

Development and characterization of genomic and expressed SSRs for levant cotton (*Gossypium herbaceum* L.)

Satya Narayan Jena · Anukool Srivastava · Krishan Mohan Rai · Alok Ranjan · Sunil K. Singh · Tarannum Nisar · Meenal Srivastava · Sumit K. Bag · Shrikant Mantri · Mehar Hasan Asif · Hemant Kumar Yadav · Rakesh Tuli · Samir V. Sawant

Received: 27 April 2011 / Accepted: 9 October 2011 / Published online: 30 October 2011
© Springer-Verlag 2011

Abstract Four microsatellite-enriched genomic libraries for CA(15), GA(15), AAG(8) and ATG(8) repeats and transcriptome sequences of five cDNA libraries of *Gossypium herbaceum* were explored to develop simple sequence repeat (SSR) markers. A total of 428 unique clones from repeat enriched genomic libraries were mined for 584 genomic SSRs (gSSRs). In addition, 99,780 unigenes from transcriptome sequencing were explored for 8,900 SSR containing sequences with 12,471 expressed SSRs. The present study adds 1,970 expressed SSRs and 263 gSSRs to the public domain for the use of genetic studies of cotton. When 150 gSSRs and 50 expressed SSRs were tested on a panel of four species of cotton, 68 gSSRs and 12 expressed SSRs revealed polymorphism. These 200 SSRs were further deployed on 15 genotypes of levant cotton for the genetic diversity assessment. This is the first report on the successful use of repeat enriched genomic library and expressed sequence database for microsatellite markers development in *G. herbaceum*.

Communicated by P. Heslop-Harrison.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1729-y) contains supplementary material, which is available to authorized users.

S. N. Jena · A. Srivastava · K. M. Rai · A. Ranjan · S. K. Singh · T. Nisar · M. Srivastava · S. K. Bag · M. H. Asif · H. K. Yadav · S. V. Sawant (✉)
Plant Molecular Biology and Genetic Engineering Laboratory,
National Botanical Research Institute, Rana Pratap Marg,
Lucknow, India
e-mail: samirsawant@nbri.res.in

S. Mantri · R. Tuli
National Agri-Food Biotechnology Institute,
Mohali, Punjab, India

Introduction

Cotton (*Gossypium* spp.) is an economically important crop and its contribution to the global fiber industry is continuously growing. Presently, cotton is under cultivation in more than 70 countries globally with India ranking first in acreage (8.53 m ha) and third in production (16 m bales) after China and USA, respectively. About 90% of the commercially produced cotton comes from the species *G. hirsutum* L. followed by *G. barbadense* L. (8%). The rest of the 2% belongs to *G. herbaceum* L. and *G. arboreum* L. grown mainly in South and Southeast Asia. *G. herbaceum*, commonly known as “levant cotton”, is usually grown in rain-fed areas and found to be tolerant to various abiotic stresses like salinity/sodicity, drought/water and wind. Many efforts have been undertaken to develop Simple Sequence Repeat (SSR) markers both from genomic (Blenda et al. 2006; Nguyen et al. 2004) and EST sequences (Han et al. 2004; Qureshi et al. 2004; Zhang et al. 2007) for cotton. SSRs are the comparatively most efficient class of molecular markers that show a high level of polymorphism (Tautz and Renz 1984) with variation in their abundance in different taxon and reported to be higher in the non-coding regions in comparison to coding region (Hancock 1995). The SSRs present in the coding regions are reported to be less polymorphic in comparison to those in the non-coding regions (Cuadrado and Schwarzacher 1998). The SSR markers have been successfully applied to a variety of genetic studies, including the construction of the genetic maps and quantitative trait loci (QTL) gene tagging (Korzun et al. 1999; Nguyen et al. 2004; Ramsay et al. 2000; Taramino et al. 1997), assessment of genetic diversity (Doldi et al. 1997; Gao et al. 2003), MAS (Gao et al. 2003; Guo et al. 2006), cultivar identification and pedigree studies (Sefc et al. 1997). EST–SSR markers have

been preferred over gSSR markers for various genetic improvement programs owing to their higher inter-specific transferability (Guo et al. 2006) and they were assessed as better candidates for gene tagging (Rong et al. 2004). A saturated genetic linkage map with the whole genome coverage is crucial in the analysis of QTL for important agronomic traits, map-based cloning and comparative genomics studies. Thus, many researchers developed mapping populations and constructed genetic linkage maps in cotton (Guo et al. 2007; Rong et al. 2004). However, most of the linkage mapping and QTLs studies carried out in cotton are based on *G. hirsutum* × *G. barbadense* progenies. Thus, majority of the microsatellite markers developed so far belongs to AD-genomes (*G. hirsutum* and *G. barbadense*) and only few from A-genome such as *G. arboreum*. Some of these markers reported earlier have failed to integrate in the genetic maps of the diploid A-genome species, such as *G. herbaceum* and *G. arboreum* (Rong et al. 2004). However, evolutionary pattern, polyploidy nature and complex genome structure of various *Gossypium* species made the transferability of SSRs high amongst the *Gossypium* species and even in allied genera, resulting in a small number of amplification failure and non-specific amplification (Guo et al. 2006). The high degree of cross-species transferability of SSRs would be useful and valuable attribute in characterizing the species relationships, trait specific introgression and mining desirable alleles from the wild germplasm pools in *Gossypium*. Thus, there is a great need to discover and develop more microsatellite markers for genetic and linkage analyses in levant cotton with a valuable attribute of cross-species transferability.

Therefore, the present investigation was undertaken to: (1) develop a comprehensive set of genomic and expressed SSRs in levant cotton; (2) analyze the frequency of newly developed SSRs in the levant cotton genome; (3) evaluate the cross-species/genera transferability and their polymorphism in a set of 20 cotton accessions belonging to four species and (4) study genetic diversity among 15 accessions of levant cotton using a large dataset of gSSRs and expressed SSRs.

Materials and methods

Plant materials

Twenty elite genotypes including five each of *G. herbaceum*, *G. arboreum*, *G. hirsutum* and *G. barbadense* and one each from four different genera of family Malvaceae i.e. *Hibiscus rosa sinensis* L., *Thespesia populnea* (L.) Soland. ex Correa., *Abelmoschus esculentus* L. and *Kydia calycina* Roxb. were used in the present study (Table 1).

The leaf samples of each genotype were collected from the germplasm bank maintained at the Plant Molecular Biology Division, National Botanical Research Institute, Lucknow, India and the voucher specimens of all the genotypes were deposited in the Herbarium (LWG) of the National Botanical Research Institute, Lucknow, India.

Construction of genomic libraries and sequencing

Genomic DNA was isolated from *G. herbaceum* cultivar VAGAD following the modified C-TAB protocol (Jena et al. 2004). The microsatellite capture and enrichment was carried out following (Jones et al. 2000) with little modifications. Briefly, genomic DNA was partially digested with a cocktail of seven blunt-end restriction enzymes (*Bsr*BI, *Eco*RV, *Hae*III, *Pvu*II, *Rsa*I, *Sca*I and *Stu*I). 300–750 bp fragments were ligated to a 20 bp adapter, which contained a *Hind*III site at the 5' end, and then subjected to magnetic bead capture with 5'-biotinylated probe using Biotin-CA(15), Biotin-GA(15), Biotin-AAG(8) and Biotin-ATG(8) as the capture molecules for the four genomic libraries. The captured molecules of the four genomic libraries were amplified using a primer complementary to the adapter and digested with *Hind*III to remove the adapter sequences. Each library was ligated into the *Hind*III site of pUC19. The plasmids were then transformed into *Escherichia coli* DH5 α . Recombinant clones were chosen arbitrarily for sequencing on an ABI's DNA Analyzer 3730xl, using the ABI Big Dye terminator cycle sequencing v3.1 kit. Sequence analysis was performed on the Sequence Analyzer version 3.0. The sequences of 263 clones with SSRs were deposited to NCBI genebank (Accession number HQ524320–HQ524575 and JF495492–JF495686).

Transcriptome sequencing and assembly

The total RNA from the leaf and the root tissues of the four genotypes of *G. herbaceum* (RAHS 14, RAHS IPS 187, VAGAD and GUJCOT 21) was isolated using the Spectrum plant total RNA Kit (Sigma-Aldrich). Plant tissues were ground to a fine powder in liquid nitrogen and lysed in a lysis solution that releases RNA and at the same time inactivates ribonucleases and interfering secondary metabolites, such as polyphenolic compounds. After the removal of cellular debris by centrifugation and filtration steps, RNA was captured onto a binding column using a unique binding solution, which effectively prevents polysaccharides as well as genomic DNA from clogging the column. Residual impurities and most residual genomic DNA were removed by wash solutions and purified RNA was eluted in RNase-free water. After the DNaseI (Ambion) treatment, integrity and quantity of RNA was checked using Bioanalyzer 2100 (Agilent Inc., Palo Alto, CA,

Table 1 Details of the 20 genotypes of cotton belonging to four different species cultivated in India and four other allied genera of family Malvaceae

Sl. no.	Genotype name	Species	Source of collection	Accession number ^f	Origin
1	579	<i>G. arboreum</i>	UAS-D, Dharwad, Karnataka ^a	250324	Old World
2	551	<i>G. arboreum</i>	UAS-D, Dharwad, Karnataka ^a	250326	Old World
3	557	<i>G. arboreum</i>	UAS-D, Dharwad, Karnataka ^a	250320	Old World
4	221567	<i>G. arboreum</i>	GAU, Banaskantha, Gujarat ^b	221567	Old World
5	221566	<i>G. arboreum</i>	GAU, Banaskantha, Gujarat ^b	221566	Old World
6	249001	<i>G. barbadense</i>	Guwahati, Assam	249001	New World
7	249003	<i>G. barbadense</i>	Dakuapara, Assam	249003	New World
8	249004	<i>G. barbadense</i>	Dakuapara, Assam	249004	New World
9	249002	<i>G. barbadense</i>	Bhagdabari, Assam	249002	New World
10	Suvin	<i>G. barbadense</i>	TNAU, Tamil Nadu ^c	250345	New World
11	VAGAD	<i>G. herbaceum</i>	UAS-D, Dharwad, Karnataka ^a	250311	Old World
12	GUJCOT21	<i>G. herbaceum</i>	UAS-D, Dharwad, Karnataka ^a	250314	Old World
13	RAHS14	<i>G. herbaceum</i>	UAS-D, Dharwad, Karnataka ^a	250332	Old World
14	IPS187	<i>G. herbaceum</i>	UAS-D, Dharwad, Karnataka ^a	250313	Old World
15	RAHS132	<i>G. herbaceum</i>	UAS-D, Dharwad, Karnataka ^a	250312	Old World
16	JKC703	<i>G. hirsutum</i>	JK Agri., Hyderabad, Andhra Pradesh ^d	250304	New World
17	JKC771	<i>G. hirsutum</i>	JK Agri., Hyderabad, Andhra Pradesh ^d	250346	New World
18	JKC777	<i>G. hirsutum</i>	JK Agri., Hyderabad, Andhra Pradesh ^d	250335	New World
19	LRA5166	<i>G. hirsutum</i>	JK Agri., Hyderabad, Andhra Pradesh ^d	250341	New World
20	AS2	<i>G. hirsutum</i>	UAS-D, Dharwad, Karnataka ^a	250347	New World
21	Ae	<i>Abelmoschus esculentus</i>	NBRI, Lucknow, Uttar Pradesh ^e	252132	–
22	Hr	<i>Hibiscus rosasinensis</i>	NBRI, Lucknow, Uttar Pradesh ^e	252133	–
23	Cc	<i>Kydia calycina</i>	NBRI, Lucknow, Uttar Pradesh ^e	252131	–
24	Tp	<i>Thespesia populnea</i>	NBRI, Lucknow, Uttar Pradesh ^e	249679	–

^a University of Agricultural Sciences, Dharwad, Karnataka

^b Gujarat Agricultural University, Banaskantha, Gujarat

^c Tamil Nadu agricultural university, Tamil Nadu

^d JK Agri Genetics, Hyderabad, Andhra Pradesh

^e National Botanical Research Institute, Lucknow, UP

^f Accession number assigned at National Botanical Research Institute (NBRI) herbarium, Lucknow

USA). Three micrograms of the total RNA was reverse transcribed using a T7-Oligo (dT) primer in the first-strand cDNA synthesis reaction. Following the RNase H-mediated second-strand cDNA synthesis, the double-stranded cDNA was further enriched which served as a template in the subsequent in vitro transcription (IVT) reaction. The IVT reaction was carried out using T7 RNA Polymerase for complementary RNA (cRNA). The cRNA (3 µg) was then reverse transcribed in the first-strand cDNA synthesis step of using random hexamer primer, followed by the RNase H-mediated second-strand cDNA synthesis. The samples were column-purified (QIAquick PCR purification kit, Qiagen) and sequenced using 454 GS FLX pyrosequencer following standard protocol (Jarvie and Harkins 2008). A total of 318,872 reads generated were assembled into 147,510 unigenes (106,960 singletons and 40,550 contigs). The assembly was done using CAP3 assembler

(Huang and Madan 1999) with the criteria of minimum of 40 bases overlap and 90% identity. The expressed sequences were deposited to NCBI database (GenBank accession SRX0368495-99).

SSR mining, PCR amplification and fragment analysis

The identification and localization of the microsatellites in the sequences derived from SSR enriched libraries and transcriptome pyrosequencing was performed using the simple sequence repeat identification tool (SSRIT), downloaded from the Cornell University web link <http://gramene.agrinome.org/db/searches/ssrtool> and the microsatellite (MISA) searching tool (<http://pgrc.ipk-gatersleben.de/misa/misa.html>). The criteria for the SSR search were repeat stretches having a minimum of ten repeat units for mononucleotide repeats (MNRs), five repeat units for dinucleotide

(DNRs), trinucleotide (TNRs), tetranucleotide (TtNRs), pentanucleotide (PNRs) and hexanucleotide repeats (HNRs). The maximal number of interrupting bases in a compound microsatellite was set to 100. Primer pairs were designed for the SSR containing sequences with a minimum of five repeats for all the SSRs using PRIMER3 (<http://frodo.wi.mit.edu/primer3/>) with threshold criteria of 20–28 nucleotide length, T_m of 55–65°C and GC content of 45–65%. The primers were commercially synthesized (ABI, Gene Identification Pvt. Ltd, USA and MWG Pvt. Ltd, Bangalore, India) with forward primers having the fluorescent label FAM/HEX/VIC/NED. The details of these new markers viz., locus designation, primer sequences, repeat motifs and GenBank accession numbers are summarized in Supplemental data S1. The primer pairs were standardized and PCR was performed. The PCR reactions included 20 ng of template DNA, 10 μ l of PCR master mix (Fermentas, USA) and 0.2 μ M of each primer in a 20 μ l reaction. The PCR protocols included: initial denaturation of 5 min at 95°C followed by 40 cycles with 50 s at 95°C, 50 s at 58–64°C (primer specific), 90 s at 72°C and a final extension of 10 min at 72°C. The amplified products were run on a capillary-based 3730xl DNA Analyzer (ABI, USA) and the products were precisely sized for major, comparable and conspicuous peaks using GeneMapper v4.0 (ABI, USA), with default parameters.

Genomic SSRs and expressed SSRs redundancy analysis

All the developed SSRs, including genomic and expressed SSRs, were compared with the entire cotton marker database (CMD, <http://www.cottonmarker.org>) to assess the novelty/newness of these SSR sequences. All the cotton SSR sequences were downloaded from CMD as of 25 April 2011. The whole sequences with SSR repeat motif were searched for their match with BLASTN search against Cotton Marker Database. In addition, the forward flanking region (50 bp immediate upstream the repeat motif) and reverse flanking region (50 bp immediate downstream the repeat motif) were aligned with said marker database at $\geq 90\%$ identity to test their novelty. Besides these verifications, the primer pairs from all the SSR containing sequences were also used to compare with the downloaded reference sequences for the duplications or redundancies with the reported SSRs in public database as per the earlier literature for redundancy verification (Xiao et al. 2009).

Statistical and genetic analyses

The allelic data for each of the five genotypes of *G. herbaceum*, *G. arboreum*, *G. hirsutum* and *G. barbadense*

were used to calculate the different statistical and genetic parameters using Power Marker (v 3.25), Popgene (v1.32) and Arlequin (v3.1). The effective allele number is estimated as the reciprocal of homozygosity (Hartl and Clark 1989) and the Shannon index is measured for gene diversity (Shannon and Weaver 1949). The observed heterozygosity (H_o) was calculated as a fraction of the heterozygous genotypes over the total number of genotypes. The expected heterozygosity (H_e) was calculated based on the probability that two individuals taken at random from a given sample would possess different alleles at a locus (Nei and Li 1979) and according to the following formula:

$$H_e = (n/n - 1) \left(1 - \sum p_i^2 \right)$$

The polymorphism information content (PIC) value for each SSR marker was calculated following (Botstein et al. 1980) as:

$$PIC = 1 - \sum p_i^2 - \sum_n \sum 2p_i^2 p_j^2$$

where n is the total number of alleles detected for a microsatellite marker, p_i is the frequency of the i th allele and p_j is the frequency of the $(i + 1)$ th allele in the set of the analyzed genotypes. The cross-taxa transferability was calculated as a proportion of the primers showing successful amplification in relation to all the tested primers. For *G. hirsutum*, the markers that showed invariable presence of ‘double alleles’ across the tested germplasm were considered as independent amplifications from the duplicated loci present in the two distinct copies. These markers were excluded from the analysis for the allelic attributes described above.

The allelic data from 15 genotypes of *G. herbaceum* were used to ascertain the genetic relationships/affinities among them using cluster analysis based on the Nei’s genetic distance measure with 100 bootstraps. Further, UPGMA trees were generated separately for each matrix using Mega 4.1 by ‘neighbor’ command, which was followed by the generation of consensus trees, one each for the cultivated germplasm and inter-species relationships.

Cross-species/genera SSR transferability

Amplification products of five each of expressed and gSSRs which showed amplification in all four species of *Gossypium* and four allied genera of Malvaceae were sequenced to examine their conservedness and cross-species/genera transferability. The final edited sequences belonging to each locus were compared with the respective original expressed SSR and gSSR sequences using CLUSTALX (<http://www.ftp-igbmc.u-strasbg.fr/pub/ClustalX>) for ascertaining the target domain/SSR conservation.

Results

Identification and characterization of genomic SSRs derived from enriched libraries

A total of 750 recombinant colonies was identified and 500 clones were chosen arbitrarily for sequencing. Out of the 500 sequences, 60 were found to be redundant, 12 were without repeat or repeat with motif frequency less than five and rest 428 sequences were unique. These 428 sequences were searched for SSRs which resulted in identification of 584 gSSRs (Table 2). In the 428 SSR containing sequences, 263 (with 373 gSSRs) were found suitable for primer designing covering a total of 319 gSSRs (Supplemental Table S1 for gSSR primer details) and rest 165 sequences (with 211 gSSRs) were found to be unsuitable and hence not taken up for further analysis. Out of these 373 gSSRs, 48 gSSRs (12.9%) represented non-targeted MNRs, while 77 gSSRs (20.6%) represented DNRs (58 targeted and 19 non-targeted), 243 gSSRs (65.1%) represented TNRs (207

targeted and 36 non-targeted), 3 gSSRs represented non-targeted TtNRs and 2 gSSRs represented non-targeted HNRs from the genomic enrichment library.

A total of 38 DNRs out of 77, and 177 TNRs out of 243 had perfect repeat motifs. In addition, each single TtNR and HNR had a perfect repeat motif. A total of 46 were compound repeats with 102 gSSRs in which 36 were compound imperfect while 10 were compound perfect (Supplemental Table S1 for repeat details in different SSRs). The 428 gSSRs containing sequences represented approximately 192.6 kb (avg length 0.45 kb) of the cotton genome (Table 2) and an average frequency of gSSR as $\sim 1/0.33$ kb and/or 1.36 gSSR per clone in the cotton enriched library.

Amongst the targeted motifs of gSSRs, the TNR motif AAG/CTT was most abundant with a frequency of 46.6% (Fig. 1), while the DNR motif AG/CT had a frequency of 8.6% (Fig. 1). In addition, a few non-targeted DNRs (AT/AT), TNRs (AAT/ATT, ACC/GGT, AGC/CGT, AGG/CCT, AGT/ATC), TtNRs (AAAG/CTTT, AAGG/CCTT)

Table 2 Summary statistics of screening of the small-insert partial genomic library of cotton (*G. herbaceum*) for putative SSR recombinant clones/sequences and genomic SSRs

Summary of screening/sequencing	
Total Number of clones screened	750
Number of clones selected and sequenced after screening	500
Number of redundant clones	60
Number of clones without repeat/repeat with less number of motif	12
Number of SSR containing sequences used for primer design/synthesis	263 (containing 373 SSRs)
Number of SSR containing sequences unable to design primer	165 (containing 211 SSRs)
Total number of SSR containing clones	428
Number of sequences containing more than 1 SSR	87 (from successful designed sequences) and 74 (from failure sequences)
Number of sequences containing compound SSRs	46 (from successful designed sequences) and 22 (from failure sequences)
Average size of the cloned/sequenced inserts	0.45 Kb
Haploid genome size of <i>G. herbaceum</i> (A)	1667 Mb
Summary of SSRs identified	
	In the successful 263 sequences
Number of non-targeted MNRs of minimum 10-mer length (a)	48
Number of targeted DNRs having a minimum length of 10 (b)	58
Number of non-targeted DNRs having a minimum length of 10 (c)	19
Total number of DNRs (b + c)	77
Number of targeted TNRs having a minimum length of 15 (d)	207
Number of non-targeted TNRs having a minimum length of 15 (e)	36
Total number of TNRs (d + e)	243
Number of non-targeted TtNRs of minimum length of 25(f)	3
Number of non-targeted PNRs of minimum length of 25 (g)	0
Number of non-targeted HNRs of minimum length of 30 (h)	2
Total Number of SSRs (a + b + c + d + e + f + g + h)	373

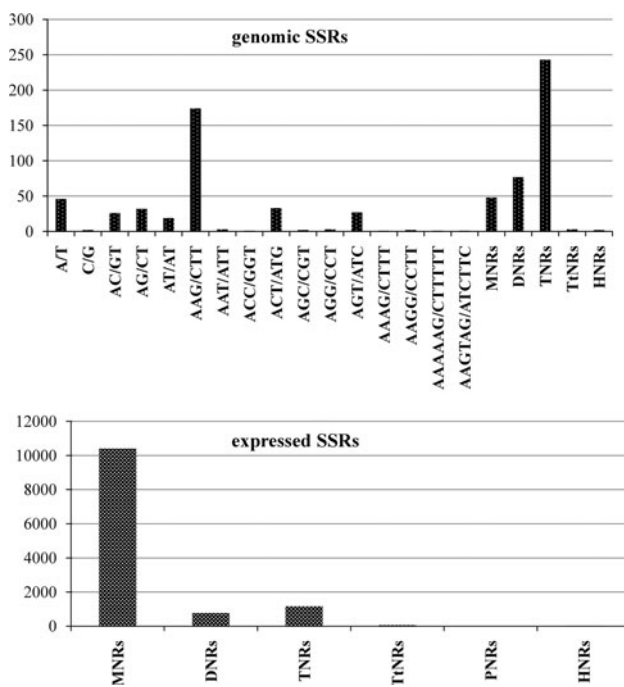


Fig. 1 Frequencies of genomic SSRs and expressed SSRs in levant cotton *G. herbaceum* cultivar VAGAD. The gSSRs and expressed SSRs frequencies were calculated in SSR counts in repeat enriched genomic library sequences and expressed sequences database, respectively

and HNRs (AAAAAG/CTTTTT, AAGTAG/ATCTTC, ACATAT/ATATGT) motifs were also recovered. Contaminations of reverse motifs were also obtained with the targeted motifs as expected. 14 AG/CT motifs were obtained in a library captured for the GA/TC motif and 17 AC/GT motifs in the libraries captured for the CA/TG motif. In addition, 27 AGT/ATC, 3 AAT/ATT, 1 ACC/GGT, 2 AGC/CGT and 3 AGG/CCT motifs were also obtained in the library captured for AAG/CTT and ATG/CAT motifs.

Identification and characterization of expressed SSRs derived from transcriptome sequences

The transcriptome sequencing of *G. herbaceum* produced 318,872 reads those were assembled into 147,510 unigenes, out of which 40,550 were contigs and 106,960 were singleton. A total of 99,780 unigenes, with a read length greater than 180 bp, were analyzed for identification of expressed SSRs. In these 99,780 unigenes, 8,900 were found to be SSR containing and represented a total number of 12,471 expressed SSRs. These 12,471 expressed SSRs include 10,407 MNRs, 776 DNRs, 1,168 TNRs, 67 TtNRs, 18 PNRs and 35 HNRs (Table 3). Out of 8,900, 1,918 expressed SSR containing sequences (with 2,369 expressed SSRs) were found suitable for primer designing. Separate primers were designed for those SSRs which were distant

by more than 100 bp in a single sequence. Thus, a total of 1,970 primers were designed covering 2,029 expressed SSRs in 1,918 sequences (Supplemental Table S2 for expressed SSR primer details). Rest 6,982 SSR containing sequences (with 10,102 expressed SSRs) were found to be unsuitable and hence not taken up for further analysis (Table 3). A total of 65 different SSR motifs were identified in the unigene set. The top 15 motifs (any two complementary sequences were considered one motif) represented 99% of the expressed SSRs, while the remaining 50 motifs accounted for only 1% of the expressed SSRs detected. Among the DNRs, the AT motif was the most common (3.8%) followed by the AG (1.85%) and AC (0.6%) motifs. Similarly, among the TNRs, the motif AAG was the most common (2.8%) followed by the motifs ACC (1.3%) and AGT (1.1%). However, the expressed SSRs data in the present study represented an extremely less number of TtNRs and PNRs (Fig. 2).

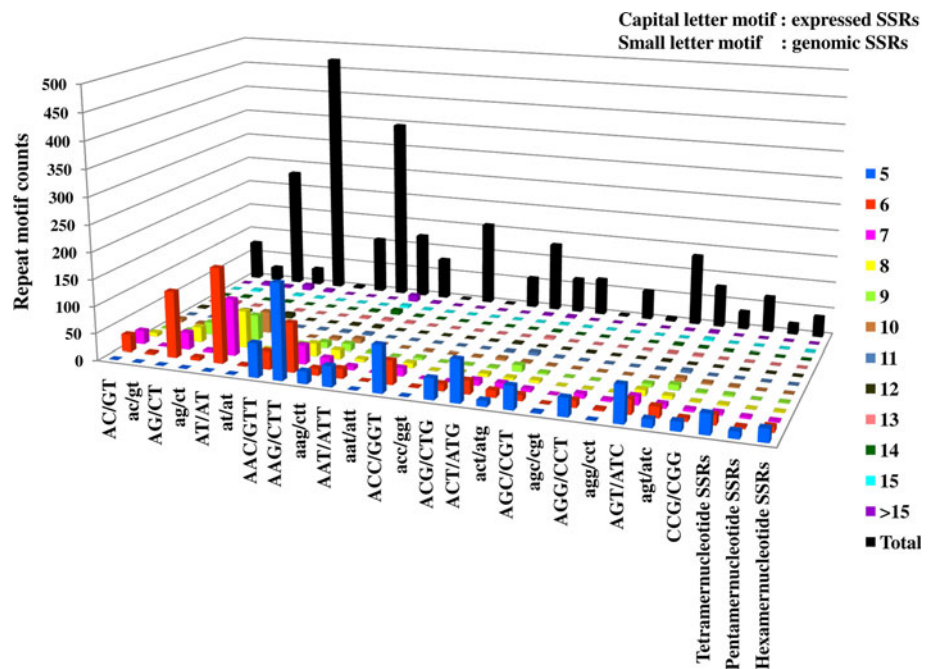
Of the total 1,918 expressed SSR containing sequences those were successful in primer designing, 1,865 (97.2%) contained simple repeat motifs, while 53 (2.8%) were of a compound type (Supplemental Table S2). Among all motifs, DNR motif AT/AT were most abundant with frequency of 5.8% followed by TNR motifs AAG/CTT and DNR motifs AG/CT with frequencies of 3.5 and 2.9%, respectively. The 99,780 sequences represented approximately 29.9 Mb of cotton transcriptome having 12,471 expressed SSRs, thereby suggesting an average frequency of SSR as $\sim 1/2.4$ kb and/or 1/8 expressed sequences in the cotton transcriptome analyzed (Table 3). Furthermore, most of these SSRs represented smaller repeat-unit size.

Marker transferability and genetic relationship across species/allied genera

To test cross-species/genera transferability 200 SSRs were tested on a panel of four species of cotton *i.e.* *G. hirsutum*, *G. herbaceum*, *G. barbadense* and *G. arboreum* with five elite genotype each and four allied species (*Hibiscus rosasinensis*, *Thespesia populnea*, *Abelmoschus esculentus* and *Kydia calycina*) with one genotype each. The newly developed 200 SSRs from *G. herbaceum* showed 89.0% amplification (178) in *G. hirsutum*, 82.0% (164) in *G. barbadense*, 93.0% (186) in *G. herbaceum* and 87.0% (174) in *G. arboreum* (Table 4). A maximum of 76.0% cross-genera transferability noticed with *Hibiscus* followed by *Thespesia* (71.5%), *Kydia* (66.0%) and *Abelmoschus* (58.0%). Overall, an average transferability of $\sim 88\%$ was observed within different *Gossypium* species (across species), which was higher than other genera of the family (across genera), *i.e.* *Hibiscus* spp. ($\sim 76\%$), *Thespesia* spp. ($\sim 71.5\%$), *Kydia* spp. ($\sim 66\%$) and *Abelmoschus* spp. ($\sim 58\%$).

Table 3 Summary statistics of screening of cotton (*G. herbaceum*) transcriptome for putative contigs and expressed SSRs

Summary of screening/sequencing	
Total Number of reads in transcriptome sequencing	318,872
Number of unigenes after assembly	147,510 (40,550 contigs and 106,960 singletons)
Number of transcripts analyzed with greater than 180 bp	99,780
Number of SSR containing sequences	8,900 (containing 12,471 SSRs)
Number of sequences containing more than 1 SSR	1,948
Number of SSR containing sequences used for primer designing	1,918 (containing 2,369 SSRs)
Number of SSR containing sequences unsuitable for primer designing	6,982 (containing 10,102 SSRs)
Number of SSR containing sequences used for primer synthesis	50
Average size of the contigs	0.3 Kb
Estimated transcriptome screened (number of contigs analyzed × average contigs size)	29.9 Mb
Summary of expressed SSRs identified in the cotton ESTs	
	In the successful 1,918 sequences
Total number of MNRs of minimum 10-mer length (a)	1,842
Total number of DNRs of minimum 10-mer length (b)	241
Total number of TNRs of minimum 15-mer length (c)	263
Total number of TtNRs of minimum 20-mer length (d)	9
Total number of PNRs of minimum 25-mer length (e)	5
Total number of HNRs of minimum 30-mer length (f)	9
Total Number of SSRs (a + b + c + d + e + f)	2,369

Fig. 2 Frequencies of different subclasses of expressed microsatellites measured in repeat motif counts. The total SSR frequency is sub-classified into dimer, trimer, tetramer, pentamer and hexamer SSRs

The verification of redundancy against existing sequence data were performed by deploying three different strategies using BLASTN search. In the first strategy, we performed BLAST hits of the complete sequence against the CMD database, in second, we used 50 bp immediate upstream and downstream of repeat motif and in third we used forward and reverse primers. A total of 99 non-

redundant novel gSSRs was obtained in the first strategy, whereas 171 and 119 gSSRs were obtained in the second and third strategy, respectively. Similarly, a total of 1,355, 1,662 and 1,491 expressed SSRs were found to be non-redundant in first, second and third strategy, respectively (Supplemental Table S3a and b). When the non-redundant SSRs (gSSRs and expressed SSRs), obtained in three

Table 4 Cross-species transferability of *G. herbaceum*-derived genomic SSRs and expressed SSRs among different genomes in *Gossypium* and other four allied genera

Genome/ species	No. of SSRs amplified in all genotypes	% of SSRs amplified in all genotypes	No. of SSRs amplified only in few genotypes	No. of null amplified SSRs
A1 (<i>G. herbaceum</i>)	186	93.0	14	0
A2 (<i>G. arboreum</i>)	174	87.0	26	0
AD1 (<i>G. hirsutum</i>)	178	89.0	22	0
AD2 (<i>G. barbadense</i>)	164	82.0	36	0
<i>Hibiscus</i> sp.	152	76.0	nd	48
<i>Thespesia</i> sp.	143	71.5	nd	57
<i>Kydia</i> sp.	132	66.0	nd	68
<i>Abelmoschus</i> sp.	116	58.0	nd	84

nd Not determined as each species is represented by single accession, so the number of partial amplified SSRs could not be estimated

strategies were compared each other, we came in a conclusion that 79 gSSRs and 1,308 expressed SSR, common in all three strategies were really novel (Supplemental figure S4).

Further, PCR amplicon of five gSSRs with four cotton species were sequenced to check the cross-species conservation and transferability. These sequence files were analyzed and confirmed unequivocally the cross-species conservation and transferability and three out of five revealed clear distinctions (Fig. 3). In all the cases, the sequenced alleles from the different species were homologous to the original locus from which the marker was developed.

For ascertaining the useful attributes of the genetic markers, a total of 200 SSRs (150 genomic and 50 randomly selected expressed SSRs) were tested on a panel of five elite genotypes of each of four species of *Gossypium* i.e. *G. herbaceum*, *G. arboreum*, *G. hirsutum* and *G. barbadense* (Supplemental Table S5 for details of panel data) and four allied taxa (*Hibiscus rosasinensis*, *Thespesia populnea*, *Abelmoschus esculentus* and *Kydia calycina*) each with one genotype. In general, these 200 newly developed markers revealed low to medium allelic diversity (Supplemental figure S6 a, b), and notably 72 of them resulted in single allele in the case of all the four species of *Gossypium* studied. Overall, a maximum of four and six alleles with an average of 2.7 and 3.8 alleles/marker were obtained for diploid and tetraploid genotypes, respectively. The distribution of the number of alleles amplified by each polymorphic marker was highly skewed for *G. herbaceum*. The mean PIC value for *G. herbaceum* was 0.32 (range 0.16–0.80), which was higher than 0.20 (range 0.16–0.80) observed for *G. arboreum*, and 0.23 and 0.24 (range 0.16 to 0.80) observed for *G. barbadense* and *G. hirsutum*, respectively (Table 5). The above mentioned SSR allelic data, when used to calculate the heterozygosity estimates, revealed highly significant differences between the observed and expected heterozygosity for *G. herbaceum* (mean H_o : 0.27 and mean H_c : 0.26; paired t value = 3.64; $P = 0.01$) as well as for *G. arboreum* (mean H_o : 0.24;

mean H_c : 0.26; paired t value = -2.54 ; $P = 0.01$). Thus, the results suggested significant heterozygote deficiency in both the germplasm sets (Table 5).

The UPGMA based clustering showed the grouping of all the four species of *Gossypium* in one (cotton cluster) and other four genera in one group (out-group cluster). In the cotton cluster, *G. herbaceum* and *G. arboreum* were more closely related and formed one sub-cluster than the other sub-cluster of *G. hirsutum* and *G. barbadense*. Amongst all the allied species, *Thespesia populnea* stood closer to the cotton cluster.

Generic affinities and population stratification within cultivated *G. herbaceum* germplasm

The 150 gSSRs and 50 expressed SSRs data were also examined for their potential use in genetic diversity analysis among 15 cultivated genotypes of levant cotton along with *Thespesia populnea* as out-group. The Nei's genetic distance (Nei et al. 1983) values ranged from 0.13 (*G. herbaceum* RAHS127 vs. *G. herbaceum* RAHS IPS 187) to 0.30 (*G. herbaceum* VAGAD vs. *G. herbaceum* H17) with an average value of 0.23 ± 0.06 (Fig. 4). The out-group *Thespesia populnea* showed a relatively larger amount of average genetic distance (0.49) to all the genotypes of *G. herbaceum*. Further, the UPGMA phenetic tree generated revealed two major clusters including seven genotypes in each clusters and GUJCOT 21 standalone (Fig. 4).

Discussion

Distribution and abundance of SSR motifs

Expressed sequences available in the public databases are invaluable resources for identifying the SSRs in addition to the genomic resources like BAC-end sequences, repeat enriched clonal sequences, etc. The cotton expressed sequences available in public database are mostly derived

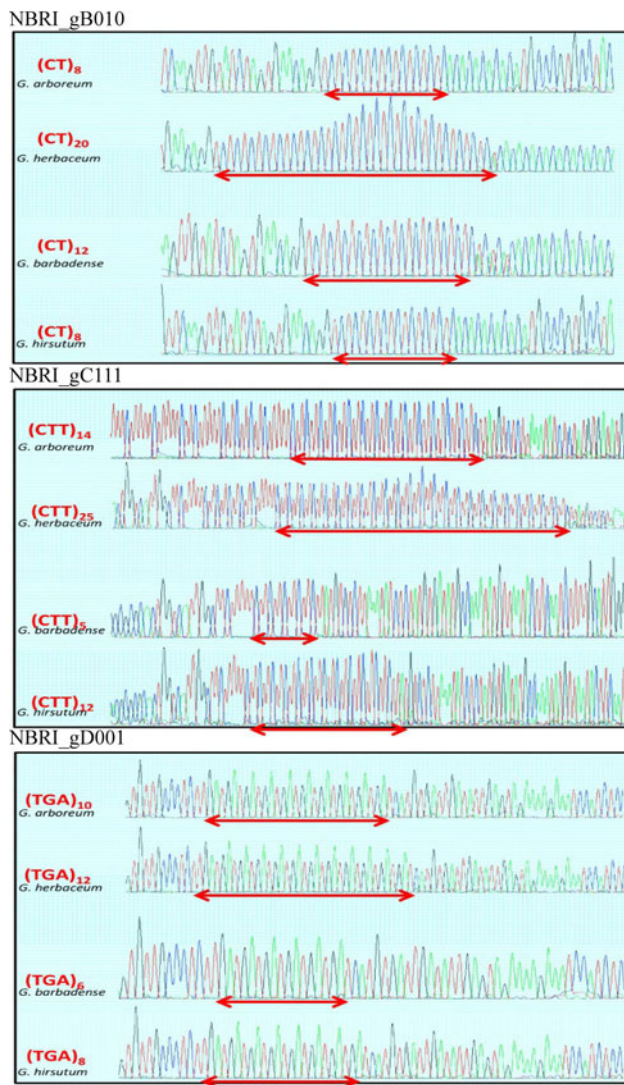


Fig. 3 Comparative electropherogram analysis of three SSR loci (NBRI_gB010, NBRI_gC111 and NBRI_gD001) among four species of cotton

from *G. arboreum*, *G. hirsutum* and *G. barbadense*. However, *G. herbaceum*-derived expressed sequences are lacking in the present public domain. The EST sequences of several crop plants have been widely exploited for development, characterization and deployment of SSRs for

various genetic studies (Varshney et al. 2005). The frequency, distribution and abundance of SSRs are reported to be highly variable depending on the SSR search criteria, the size of the dataset and the database mining tool (Varshney et al. 2005). In various studies conducted earlier, using the same search criterion as opted in the present study, the average frequency of EST–SSRs was reported to be 1/13.05 Kb in *G. arboreum* and *G. hirsutum* (Han et al. 2004; 2006), 1/14.8 kb in *G. raimondii* (Wang et al. 2006) and 1/10.4 kb in *G. barbadense* (Lü et al. 2010). However, the average frequency of expressed SSRs in the present investigation was found to be 1/2.8 Kb of genome of *G. herbaceum*. Thus, the occurrence of SSRs in the present study was more than fourfold higher than the earlier studies in different species of *Gossypium*. This insights our present study strength especially the large amount of dataset of transcriptome sequencing with 454 GS FLX.

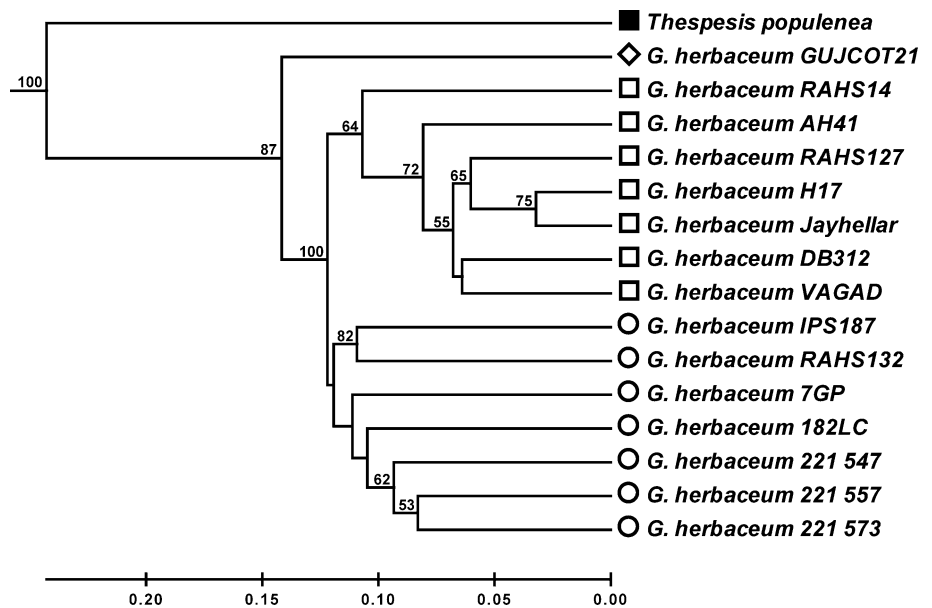
Higher frequency of TNR–SSRs have been reported in ESTs in most of the plants, followed by the DNR and TtNR types (Cardle et al. 2000; Kota et al. 2001; Varshney et al. 2005). Varshney et al. (2002) reported that among cereal species, TNRs were the most frequent (54–78%) followed by DNRs (17.1–40.4%) and TtNRs (3–6%). Frequencies and distribution of different repeat motifs varied substantially in studies by Morgante et al. (2002), Gao et al. (2003) and Kantety et al. (2002). However, amongst the plants, there is a different type of abundance of trinucleotide motif (Gao et al. 2003). Kantety et al. (2002) showed that the (CCG)_n repeat motif is the most abundant in wheat and sorghum, while Gupta et al. (1996) found that the (AAG)_n repeat was the most abundant motif in the trinucleotide repeat. The EST–SSR data obtained from *G. arboreum*, *G. raimondii* and *G. hirsutum* in previous studies have shown that the (AAG)_n repeat motif was the most abundant in the genus *Gossypium* (Guo et al. 2006). The present study on *G. herbaceum* produced similar results. Amongst these expressed SSRs, the most abundant repeat type after MNR was DNR (56.6%). Thus, there are differences in different views in repeat type abundance in different plant taxa. These apparent differences in the relative abundance of the DNRs and TNRs can be attributed to the differences in the SSR search criteria used for EST database mining in

Table 5 Comparative basic statistical measures among four cotton species

Species	Average n_a	Average n_e	Average I	Average H_o	Average H_e	Average PIC
<i>G. herbaceum</i>	1.98	1.59	0.42	0.27	0.26	0.32
<i>G. arboreum</i>	1.87	1.60	0.40	0.24	0.26	0.20
<i>G. hirsutum</i>	1.97	1.66	0.46	0.33	0.31	0.24
<i>G. barbadense</i>	1.96	1.65	0.45	0.35	0.31	0.23

n_a Allele number, n_e effective allele number, I Shannon index, H_o observed heterozygosity, H_e expected heterozygosity, PIC polymorphic information content

Fig. 4 UPGMA tree of 15 genotypes of *G. herbaceum* based on Nei's genetic distance using 200 SSRs



the different studies. It was invariably noted that in most of the earlier studies which showed abundance of TNRs, the minimum number of repeat units was considered higher for DNRs (6–10 repeats) in comparison to TNRs (5–6 repeats). However, in the present study, the same number of minimum repeat units (5) was considered for all types of SSRs (DNRs, TNRs, TtNRs, PNRs and HNRs) except MNRs (minimum 10 repeats). Interestingly, when this criterion was changed to six repeat units for DNRs and five repeat units for TNRs, TtNRs, PNRs and HNRs, a higher abundance of TNRs (56.6%) in comparison to DNRs (37.6%) (data not shown) was obtained, as reported in many earlier studies. Thus, the obtained results of the present study demonstrate that the SSR search criteria used for EST database mining can significantly alter the relative estimates of the frequency/distribution of expressed SSRs, thereby supporting the earlier opinion (Varshney et al. 2005). Subsequently, these data suggest the need for formulating a universally acceptable definition of SSR to obtain more meaningful estimates and avoid discrepancies in the absolute values in all the future comparative studies. Furthermore, it was significant to note that in general, the GC-rich SSR motifs were less frequent in *G. herbaceum* sequences. This was most evident in the relative abundance of AG/AAG and the deficiency of CG/CCG repeats motifs amongst the DNRs/TNRs, respectively identified in the present study. Interestingly, a similar difference in the SSR motif in ESTs has been reported earlier (Cardle et al. 2000; Gao et al. 2003).

Development of new SSR markers

Until 25 April 2011, approximately 16,162 SSR markers have been reported in the Cotton Marker Database

(<http://www.cottonmarker.org/Downloads.shtml>). In spite of the availability of quite a good number of markers, there is lack of dense and saturated genetic map in cotton. Thus, a continuous effort is needed to develop efficient genetic markers for this crop to facilitate the understanding of genetics of complex traits and marker-assisted breeding for development of desired plant types. In this context, a set of 1,308 novel expressed and 79 gSSR markers identified in the present study are expected to be a significant addition to the presently available repertoire of microsatellite markers. The expressed SSR markers, in addition to the merits of the conventional (genomic) SSR markers, are also expected to improve the detection of the marker-trait associations as they are a part of the transcribed domain(s) of the genome. In fact, in recent years emphasis is slowly shifting towards the development of the functional molecular markers instead of the anonymous markers (Andersen and Lubberstedt 2003) as they have the potential for assaying the functional diversity in the germplasm collection or natural population. In addition, they may prove to be more useful for the marker-assisted selection, if found to be associated with a gene/QTL of interest.

Level of allelic polymorphism and genetic diversity

Various genetic parameters viz., allelic diversity, PIC, H_o , H_e calculated for all the newly developed SSRs demonstrated their utility as genetic markers. The markers revealed low to moderate allelic/genetic diversity, which was comparable with the earlier reported cotton gSSRs. In addition, the genetic diversity of gSSRs was invariably higher than the expressed SSRs, as expected. The total number of alleles amplified by the different markers in the tested *G. herbaceum* and *G. arboreum* was almost similar.

However, the markers were found significantly more informative with higher PIC values for *G. herbaceum*. The heterozygosity measures (H_o , H_e) indicated significant heterozygote decay (deficiency) in the tested germplasm.

Cross-species/generic transferability

Although, gSSRs generally show low transferability in comparison to SSRs developed from expressed sequences, but reported to be more polymorphic. Genomic SSRs derived from wheat showed 17% cross-species transferability to rye (Kuleung et al. 2004), and 84% cross-species transferability from wheat to 18 *Triticum-Aegilops* species (Bandopadhyay et al. 2004). SSRs derived from Tall fescue (*Festuca arundinacea* Schreb.) showed 92% transferability to seven grass species (Saha et al. 2004). SSRs derived from *G. arboreum* derived ESTs also showed 96.5% of transferability in 22 diploid species of *Gossypium* (Guo et al. 2006). Similarly, Scott et al. (2000) tested the transferability of ten *Vitis* EST–SSRs among grape cultivars, other grape species and related genera, and found over 60% of of transferability. The EST–SSRs developed from grape and apricot EST sequences (Decroocq et al. 2003) were investigated for transferability across 46 related grape species and 29 members of the *Rosaceous* and found that the grape derived SSRs were transferable to most *Vitaceae* accessions tested while the apricot SSRs were found to be most useful within the subgenus *Prunophora*. Thus, all the earlier studies indicate that EST–SSRs can often be transferred across relatively large taxonomic distances, spanning not just species within a genus, but in some instances multiple genera within a family.

In the present study, high cross-species transferability of approximately 88% was found within diploid as well as tetraploid species of cotton with 150 genomic and 50 expressed SSRs derived from *G. herbaceum*. The transferability of the expressed SSRs was found to be higher than gSSR but with low polymorphism. This indicated that the expressed SSRs had higher conservation in genome across species. The higher transferability of expressed SSRs across diploid and tetraploid cotton species could be related with evolutionary trends of different genome of cotton as A and D sub-genome species diverged from a common ancestor, about 6–11 million years ago and the putative ‘A’ × ‘D’ polyploidization event occurred in the New World, about 1.1–1.9 million years ago (Wendel 1989). However, changes in the SSR repeat number in protein-coding, 5’UTR and 3’UTR regions could lead to a gain or loss of gene function or disrupt other cellular functions (Li et al. 2004). The newly developed SSRs revealed robust cross-species/genera amplifications with alleles of comparable sizes. The cross-species markers transferability found in the present study are comparable

with the earlier report on cotton thereby indicating a high level of transferability amongst the *Gossypium* species.

Diversity analysis and genetic relatedness within/ between cotton and allied species

Despite a relatively low level of polymorphism, the gSSRs and expressed SSRs described here were able to individualize all the 15 genotypes of levant cotton, as well as characterize the 15 genotypes of the other three cultivated cotton species. The UPGMA tree based on the allelic diversity clustered the tested genotypes as per their species status and broadly conformed to their ploidy level. The genetic diversity was higher within the 15 *G. herbaceum* in comparison with the *G. arboreum*, *G. hirsutum* and *G. barbadense* genotypes.

To summarize, the present study adds 1,308 novel expressed SSRs and 79 gSSRs to the public cotton domain. Out of 1,308, 50 randomly selected SSRs (~2.5%) have also been tested for their potential use as expressed SSR markers. In addition, more than 57% gSSRs were successfully validated. Thus, the present study provides expressed SSR and gSSR markers not only for the cultivated cotton species but also for the genetic studies involving related allied species that constitute the important secondary gene pool for the improvement of cotton.

Acknowledgments The present work was financially supported by the Council of Scientific and Industrial Research, New Delhi (CSIR under NMITLI, SIP 005).

References

- Andersen JR, Lubberstedt T (2003) Functional markers in plants. *Trends Plant Sci* 8:554–560
- Bandopadhyay R, Sharma S, Rustgi S, Singh R, Kumar A, Balyan HS, Gupta PK (2004) DNA polymorphism among 18 species of *Triticum-Aegilops* complex using wheat EST–SSRs. *Plant Sci* 166:349–356
- Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, Ficklin S, Staton M, Eshelman R, Ulloa M, Saha S, Burr B, Liu S, Zhang T, Fang D, Pepper A, Kumpatla S, Jacobs J, Tomkins J, Cantrell R, Main D (2006) CMD: a cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics* 7:132–142
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156:847–854
- Cuadrado A, Schwarzacher T (1998) The chromosomal organization of simple sequence repeats in wheat and rye genomes. *Chromosoma* 107:587–594
- Decroocq V, Favé MG, Hagen L, Bordenave L, Decroocq S (2003) Development and transferability of apricot and grape EST

- microsatellite markers across taxa. *Theor Appl Genet* 106:912–922
- Doldi ML, Vollmann J, Lelley T (1997) Genetic diversity in soybean as determined by RAPD and microsatellite analysis. *Plant Breed* 116:331–335
- Gao L, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol Breed* 12:245–261
- Guo W, Wang W, Zhou B, Zhang T (2006) Cross-species transferability of *G. arboreum*-derived EST–SSRs in the diploid species of *Gossypium*. *Theor Appl Genet* 112:1573–1581
- Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Lu K, Shi B, Zhang T (2007) A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*. *Genetics* 176:527–541
- Gupta PK, Balyan HS, Sharma PC, Ramesh B (1996) Microsatellites in plants: a new class of molecular markers. *Curr Sci* 70:45–54
- Han ZG, Guo WZ, Song XL, Zhang TZ (2004) Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. *Mol Genet Genomics* 272:308–327
- Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST–SSRs in allotetraploid cotton. *Theor Appl Genet* 112:430–439
- Hancock JM (1995) The contribution of slippage-like processes to genome evolution. *J Mol Evol* 41:1038–1047
- Hartl D, Clark A (1989) Principles of population genetics, 2nd edn. Sinauer Associates, Sunderland
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9:868–877
- Jarvie T, Harkins T (2008) Transcriptome sequencing with the Genome Sequencer FLX system. *Nat Methods* 5:6–8
- Jena S, Sahu P, Mohanty S, Das A (2004) Identification of RAPD markers, in situ DNA content and structural chromosomal diversity in some legumes of mangrove flora of Orissa. *Genetica* 122:127–226
- Jones KC, Levine KF, Banks JD (2000) DNA-based genetic markers in black-tailed and mule deer for forensic applications. *Calif Fish Game* 86:115–126
- Kantety RV, Rota ML, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48:501–510
- Korzun V, Röder MS, Wendehake K, Pasqualone A, Lotti C, Ganal MW, Blanco A (1999) Integration of dinucleotide microsatellites from hexaploid bread wheat into a genetic linkage map of durum wheat. *Theor Appl Genet* 98:1202–1207
- Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas* 135:145–151
- Kuleung C, Baenziger PS, Dweikat I (2004) Transferability of SSR markers among wheat, rye, and triticale. *Theor Appl Genet* 108:1147–1150
- Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 21:991–1007
- Lü YD, Cai CP, Wang L, Lin SY, Zhao L, Tian LL, Lü JH, Zhang TZ, Guo WZ (2010) Mining, characterization, and exploitation of EST-derived microsatellites in *Gossypium barbadense*. *Chin Sci Bull* 55:1889–1893
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with non repetitive DNA in plant genomes. *Nat Genet* 30:194–200
- Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* 76:5269–5273
- Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *J Mol Evol* 19:153–170
- Nguyen TB, Giband M, Brottier P, Risterucci AM, Lacape JM (2004) Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *Theor Appl Genet* 109:167–175
- Qureshi SN, Saha S, Kantety RV, Jenkins JN (2004) EST–SSR: a new class of genetic markers in cotton. *J Cotton Sci* 8:112–123
- Ramsay L, Macaulay M, Degli Ivanissevich S, MacLean K, Cardle L, Fuller J, Edwards KJ, Tuveesson S, Morgante M, Massari A (2000) A simple sequence repeat-based linkage map of barley. *Genetics* 156:1997–2005
- Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* 166:389–417
- Saha MC, Mian MAR, Eujayl I, Zwonitzer JC, Wang L, May GD (2004) Tall fescue EST–SSR markers with transferability across several grass species. *Theor Appl Genet* 109:783–791
- Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* 100:723–726
- Sefc KM, Steinkellner H, Wagner HW, Glössl J, Regner F (1997) Application of microsatellite markers to parentage studies in grapevine. *Vitis* 36:179–183
- Shannon CE, Weaver W (1949) A mathematical model of communication. University of Illinois Press, Champaign
- Taramino G, Tarchini R, Ferrario S, Lee M, Pe ME (1997) Characterization and mapping of simple sequence repeats (SSRs) in *Sorghum bicolor*. *Theor Appl Genet* 95:66–72
- Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res* 12:4127–4138
- Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* 7:537–546
- Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23:48–55
- Wang C, Guo W, Cai C, Zhang T (2006) Characterization, development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich. *Chin Sci Bull* 51:557–561
- Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. *Proc Natl Acad Sci USA* 86:4132–4136
- Xiao J, Wu K, Fang DD, Stelly DM, Yu J, Cantrell RG (2009) New SSR markers for use in cotton (*Gossypium* spp.) improvement. *J Cotton Sci* 13:75–157
- Zhang YX, Lin ZX, Li W, Tu LL, Nie YC, Zhang XL (2007) Studies of new EST–SSRs derived from *Gossypium barbadense*. *Chin Sci Bull* 52:2522–2531