Indian J. Genet., 77(1): 126-133 (2017)
DOI: 10.5958/0975-6906.2017.00017.7

Discovery of single nucleotide polymorphism in *Gossypium hirsutum* and *G. barbadense* through next generation sequencing approach

Reddy S. Sudakara, I. S. Katageri*, N. V. Mohan Kumar, M. P. Jadhav, Sateesh Adiger, Navajeet Chakravartty[1], H. M. Vamadevaiah and V. B. Reddy[1]

University of Agricultural Sciences, Agricultural Research Station, Dharwad Farm, Dharwad 07; [1]SciGenom Labs Pvt. Ltd., Hyderabad

(Received: December 2016; Revised: December 2016; Accepted: January 2017)

Abstract

Single nucleotide polymorphism was identified using ddRAD and pooled amplicon sequencing between the allotetraploid cotton genotypes of *G. hirsutum* var. DS-28 (POOL-H) and *G. barbadense* var. SBYF-425 (POOL-B). The 471270 (470161 SNP and 1109 INDELs), 2723 (2621 SNP and 102 INDELs) and 780 (759 SNP and 21 INDELs) variant calling were predicted at read depth of 10 in the genotypes by using *G. raimondii* (D_5), *G. arboreum* (A_2) and *G. hirsutum* (AD) as reference genomes, respectively. Further, filtered out false positive homoeologous SNPs between parents (DS-28 and SBYF-425) at read depth of 10, identified 245 and 45 homozygous polymorphic SNPs between parents based on pooled amplicon sequencing and ddRAD sequencing by using *G. raimondii*, *G. arboreum* and *G. hirsutum* as a reference genomes. Based on gene function, 5 SNPs were converted into KASPar markers and genotyped in 21 RILs for fibre quality traits. Identified polymorphic SNPs and INDELs can be used for genotyping and QTL mapping after converting into functional markers.

Key words: *G. hirsutum, G. barbadense,* SNPs and INDELs, *G. raimondii* and KASPar marker

Introduction

Cotton (*Gossypium* spp.) is called as king of fibre, white gold, produces natural raw fibre and is an economically important agriculture commodity benefit farmers and textile industries. Cotton is one of the most intensively cultivated fibre crops worldwide, in more than 80 countries and in varying climatic conditions. The genus *Gossypium* is also an important model species for polyploidy and the biological processes of cell wall elongation and cellulose biosynthesis in fibre cells. This clade consists of approximately 45 diploid species and five allotetraploid species. Genomes of the allotetraploid species having 52 chromosomes (2n = 4x = 52) and are believed to be originated from a single polyploidization event between, diploid A and D genomes (n = 2x = 26) approximately 1-2 million years ago (Wendel et al. 2009) and they share basic AD genome architecture. Chromosomes of the *G. hirsutum* genome (AD_1) have been numbered according to their evolutionary origins and meiotic pairing relationships. Chromosomes 1-13 comprise the "A" sub-genome (At) that originated from the extinct A genome diploid ancestor and chromosomes 14-26 comprise the "D" sub-genome (Dt) that originated from the extinct D-genome diploid ancestor. There are four major species cultivated worldwide, two diploid (2n=2x=26) *G. arboreum* (A_2 genome) and *G. herbaceum* (A_1) and two allotetraploids (2n=4x=52) *G. hirsutum* L. or upland cotton (AD_1) and *G. barbadense* L. (AD_2), extra long staple Pima, Egyptian cotton or Sea Island cotton. Upland cotton cultivation represents over 95% of the fibre produced worldwide due to its high yield, but generally Pima cotton the next most cultivated cotton, exhibits longer, stronger, and finer fibre.

Many of the mapping efforts in cotton consist of interspecific biparental populations of *G. hirsutum* × *G. barbadense* which offers a higher polymorphism rate and segregation for superior fibre quality characteristics than intraspecific crosses. Moderate density linkage maps have been created using restriction fragment length polymorphisms (RFLPs) (Reinisch et al. 1994; Shappely et al. 1998), amplified

*Corresponding author's e-mail: katageriis@uasd.in
Published by the Indian Society of Genetics & Plant Breeding, F2, First Floor, NASC Complex, PB#11312, IARI, New Delhi 110 012
Online management by indianjournals.com; http://epubs.icar.org.in/journal/index.php/IJGPB

fragment length polymorphisms (AFLPs) (Lacape et al. 2003) and simple sequence repeats (SSRs) (Yu et al. 2012; Gao et al. 2007). SSRs have also been used for wide-cross whole-genome radiation hybrid (WWRH) mapping for production of syntenic groups (Gao et al. 2004, 2006). Because of its codominant nature and high reproducibility, SSR markers became "markers of choice" for plant breeders. But in recent years this has changed after the emergence and evolution of reliable and high-throughput genotyping platform such as, next-generation sequencing (NGS), is capable of simultaneously assaying hundreds of thousands of SNPs (McCouch et al. 2010; Davey et al. 2011; Feuillet et al. 2011; Morrell et al. 2012). Single nucleotide polymorphism (SNP) represents the most prevalent category of polymorphism available within the genome. Discovery of the SNP marker in a polyploid species like cotton is a very tough task, because of multiallele combinations in cotton. Many of the seed industries are now using SNP markers as a marker of choice in MAS in corn and other crops. However, there is no such information available in public cotton databases. An et al.(2008) discovered a strategy to identify SNP markers in tetraploid cotton, a few SNP markers were identified while studying R2R3-MYB transcription factors. The first large-scale SNP discovery in cotton has been done by Van Deynze in 2009. They developed about 1,000 SNPs and 300 INDELs by re-sequencing the ESTs of 24 upland cotton genotypes. In spite of this, expressed sequence tags (ESTs) have been mined for large-scale SNP discovery in plants including *Arabidopsis*, barley, maize, sugarcane, and tomato. SNPs mined from ESTs have the potential to be functional markers if, the particular EST or gene is responsible for phenotypic variations. Several methods for identifying SNPs from ESTs have been reported and numerous cotton ESTs are available in public databases, providing important foundations for the development of EST-based cotton SNP markers. Few studies have developed and mapped SNPs in cotton (Yu et al. 2012; Bayers et al. 2012). SNP development efforts to-date has produced relatively few numbers of SNPs using different genome reduction methods in cultivated species (Bayers et al. 2012; Van Deynze et al. 2009; Rai et al. 2013; Buriev et al. 2010; An et al. 2008). Therefore, the present study was focused mainly on discovery of SNPs between allotetraploid cotton genotypes DS-28 (*G. hirsutum*) and SBYF-425 (*G. barbadense*).

Materials and methods

Plant material

G. hirsutum var. DS-28 and *G. barbadense* var. SBYF-425 (parents of interspecific hybrid DCH-32) were taken for the study. DS-28 is a recipient parent, superior for its agronomic traits and SBYF-425 is a donor, superior for its fibre quality traits.

DNA extraction

DNA was extracted from the leaves of 4-5 week old plants of *G. hirsutum* var. DS-28 and *G. barbadense* var. SBYF-425 by using C-TAB method with minor modifications as suggested by Sanganavar et al. (2013). Extracted DNA was diluted in DNase free Nano pure water and quantified using Nano-drop spectrometer (ND1000, Nano drop technologies).

Library preparation and DNA (ddRAD) sequencing

Genome complexity was reduced using ddRAD method as described by Peterson et al. (2012). Total genomic DNA was double digested (1microgram) by using *Sphl and Mluc1* restriction enzymes and the product was cleaned using ampure beads. Digested fragments were ligated to a *Sphl and Mluc1* site specific P1 (Barcoded) and P2 adaptors using T4 DNA ligase and ligated product was pooled and cleaned. Fragments with approximate 300-400 bp size were selected. Resulting fragments were PCR amplified with adaptor specific primers to enrich and add Illumina specific adapters and flow cell annealing sequences. Quality check was performed on Bio-analyzer and final pooling and sequencing was done for selected fragments. Sequencing was carried out using Illumina True-Seq chemistry on Illumina Hi-Seq 2000 platform.

Pooled amplified RAPD amplicon sequencing

Pooled amplified RAPD amplicon DNA sequencing was done by physical shearing with covaries. Ligation of P1 (barcoded) and P2 adapters was done using T4 DNA ligase. Pooling and cleaning the product was carried out and size selection was done after 2% agarose gel electrophoresis. PCR with adopter specific primers was done to enrich and add the Illumina specific adapters and flow cell annealing sequences. Quality check was carried on bio- analyser. Final pooling and sequencing of precisely selected amplicons was done.

Fastq pre-processing and alignment

The samples were demultiplexed first to obtain reads

for each sample. Up to one mismatch was allowed to demultiplex the sample data. The enzyme cut sites overhang bases were trimmed from raw fastq files for further processing. The cut site pattern at 5' end is CATGC, whereas for 3' end it is TTAA. The low quality bases and regions showing base bias at the start or end regions were removed from the reads. The Illumina 5' and 3' adapter sequences were removed from the reads for downstream analysis. The processed reads were aligned on the reference *G. raimondii, G. arboreum* and *G. hirsutum* genome downloaded from ftp://public.genomics.org.cn/BGI/cotton/Assembly/G.raimondii.chromosome.fasta.gz, http://cgp.genomics.org.cn/page/species/download.jsp?category=arboreum and ftp://public.genomics.org.cn/BGI/cotton/Gossypium_hirsutum/Gossypium_hirsutum_v1.0.gz. The paired-end alignment was performed using Bowtie2 (version 2.0.5) program using default parameters.

Identification of variants and variant comparison

Accurate identification of SNPs in tetraploid sequence data was a challenge due to the potential co-assembly of homozygous alleles. The term SNP in the present study refers only to allelic single nucleotide difference but not the difference that distinguish the two resident genomes of the allotetraploid (Genome specific SNPs). In RRS (ddRAD seq) assemblies, identification of these genome specific SNPs in separately assembled homeolog was performed using Samtools programme (Samtools version 0.1.18). The identified SNPs and INDELs (Variants) were compared with the variants identified from the whole genome sequence of reference genomes. Variant annotation was done for identified short listed variants of each genotypes (DS-28 and SBYF-425) with reference to *G. raimondii, G.arboreum*, and *G. hirsutum* genome sequences using BLASTn programme (http://blast.ncbi.nlm.nih.gov/ LINK_LOC=blasthome).

SNP assay design and genotyping

Assay design was performed using Kraken™ software for selected SNPs based on gene function in cotton metabolic pathway. The KASP Assay mix contains three assay specific non-labelled primers: two allele specific forward primers plus one common reverse primer. The sequences for these primers were generated with Kraken™ software. The KASP Master mix contains the universal FRET cassettes, ROX™ passive reference dye, KASPTaq™ DNA polymerase, free nucleotides and $MgCl_2$ in an optimised buffer solution. Genotyping was carried out using these developed KASPar markers in RILs (Recombinant Inbreed Lines) derived from DCH-32.

Results

Sequencing and alignment of pooled amplified and ddRAD reads

The sequencing statistics for cotton genotypes, DS-28, and SBYF-425 was performed by following standard procedure. The two approaches considered for sequencing, includes ddRAD sequencing and pooled amplicon sequencing. A total of 2, 39, 230 and 11, 91, 226 raw reads were obtained from DS-28 and SBYF-425, respectively through ddRAD sequencing approach, while in case of pooled amplicon sequencing approach it was 4,015,536 and 2,829,942, respectively (Table 1). Then reads were pre-processed (total number of reads after filtering and pre-processing) and aligned to the available closely related cotton D_5 genome (*Gossypium raimondii*), AD genome (*Gossypium hirsutum*) and A_2 genome (*Gossypium arboreum*). The results obtained from paired end read alignment for each genotype were filtered for total aligned reads. The filtered aligned reads of DS-28 and SBYF-425

Table 1. Preprocessed statistics

Samples	ddRAD sequencing		Pooled amplicon sequencing	
	DS-28	SYBF-425	SYBF-425 (POOL-B)	DS-28 (POOL-H)
Total reads	2392230.00	1191226.00	4015536.00	28229942.00
No. of passed reads	224848.00	1168070.00	4014762.00	2829464.00
Total passed reads (%)	93.99	98.05	99.98	99.98
No. of failed reads	14382.00	23156.00	774.00	478.00
Per cent failed reads	6.01	1.94	0.20	0.20

were mapped on particular chromosome position with reference to the reference cotton genome (D_5, AD and A_2). The total number of mapped reads and uniquely mapped reads for each genotype, DS-28, and SBYF-425 with each reference cotton genome (D_5, AD and A_2) obtained from sequencer are presented in Table 2.

SNPs were identified between parents (DS28 and SBYF425) at read depth cutoff 10 including false positives SNP and 44 SNP were identified after filtering out false positive homeologous SNP between parents at read depth of 10 by using *G. raimondii* (D_5), *G. arboreum* (A_2) as a reference genome, but only one

Table 2. Overall paired end alignment statistics of DS-28 and SBYF-425 with *G. arboreum*, *G. raimondii* and *G. hirsutum* as reference

Samples	ddRAD sequencing		Pooled amplicon sequencing	
	DS-28	SYBF-425	SYBF-425 (POOL-B)	DS-28 (POOL-H)
Total reads	2392230.00	1191226.00	4015536.00	28229942.00
Total number of reads after filtering	224848.00	1168070.00	4014762.00	2829464.00
Gossypium arboreum				
Overall alignmnent (%)	69.15	70.97	60.13	61.61
No. of alignmnents	155482.392	828979.279	2414076.391	1743232.77
No. of unique alignments	130467.00	681149.00	1842721.00	1287478.00
Per cent unique alignment	83.91	82.17	76.33	73.86
Gossypium raimondii				
Overall alignmnent (%)	40.99	43.90	40.95	40.17
No. of alignmnents	92165.1952	512782.73	1644045.039	1136595.689
No. of unique alignments	77311.00	429137.00	1225524.00	837994.00
Per cent unique alignment	83.88	83.69	74.54	73.73
Gossypium hirsutum				
Overall alignmnent (%)	85.74	87.09	80.89	85.41
No. of alignmnents	192784.6752	1017272.163	3247540.982	2416645.202
No. of unique alignments	171215.00	863549.00	2480050.00	1811341.00
Per cent unique alignment	88.81	84.89	76.37	74.95

SNP discovery

The chromosome wise variant call (includes SNPs and INDELs) summary results at different variant high quality depth cut-off 10 are given in Table 3. The total number of variants were identified with reference to *G. raimondii* are 471270 and 2723 with reference to *G. arboreum* at high quality read depth cut off 10 for both the genotypes. Considering *G. hirsutum* as reference genome the total number of variants were, 780 at high quality read depth cut off 10. The identified SNPs and INDELs at cut off depth 10 are more informative than identified SNPs and INDELs at depth cut off 2 and 5 (data not shown). Because, the probability of getting errors or ambiguity of errors is more at depth cut off 2 and 5 as compared to the depth cut off 10. Based on ddRAD sequencing, 69 homozygous polymorphic homozygous polymorphic SNP was identified between parents when *G. hirsutum* (AD) used as a reference genome at read depth of 10. Based on pooled amplicon sequencing, 425 homozygous polymorphic SNPs were identified between DS28 and SBYF425 at read depth cut off of 10 and 239 SNP were identified after filtering out false positive homoeologous SNP between DS-28 and SBYF-425 at depth cutoff 10 by using *G. raimondii* (D_5), *G. arboreum* (A_2) as a reference genome, six homozygous polymorphic SNP were detected between parents when *G. hirsutum* (AD) used as a reference genome at read depth of 10 (Tables 6 and 7). The Chromosome wise statistics for SNP identified in different cotton species is shown in Table 3.

Table 3. Chromosme-wise summary of variants at high quality depth cut of 10 for *G. raimondii G. arboreum* and *G. hirsutum*

Chromosome	*G. raimondii*			*G. arboretum*			*G. hirsutum*								
SNPs	INDELs	Total	Chr.	SNPs	INDELs	Total	Chr.	SNPs	INDELs	Total	Chr	SNPs	I	NDELs	Total
gi|432003306|gb|CM001751.1|	59878	168	60046	CA_chr1	134	2	136	A01	178	0	178	D01	10	0	10
gi|432003307|gb|CM001751.1|	35216	70	35286	CA_chr2	77	2	79	A02	10	0	10	D02	10	0	10
gi|432003308|gb|CM001750.1|	66818	152	66970	CA_chr3	92	8	100	A03	2	0	2	D03	0	0	0
gi|432003309|gb|CM001749.1|	66559	118	66677	CA_chr4	96	4	100	A04	2	0	2	D04	0	0	0
gi|432003310|gb|CM001748.1|	77	4	81	CA_chr5	48	2	50	A05	4	0	4	D05	2	1	2
gi|432003311|gb|CM001747.1|	24	0	24	CA_chr6	106	18	124	A06	44	2	46	D06	18	0	18
gi|432003312|gb|CM001746.1|	26	2	28	CA_chr7	159	12	171	A07	12	0	12	D07	1	1	2
gi|432003313|gb|CM001745.1|	59	0	59	CA_chr8	384	10	394	A08	58	2	60	D08	8	0	8
gi|432003314|gb|CM001744.1|	13858	40	13898	CA_chr9	199	14	213	A09	10	0	10	D09	28	0	28
gi|432003315|gb|CM001743.1|	62315	141	62456	CA_chr10	134	6	140	A10	181	6	187	D10	15	0	15
gi|432003316|gb|CM001742.1|	45299	92	45391	CA_chr11	869	6	875	A11	14	2	16	D11	2	0	2
gi|432003317|gb|CM001741.1|	66204	144	66348	CA_chr12	90	0	90	A12	6	2	8	D12	0	0	0
gi|432003318|gb|CM001740.1|	53828	178	54006	CA_chr13	96	8	104	A13	10	0	10	D13	18	2	20
Scaffolds	-	-	-	-	137	10	147	-	-	-	-	-	116	4	120
Total	470161	1109	471270	-	2621	102	2723	-	-	-	-	-	759	21	780

Chr.= Chromosome

Variant calling and SNP assay

Among the total polymorphic SNPs, 35 SNPs were selected and bioinformatics analysis was carried out for filtered variants, flanking sequences between DS-28 and SBYF-425 and homology search (BLASTn) for flanking sequences to identify the similarity, alignment length, mismatches and E-valve with reference to closely available D_5 cotton genome in NCBI nucleotide database. Of selected 35 SNPs, only 8 SNPs were selected for KASPar marker development based on their function in cotton metabolic pathways for fibre traits. Those identified novel genes obtained between DS-28 and SBYF-425 were coding or governing for improvement of fibre traits in cotton. Of which several transcripts encoding MYB family, *Theobroma cacao* Breast cancer associated RING 1, putative isoform 4 (TCM_003952), *G. hirsutum* clone MX008C17, *G. hirsutum* clone NBRI_GE54083 microsatellite sequence, Replication factor C/DNA polymerase III gamma-tau subunit, *G. hirsutum* mono-functional lysine-ketoglutarate reductase 2 and *Theobroma cacao* Phosphoribosyl transferase family protein isoform 1 and 2 (TCM_042341) transcription factors were differentially expressing in calcium (Ca^{2+}) and phytohormone-mediated signaling pathways play a crucial regulatory role in fibre cell initiation and differentiation. Some of the identified genes performing a function such as protein with binding function or cofactor requirement, cellular transport, transport facilities and transport routes and functioning in herbicide control cotton event, trichomes development and functions in the tip growth of root hair cells. Calcium mediated signaling pathway plays an important role in cell division, differentiation and root hair elongation. Preferential expression of genes encoding calcium binding proteins involved in Ca^{2+}-mediated signaling pathways during fibre initiation and

Table 4. Assay designs for KASPar markers development

ID	Allele FAM	Allele HEX	CG%_ FAM	CG%_ HEX	CG%_ Common	Back-ground
SRCF-1	A	G	45.8	47.8	52.1	ROX
SRCF-2	C	T	54.5	52.2	42.6	ROX
SRCF-5	G	T	61.9	52.2	48	ROX
SRCF-6	A	G	47.8	52.2	37.9	ROX

Table 5. List of RILs homozygous for FAM allele

RILs no.	Marker name	Fibre length ngth (mm)	Fibre stre-(µg/inch) g/tex)	Micro-naire (%)	Unifor-mity ratio	Matu-rity (%)	Elon-gation
RIL-2	SRCF-7	27.9	23.5	3.4	48	0.6	5.6
RIL-6	SRCF-7	25.6	21.2	3.6	51	0.63	5.4
RIL-16	SRCF-7	28.2	22.1	3.5	48	0.62	5.5
RIL-21	SRCF-1	25.2	18.7	4.1	49	0.67	5.1
P1 (DS-28)		28	20	4.2	40	0.5	5.0
P2 (SBYF-425)		35	26	3.2	100	0.7	7.5

Table 6. Polymorphic homozygous matrix between the samples at read depth =10 before filtering homeologous SNPs

	G. arbo-reum	*G. rai-mondii*	Pool B	Pool H	SYBF-425	DS-28
G. arboreum	0					
G. raimondii	3719	0				
Pool B	1516	1325	0			
Pool H	1105	797	425	0		
SYBF-425	199	213	55	42	0	
DS-28	202	136	40	31	69	0

Table 7. Polymorphic homozygous matrix between the samples at read depth 10 after filtering homeologous SNPs

	G. arbo-reum	*G. rai-mondii*	Pool B	Pool H	SYBF-425	DS-28
G. arboreum	0					
G. raimondii	0	0				
Pool B	810	810	0			
Pool H	512	512	239	0		
SYBF-425	127	127	23	15	0	
DS-28	68	68	19	12	44	0

Table 8. Polymorphic homozygous matrix between the samples at read depth 10

	G. hirsutum	DS-28	Pool B	Pool H	SYBF-425
G. hirsutum	0				
DS-28	4	0			
Pool B	336	1	0		
Pool H	29	0	6	0	
SYBF-425	2	1	0	1	0

elongation. Phospholipids are major structural components of plasma membrane (PM) and involved in lipid signaling pathway. In addition, transcripts encoding integrase-type DNA-binding super family proteins involved in defense mechanism and fibre development. The detail information of genes function in metabolic pathways for cotton fibre improvement was mentioned in Table 1. Of the selected 8 SNPs, 5 were converted in to KASPar markers and were selected based on allelic polymorphism between DS-28 and SBYF-425.

Genotyping

These developed KASPar markers for fibre traits were used for genotyping in 21 RILs. Among the 21 RILs only four RILs showing homozygous for designed FAM allele. Of which RIL number 21 was shown homozygous for FAM 'A' allele, remaining all RILs were shown heterozygous for KASPar marker SRCF-1 and RIL number 16, 6, 2 shown homozygous for FAM 'C' allele of KASPar marker SRCF-7. The KASPar marker SRCF-2, SRCF-5, SRCF-6 are shown heterozygous in all 21 RILs population. The four RILs 21, 16, 6 & 2 shown good fibre quality traits in comparison with phenotypic data (Unpublished data).

Discussion

Single nucleotide polymorphism (Jordan and Humphries 1994) represent the most common variations across a genome, occurs at a frequency of one in 1000 nucleotides in the genomic DNA and they can be used directly to detect the alleles responsible for a trait of interest. Among the several techniques available in public domain to study single nucleotide difference, ddRAD sequencing, a modified version of RAD sequencing (Peterson et al. 2012), is highly advanced and useful technology for genotyping, mapping and to identify the difference in plants and other living organisms. The technique help cutting the

fragments in appropriate size with both enzymes facilitating the sequencing. It increases the probability of the same genomic regions will be sequenced across individuals and eliminates random shearing, end repair and too high DNA losses.

The present study discovered novel SNPs between *G. hirsutum* var. DS-28 and *G. barbadense* var. SBYF-425 through 'ddRAD' sequencing and pooled amplicon sequencing. As *G. barbadense* is a good source of superior fibre quality to *G. hirsutum,* the identification of SNPs between these two genotypes enhances the process of introgression between these two species. By eliminating the intergenomic SNPs, the true allelic SNPs were identified between DS-28 and SBYF-425. Discovery of SNPs between DS-28 and SBYF-425 was made through ddRAD sequencing and genotyping with KASPar marker. In the present study, 24.16 Mb (DS-28) and 120.32Mb (SBYF-425) of raw reads were sequenced based on ddRAD sequencing and 1204.44 (DS-28) and 848.98 Mb (SBYF-425) of raw reads were sequenced based on pooled amplicon sequencing. ddRAD sequences and pooled amplified sequences of these two genotypes were aligned by using paired-end alignment and it was performed by using Bowtie2 (version 2.0.5) program with reference genome of *G. raimondii, G. arboreum* and *G. hirsutum* (Wang et al. 2012). *G. raimondii* is the progenitor of both *G. hirsutum* and *G. barbadense*, contributing D genome to these two species. Therefore, the samples got from ddRAD sequencing were aligned to reference genomes *G. raimondii, G. arboreum and G. hirsutum*, the alignment percentage recorded was 40.99, 69.15, 85.74, respectively in DS-28 and 43.99, 70.97, 87.09 in SBYF-425, respectively. The samples got from pooled amplicon sequencing were aligned to reference genomes (*G. raimondii, G. arboreum* and *G. hirsutum*) and the alignment percentage was 40.95, 60.13, 80.89 and 40.17, 61.61, 85.41 DS-28 and SBYF-425, respectively. There are 471270, 2723 and 780 variants predicted at depth cut-off 10 with reference genome of *G. raimondii, G. arboreum* and *G. hirsutum,* respectively. Similarly, Byers et al. (2014) used 8x minimum coverage depth to identify allelic SNP's in allotetraploid cotton. A total number of 471270 variants were identified with reference *to G. raimondii* genome sequence at minimum depth cutoff of 10. SNP calling was performed using samtools (samtools version 0.1.18). The total 35 SNPs were selected (based on the various criteria) and bioinformatics analysis for filtered variants, flanking sequences obtained between DS-28 and SBYF-425 and homology search (BLASTn) for flanking sequences was carried out to find out identity percentage, alignment length, mismatches and E-valve with D5 cotton genome (known for trait of interest) in the NCBI nucleotide database. Similarly, SNP discovery was made between *G. hirsutum* and *G. barbadense* by using a genome reduction experiment and then KBioscience KASPar assays were designed for a portion of the intra-specific *G. hirsutum* SNPs as per Bayer et al. (2012). The KASPar assay was designed for 5 selected SNPs of a target site of interest. It is suggested that the polymorphic SNPs and INDELs can further be used for genotyping and QTL mapping after converting into functional markers using other genotyping platforms for fibre quality traits improvement in MAS breeding programme.

Authors' contribution

Conceptualization of research (ISK); Designing of the experiments (ISK, HMV); Contribution of experimental materials (ISK); Execution of field/lab experiments and data collection (SR, ISK); Analysis of data and interpretation (MPJ, NC, VBR); Preparation of manuscript (SR, MVM, SA, ISK).

Declaration

The authors declare no conflict of interest.

References

An C., Saha S., Jenkins J. N., Ma D. P., Scheffler B. E., Kohel R. J., John Z. Y. and Stelly D. 2008. Cotton (*Gossypium* spp.) R2R3-MYB transcription factors SNP identification, phylogenomic characterization, chromosome localization, and linkage mapping. Theor. Appl. Genet., **116:** 1015-1026.

Buriev Z. T., Saha S., Abdurakhmonov I. Y., Jenkins J. N., Abdukarimov A., Scheffler B. E. and Stelly D. M. 2010. Clustering, haplotype diversity and locations of MIC3: a unique root-specific defense-related gene family in Upland cotton (*Gossypium hirsutum* L.). Theor. Appl. Genet., **120:** 587 - 606.

Byers R. L., Harker D. B., Yourstone S. M., Maughan P. J. and Udall J. A. 2012. Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet., **124:** 1201-1214.

Gao W., Chen Z. J, Yu J. Z., Kohel R. J., Womack J. E. and Stelly D. M. 2006. Wide-cross whole-genome radiation hybrid mapping of the cotton (*Gossypium barbadense* L.) genome. Mol. Genet. Genomics, **275:** 105-113.

Gao W., Chen Z. J., Yu J. Z., Raska D., Kohel R. J., Womack J. E. and Stelly D. M. 2004. Wide-cross whole-genome radiation hybrid mapping of cotton

(*Gossypium hirsutum* L.). Genetics, **167:** 1317-1329.

Guo W., Cai C., Wang C., Han Z., Song X., Wang K., Niu X., Wang C., Lu K. and Shi B. 2007. A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*. Genetics, **176:** 527-541.

Islam M. S., Thyssen G. N. and Fang D. D. 2014. Detection, Validation and Application of Genotyping-by-Sequencing Based Single Nucleotide Polymorphisms in Upland Cotton (*Gossypium hirsutum* L.). The Plant Genome, **8**(1): 1-10.

Jordan S. A. and Humphries P. 1994. Single nucleotide polymorphism in exon 2 of the BCP gene on 7q31-q35. Hum. Mol. Genet., **3:** 1909-1915.

Lacape J. M., Nguyen T. B., Thibivilliers S., Bojinov B., Courtois B., Cantrell R. G., Burr B. and Hau B. 2003. A combined RFLP-SSR-AFLP map of tetraploid cotton based on a *Gossypium hirsutum* x *Gossypium barbadense* backcross population. Genome, **46:** 612-626.

Peterson B. K., Weber J. N., Kay E. H., Fisher H. S. and Hoekstra H. E. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE, **7:** e37135.

Rai K. M., Singh S. K., Bhardwaj A., Kumar V., Lakhwani D., Srivastava A., Jena S. N., Yadav H. K., Bag S. K. and Sawant S. V. 2013. Large-scale resource development in *Gossypium hirsutum* L. by 454 sequencing of genic-enriched libraries from six diverse genotypes. Plant Biotechnol. J., **11:** 953-963.

Reinisch A. J., Dong J. M., Brubaker C. L., Stelly D. M., Wendel J. F. and Paterson A. H. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics, **138:** 829-847.

Shappley Z. W., Jenkins J. N., Meredith W. R. and McCarty J. C. 1998. An RFLP linkage map of Upland cotton, *Gossypium hirsutum* L. Theor. Appl. Genet., **97:** 756-761.

Van Deynze A., Stoffel K., Lee M., Wilkins T. A., Kozik A., Cantrell R. G., Yu J. Z., Kohel R. J. and Stelly D. M. 2009. Sampling nucleotide diversity in cotton. BMC Plant Bio., **9**: 125-136.

Wang K., Wang Z., Li F., Ye W., Wang J. et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii.* Nature Genet., **44:** 1098-1104.

Wendel J. F., Brubaker C., Alvarez I., Cronn R. and Stewart J. M. 2009. Evolution and Natural History of the Cotton Genus. In: Genetics and Genomics of Cotton (Ed. A. H. Paterson) Volume 3. New York: Springer. 3-22.

Yu J. Z., Kohel R. J., Fang D. D., Cho J., Van Deynze A., Ulloa M., Hoffman S. M., Pepper A. E., Stelly D. M., Jenkins J. N., Saha S., Kumpatla S. P., Shah M. R., Hugie W. V. and Percy R. G. 2012. A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. G3 (Bethesda), **2:** 43-58.