expressed more abundantly in wild type ovules (cv. Xu-142, 5 DPA) than in its *fl* (fuzzless/lintless) mutant [6]. Shortly thereafter, 172 fiber specific genes from the same cultivar Xu-142 were found to be up-regulated in 10 DPA elongating fibers on microarrays [22]. Similarly, using spotted oligo-gene arrays, 91 genes were differentially expressed in 0 DPA ovules compared to 3 DPA; moreover, 30 (0 DPA) and 117 (3 DPA) genes were differentially

Abbreviations: NAC, (NAM/ATAF/CUC); ACC, Aminocyclopropane-1-carboxylic acid; SCP, Serine carboxypeptidase; FDH, Fiddlehead-like protein; MAPK, Mitogen-activated protein kinase; NAK, NF-κB-activating kinase; WBC, White/brown complex; PCR, Polymerase chain reaction; DGE, Digital gene expression.

* Corresponding author. Department of Biochemistry and Molecular Biology, College of Life Science, Nanjing Agricultural University, Nanjing 210095, China.
E-mail address: zmyang@njau.edu.cn (Z.M. Yang).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.08.009

Table 1
Categorization and abundance of tags. Clean tags are tags after filtering dirty tags (low quality tags) from raw data. Distinct tags are different kinds of tags. Unambiguous tags are the reminder clean tags after removing tags mapped to reference sequences from multiple genes.

Summary		WT1	WT2	M1	M2
Raw tag	Total	3664152	3749168	3567539	3761744
	Distinct tag	1048862	850107	996804	708544
Clean tag	Total number	2973104	3139306	2943654	3392103
	Distinct tag number	357852	280787	372952	382503
All tag mapping to gene	Total number	1661747	2235245	1597619	1498186
	Total % of clean tag	55.89%	71.20%	54.27%	44.17%
	Distinct tag number	82707	88891	79653	59622
	Distinct tag % of clean tag	23.11%	31.66%	21.36%	15.59%
Unique tag mapping to gene	Total number	898235	985062	857748	730904
	Total % of clean tag	30.21%	31.38%	29.14%	21.55%
	Distinct tag number	54756	55578	52706	39093
	Distinct tag % of clean tag	15.30%	19.79%	14.13%	10.22%
All tag-mapped genes	Number	42935	44361	42682	37490
	Percentage of reference genes	60.76%	62.77%	60.40%	53.05%
Unique tag-mapped genes	Number	23854	24442	23497	19957
	Percentage of reference genes	33.76%	34.59%	33.25%	28.24%
Unknown tag	Total number	1301264	893423	1333772	1878736
	Total % of clean tag	43.77%	28.46%	45.31%	55.39%
	Distinct tag number	274756	191566	292887	322453
	Distinct tag % of clean tag	76.78%	68.22%	78.53%	84.30%

expressed between the naked seed mutant (*N1N1*) and its isogenic line TM-1, respectively [10]. Furthermore, a global analysis of 32789 ESTs from the earlier stage (−3 to 3 DPA) fiber ovules of *G. hirsutum* revealed a significant increase in the percentage of genes encoding putative transcription factors such as MYB and WRKY, and genes coding proteins involved in auxin, brassinosteroid (BR), gibberellic acid (GA), abscisic acid (ABA), and ethylene signaling and other metabolism pathways [2,25]. These results indicate that genome-wide analysis of the cotton fiber model can assist functional annotation of genes associated with cotton fiber development. Indeed, microarray technology has provided rich information on gene expression models, but the technique becomes even more challenging when dealing with the more inventories of gene species and more complex genomes. Moreover, microarrays have several other inherent limitations, as such that the genes represented by unspecific probe sets and with low expression levels can not be reliably detected. Recently, the high-throughput or deep-sequencing technology has become a powerful tool to permit the concomitant sequencing of millions of signatures to the genome and identify specific genes and the abundance of gene expression in a sample tissue [26,27]. This approach has highlighted the benefits of providing more thorough qualitative and quantitative description of gene expression than the previous microarray-based assay.

Taking this advantage, we present here the first genome-wide analysis of the gene expression during the early two stages (−2 to 1 DPA, here defined as stage I, and 2–8 DPA, stage II) of cotton fiber development using massively parallel deep-sequencing developed by Solexa Illumina. As a result of the assay, we annotated thousands of signatures matching predicted genes and quantified the transcript abundance in the developing fibers. In addition, we have profiled the gene expression between the wild type (WT) and fuzzless/lintless mutant (*fl* mutant in the WT background) cotton ovules and found great changes in gene expression in the mutant. Because these data are based on the direct sequencing, they comprise a rich resource that will be used for annotation of any fiber genomic sequences produced in the future.

2. Results

2.1. Transcriptome profiling of ovule libraries

To obtain a global view of transcriptome relevant to cotton fiber development, we used Solexa Genome Analyzer to perform high-throughput tag-sequencing (Tag-seq) analysis on poly(A)-enriched RNAs from four cotton ovule libraries including two from cultivar Xuzhou 142 (WT1: stage I; WT2: stage II) and two from *fl* mutant in Xuzhou 142 background (M1: stage I; M2: stage II). We identified expressed genes in initial (−2 to 1 DPA) and elongating fibers (2–8 DPA), because at the two stages, the different sets of transcripts responsible for fiber initiation and elongation may be enriched. The total number of tags per library ranged from 3.5 to 3.8 million, and the number of tags entitled with distinct sequences ranged from 0.7 to 1.0 million (Table 1). The distribution of total and distinct tag counts over different tag abundance categories showed very similar tendencies for all libraries (Fig. 1). Among the distinct tags, less than 1% had the copy number higher than 100 counts, whereas 23% of the tags were present between 5 and 50 copies, and more than 75% of the transcripts were 2–5 copies. After removal of low quality tags, we obtained a total of 2,973,104, 3,139,306, 2,943,654, and 3,392,103 clean tags that corresponded to 357,852, 280,787, 372,952, and 382,503 distinct tags for WT1, WT2, M1, and M2, respectively (Table 1). The four data sets represent the expressed sequences or transcriptome for each library. Matching the tags to genes is an important step to annotate sequences and can reveal the molecular events behind the gene expression [28]. In this study, all clean tags were aligned to the reference cotton database (TIGR, http://www.tigr.org). Approximately 15% of the distinct tags can be uniquely mapped to the reference

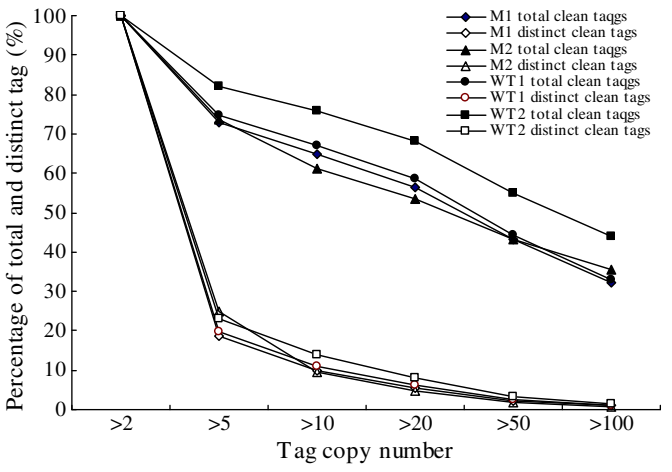

Fig. 1. Distribution of total tag (filled) and distinct tag (open) counts over different tag abundance categories from the four libraries.

sequences, and 31.4% of existing transcripts were matched by tags. Finally, the tags mapping to the database generated 23,854, 24,442, 23,497, and 19,957 tag-mapped transcripts for WT1, WT2, M1, and M2 libraries, respectively (Supplemental Table S1). However, a vast majority of distinct tags (about 77%) cannot be aligned to the reference genes due to the incomplete sequences.

In order to estimate whether the sequencing depth was sufficient for the transcriptome coverage, the sequencing saturation was analyzed for the four libraries. Genes mapped by unambiguous tags increased with the total number of tags. However, when the sequencing counts reached up to 1 M (1 M is equivalent to 10^6 tags) or higher, the number of detected genes was saturated (Supplemental Table S1). We also analyzed the distribution of ratio of distinct tag copy number in each pair of libraries and found that more than 98% distinct tags had a ratio within 5 folds (Supplemental Fig. S2). Antisense genes play an important role on gene expression and regulation. Sequencing tags mapped to the complementary strand of the sense gene suggest that its antisense strand has transcripts and that this gene may use bidirectional regulation. In this study, about 18.2, 15.3, 17.5, and 11.8% of the genes were matched by the distinct tags in antisense orientations (Supplemental Tables S2 and S3). In general, the tag frequencies were much lower on the antisense strand than on the sense strand.

2.2. Differential gene expression in the libraries

To profile the gene expression, we first mapped the unique tags from each library to the existing gene sequences. Analysis of the four libraries revealed 19,267, 20,702, 19,018, and 16,374 tag-mapped sense transcripts for WT1, WT2, M1, and M2, respectively (Supplemental Tables S1 and S3). The data sets implicate two stages corresponding to fiber initiation and elongation and two types of cell fibers (WT and *fl* mutant). There were 19,267 and 20,702 transcripts detected from WT1 (stage I) and WT2 (stage II), respectively. Among them 16,186 existed in both libraries, and 3076 and 4516 were present only in WT1 and WT2, respectively (Fig. 2). We also found a great change in the number of transcripts between WT1 and M1 or WT2 and M2. The varied transcripts between the two libraries suggest that the gene expression is most likely reprogrammed.

Differentially expressed genes between the two WT libraries or WT vs. M pairs give a clue to molecular events related to the fiber development. The expression abundance of tag-mapped genes in the data sets was analyzed by counting the number of transcripts per million (TPM) clean tags. We first normalized the read density measurement and then used FDR (false discovery rates) <0.001 and the absolute value of $|\log_2{}^{\text{Ratio}}| \geq 1$ as a threshold to judge the statistical significance of gene expression. From the four data sets, we found plenty of genes differentially expressed between two libraries (Supplemental Fig. S3). There are 2807 genes that were differently expressed between WT1 and WT2 (Table 2). Of these, 1708 were up-regulated and 1099 were down-regulated in expression; more genes were expressed preferentially in WT2 (296) than in WT1 (75), suggesting that many genes were enriched during the transition of fiber initiation to elongation. In contrast, 448 transcripts were detected in M1, while only 160 transcripts were found in M2; also, most of the genes (2730) were down-regulated in M2. While the total number of genes (572) in WT1 vs. M1 pair was much lower, the most amount of genes (4277) were differentially expressed in the WT2 vs. M2 pair. Compared to the WT1 vs. M1 data set, more genes from the WT2 and M2 pair were down-regulated in expression.

We presented the 20 most differentially expressed genes in WT1/M1 and WT2/M2 libraries, respectively. The relative abundance was expressed as a TPM ratio of WT1:M1 or WT2:M2. As shown in Table 3, the most differentially expressed genes in WT1/M1 are chalcone synthase 1, peptidase, and tRNA synthetase. Unexpectedly, most of the transcripts were found to be enzymes such as cellulose synthase, phosphatase, and dehydrogenase, all of which are most likely involved in the fiber cell development. There are some transcriptional regulator proteins (e.g. zinc finger and NAC domain-containing proteins) involving the early fiber initiation. In addition, several other transcripts coding specific and functional proteins such as outward rectifying potassium channel, ubiquitin, mitogen-activated kinase, and NAK-type protein kinase were found to be differentially expressed between the two libraries. We also identified many transcripts in WT2/M2 samples, which code transcript factors (e.g. MYB, NAC domain-containing protein, and WRKY transcription factor), enzymes related to fiber cell wall synthesis, hormone, and others (Table 3). Notably, all these transcripts were depressed in the mutants at stage II.

To test whether the transcription patterns had coverage of the well-defined genes, we compared our data sets to those that were reported previously. Table 4 illustrates 27 genes that are present in

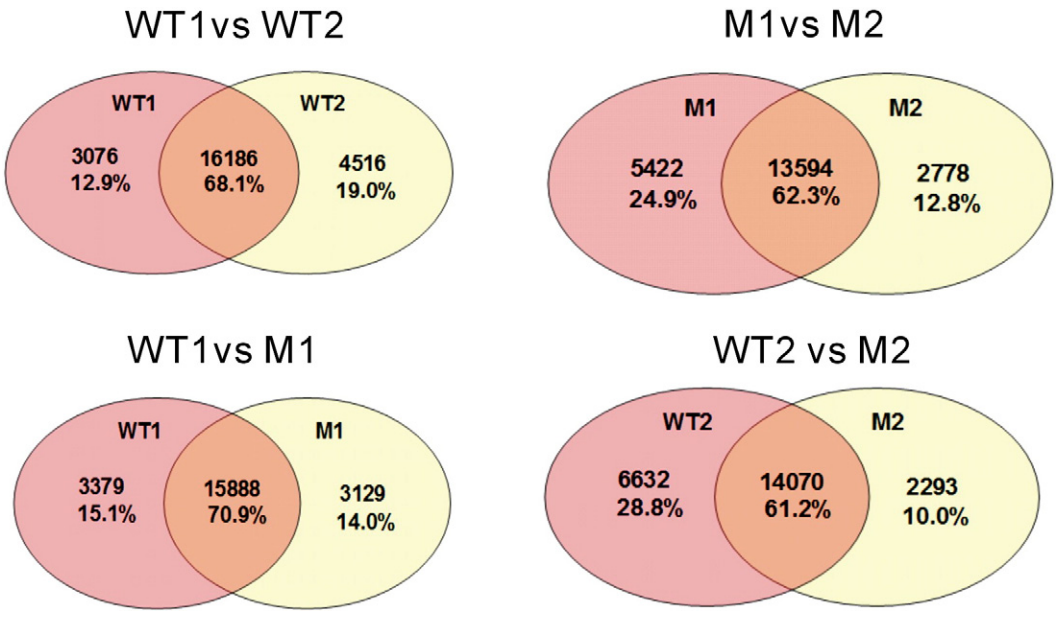

Fig. 2. Analysis of tag-mapped transcripts between two libraries. Venn diagram of the quantitative gene classification into those specifically expressed in one library or those expressed in both libraries.

Table 2
Differentially expressed genes between libraries. All the genes mapped to the genome were examined for their expression differences between the different libraries. Numbers of differentially expressed genes represent across sense transcripts, using threshold values FDR ≤ 0.001 and $|\log_2^{Ratio}| \geq 1$ for controlling false discovery rates.

	WT1 vs. WT2	M1 vs. M2	WT1 vs. M1	WT2 vs. M2
Total	2807	3571	572	4277
Up-regulated	1708	841	130	538
Down-regulated	1099	2730	442	3739
Expressed only in former	75	448	51	920
Expressed only in latter	296	160	15	84

our tag-mapped data set. These genes coding a variety of proteins or enzymes showed diverse expression abundance in the libraries. Several genes such as *GhFDH* (coding fiddlehead-like protein) [6,20,23] and *GhSCP* (coding serine carboxypeptidase precursor) [6,23] were expressed at higher levels in WT ovules, while their expression was depressed in the mutant. However, apart from the genes listed here, there were a few genes that have not been matched by our tag sequences. We further analyzed the relative expression abundance between the two libraries and found that 35 out of 108 library pairs showed significantly differential expression from one library to another (Supplemental Table S4). For instance, *GaMYB2* (coding a GL1-like MYB domain-containing protein) [8] and *GhMAPK* (mitogen-activated protein kinase) [22] were differentially expressed only in the WT2 vs. M2 library. However, genes in most of the library pairs showed no significantly differential expression.

We finally presented the 20 most abundant genes expressed in WT1 and WT2, along with their counterparts M1 and M2, respectively. The most abundant transcript expressed in both WT1 and WT2 was a peroxidase precursor (Supplemental Table S5). Transcripts such as flavonoid 3′,5′-hydroxylase, catalase, phospholipase D delta isoform 1b, and ovule/fiber cell elongation protein Ghfe1 were also found to be abundantly expressed in the two libraries. Interestingly, 11 genes were coordinately expressed in both WT1 and WT2, and only 9 distinct genes were preferentially expressed either in WT1 or WT2. The differences of gene expression levels were also found between WT and mutants, particularly between WT2 and M2.

Table 3
Top 20 most differentially expressed annotated genes between the WT1/M1 and WT2/M2 libraries based on the expressed tag frequency.

Gene ID	Relative abundance WT1/M1 (TPM ratio)	Functional annotation
DW506069	336	Chalcone synthase 1
DW235987	336	Peptidase
TA38279	336	tRNA synthetase
DR456425	269	Ubiquitin-like protein
TA38814	235	β-1,3-glucanase-like protein
DW240553	235	Short-chain dehydrogenase/reductase
TA38808	202	Putative lysophospholipase
TA24626	202	L-lactate dehydrogenase
TA24106	202	Microtubule-associated protein
DR458380	202	Serine/threonine-protein phosphatase
TA43990	168	Phosphatase
TA42092	168	7-keto-8-amino pelargonic acid synthase
TA40208	168	Cellulose synthase
TA29976	168	Alcohol dehydrogenase A
DT565671	168	NAK-type protein kinase
DR452673	168	Mitogen-activated protein kinase 1
DT558243	168	Zinc finger protein
TA43506	135	Histone acetyltransferase
TA41985	135	NAC domain-containing protein
TA32783	135	Outward rectifying potassium channel
Relative abundance WT2/M2 (TPM ratio)		
TA23503	18475	Long-chain fatty acid elongation enzyme
TA33135	4428	Amidase
TA38812	3026	3-ketoacyl-CoA synthase
AJ513871	2007	ACC oxidase 1
TA24393	1879	Short-chain dehydrogenase/reductase
TA21025	1625	Pectate lyase
CA992692	1529	Putative NAC domain protein NAC2
TA21760	1338	Glutamate decarboxylase
TA33302	892	P450 monooxygenase
DR455678	669	WRKY transcription factor 45
TA38623	637	MYB19
TA41859	542	Carbonic anhydrase
DW507252	542	Pectinesterase
DN827267	542	Chitinase
TA27591	510	Phosphoglycerate dehydrogenase-like protein
AI729907	478	Fiber lipid transfer protein
AI728825	478	2-oxoglutarate dehydrogenase
TA20105	446	Tubulin β-1 chain
DW503843	414	Putative glutamine synthetase
AI731980	382	Nonspecific lipid transfer protein precursor

2.3. Confirmation of tag-mapped genes by qRT–PCR

To confirm the tag-mapped genes in cotton ovules or developing fibers, nine genes were selected at random for quantitative RT–PCR assays. Expression of 6 genes (TA21337, AT22298, TA21774, TA20417, TA24519, and DT046968) by qRT–PCR fitted in well with the pattern of Tag-seq analysis (Fig. 3). Two genes (TA20379 and TA21004) behaved similarly between qRT–PCR and deep-sequencing methods. Only one gene (AY039663: predicted putative β-mannosidase) did not show consistent expression between qRT–PCR and Tag-seq data sets (Supplemental Fig. S4).

2.4. Analysis of tag-mapped genes potentially targeted by miRNAs

MicroRNAs may be involved in gene regulation of cotton fiber development [43,44]. Recently, we have deep-sequenced the small RNAs from the same ovules/fibers and found at least 111 conserved/known miRNAs representing 22 miRNA families [45]. In addition, 99 potential targets were predicted for most of the conserved miRNA families. Using the miRNA-target genes, comparison was made to the tag-mapped sequences from this study. We observed that 62 (62%) out of 99 genes potentially targeted by 43 miRNAs (18 families) can be matched by tag-mapped genes (Supplemental Table S6). We analyzed their expression change between the wild type and mutant libraries. Within the 82 miRNA-target pairs, 40 (~50%) were negatively related (Supplemental Table S7).

3. Discussion

A global analysis of transcriptome will facilitate identification of systemic gene expression and regulatory mechanisms for fiber development [10,20]. In this study, we performed transcriptome profiling of fiber-bearing and fiberless ovules to identify genes that are differentially expressed in initial and developing fibers. Using the tag-based deep-sequencing [46], we could obtain a direct digital readout of cDNAs and achieve an essentially dynamic range of genes from the libraries. Thus, the present study has represented the most comprehensive analyses of the cotton fiber transcriptome. Approximately 37,000–45,000 tag-mapped genes were identified for each library. This is very similar to the estimated number of genes from the cotton genome [47], suggesting that the number of tag-matched genes is nearly saturated. Unfortunately, sequencing of the upland cotton genome is incomplete, and many other genomic survey sequences may not be publicly available; there are still a large proportion of unique tags that cannot match with the cotton genomic sequences. These unmatched unique sequences represent the novel genes to be identified in the future.

Fiber development undergoes several consecutive processes throughout fiber initiation and development. Understanding of the temporally regulated genes will give insights into fiber initiation and

Table 4
Expression abundance of several fiber development related genes identified previously and in this study. TPM, transcripts per million clean tags.

Genes	Accession	TPM				Annotation	References
		M1	M2	WT1	WT2		
GhMYB25	TA23109	0.01	0.01	5.38	11.9	Transcription factor	[10,20,30]
GhMYB2	TA30693	10.53	20.34	10.43	3.82	MYB-like DNA-binding domain proteins (MYB2)	[23,31]
GhMYB3	TA36543	24.8	22.4	18.16	27.08	MYB-like DNA-binding domain proteins (MYB3)	[23,31]
GhMYB6	TA28979	26.5	22.99	24.55	45.55	MYB-like DNA-binding domain proteins (MYB6)	[23,31]
GhHD1	TA25306	136.9	28.3	179.61	68.49	Homeodomain-leucine zipper transcription factor	[20,24]
GhTUB1	TA20069	1.36	0.88	3.03	13.38	β tubulin 1	[6,23]
GhACT1	TA20417	48.24	36.56	37.33	113.08	Actin (ACT1)	[32]
GhKCH1	TA35995	1.7	0.01	1.01	0.64	Kinesin (KCH1)	[33]
GhKCBP	TA30985	19.36	8.25	18.5	14.33	Kinesin-like calmodulin binding protein	[22]
GhMAPK	TA26640	0.01	0.01	0.67	5.42	MAP kinase-like protein	[22]
GhRac1	TA30229	81.19	47.46	59.53	69.12	RAC-like G-protein Rac 1	[29]
GhCPK1	TA29723	26.16	16.21	24.22	36.95	Calcium-dependent protein kinase	[34]
GhSMT	TA21004	150.83	67.21	112.68	130.28	Sterol-C-methyltransferase	[6]
GhGlcAT1	TA21461	0.01	0.01	0.01	0.64	Glycuronosyltransferase-like protein	[20]
GhSuSy	TA20762	1.7	0.01	2.35	0.01	Sucrose synthase	[5,35,36]
GhCesA-5	TA21774	40.77	31.25	32.63	106.07	Cellulose synthase catalytic subunit (RWS1) (*A. thaliana*)	[6]
GhGA20ox1	TA38486	3.74	6.78	1.68	4.14	Gibberellin 20-oxidase	[37]
GhPEL	TA21025	0.01	0.01	0.01	16.25	Pectate lyase	[38]
GhPOX1	TA26510	2.38	2.06	0.67	8.28	Peroxidase	[39]
GhXTH1-1/2	TA20227	0.01	4.72	0.01	35.36	Xyloglucan endotransglycosylase	[40]
GhSCP	TA22298	97.84	81.66	142.95	408.69	Putative serine carboxypeptidase precursor	[6,23]
GhWBC1	TA24519	8.15	4.13	18.5	75.49	ABC transporter	[41]
ON035N9	TA25655	0.01	2.95	1.68	4.14	Fatty acid elongase	[24]
ON033M19	TA21267	17.67	52.18	65.59	597.58	Lipid transfer protein	[24]
GhEF1A	TA20426	3.06	0.01	1.68	2.23	Translation elongation factor 1A-5	[42]
GhRDL	TA20596	0.01	0.88	0.01	0.01	Dehydration-induced protein RD22-like protein	[6,20,23]
GhFDH	TA21337	6.11	7.08	9.42	117.86	Fiddlehead-like protein	[6,20,23]

elongation. At least 2807 genes (in WT1 vs. WT2 pair) were enriched individually at the two stages (Table 2). The differentially expressed genes are expected to function specifically at each defined phase. Our study showed that 1708 transcripts were highly expressed and 1099, down-regulated. Notably, 75 genes were expressed only in WT1, and 296 genes were preferentially expressed in WT2, suggesting that the expressed genes varied at stage I and stage II and that the fiber at these stages undergo a critical step of rapid transition of gene expression or regulation. More importantly, there were substantial genes whose expression was modulated in the mutant, and most of them (77.3% for M1/WT1 and 87.4% for M2/MT2) were down-regulated at the two stages (Table 2). These results indicate that expression of most fiber developmental genes is impaired by mutation.

The Tag-seq based assay provides an inventory of transcript population in fibers that can be sorted based on its abundance. We chose 20 genes that show the most differential expression in WT1/M1 and WT2/M2 pairs. The majority of the genes in WT1/M1 are involved in the transcriptional regulation, signaling, and metabolic pathways. These genes may be responsible for the fiber initiation and rapid extension. Mitogen-activated protein kinase (MAPK) and serine/threonine-protein phosphatase constitute the cascade of signaling pathway regulating many downstream biochemical events. Evidence was provided previously from the Xu-142 wild type in which the mRNA coding MAPK increased 51-fold higher in 10 DPA fiber cells than in 0 DPA ovules [22]. A similar result was obtained in this study that the gene was activated with the fiber cell development from early initiation to rapid elongation stage (Table 4). *GhFDH* was initially isolated from Xu-142 from cDNAs derived from ovules at 1, 3, 5, 9, 14, and 18 DPA and identified to encode a protein with a 75% amino acid sequence identity to *Arabidopsis* FDH protein [6]. The *Arabidopsis* FDH is a putative β-ketoacyl-CoA synthase which is involved in the biosynthesis of long-chain lipids in the cuticle [48]. It is constitutively expressed in the developing fiber cell, particularly at the late stage of fiber elongation. *GhFDH* was also sequenced from this study. Experimental validation with digital gene expression (DGE) (Table 4) and qRT–PCR (Fig. 3) showed an expression pattern similar to the previous report. *GhSCP* encodes a 507 amino acid protein with N-terminal transit peptide of 25 amino acid residues, and one of its protein functions is involved in brassinolide signaling and acting as acyltransferases in plant secondary metabolism [6]. These genes had relatively higher transcript abundance in developing fibers, particularly at the late stage of fiber development (Fig. 3). *GaMYB2* coding a cotton R2R3 MYB transcription factor has been well identified as a regulator of fiber development and predominantly expressed in early developing fiber cells [8]. *GaMYB2* is a functional homologue of *Arabidopsis* GLABRA1 (GL1), an essential component regulating *Arabidopsis* trichome formation. In this study, a *GhMYBL2* (coding MYB-like DNA-binding domain proteins) was detected from *G. hirsutum*. This gene may be also involved in the regulation of fiber development. We additionally analyzed the genes that were expressed most abundantly in WT1 and WT2. The majority of the genes are involved in cell wall solidification. Notably, expression of all the 20 genes from WT2/M2 sample was depressed, suggesting that expression of these genes is impaired in the mutant.

In our data set, at least 27 genes that have been isolated or characterized before were detected. Statistic analysis revealed that these genes were differentially expressed in the libraries. However, nine genes such as *GhKCH1* and *GhCBP* showed no significantly differential expression in any of the library pairs. Moreover, there are some genes that were reported previously but could not be recovered in this study (data not shown). The inconsistence may be the result of the following: (1) the present study used only TIGR database rather than NCBI to explore tag-mapped genes. As a result, some genes were most likely to be missed; (2) some transcripts or even highly transcribed genes are unmatched by any of the signatures, possibly due to the lack of CATG site in the sequences; (3) some distinct genes might be matched by the same tags and were possibly removed from the data sets; and (4) some data sets from this study and others cannot be comparable because transcripts from different DPA ovules or cultivars were used.

Previous studies demonstrated that PCR-based cDNA screening and microarrays identified a set of genes differentially expressed in the fiberless mutants [6,22]. Now, we found that more novel genes in

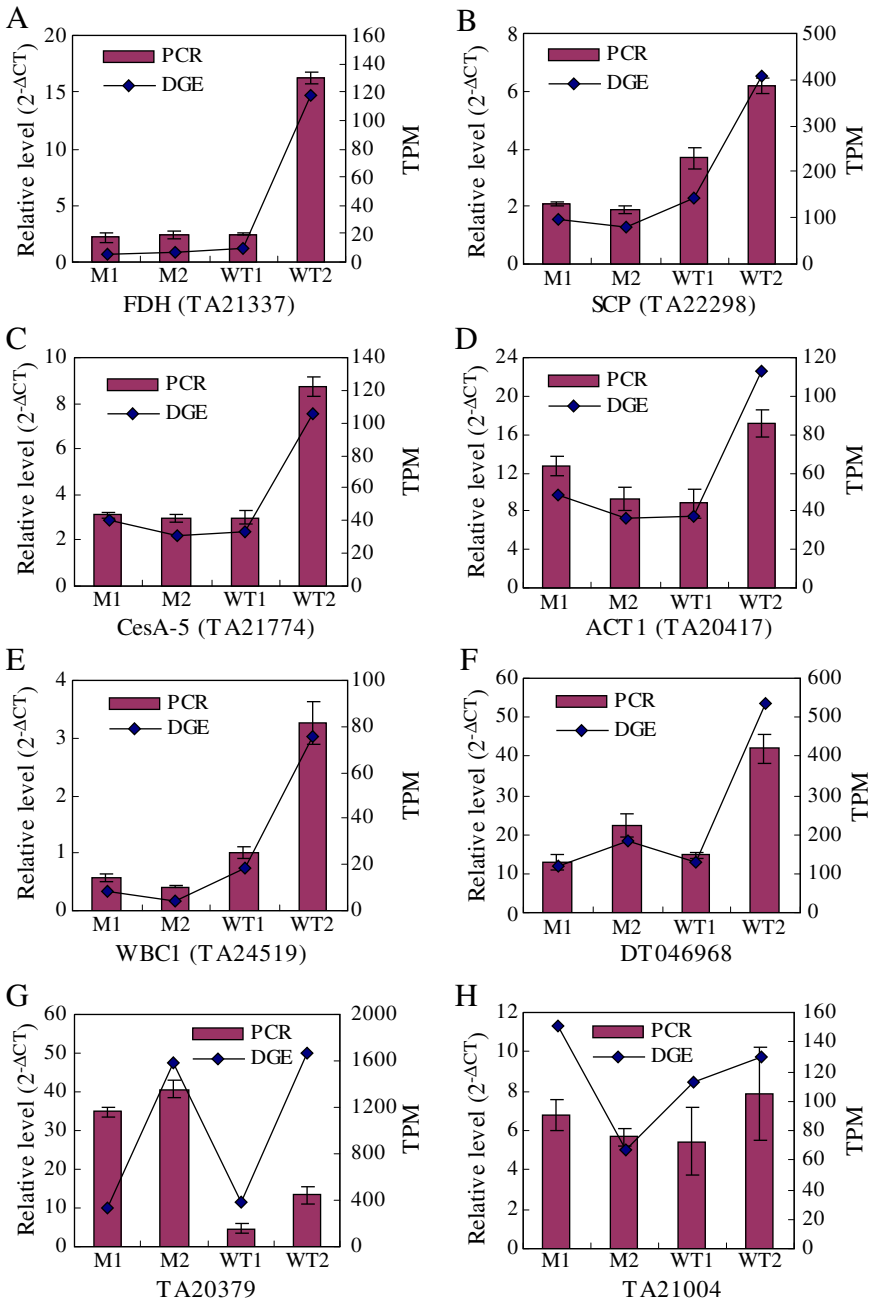

Fig. 3. Quantitative RT–PCR validation of tag-mapped genes from cotton ovules and developing fibers. (A)–(E), genes have been identified or reported before, including FDH (TA21337) (A), SCP (TA22298) (B), CesA-5 (TA21774) (C), ACT (TA20417) (D), and WBC1 (TA24519) (E). Reported functions and digital expression of these genes are also presented in Table 3. The other tag-mapped genes included DT046968 (F, predicted flavonoid 3′,5′-hydroxylase), TA20379 (G, predicted peroxidase), and TA21004 (H, sterol-C-methyltransferase). TPM, transcripts per million mapped reads.

the fiberless mutant were disturbed at the transcriptional level. One of the remarkable features relevant to gene expression is that most of the genes were down-regulated in expression in M1 and M2 than those in WT1 and WT2. Moreover, there were fewer genes that were specifically expressed in M1 and M2 than those in WT1 and WT2. These results support the idea that the fiberless mutants suffer from the loss of functional genes which are responsible for the fiber development. The precise loci impaired by the mutation remain to be elucidated, but our data show very low expression of several major genes in the mutant. Among these genes is *GhMYB25*, which is coding a putative R2R3 MYB protein. Using cDNA and oligo microarray techniques, *GhMYB25* was identified to be enriched in the fiber initial cells relative to the non-fiber ovular epidermal cells [10,20]. Recent study has demonstrated that *GhMYB25*-silenced cotton had altered the timing of rapid fiber elongation, short fibers, and reduced seed production, whereas ectopic overexpression of *GhMYB25* increased cotton fiber initiation and leaf trichome number [30]. Overexpression of *GhMYB25* increased branches of leaf trichomes in transgenic tobacco [20]. The present study showed that the mutant had a very low level of *GhMYB25* expression (Table 4). Possibly, *GhMYB25* loss of function contributes to the abnormal fiber phenotype. Several other genes such as *GhWBC1* were also expressed at low levels in M2 than WT2. The *GhWBC1* expression was found weak in the fiber cells of an *li* (*ligon-lintless*) mutant [41]. The down-regulated gene expression in the mutant is most likely responsible for the abnormal phenotype of cotton fibers. Finally, one transcript of interest was the peroxidase precursor (TA20379), which was identified in the fiber using deep-sequencing and qRT–PCR. Peroxidase catalyzes the conversion of H_2O_2

into water by oxidizing various hydrogen donor molecules such as phenolic compounds and auxin metabolites. The heme-containing group III peroxidases have complex relations with H_2O_2 [49]. In normal conditions, cell wall loosening or cross-linking depends on the status of growth and can be regulated by cell wall-linked peroxidases. Consumption of hydrogen peroxide by the active peroxidase should confer the cross-linking of several compounds and may facilitate fiber cell wall construction. Interestingly, several recent studies provided evidence that the reactive oxygen species, along with peroxidase, are involved in the mediation of the fiber cell initiation and development [42,50].

4. Materials and methods

4.1. RNA isolation

Upland cotton plants (*G. hirsutum* cv. Xuzhou 142; wild type, WT) with superior fiber quality and fuzzless/lintless (in Xuzhou 142 background; mutant, M) were field grown at the Jiangsu Agricultural Academy of Sciences, Nanjing. Flowers were tagged before the days of anthesis. The samples were harvested from −2 to 8 days post anthesis (DPA) ovules and dissected. The excised ovules were immediately frozen in liquid nitrogen and stored at −80 °C. Total RNA was isolated from −2, −1, 0, 1, 2, 3, 5, and 8 DPA wild type and mutant ovules. To profile the genes from fiber initials (here defined as stage I) and developing fibers (at stage II), four libraries were constructed. The WT1 (wild type at stage I) and M1 (mutant at stage I) libraries were generated from the original RNA pool derived from −2, −1, 0, and 1 DPA ovules, and the WT2 (wild type at stage II) and M2 (mutant at stage II) libraries were created from the original RNA pool from 2, 3, 5, and 8 DPA ovules.

Ovular RNA was extracted using column plant RNAout Kit (Tiandz) according to the manufacturer's instructions. Ten μg RNA samples were incubated for 15 min at room temperature with 1 unit of DNase I (Takara) to remove residual genomic DNA and incubated for 10 min at 65 °C with 1 μL of 20 mM EDTA for DNase I inactivation. One percent agarose gel (Biowest) buffered by Tris–acetate–EDTA was run to indicate the integrity of RNA. All RNA samples were quantified and examined for protein contamination (A260/A280 nm ratios) and reagent contamination (A260/A230 nm ratios) by a Nanodrop ND-1000 spectrophotometer. Extracted RNA samples were selected based on 28S/18S rRNA band intensity (2:1) and spectroscopic A260/A280 nm readings between 1.8 and 2.0, A260/A230 nm readings greater than 1.5.

4.2. Library construction and sequencing

The 3′-tag digital gene expression libraries were prepared using Illumina Gene Expression Sample Preparation Kit according to the manufacturer's instructions. Briefly, 6 μg total RNA was used for mRNA capture with magnetic oligo(dT) beads. The first and second cDNA strands were synthesized, and the stranded bead-bound cDNA was subsequently digested with *Nla*III. The 3′-cDNA fragments attached to the oligo(dT) beads were ligated to the Illumina GEX *Nla*III Adapter 1, which contained a recognition site for the endonuclease *Mme*I for cutting 17 bp downstream of the recognition site (CATG) to produce tags with adapter 1. After removing 3′ fragments with magnetic beads precipitation, an Illumina GEX adapter 2 was introduced at the site of *Mme*I cleavage. The resulting adapter-ligated cDNA tags were amplified using PCR-primers that were annealed to the adaptor ends for 15 cycles. The 85 base fragments were purified and recovered by 6% polyacrylamide Tris–borate–EDTA gel. The final quality of tagged sequences was checked by Agilent 2100 Bioanalyzer. The four constructed tag libraries underwent Illumina proprietary sequencing chip (flow cell) for cluster generation through situ amplification and were deep-sequenced using Illumina Genome Analyzer. Image analysis, base calling, and quality calibration were performed using the Solexa Automated Pipeline, after which the raw data (tag sequences and counts) were produced.

4.3. Data processing and digital tag profiling

Raw sequence reads were filtered by the Illumina pipeline. The 3′ adaptor sequence was removed from raw sequences. All low quality tags such as short tags (<21 nt), empty reads, and singletons (tags that occurred only once) were removed. The remaining high quality sequences were mapped to cotton TIGR reference sequences (TIGR, http://www.tigr.org) by SOAP. For monitoring the mapping events on both strands, the sense and antisense sequences were included in the data collection. We mapped all expressed tags onto a preprocessed database of 17 base-long sequences of TIGR located next to the *Nla*III restriction site, and only one mismatch was allowed. Tags mapped to more than one transcript were excluded from our analysis. When multiple types of tags were aligned to the different positions of the same gene, the gene expression levels were represented by the summation of all.

4.4. Quantitative RT–PCR analysis

To ensure maximum specificity and efficiency for PCR amplification of cDNAs under a standard set of reaction conditions, a stringent set of criteria was used for primer design, including predicted melting temperatures of 58 ± 2 °C, limited self-complementarity, a primer length between 18 and 24 nt, and PCR amplicon lengths of 50–150 bp. Total RNA was extracted from tissues using column plant RNAout Kit (Tiandz). The first cDNA strand was synthesized from 1.0 μg total RNA by Moloney murine leukemia virus reverse transcriptase (Promega) using oligo(dT) primers. The quantitative RT–PCR was performed with a MyiQ Single Color Real-Time PCR system (Bio-Rad) in a final volume of 20 μL containing 2 μL of a 1/10 dilution of cDNA in water, 10 μL of the 2× SYBR Premix Ex Taq (TaKaRa), and 200 nM of forward and reverse primers (Supplemental Table S8). The thermal cycling conditions were 40 cycles of 95 °C for 5 s for denaturation and 60 °C for 30 s for annealing and extension. All reactions were run in triplicate by monitoring the dissociation curve to control the dimers. PCR efficiency was determined by a series of 2-fold dilutions of cDNAs. The calculated efficiency of all primers was 0.9–1.0. Cotton gene gb polyubiquitin-2 (EE592464) was used as a normalizer, and the relative expression levels of genes were presented by $2^{-\triangle CT}$ ($\triangle$CT is the difference of CT between the control gb polyubiquitin-2 products and the target gene products).

4.5. Statistical evaluation

Statistical analysis was performed to identify differentially expressed genes between the libraries using a rigorous algorithm described previously [51]. The gene expression was normalized to transcripts per million clean tags. For gene expression variance, the statistical *t*-test was used to identify genes expressed between libraries. *P* values were adjusted by the multiple testing procedures described by Benjamini and Yekutieli [52], by controlling the false discovery rate (FDR). In this study, we used stringent value FDR<0.001 and the absolute value of $|\log_2{}^{Ratio}| \leq 1$ as the threshold to judge the significant difference of gene expression. The correlation of the detected count numbers between parallel libraries was statistically assessed by the calculation of Pearson correlation coefficient.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2010.08.009.

Acknowledgments

This research was financially supported by the National Basic Research Program of China, 863 Project of China (2008AA10Z128).

References

[1] C.H. Haigler, D.H. Zhang, C.G. Wilkerson, Biotechnological improvement of cotton fibre maturity, Physiol. Plant. 124 (2005) 285–294.
[2] S.S. Yang, et al., Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fibre cell development in allotetraploid cotton, Plant J. 47 (2006) 761–775.
[3] R. Hovav, et al., A majority of cotton genes are expressed in single-celled fibre, Planta 227 (2008) 319–329.
[4] J.J. Lee, A.W. Woodward, Z.J. Chen (2007) Gene expression changes and early events in cotton fibre development. Annual Bot. 100 (1984) 1391–1401.
[5] Y.L. Ruan, P.S. Chourey, A fibreless seed mutation in cotton is associated with lack of fibre cell initiation in ovule epidermis and alterations in sucrose synthase expression and carbon partitioning in developing seeds, Plant Physiol. 118 (1998) 399–406.
[6] C.H. Li, et al., Isolation of genes preferentially expressed in cotton fibres by cDNA filter arrays and RT–PCR, Plant Sci. 163 (2002) 1113–1120.
[7] M. Hülskamp, Plant trichomes: a model for cell differentiation, Nat. Rev. Mol. Cell. Biol. 5 (2004) 471–480.
[8] S. Wang, et al., Control of plant trichome development by a cotton fibre MYB gene, Plant Cell 16 (2004) 2323–2334.
[9] Y.L. Ruan, D.J. Llewellyn, R.T. Furbank, P.S. Chourey, The delayed initiation and slow elongation of fuzz-like short fibre cells in relation to altered patterns of sucrose synthase expression and plasmodesmata gating in a lintless mutant of cotton, J. Exp. Bot. 56 (2005) 977–984.
[10] J.J. Lee, et al., Developmental and gene expression analyses of a cotton naked seed mutant, Planta 223 (2006) 418–432.
[11] X.X. Shangguan, B. Xu, Z.X. Yu, L.J. Wang, X.Y. Chen, Promoter of a cotton fibre MYB gene functional in trichomes of *Arabidopsis* and glandular trichomes of tobacco, J. Exp. Bot. 59 (2008) 3533–3542.
[12] J.A. Humphries, A.R. Walker, J.N. Timmis, S.J. Orford, Two WD-repeat genes from cotton are functional homologues of the *Arabidopsis thaliana* TRANSPARENT TESTA GLABRA1 (TTG1) gene, Plant Mol. Biol. 57 (2005) 67–81.
[13] O.G. Oppenheimer, P.L. Hermn, S. Sivakumaran, J. Esch, D.M. Marks, A myb gene required for leaf trichome differentiation in *Arabidopsis* is expressed in stipules, Cell 67 (1991) 483–493.
[14] J.C. Larkin, D.G. Oppenheimer, A.M. Lloyd, E.T. Paparozzi, M.D. Marks, Roles of the GLABROUS1 and TRANSPARENT TESTA GLABRA genes in *Arabidopsis* trichome development, Plant Cell 6 (1994) 1065–1076.
[15] T. Payne, J. Clement, D. Arnold, A. Lloyd, Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in *Nicotiana tabacum*, Development 126 (1999) 671–682.
[16] S. Schellmann, M. Hülskamp, Epidermal differentiation: trichomes in *Arabidopsis* as a model system, Int. J. Dev. Biol. 49 (2005) 579–584.
[17] A.B. Arpat, et al., Functional genomics of cell elongation in developing cotton fibres, Plant Mol. Biol. 54 (2004) 911–929.
[18] T.A. Wilkins, A.B. Arpat, The cotton fibre transcriptome, Physiol. Plant. 124 (2005) 295–300.
[19] Y.H. Shi, et al., Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fibre cell elongation, Plant Cell 18 (2006) 651–664.
[20] Y. Wu, A.C. Machado, R.G. White, D.J. Llewellyn, E.S. Dennis, Expression profiling identifies genes expressed early during lint fibre initiation in cotton, Plant Cell Physiol. 47 (2006) 107–127.
[21] W.L. Xu, X.L. Wang, H. Wang, X.B. Li, Molecular characterization and expression analysis of nine cotton GhEF1A genes encoding translation elongation factor 1A, Gene 389 (2007) 27–35.
[22] S.J. Ji, et al., Isolation and analyses of genes preferentially expressed during early cotton fibre development by subtractive PCR and cDNA array, Nucleic Acids Res. 31 (2003) 2534–2543.
[23] E.W. Taliercio, D. Boykin, Analysis of gene expression in cotton fibre initials, BMC Plant Biol. 7 (2007) 22.
[24] Y. Wu, et al., Laser capture microdissection and cDNA microarrays used to generate gene expression profiles of the rapidly expanding fibre initial cells on the surface of cotton ovules, Planta 226 (2007) 1475–1490.
[25] L. Li, X.L. Wang, G.Q. Huang, X.B. Li, Molecular characterization of cotton GhTUA9 gene specifically expressed in fibre and involved in cell elongation, J. Exp. Bot. 58 (2007) 3227–3238.
[26] A. Iandolino, K. Nobuta, F.G. da Silva, D.R. Cook, B.C. Meyers, Comparative expression profiling in grape (*Vitis vinifera*) berries derived from frequency analysis of ESTs and MPSS signatures, BMC Plant Biol. 8 (2008) 53.
[27] P. Vizoso, et al., Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with peach fruit quality, BMC Genomics 10 (2009) 423.
[28] Z. Hegedus, et al., Deep sequencing of the zebrafish transcriptome response to mycobacterium infection, Mol. Immunol. 46 (2009) 2918–2930.
[29] H.J. Kim, B.A. Triplett, Cotton fibre growth in planta and in vitro: models for plant cell elongation and cell wall biogenesis, Plant Physiol. 127 (2001) 1361–1366.
[30] A. Machado, Y. Wu, Y. Yang, D.J. Llewellyn, E.S. Dennis, The MYB transcription factor GhMYB25 regulates early fibre and trichome development, Plant J. 59 (2009) 52–62.
[31] L.L. Loguercio, J.Q. Zhang, T.A. Wilkins, Differential regulation of six novel MYB-domain genes defines two distinct expression patterns in allotetraploid cotton (*Gossypium hirsutum* L.), Mol. General Genetics 261 (1999) 660–671.
[32] X.B. Li, X.P. Fan, X.L. Wang, L. Cai, W.C. Yang, The cotton ACTIN1 gene is functionally expressed in fibres and participates in fibre elongation, Plant Cell 17 (2005) 859–875.
[33] M.L. Preuss, et al., A plant-specific kinesin binds to actin microfilaments and interacts with cortical microtubules in cotton fibres, Plant Physiol. 136 (2004) 3945–3955.
[34] Q.S. Huang, H.Y. Wang, P. Gao, G.Y. Wang, G.X. Xia, Cloning and characterization of a calcium dependent protein kinase gene associated with cotton fibre development, Plant Cell Rep. 27 (2008) 1869–1875.
[35] K.D. Notle, D.L. Hendrix, J.W. Radin, K.E. Koch, Sucrose synthase localization during initiation of seed development and trichome differentiation in cotton ovules, Plant Physiol. 109 (1995) 1285–1293.
[36] Y.L. Ruan, D.J. Llewellyn, R.T. Furbank, Suppression of sucrose synthase gene expression represses cotton fibre cell initiation, elongation, and seed development, Plant Cell 15 (2003) 952–964.
[37] Y.H. Xiao, et al., Gibberellin 20-oxidase promotes initiation and elongation of cotton fibres by regulating gibberellin synthesis, J. Plant Physiol. 167 (2010) 829–837.
[38] H.H. Wang, et al., The essential role of GhPEL gene, encoding a pectate lyase, in cell wall loosening by depolymerization of the de-esterified pectin during fibre elongation in cotton, Plant Mol. Biol. 72 (2010) 397–406.
[39] W. Mei, Y. Qin, W. Song, J. Li, Y. Zhu, Cotton *GhPOX1* encoding plant class III peroxidase may be responsible for the high level of reactive oxygen species production that is related to cotton fibre elongation, J. Genet. Genomics 36 (2009) 141–150.
[40] G. Michailidis, A. Argirioua, N. Darzentas, A. Tsaftaris, Analysis of xyloglucan endotransglycosylase hydrolase (XTH) genes from allotetraploid (*Gossypium hirsutum*) cotton and its diploid progenitors expressed during fibre elongation, J. Plant Physiol. 166 (2008) 403–416.
[41] Y.P. Zhu, K.X. Xu, B. Luo, J.W. Wang, X.Y. Chen, An ATP-binding cassette transporter GhWBC1 from elongating cotton fibres, Plant Physiol. 133 (2003) 580–588.
[42] Z. Xu, et al., Gene-rich islands for fibre development in the cotton genome, Genomics 92 (2008) 173–183.
[43] M. Pang, et al., Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fibre development in allotetraploid cotton (*Gossypium hirsutum* L.), Genome Biol. 10 (2009) R122.
[44] M.B. Ruan, Y.T. Zhao, Z.H. Meng, X.J. Wang, W.C. Yang, Conserved miRNA analysis in *Gossypium hirsutum* through small RNA sequencing, Genomics 94 (2009) 263–268.
[45] P.B. Kwak, Q.Q. Wang, X.S. Chen, C.X. Qiu, Z.M. Yang, Enrichment of a set of microRNAs during the cotton fibre development, BMC Genomics 10 (2009) 457.
[46] A.S. Morrissy, et al., Next-generation tag sequencing for cancer gene expression profiling, Genome Res. 19 (2010) 1825–1835.
[47] P.D. Rabinowicz, et al., Differential methylation of genes and repeats in land plants, Genome Res. 15 (2005) 1431–1440.
[48] R.E. Pruitt, J.P. Vielle-Calzada, S.E. Ploense, U. Grossniklaus, S.J. Lolle, FIDDLEHEAD, a gene required to suppress epidermal cell interactions in *Arabidopsis*, encodes a putative lipid biosynthetic enzyme, Proc. Natl Acad. Sci. USA 97 (2000) 1311–1316.
[49] Y.J. Xue, L. Tao, Z.M. Yang, Aluminum-induced cell wall peroxidase activity and lignin synthesis are differentially regulated by jasmonate and nitric oxide, J. Agri Food Chem. 56 (2008) 9676–9684.
[50] H.B. Li, et al., A cotton ascorbate peroxidase is involved in hydrogen peroxide homeostasis during fibre cell development, New Phytol. 175 (2007) 462–471.
[51] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1997) 986–995.
[52] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Stat. 29 (2001) 1165–1188.