RESEARCH ARTICLE

Microarray-based large scale detection of single feature polymorphism in *Gossypium hirsutum* L.

ANUKOOL SRIVASTAVA, SAMIR V. SAWANT and SATYA NARAYAN JENA*

Plant Molecular Biology Laboratory, CSIR-National Botanical Research Institute, Lucknow 226 001, India

Abstract

Microarrays offer an opportunity to explore the functional sequence polymorphism among different cultivars of many crop plants. The Affymetrix microarray expression data of five genotypes of *Gossypium hirsutum* L. at six different fibre developmental stages was used to identify single feature polymorphisms (SFPs). The background corrected and quantile-normalized $\log_2$ intensity values of all probes of triplicate data of each cotton variety were subjected to SFPs call by using SAM procedure in R language software. We detected a total of 37,473 SFPs among six pair genotype combinations of two superior (JKC777 and JKC725) and three inferior (JKC703, JKC737 and JKC783) using the expression data. The 224 SFPs covering 51 genes were randomly selected from the dataset of all six fibre developmental stages of JKC777 and JKC703 for validation by sequencing on a capillary sequencer. Of these 224 SFPs, 132 were found to be polymorphic and 92 monomorphic which indicate that the SFP prediction from the expression data in the present study confirmed a ∼58.92% of true SFPs. We further identified that most of the SFPs are associated with genes involved in fatty acid, flavonoid, auxin biosynthesis etc. indicating that these pathways significantly involved in fibre development.

[Srivastava A., Sawant S. V. and Jena S. N. 2015 Microarray-based large scale detection of single feature polymorphism in *Gossypium hirsutum* L. *J. Genet.* **94**, 669–676]

Introduction

Genetic mapping and marker-assisted breeding in any crop requires a best marker system with increased throughput, decreased cost per data-point and greater map resolution (Kumar *et al*. 2007; Gupta *et al*. 2008). In recent years, molecular marker technology in higher plants has shifted from the random DNA markers (RDMs) that developed in the past arbitrarily from genomic DNA to the molecular markers representing the coding sequences. Microarrays-based molecular markers commonly called as single feature polymorphisms (SFPs) that detected by hybridization of DNA or cRNA to oligonucleotide probes is one of them, mostly used to identify genetic polymorphisms (Hazen *et al.* 2005).

Several methods have been reported for SFP discovery in a variety of plant species. All the methods are based on the idea that variation in a target sequence lowers the probe hybridization signal intensity on an array. However, there are differences in methodologies to detect this decreased intensity. Winzeler *et al.* (1998) first reported SFP detection in yeast (*Saccharomyces cerevisiae*) genomic DNA was hybridized on a high-density oligonucleotide expression array in which a linear regression model is used to fit probe log(PM) intensities and then followed by F-test to detect the probes that have significant differences in intensities between two yeast strains. The same method has been tested on genomic and transcriptome data of barley (*Hordeum vulgare* L.) (Cho *et al.* 2006). Besides, there are also several other methods like significance analysis of microarray (SAM) (Tusher *et al*. 2001), robustified projection pursuit (RPP) (Cui *et al.* 2005) linear model by using SAS/JMP (Ronald *et al.* 2005), positional-dependent-nearest-neighbouring model (PDNN) (Zhang *et al.* 2003) and probe affinity outlier pursuit (PAOP) algorithm (Cui *et al.* 2005). West *et al.* (2006) proposed a new method to detect SFPs in *Arabidopsis thaliana* by examining probe intensity among all the 11 probes of a probe set in parallel and calculating a value called SFPdev, which is the hybridization signal difference between one probe and the average of the other 10 probes divided by the individual probe signal. Thus, SFPs based on expressed sequences are an efficient source of large number of genic markers to moderate-throughput and usually nongenic marker systems such as RAPDs, AFLPs and SSRs.

The draft genome of tetraploid cotton (*Gossypium hirsutum*) has just been published, overcoming the formidable task of quality assembly due to highly homeologous subgenomes

*For correspondence. E-mail: satyanarayan@nbri.res.in.

Keywords. cotton; microarray; single feature polymorphism; single-nucleotide polymorphism; statistical analysis of microarray.

(Li *et al.* 2015; Zhang *et al.* 2015). The sequencing initiative of *G. hirsutum* was awaiting for a long time due to the huge genome size (2.5 Gb) and large amount of (~80%) repetitive sequences (Hendrix and Stewart 2005). Species *G. hirsutum* is allotetraploid with 52 chromosomes consisting of 13 homoeologous chromosomes from each of two ancestral genomes (A, D). Although, upland cotton (*G. hirsutum* L.) is a major world fibre crop, there is no report yet on SFP discovery in cotton. Microarray analyses of various plant species have provided a fundamental platform for the discovery of genic markers which can be finally identified for candidate genes governing the trait. A number of various studies have reported the use of oligonucleotide arrays, for large number of genomewide functional SFP prediction, in wheat between near-isogenic lines contrasting in stripe rust resistance using the Affymetrix GeneChip (Coram *et al.* 2008), in maize (Gore *et al.* 2007), in cowpea using soyabean genome array (Das *et al.* 2008), in rice (Kumar *et al.* 2007) and in barley (Rostoks *et al.* 2005).

The objective of the present study was to identify the genomewide SFPs by using inhouse microarray expression data of *G. hirsutum* genotypes. Here we demonstrate that hybridization of cRNA to cotton whole genome expression array (Affymetrix, Santa Clara, USA) is quite sensitive in predicting SFPs prior of their sequence information. The statistical algorithm presented here allowed us to distinguish genotype-dependent hybridization differences at the probe level and identified several thousand SFPs in five genotypes that used in this study.

Materials and methods

Plant materials used and experimental conditions

Five genotypes of *G. hirsutum* (two superior JKC777 and JKC725 and three inferior JKC703, JKC737 and JKC783, with respect to fibre quality) are used in this study. These cotton varieties are parents of two mapping population developed by JK Agri Genetics, Hyderabad, India.

Total RNA isolation from cotton fibres and microarray hybridization

Fibres were manually striped from the ovules (freshly harvested) of different developmental time points. Additionally, ovules as well as fibres were inspected visually for any cell damage or other contaminating tissue. Total RNA was extracted using the Spectrum plant total RNA Kit (Sigma St Louis, USA) according to the manufacturer's instructions. All the samples were quantified on a Nanodrop spectrophotometer and electrophoresed on 1% agarose gel to test the integrity and purity. Elution was done with nuclease-free water after DNaseI treatment (Ambion, Thermo Fisher Scientific, USA), quantified and checked for integrity using a Bioanalyzer 2100 (Agilent, Palo Alto, USA). Direct labelling procedure was used with 1 μg of total RNA sample, double-stranded cDNA was synthesized with a T7 promoter-containing oligo (dT) primer using a GeneChip one-cycle cDNA synthesis kit (Affymetrix, Santa Clara, USA), followed by *in vitro* transcription using a GeneChip IVT labelling kit (Affymetrix, Santa Clara, USA). cRNA was fragmented for hybridization to Affymetrix cotton genome arrays, incubated at 50°C temperature for 16 h at 60 rpm in hybridization oven. Arrays were washed and stained on an Affymetrix Fluidics Station 450. Then they were scanned using GeneChip Scanner 3000. A summary of the image signal data for every gene interrogated on the array was generated using the Affymetrix statistical MAS 5.0 (GCOS v1.3) algorithm.

Microarray quality control and data processing

The scanned image of each GeneChip was visually inspected for artifacts and standard quality control parameters were checked in accordance with the manufacturer's recommendations (GeneChip Expression Analysis Data Analaysis Fundamentals; www.affymetrix.com). For each image signal, the mean value of three biological repeats was taken. Signals lower than 2000 were seen as low expression and were omitted from further statistical analysis. The microarray experiments on cRNA hybridization were conducted in replicates for all accessions for the reproducibility analysis.

SFP prediction in gene expression data

All statistical analyses of '.CEL' files were performed using the Array-assist software 5.2.2 (Agilent Technologies, Santa Clara, USA). Raw '.CEL' files were background corrected and normalized. Subsequently, only the 11 perfect match (PM) features from each of 263,551 probe sets were fit with the linear model:

$$\log(Ytgrp) = \mu + \text{tissue} + \text{genotype} + \text{genotype} \times \text{tissue} + \text{probe} + \text{error}.$$

Where Y is the background corrected normalized intensity of t, g, r and p in a probe set. μ is the mean probe intensity, while tissue has six developmental stages. A false discovery rate (FDR) of 0% median FDR was employed to control the familywise error rate. The background corrected and normalized log2 intensity values of all PM features of triplicate data of five cotton genotypes were subjected to SFPs call by using the MeV4.1 package (www.tm4.org/mev.html) and SAM (Saeed *et al.* 2003). SAM procedure allows the users to choose the delta value, a threshold for the SAM d-statistics, so as to get a balanced number of significant genes or SFPs as in the present study with a tolerable FDR estimated by 100 permutation. In terms of accuracy, the median FDR was superior to the mean FDR when the proportion of differentially expressed genes was large. Thus, in the present study, median FDR was taken into consideration for the significance SFP calls.

SFP confirmation and their functional gene annotation

Primer pairs for 51 genes were designed using Primer3 software targeting 224 SFP probes and synthesized. All the 51 genes were amplified using each genomic DNA as template. Subsequently, the same was gel-eluted and sequenced in ABI DNA Analyzer 3730xl (ABI, USA) following standard protocol. All sequences generated for all probe sequences were included for comparison between two different genotypes. Polymorphisms were called for SNP in aligned file of same gene with manual examination. Probe with multiple SNPs were allocated to a single group. In addition to these SNPs, insertions and deletions were scored as polymorphisms. Besides the sequence polymorphism confirmation, all the unique genes comprising targeted SFPs were annotated with TAIR10 protein database and the same was subjected to KOBAS 2.0 (http://kobas.cbi.pku.edu.cn/home.do) to know the significant pathways with $P \leq 0.05$.

Results

Hybridization and data quality

The five genotypes such as JKC703, JKC725, JKC737, JKC777 and JKC783 were used in the present study. To assess our microarray intensity data, the raw intensity data of only PM probes/features of cotton cultivars were log2 transformed and studied by density plots, and pairwise scattered plot (Borevitz *et al.* 2007) for 10,000 randomly selected probes. The results showed no major deviations among biological replicates of cotton cultivars and biological replicates were highly correlated (see figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Thus, the probes falling above or below diagonal lines indicate their differential hybridization and hence considered as SFPs (figure 1). The number of such features showing differential hybridization was least in JKC777 and JKC783 when compared with other pairs, while JKC725 and JKC783 showed maximum SFPs (table 1).

SFP identification from cotton GeneChip expression data

Gene expression data of two superior (JKC777 and JKC725) and three inferior (JKC703, JKC737 and JKC783) cotton genotypes was generated in six different fibre developmental stages. We compared superior JKC777 and JKC725 with JKC703, JKC737 and JKC783 in pairwise combination for SFP discovery. The resulting data matrix of 24,132 probe sets with 11 PM probes were analysed individually for six different developmental stages for SFP prediction using MeV 4.1 software and SAM procedure. Based on 10% S0 and 0% FDR (table 1 in electronic supplementary material), we selected a total of 4202 SFPs in JKC777 and JKC703 pair: 752 in 0DPA, 375 in 6DPA, 620 in 9DPA, 328 in 12DPA, 1741 in 19DPA and 386 in 25DPA at 0% estimated median FDR. Similarly, 3531 SFPs for JKC777 and JKC737 pair; 2095 SFPs for JKC777 and JKC783 pair; 7767 SFPs for JKC725 and JKC703; 9182 SFPs for JKC725 and JKC737 pair and 10696 SFPs for JKC725 and JKC783 pair were identified at ~0% median FDR. Thus, a total of 37,473 SFPs were resulted in six different fibre developmental stages in six pair cotton genotype combinations (table 1; table 2 in electronic supplementary material). Among the six cotton genotype-pair combination, the pair JKC725 and JKC783 showed maximum number of significant SFPs (figure 2). Subsequently, we were interested to study the SFP distribution across all the probes for genes. We identified that the SFPs were distributed almost equally in all the 11 probes and there was no significant enrichment of SFPs as per the probe distribution (table 2).

Confirmation of SFPs by sequencing

To investigate whether SFPs were due to sequence polymorphism in the probes, we decided to sequence genes showing SFPs between JKC703 and JKC777. We randomly selected 51 genes showing 224 SFPs, these regions were amplified and sequenced. All the 51 genes amplified in both genotypes showed polymorphism, thus confirming that SFPs were actually due to sequence polymorphism between genotypes (table 3; figure 3). In addition to these SNPs, there were 10 indels (insertion/deletion) in 10 gene sequences (figure 3).

Different biological pathways operating in superior and inferior genotypes

Following which we were interested in identifying the genes and their representative metabolic pathways in which SFPs were found. Thus, genes representing SFPs for each pair was analysed for the identification of biological pathways using KOBAS software (http://kobas.cbi.pku.edu.cn/home.do). It was noteworthy that many metabolic/biological pathways like flavone and flavonol biosynthesis, elongation and metabolism of fatty acid, glutathione metabolism, cutin, suberin and wax biosynthesis, starch, sucrose and tryptophan metabolism were most significant in JKC777 and JKC703 (figure 4) and JKC725 and JKC703 (figure 2 in electronic supplementary material). Interestingly, the most significant pathways in JKC777 and JKC737 (figure 3 in electronic supplementary material) and JKC725 and JKC737 (figure 4 in electronic supplementary material) were fatty acid elongation, phagosome, snare interaction, ribosome, glycolysis/gluconeogenesis and tyrosine metabolism. Whereas in the case of JKC777 and JKC783 (figure 5 in electronic supplementary material) and JKC725 and JKC783 (figure 6 in electronic supplementary material) genotypes fatty acid metabolism, cutin suberin and wax biosynthesis, flavonoid biosynthesis, starch and sucrose biosynthesis, tryptophan metabolism, vitamin B6 and carbon fixation in photosynthetic organism were found to be most significant pathways.

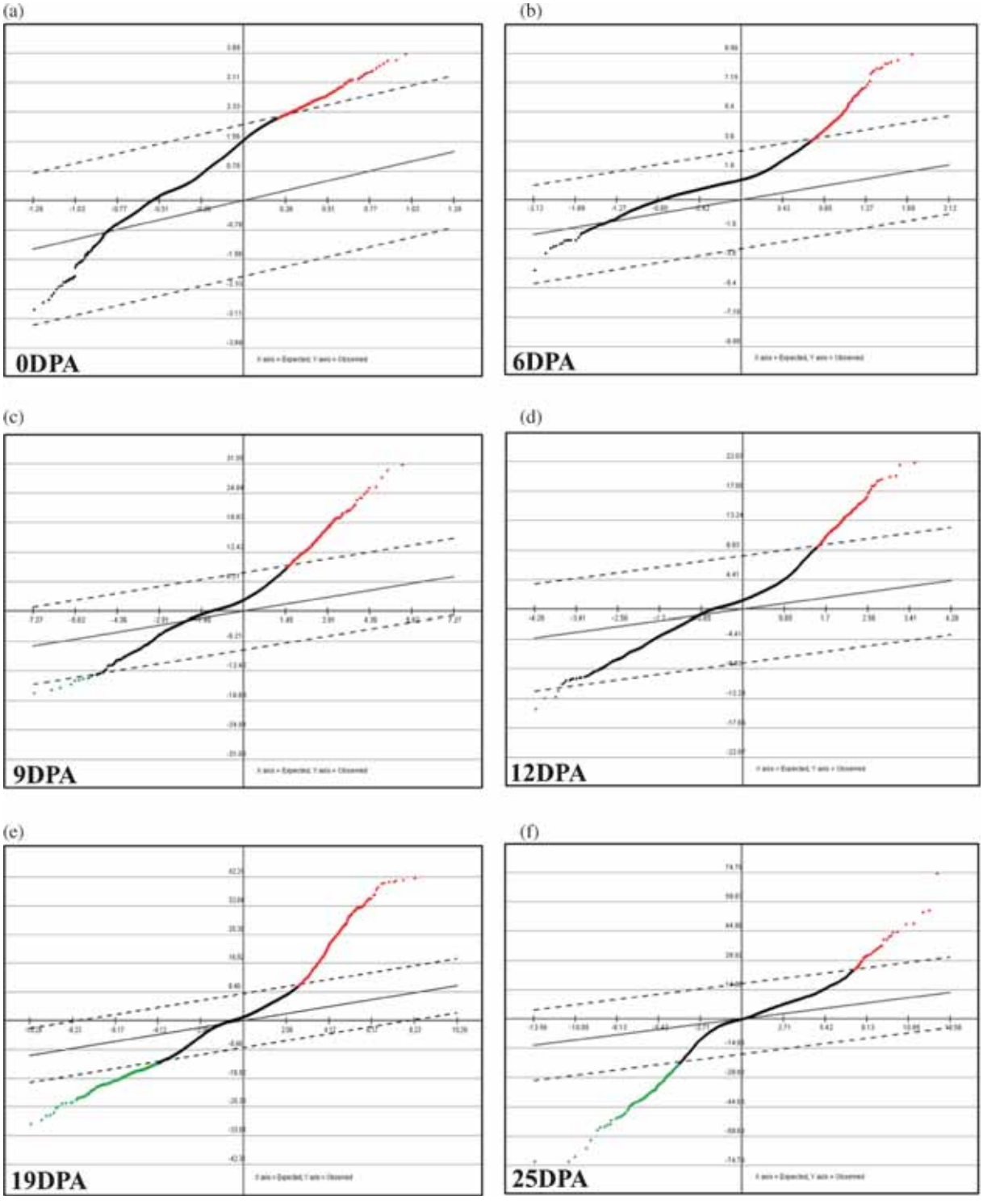

Figure 1. SAM plot of normalized data for JKC777 and JKC703: (a) 0DPA, (b) 6DPA, (c) 9DPA, (d) 12DPA, (e) 19DPA and (f) 25DPA. Observed d-statistics (*y*-axis) is plotted against the expected d-statistics (*x*-axis) as determined by permutations and SFPs exceeding the threshold are shown in green/red. The sign (+/−) with SFPs indicates direction of polymorphism. In the (−) sign (i.e. JKC777-SFP) indicates polymorphism in JKC703 (i.e. JKC777 and JKC703) and (+) sign (i.e. JKC703-SFP) polymorphism in JKC777 (i.e. JKC703 and JKC777).

Discussion

In the present study, we tested the feasibility of gene-chip-based approach for polymorphism discovery in upland cotton, *G. hirsutum*. Our results clearly indicate that Affymetrix cotton Gene Chips designed for cotton gene expression analysis can be utilized for genomewide identification of sequence polymorphism as also shown in other crops

Table 1. Details of SFP called in six different genotype-pairs at six different fibre developmental stages (0DPA, 6DPA, 9DPA, 12DPA, 19DPA and 25DPA).

Genotype pair	0DPA	6DPA	9DPA	12DPA	19DPA	25DPA	Total SFP
JKC777 & JKC703	752	375	620	328	1741	386	4202
JKC777 & JKC737	1100	117	486	896	501	431	3531
JKC777 & JKC783	440	213	147	343	199	753	2095
JKC725 & JKC703	1689	376	1707	220	2869	906	7767
JKC725 & JKC737	1220	1587	1872	1550	1085	1868	9182
JKC725 & JKC783	1778	4095	1813	900	505	1605	10696
	6979	6763	6645	4237	6900	5949	37473

0–6DPA, fibre initiation; 6–9DPA, penultimate stage for fibre initiation and elongation; 9–12DPA, fibre elongation; 12–19DPA, penultimate stage for fibre elongation and maturation; 19–25DPA, fibre maturation stage.

Table 2. Distribution of gene-chip predicted SFPs among polymorphic probe sets.

		1	2	3	4	5	6	7	8	9	10	11	Total
JKC777 and JKC703													
Six DPAs	(−)SFP-PS	62	79	74	69	83	68	72	66	72	64	70	779
	(+)SFP-PS	244	290	284	304	330	312	300	361	347	323	328	3423
Total		306	369	358	373	413	380	372	427	419	387	398	4202
JKC777 and JKC737													
Six DPAs	(−)SFP-PS	126	120	124	114	142	142	153	144	140	154	140	1499
	(+)SFP-PS	144	137	174	153	134	162	202	198	219	221	288	2032
Total		270	257	298	267	276	304	355	342	359	375	428	3531
JKC777 and JKC783													
Six DPAs	(−)SFP-PS	32	32	23	30	31	41	41	45	40	30	40	385
	(+)SFP-PS	111	135	136	144	139	136	140	166	211	183	209	1710
Total		143	167	159	174	170	177	181	211	251	213	249	2095
JKC725 and JKC703													
Six DPAs	(−)SFP-PS	100	126	123	116	112	119	132	119	144	146	125	1362
	(+)SFP-PS	530	530	542	558	580	593	555	611	618	634	654	6405
Total		630	656	665	674	692	712	687	730	762	780	779	7767
JKC725 and JKC737													
Six DPAs	(−)SFP-PS	567	665	621	588	612	624	652	630	630	623	634	6846
	(+)SFP-PS	204	218	227	210	214	216	209	224	209	199	206	2336
Total		771	883	848	798	826	840	861	854	839	822	840	9182
JKC725 and JKC783													
Six DPAs	(−)SFP-PS	837	923	885	833	850	809	804	786	781	736	670	8914
	(+)SFP-PS	167	186	169	158	171	160	156	143	159	149	164	1782
Total		1004	1109	1054	991	1021	969	960	929	940	885	834	10696

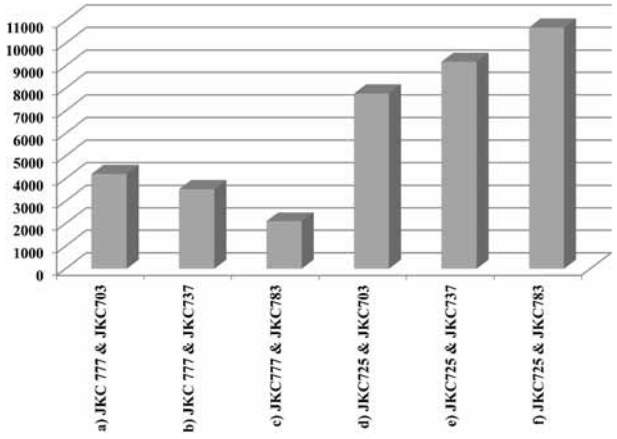

Figure 2. Detailed SFPs identified in all six pair combination of cotton genotypes. JKC725 and JKC783 resulted higher number of SFPs followed by JKC725 and JKC737.

Table 3. Sequencing and validation of SFPs for JKC777 and JKC703.

SFPs selection/confirmation details	Number
Randomly selected genes showing SFPs	51
No. of SFPs identified	224
No. of genes show polymorphism (with sequencing)	51
No. of genes do not show polymorphism (with sequencing)	0
No. of SNPs identified	122
No. of indels identified	10

(Cho *et al.* 2006; Choi *et al.* 2007). The PM estimates of three biological replicates of all the genotypes are more or less same without any significant variation (figure 1 in electronic supplementary material) which indicates the PM estimates in all the genotypes from the normalized data of Affymetrix cotton Gene Chips are genuine and realistic to

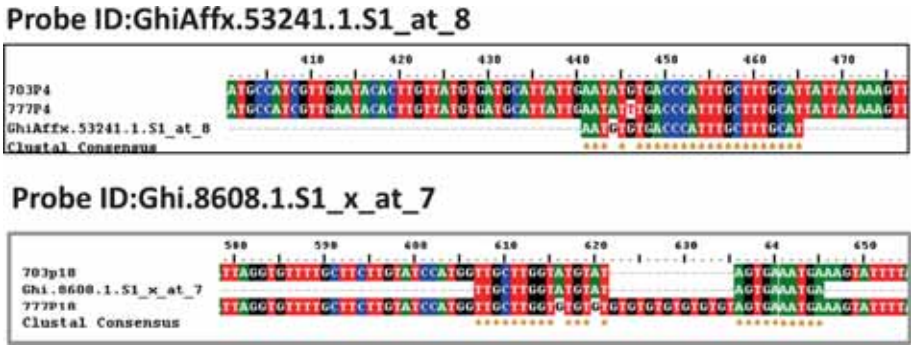

Figure 3. Confirmation of two SFPs in aligning sequences of JKC703 and JKC777 with respective Affymetrix polymorphic probe sequences.

measure the differential expressional variation in six dataset combinations.

The upland cotton genotypes such as JKC777 and JKC725 are superior cotton cultivars, while JKC703, JKC737 and JKC783 are inferior cotton cultivars with respect to the fibre quality. The genetic differences between two superior upland cotton genotypes and among three inferior upland cotton genotypes are very small as compared to high degree of polymorphism observed among diploid species of cotton like *G. herbaceum* and *G. arboreum* (Jena *et al.* 2012). However, these genotypes do show contrasting fibre traits (data published elsewhere) and these genotypes were

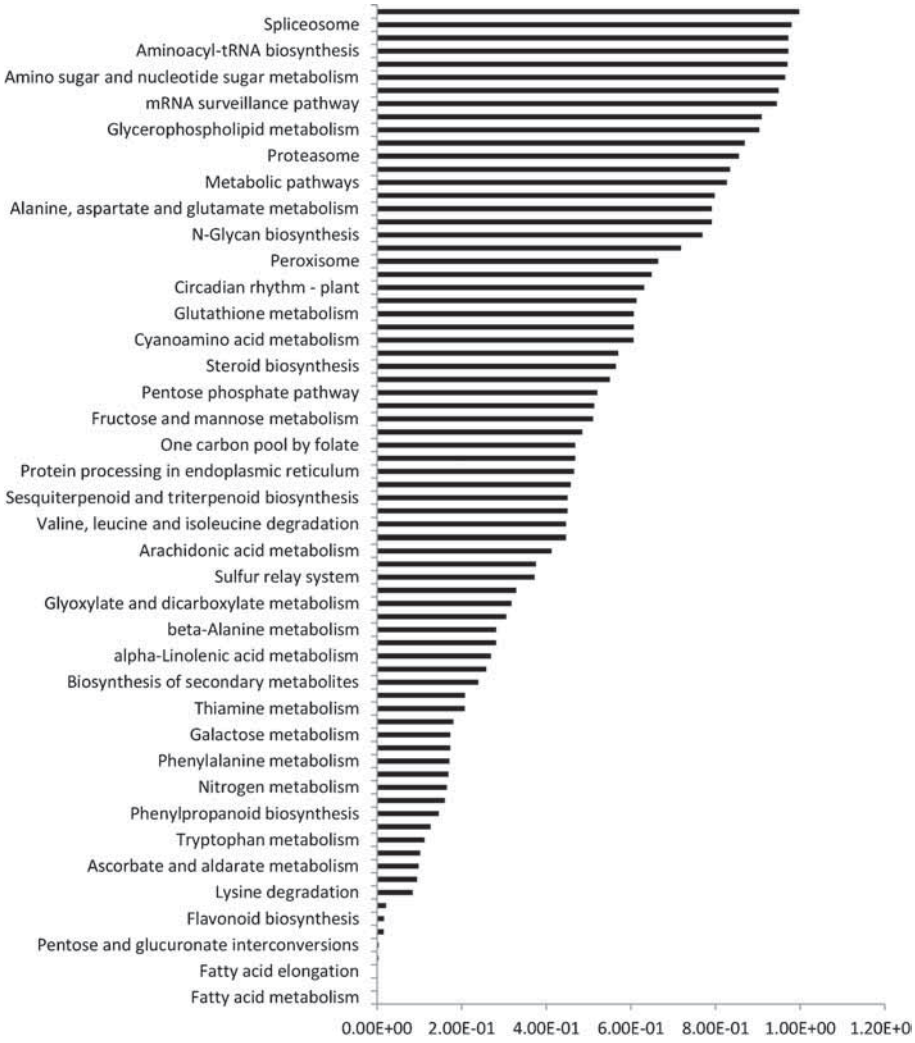

Figure 4. Analysis of differential expressed genes harbouring SFPs by KOBAS in JKC777 and JKC703.

used for development of mapping populations. However, in our earlier study using AFLP markers we identified some polymorphism between superior and inferior genotypes (Jena *et al.* 2012). Narrow genetic backgrounds have been earlier reported to result in few SFPs in previous studies (Gore *et al.* 2007; Kumar *et al.* 2007). On the basis of SFP calls, the most diverse genetic background is between JKC725 and JKC783 followed by JKC725 and JKC737 and least in between JKC777 and JKC783 (figure 2) which supported the same genetic distance among the genotypes in our earlier study (Jena *et al.* 2012).

The number of SFPs varies in different fibre developmental stages as well as in different genotype pair combinations (table 1). Further, there were more SFP calls, when the expression level variation was high between the genotypes (table 1). Therefore, in the present study, the SFP calls depend on the fibre developmental stage and genotype-dependent differential gene expression.

From verification of sequence information of predicted SFP conducted in one dataset namely, JKC777 and JKC703, we found that SFPs can reliably be predicted in upland cotton with ∼59% (132 indel plus SNPs out of 224 SFP calls) detection sensitivity (table 3). The sequence polymorphism in sequenced probes were relatively less as compared to that in *Arabidopsis* (Borevitz *et al.* 2003; Werner *et al.* 2005) and Barley (Rostoks *et al.* 2005). This may be due to complex and huge genome size of *G. hirsutum* or may be due to methodology adopted for SFP detection. Since in SAM analysis, the significant differences in the probe intensity between genotypes resulted in a high number of SFP calls. To minimize the false positive, the differences in probe intensities can be subtracted by the RMA (robust multichip analysis) normalized-expression index as suggested for complex genome of *Triticum aestivum* (Coram *et al.* 2008). Thus, selecting proper FDR and background corrections using RMA is important to minimize false positive.

In KOBAS analysis, several metabolic/biological pathways were found to be significant in comparison to superior genotypes against the inferior fibre genotypes (figures 2 to 6 in electronic supplementary material). All the significant pathways like cutin, suberin, wax biosynthesis and flavonoid biosynthesis which were reported to play important role in fibre development (Al-Ghazi *et al.* 2009) and the maximum SFPs were detected in those pathways. Since, we were comparing superior genotype with inferior one, enrichment of SFPs in those pathways do reflect the differences in fibre quality. The most significant pathway like fatty acid metabolism reported earlier to promote cotton fibre development (Qin *et al.* 2007). Similarly, tryptophan metabolism which eventually leads into auxin biosynthesis and auxin has been reported to play important role in cotton fibre initiation (Zhang *et al*. 2011). Thus, SFPs identified in important genes that govern important pathways in fibre development will be crucial for future marker development programme using the mapping population developed between these genotype crosses.

Acknowledgements

The authors thank JK Agri Genetics Ltd., India for providing the upland cotton genotypes under the collaborative NMITLI project. The financial support received from the CSIR, New Delhi, under NMITLI project.

References

Al-Ghazi Y., Bourot S., Arioli T., Dennis E. S. and Llewellyn D. J. 2009 Transcript profiling during fibre development identifies pathways in secondary metabolism and cell wall structure that may contribute to cotton fibre quality. *Plant Cell Physiol.* **50**, 1364–1381.

Borevitz J. O., Liang D., Plouffe D., Chang H. S., Zhu T., Weigel D. *et al.* 2003 Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523.

Borevitz J. O., Hazen S. P., Michael T. T., Morris G. P., Baxter I. R., Hu T. T. *et al.* 2007 Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **104**, 12057–12062.

Cho S., Garvin D. F. and Muehlbauer G. J. 2006 Transcriptome analysis and physical mapping of barley genes in wheat–barley chromosome addition lines. *Genetics* **172**, 1277–1285.

Choi I. Y., Hyten D. L., Matukumalli L. K., Song Q., Chaky J. M., Quigley C. V. *et al.* 2007 A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* **176**, 685–696.

Coram T. E., Settles M. L., Wang M. and Chen X. 2008 Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. *Theor. Appl. Genet.* **117**, 401–411.

Cui X., Xu J., Asghar R., Condamine P., Svensson J. T., Wanamaker S. *et al.* 2005 Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics* **21**, 3852–3858.

Das S., Bhat P. R., Sudhakar C., Ehlers J. D., Wanamaker S., Roberts P. A. *et al.* 2008 Detection and validation of single feature polymorphisms in cowpea (*Vigna unguiculata* L. Walp) using a soybean genome array. *BMC Genomics* **9**, 107.

Gore M., Bradbury P., Hogers R., Kirst M., Verstege E., Van Oeveren J. *et al.* 2007 Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. *Crop Sci.* **47**, 135–148.

Gupta P. K., Rustgi S. and Mir R. R. 2008 Array-based high-throughput DNA markers for crop improvement. *Heredity* **101**, 5–18.

Hazen S. P., Borevitz J. O., Harmon F. G., Pruneda-Paz J. L., Schultz T. F., Yanovsky M. J. *et al.* 2005 Rapid array mapping of circadian clock and developmental mutations in *Arabidopsis*. *Plant Physiol.* **138**, 990–997.

Hendrix B. and Stewart J. M. 2005 Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* **95**, 789–797.

Jena S. N., Srivastava A., Singh U. M., Roy S., Banerjee N., Rai K. M. *et al.* 2012 Analysis of genetic diversity, population structure and linkage disequilibrium in elite cotton (*Gossypium* L.) germplasm in India. *Crop Pasture Sci.* **62**, 859–875.

Kumar R., Qiu J., Joshi T., Valliyodan B., Xu D. and Nguyen H. T. 2007 Single feature polymorphism discovery in rice. *PLoS One* **2**, e284.

Li F., Fan G., Lu C., Xiao G., Zou C., Kohel R. J. *et al.* 2015 Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**, 524–530.

Qin Y. M., Hu C. Y., Pang Y., Kastaniotis A. J., Hiltunen J. K. and Zhu Y. X. 2007 Saturated very-long-chain fatty acids promote cotton fibre and *Arabidopsis* cell elongation by activating ethylene biosynthesis. *Plant Cell* **19**, 3692–3704.

Ronald J., Akey J. M., Whittle J., Smith E. N., Yvert G. and Kruglyak L. 2005 Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* **15**, 284–291.

Rostoks N., Borevitz J. O., Hedley P. E., Russell J., Mudie S., Morris J. *et al.* 2005 Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.* **6**, R54.

Saeed Al., Sharov V., White J., Li J., Liang W., Bhagabati N. *et al.* 2003 TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378.

Tusher V. G., Tibshirani R. and Chu G. 2001 Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121.

Werner J. D., Borevitz J. O., Warthmann N., Trainer G. T., Ecker J. R., Joanne Chory J. *et al.* 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci. USA* **102**, 2460–2465.

West M. A., Van Leeuwen H., Kozik A., Kliebenstein D. J., Doerge R. W., Dina A. *et al.* 2006 High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. *Genome Res.* **16**, 787–795.

Winzeler E. A., Richards D. R., Conway A. R., Goldstein A. L., Kalman S., McCullough M. J. *et al.* 1998 Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197.

Zhang L., Miles M. F. and Aldape K. D. 2003 A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* **21**, 818–821.

Zhang M., Zheng X., Song S., Zeng Q., Hou L., Li D. *et al.* 2011 Spatiotemporal manipulation of auxin biosynthesis in cotton ovule epidermal cells enhances fibre yield and quality. *Nat. Biotechnol.* **29**, 453–458.

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

Received 8 April 2015, in final revised form 6 May 2015; accepted 20 May 2015
Unedited version published online: 6 July 2015
Final version published online: 8 December 2015