2004 BELTWIDE COTTON CONFERENCES, SAN ANTONIO, TX – JANUARY 5-9

DEVELOPMENT AND CHARACTERIZATION OF FIBER EST-DERIVED MICROSATELLITES IN COTTON (*GOSSYPIUM* SPP.)

Young-Hoon Park and Mauricio Ulloa
USDA-ARS, W.I.C.S Res. Unit
Cotton Enhancement Program
Shafter, CA
Brad A. Sicker and Thea Wilkins
Department of Agronomy and Range Science
University of California
Davis, CA

Abstract

A new set of microsatellite or simple sequence repeat (SSR) markers were developed from the sequence information of a fiber EST database from the diploid species *Gossypium arboreum*. Four hundred EST-derived SSR primer sets were designed and tested for PCR amplification. Polymorphic loci among six species of cotton were observed from 233 (73.7 %) SSR markers and 83 were monomorphic. Between *G. hirsutum* and *G. barbadense*, 103 markers were polymorphic, while the polymorphism within these two tetraploid species was low 4.7 % and 3.8 %, respectively. One hundred twenty-three SSR markers that were polymorphic among the species showed sequence similarity to the genes with known function including fiber elongation-related genes. Based on PCR amplification, the transferability of EST-derived SSR marker from the diploid species *G. arboreum* was high (89.9 %) across the six *Gossypium* species. Most of these SSR markers provided information about functional gene, demonstrated portability and strength, and may be useful for diverse genome analysis and for the improvement of the cotton crop through marker-assisted selection.

Introduction

Molecular markers are efficient tools for genome analysis in many crop species. In cotton (*Gossypium* spp.), several marker systems including restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), and microsatellites have been successfully used for mapping important traits and genetic diversity studies (Bi et al., 1999; Liu et al., 2000; Ulloa and Meredith, 2000; Ulloa et al., 2002; Lacape et al., 2003).

Microsatellites are simple sequence repeats (SSRs) that result from the mutational effects of replication slippage (Tautz, 1994). Polymerase chain reaction (PCR) amplification of simple sequence repeats (SSRs) reveals polymorphisms in length PCR product. These polymorphisms are attributed to the differences in the number of repeat units at a locus. This SSR-based marker system is very valuable because of its PCR-based, co-dominant, and multiallelic nature (Gupta et al., 1996).

Traditionally, microsatellites have been discovered by screening genomic DNA or cDNA libraries using simple repetitive oligonucleotides probes, and sequencing clones containing putative SSRs. SSR markers in cotton have been developed using these methods which are costly and time-consuming procedures (Reddy et al., 2001; Liu et al., 2000; Saha et al., 2003). Currently, more information for DNA sequences is being accumulated and made available in public databases in many economically important crops. These databases can be utilized as valuable sources for SSR mining and marker development (Thiel et al., 2003; Smulders et al., 1997; Eujayl et al., 2002; Scott et al., 2000). An expressed sequence tag (EST) database was developed from a diploid cotton species (*Gossypium arboreum* L.) from a NSF Cotton Genome Project. This EST database (http://cottongenomecenter.ucdavis.edu) contains information of about 13,000 non-redundant sequences of fiber transcripts and their homologous gene functions.

Cotton genomics is in its infancy in that despite ongoing efforts by the cotton community, a high density molecular map to facilitate marker-assisted selection in the public domain is still lacking. The cotton community has indicated that 5,000 DNA markers will be necessary to provide the requisite tools for genome analysis, evaluation of germplasm collections, and marker-assisted selection (MAS). At present, the number of DNA markers available in the public domain numbers in the hundreds, and the necessity to increase these numbers exists. Recently, cotton scientists from different parts of the U.S. met in Cary NC (Cotton Inc.) to support the development of a public Cotton Microsatellite Database (CMD) sponsored by Cotton Incorporated, which will provide more portable, publicly available, frame work markers to expedite cotton genome research. In the present on-going study, the objectives are the development of a large set of microsatellite markers from this sequence information of the EST database (*G. arboreum*) and evaluation of their potential utility for cotton genome analysis and crop improvements.

Material and Methods

PCR primer pairs flanking simple sequence repeats were designed using PRIMER3 software v. 0.2c based on the following core parameters: primer length = 18-25 bp, size of amplification product = 150-300 bp, minimum GC content 40%, GC clamp = 2, melting temperature optimum 60 °C. The range of the parameters was modified when any appropriate primer pairs were not retrieved by the software. PCR primers were synthesized by Proligo (Boulder, CO).

A DNA sample panel comprising eight different accessions of *Gossypium* was evaluated for PCR amplification as well as the level of microsatellite polymorphisms for each primer pair. The accessions included in this study were two cultivars of *G. hirsutum* [MD51ne and Fiber Max 832: allotetraploid, $(AD)_1$], two cultivars of *G. barbadense* [Pima S 6 and Pima 89590: allotetraploid, $(AD)_2$], and one sample for each of four diploid species (*G. herbaceum*: A_1, *G. arboreum*: A_2, *G. thurberi*: D_1, *G. harknessii*: $D_{2\text{-}2}$).

PCR amplification was performed in a total volume of 20 µl containing 20 ng of template DNA, 0.1 µM of each primer (forward and reverse), 1X PCR buffer, 0.2 mM of dNTPs , and 1 U of Taq polymerase (Amplitaq, Applied Biosystem, Foster city, CA) with cycling profile of 1 cycle of 2 min. at 94°C, 10 cycles of 15 sec. at 94°C, 30 sec. at 60°C (step -0.5°C/cycle for cycles 2-10), and 1min. at 72°C, 35 cycles of 15 sec. at 94°C, 30 sec. at 55°C, and 1min. at 72°C. PCR products were separated on a 3% Super Fine Resolution (SFR™) agarose (Amresco, Solon, OH) gel containing 1X TBE at 80 volts for 4-5 hrs, and visualized by Alpa imager (Alpha Innotech Corporation, San Leandro, CA) software v. 5.5 after staining with ethidium bromide.

Primer pairs that resulted in discrete PCR banding patterns (molecular marker) were scored. Each primer pair was expected to amplify DNA fragments from a SSR locus at least from *G. arboretum*. PCR products with similar sizes were considered to be alleles from the same locus. Allele names were designated by locus name (primer ID) followed by the letters a, b, c, or d, where a is the highest molecular weight allele, b is the second highest molecular weight, etc.

Results

PCR Amplification

We evaluated 511 SSR-containing EST sequences for primer design. Primer pairs flanking SSRs were successfully designed for 400 unique ESTs. Primer design was not available for the other 111 ESTs, because the SSR was located at the border of the sequenced DNA fragment, or surrounded by AT-rich repetitive DNA. From 400 microsatellites, 352 were perfect repeats and 48 were imperfect with interruption by non-repetitive bases.

PCR amplification was successful from 316 primer pairs, and these primers were included for further analysis. Two hundred eighty-four primer sets amplified DNA fragments that were scored across all eight DNA samples, and 225 primer pairs amplified DNA fragments with predicted sizes from *G. arboreum*. For the remaining 84 primer pairs, PCR amplification failed, or resulted in unexpected multiple PCR artifacts.

Analysis of SSR Polymorphism

A total of 717 alleles were revealed by 316 primer pairs (SSR markers). The number of alleles at polymorphic loci across six species ranged from two to seven, and the average number of alleles at a locus was 2.3. Two alleles were found at the majority of the polymorphic loci (48%) while more than five alleles were at only 14 loci.

The number of markers that were polymorphic within or between the species is summarized in Table 1. Two hundred thirty-three primer pairs (73.7%) revealed polymorphic loci among the six different species, and 83 primer pairs were monomorphic. Between *G. hirsutum* and *G. barbadense*, 103 markers were polymorphic. However, polymorphism was low within the two tetraploid species; 15 markers (4.7 %) within *G. hirsutum* and 12 markers (3.8 %) within *G. barbadense*.

Characterization of 316 SSRs

The repeat types comprising 316 SSRs were di-, tri-, tetra-, and pentanucleotide (Table 2). The trinucleotide motif was the most abundant (52.5%) while the other three types of motif were found at lower frequencies (10.4 to 20.9%). In the comparison of different types of repeats, the highest ratio of polymorphism was observed from the markers targeting pentanucleotide repeats (78.4%) among six species. For trinucleotide repeats, the ratio of polymorphism was the lowest across all the accessions. The number of repeats within a SSR varied from two to 27. The length of SSRs ranged from 11 to 60bp, and 183 SSRs (58%) were shorter than 19bp.

Two hundred sixty one ESTs containing these SSRs showed significant homology (e-value < 1E-20; BLASTX) to putative genes with known or unknown function, while the other 55 sequences resulted in no significant hit (Table 1). One hundred twenty-three SSR markers polymorphic among the species showed sequence similarity to genes or putative genes with known function.

Discussion

A cotton EST database comprising 13,000 sequences of fiber transcripts was feasible to develop a set of informative SSR markers. We designed 400 primer sets and amplified 316 unique SSR loci. Out of 316 SSR markers analyzed, 284 primer sets (89.9%) gave good amplification across all eight accessions of *Gossypium*, which demonstrates the high interspecific transferability of microsatellites. High cross transferability of EST-derived markers has also been demonstrated in barley (Thiel et al., 2003) and sugarcane (Cordeiro et al., 2001), and can be due to the sequence conservation of the gene-encoding region of the genome.

About 32.6 % of the SSR loci (103) analyzed were polymorphic between the two allotetraploid species, *G. hirsutum,* and *G. barbadense*. Reddy et al. (2001) developed a set of SSR-based markers from a SSR-rich genomic DNA library, and reported 49% polymorphism between the two species. One reason for relatively lower polymorphism ratio in this study could be that the loci of transcript sequences, such as ESTs, can be highly conserved among the different *Gossypium* species and, especially, among the modern cultivars. In addition, high level of sequence conservation between *G. hirsutum* and *G. barbadense* was also reported from the alignment of fiber-specific cDNA sequences (Saha et al., 2003). The polymorphism ratios within each species based on PCR amplification were very low (4.7 % for *hirsutum* and 3.8 % for *barbadense*) in this study, supporting the above statement.

More than half (52.5 %) of the total microsatellites were trinucleotide repeats. Predominantly high percentages of trinucleotide motif have been found from EST-derived SSR mining in Arabidopsis (Morgante et al., 2002) and several other crops (Morgante et al., 2002; Thiel et al., 2003). The preponderance of trimeric SSRs in the coding region can be explained by the hypothesis that selection against frameshift mutations suppresses the expansion of non-trimeric nucleotide repeats (Metzgar et al., 2000).

Sequences of 59 SSR markers polymorphic between *G. hirsutum* and *G. barbadense* were significantly homologous to the genes with known function. These genes include Cytochrome P450-like protein, Expansin, RING zinc finger protein, and ABC transporter, which are related to fiber elongation (Ji et al., 2003). Use of PCR-based SSR markers in alleles of commercially important traits, such as fiber quality and yields, will be an ideal tool for marker-assisted selection.

In conclusion, a cotton fiber EST database is a valuable source for the development of informative SSR markers. Up to date, we developed 316 EST-derived SSR markers that are highly transferable across different *Gossypium* species. These SSR markers will be useful for the development of intra- or interspecific linkage maps, evaluation of germplasm collections, and molecular tagging of important traits for marker assisted selection (MAS) in cotton. Primer sequences of these markers are being posted at websites, http://www.genome.clemson.edu and http://cottongenomecenter. ucdavis.edu for public use.

Acknowledgements

The authors would like to thank California State Support Committee and Cotton Incorporated for their support on this project. We would also like to thank Ms. Ravinder Gill for her technical assistance.

References

Bi, I.V., A. Maquet, J.-P. Baudoin, P.D. Jardin, J.M. Jacquemin, and G. Mergeai. 1999. Breeding for “low-gossypol seed and high-gossypol plants” in upland cotton. Analysis of tri-species hybrids and backcross progenies using AFLPs and mapped RFLPs. Theor. Appl. Genet. 99: 1233-1244.

Cordeiro, G.M., R. Casu, C.L. McIntyre, J.M. Manners, and R.J. Henry. 2001. Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci. 160: 1115-1123.

Eujayl, I., M.E. Sorrells, M. Baum, P. Wolters, and W. Powell. 2002. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor. Appl. Genet. 104: 399-407.

Gupta, P.K., H.S. Balyan, P.C. Sharma, and B. Ramesh. 1996. Microsatellites in plants: A new class of molecular markers. Current Sci. 70: 45-53.

Ji, S.J., Y.C. Lu, J.X. Feng, G. Wei, J. Li, Y.H. Shi, Q. Fu, D. Liu, J.C. Luo, and Y.X. Zhu. 2003. Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. Nucl. Acids Res. 31: 2534-2543.

Kohel, R. J., J. Yu, Y.-H. Park, and G.R. Lazo. 2001. Molecular mapping and characterization of traits controlling fiber quality in cotton. Euphytica 121: 163-172.

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

Liu, S., R.G. Cantrell, J.C. McCarty, and McD. Stewart. 2000. Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Sci. 40: 1459-1469.

Metzgar, D., J. Bytof, and C. Wills. 2000. Selection against frameshift mutation limits macrosatellite expansion in coding DNA. Genome Res. 10: 72-80.

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genet. 30: 194-200.

Reddy, O.U.K., A.E. Pepper, I. Abdurakhmonov, S. Saha, J.N. Jenkins, T. Brooks, Y. Bolek, and K.M. El-Zik. 2001. New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. J. Cotton Sci. 5: 103-113.

Saha, S., M. Karaca, J.N. Jenkins, A.E. Zipf, O.U.K. Reddy, and R.V. Kantety. 2003. Simple sequence repeats as useful resources to study transcribed gene of cotton. Euphytica 130: 355-364.

Scott, K.D., P. Eggler, G. Seaton, M. Rossetto, E.M. Ablett, L.S. Lee, and R.J. Henry. 2000. Analysis of SSRs derived from grapes ESTs. Theor. Appl. Genet. 100: 723-726.

Smulders, M.J.M., G. Bredemeijer, W. Ruskortekaas, P. Arens, and B. Vosman. 1997. Use of microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor. Appl. Genet. 97: 264-272.

Tautz, D. and C. Schotterer. 1994. Simple sequences. Curr. Opin. Genet. 4: 832-837.

Thiel, T., W. Michalek, R.K. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. 106: 411-422.

Ulloa M. and W.R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. J. Cotton Sci. 4: 161-170.

Ulloa, M., W.R. Meredith Jr., Z.W. Shapply, A.L. Kahler. 2002. RFLP genetic linkage maps from F2.3 populations and a joinmap of *Gossypium hirsutum* L. Theor. Appl. Genet. 104: 200-208.

Table 1. Results of PCR amplification from 400 EST-derived SSR markers and the level of polymorphism detected on an array of different type of cottons (*Gossypium* spp.).

Best hit[†]	No. of marker[‡]		Comparison of polymorphisms and no. of markers[§]											
			A & D			h vs b			h vs h			b vs b		
	T	A	*m*	*p*	*% p*	*m*	*p*	*% p*	*m*	*p*	*% p*	*m*	*p*	*% p*
G	175	128	40	88	27.8	86	42	13.3	123	5	1.6	122	6	1.9
PG	64	54	18	36	11.4	37	17	5.4	53	1	0.3	53	1	0.3
S	10	10	4	6	1.9	10	0	0.0	10	0	0.0	10	0	0.0
U	41	35	8	27	8.5	25	10	3.2	33	2	0.6	33	2	0.6
P	18	15	0	15	4.7	7	8	2.5	15	0	0.0	13	2	0.6
H	6	6	1	5	1.6	4	2	0.6	5	1	0.3	6	0	0.0
E	14	13	3	10	3.2	10	3	0.9	12	1	0.3	13	0	0.0
N	72	55	9	46	14.6	34	21	6.6	50	5	1.6	54	1	0.3
Total	400	316	83	233	73.7	213	103	32.6	301	15	4.7	304	12	3.8

[†]Database search type: BLASTX; G = gene with known function; PG = putative gene with known function; S = sequence containing similarity to known gene; U = unknown protein; P = putative protein; H = hypothetical protein; E = expressed protein; N = no significant hit.

[‡]T = no. of primer sets for which PCR amplification was tested; A = no. of primer sets for which amplified PCR fragments were analyzed for polymorphism.

[§]A & D = within six species of *Gossypium*; h vs b = between *G. hirsutum* and *G. barbadense*; h vs h = within *G. hirsutum*; b vs b = within *G. barbadense*; m = no. of monomorphic markers; p = no. of polymorphic markers; *%p* = percentage of polymorphic markers [(*p*/316)x100].

Table 2. Repeat type composition of 316 SSR markers and the level of polymorphism.

Repeat motif[†]	No. of markers scored	No. of polymorphic markers (*%p*)[‡]			
		A & D	**h vs b**	**h vs h**	**b vs b**
di-	66	46 (69.7)	23 (34.8)	5 (7.6)	4 (6.1)
tri-	166	112 (67.5)	51 (30.7)	5 (3.0)	4 (2.4)
tetra-	33	25 (75.8)	13 (39.4)	3 (9.1)	0 (0.0)
penta-	51	40 (78.4)	16 (31.4)	2 (3.9)	7.8 (40)

[†]Di = dinucleotide repeats; Tri = trinucleotide repeats; Tetra = tetranucleotide repeats; Penta = pentanucleotide repeats

[‡]A & D = within six species of *Gossypium*; h vs b = between *G. hirsutum* and *G. barbadense*; h vs h = within *G. hirsutum*; b vs b = within *G. barbadense*; m = no. of monomorphic markers; p = no. of polymorphic markers; *%p* = percentage of polymorphic markers [(no. of polymorphic loci/no. of loci scored)x100].