TITLE: **Progress and Perspectives of a Cotton Microsatellite Database as a Comprehensive Public Marker Depository for *Gossypium***

DISCIPLINE: **Genomics and Biotechnology**

AUTHORS: A. Blenda (Corresponding author)
Department of Genetics and Biochemistry
Clemson University
Biosystems Research Center
51 New Cherry Street
Clemson, SC 29634
Phone: 864-656-4643
Fax: 864-656-4293
Email: blenda@clemson.edu

M. Staton
Department of Genetics and Biochemistry
Clemson University
Biosystems Research Center
51 New Cherry Street
Clemson, SC 29634

D. Jones
R. Cantrell
Cotton Incorporated
Cary, NC 27513

D. Main
Department of Horticulture and Landscape Architecture
Washington State University, WA 99164
Phone: 509-335-2774
Fax : 509-335-8690
Email: dorrie@wsu.edu

ACKNOWLEDGEMENT: We acknowledge with thanks, Cotton Incorporated for funding this database.

ABBREVIATIONS: BAC, bacterial artificial chromosome; BC1, backcross 1st generation; BLAST, basic local alignment search tool; CAP3, contig assembly program; CIRAD - Centre International en Recherche Agronomique pour le Développement ; CIR, CIRad; CM, cotton microsatellites; CMap, comparative map; EST, expressed sequence tag; EXP, expectation value; FASTA, fast all alignment search tool; JESPR, Jenkins, El-Zik, Saha, Pepper, Reddy microsatellite repeats; MGHES, Mississippi *Gossypium hirsutum* EST-SSR; MUCS, microsatellite Ulloa complex sequence repeats; MUSB, microsatellite Ulloa simple BAC repeats; MUSS, microsatellite Ulloa simple sequence repeats; NAU, Nanjing Agricultural University; RIL, recombinant inbred line; TMB, TM-1 genetic standard BAC/BIBAC libraries microsatellite repeats.

Progress and Perspectives of a Cotton Microsatellite Database as a Comprehensive Public Marker Depository for *Gossypium*

ABSTRACT

The Cotton Microsatellite Database (CMD) (www.cottonmarker.org) was established in 2004 as a curated database resource providing centralized access to the largest collection of publicly available cotton microsatellite (SSR) markers. Over the last few years, collaboration with academic, government and industry has resulted in the collection of 8,213 markers from 12 SSR projects, with further 700 genomic SSRs pending submission to CMD. Of these projects, 5 were developed in academic institutions, 4 were released as result of academic and USDA-ARS collaborations, 2 projects were provided by USDA-ARS researchers, 1 project was submitted by an International Research Center (CIRAD, France), and the most recent project, DPL, provides the first cotton SSRs released by a private company (Delta and Pine Land). Currently, the cotton SSRs deposited in the CMD are represented by 4,814 EST-SSRs and 3,399 genomic SSRs. CMD also presents data for 5 SSR projects that were screened against the CMD standardized panel. This panel consists of 12 diverse genotypes selected from cultivated and exotic cottons. In addition, CMD provides a suite of online tools for data mining and comparative analysis. Starting 2007, quarterly newsletters provide the cotton community with information about on-going work and new additions to the database. Future development of the CMD will focus on the establishment of a standard nomenclature for cotton SSRs, visualization of original chromatograms for panel screened SSR data, inclusion of another type of cotton markers, SNPs, as well as increased collaboration with CottonDB. With cotton genome sequencing in progress, the CMD also will focus on enhanced SSR and other markers' data mining and analysis capabilities such as full sequence processing facilities. Annotation of markers will include further classification using gene ontology and KEGG terms.

KEY WORDS:

Cotton Microsatellite Database (CMD), *Gossypium*, molecular markers, simple sequence repeat (SSR), single nucleotide polymorphism (SNP), mapping.

The generation of a transportable framework of microsatellites, or simple sequence repeats (SSRs) markers, capable of being mapped in any segregating population was one of the major objectives of the International Cotton Genome Initiative (ICGI). In keeping with the proposed goal of the ICGI, the Cotton Microsatellite Database (CMD) (www.cottonmarker.org) has been initiated and funded by Cotton Incorporated. The CMD Advisory Board was formed to guide the development of CMD and to coordinate it with CottonDB (http://www.cottondb.org), the genome database serving the international cotton research community. The Advisory Board includes 5 scientists from the USDA, 2 researchers from academia, and 6 representatives from international companies.

CMD is a curated database resource providing centralized access to the largest collection of publicly available cotton SSRs (Blenda et al., 2006). Microsatellite markers can be used in various applications, including gene tagging and genome mapping, selecting progeny for a desired phenotypic trait, localizing qualitatively as well as quantitatively inherited traits, pedigree analysis, variety protection, and introgressing novel genes into breeding lines from exotic germplasm.

The novelty of the CMD is in its specific orientation toward researchers involved in molecular marker development and application to cotton breeding. It is being actively used by the international cotton community, and can be viewed as an important vehicle toward increased collaboration among academic, government and industry cotton scientists, both nationally and worldwide. The present collection of 8,213 SSRs in CMD was generated through collaboration with major cotton research groups from the USA, France and China.

The major goals of the CMD are:
(1) to collect and integrate all the publicly available cotton microsatellite data, as well as other cotton molecular marker data, in a centralized, curated, non-redundant online oracle database,

(2) to provide access to the CMD standardized panel screened data,
(3) to provide a set of comprehensive interface tools for rapid data retrieval,
(4) to provide a suite of stand-alone marker data mining tools,
(5) to provide a communication portal for collaboration within the cotton research community.

Database content and utility. The CMD website is composed of general information pages, including CMD tutorials, project pages, database query/browse interfaces and other tools such as a comparative map viewer CMap, sequence similarity (BLAST/FASTA) server, SSR server, and CAP3 server. The CMD web pages are organized such that users can easily access the data of interest regardless of the navigation starting point. A general CMD tool bar is also included in each page to aid the ease of navigation through the site. Starting in 2007, quarterly newsletters provide the cotton community with information about on-going work and new additions to the database (CMD newsletter issues are available in pdf format from the downloads page).

The CMD database is composed of 18 tables which store all the data for the microsatellite projects including information on project collaborators, SSR-containing clones, sequences, primers flanking the SSRs, repeat motif, open reading frame position, genetic markers and maps, standardized panel varieties, data homology, and publications. Of those 18 tables, four new tables were recently added to store the data for cotton SNP projects, as well as data for the visual display of the chromatograms of cotton SSRs screened against the CMD standardized panel in search for polymorphisms.

In a separate but linked database within CMD, the CMap schema consists of 16 tables including information about genetically mapped cotton SSRs. Data for cotton SSR markers and genetic maps, as well as panel screened cotton microsatellites, were submitted by

researchers and then curated for any potential errors prior to uploading to the database using scripts written in Perl version 5.8.2.

CMD currently contains 12 cotton microsatellite data projects. The first cotton SSR project BNL includes 379 genomic microsatellites presented through CMD. They were derived from *G. hirsutum* small insert genomic library enriched for (GA/CT)n and (CA/GT)n inserts (Liu et al., 2000). Other genomic SSR projects include 309 JESPR (Reddy et al., 2001), 392 CIR (Nguyen et al., 2004), 53 CM (Connell et al., 1998), 200 DPL, 750 TMB (Yu et al., 2002), and 1316 MUSB (Frelichowski et al., 2006) cotton microsatellites (TMB and MUSB SSRs are BAC-derived). 192 STV SSRs are EST-derived microsatellite markers from multiple tissues of *G. hirsutum* (Taliercio et al., 2006). 119 HAU are ESTs derived from the cDNA library of *G. barbadense*. 84 MGHES (Qureshi et al., 2004), 1169 MUSS/MUCS (Park et al., 2005) and 3250 NAU (Han et al., 2004; Han et al., 2006) microsatellites were developed by screening public databases for EST-SSRs. General statistics of the CMD SSRs and repeat analysis of the CMD SSR-containing sequences are presented in Tables 1 and 2.

Marker data, primers, microsatellite sequences and standardized panel screened data are available for download directly from each project page as well as an overall downloads page. A microsatellite information page displays the sequence along with the repeat sequence, primers, and other related information. Currently, CMD contains information on 8,213 annotated cotton microsatellites which can be viewed and downloaded. Annotation of the sequences is periodically updated so that our data reflects changes in protein records in the NCBI GenBank non-redundant protein database, SWISS-PROT database and MIPS Arabidopsis protein database.

The initial SSR search result page displays SSR identifiers. The individual SSR entry links to a page where details of the SSR are displayed with links to the corresponding project page, the top protein homolog identified through a sequence similarity search, microsatellite

sequence in GenBank, and related publications. Markers can be searched by marker name, chromosome, or cross. Other pages include a mailing group list form, so users can exchange information and be kept up to date on new developments in CMD. The message boards automatically list all the information exchanged by the mailing list. The links page contains appropriate cotton links.

In addition, a standardized panel of 12 *Gossypium* genotypes for the cotton microsatellite database (CMD) was established after extensive discussion and consultation between cotton researchers (Yu, 2004a; Yu et al., 2004b). Using this standardized genotype panel, cotton SSR markers derived from different sources or groups can be evaluated in a systematic way to minimize the potential redundancy and to determine the markers Polymorphic Information Content (PIC) values for ready applications. This genotype panel represents a balanced diversity of the core *Gossypium* germplasm including cultivated and exotic cottons. Polymorphisms arising from easily assayed variation in SSR numbers show great utility in crop genetic mapping and other applications. The amplified fragments sizes are currently available in CMD for the 375 BNL, 204 CIR, 127 JESPR, 150 STV and 200 DPL microsatellites screened against the standardized panel.

A genetically anchored physical map for cotton is being developed using cotton BAC libraries (Yu et al., 2005). Through various genetic markers, including SSRs, the cotton physical map will be anchored on the future consensus cotton genetic map (Yu et al., 2005). CMD stores and presents currently available data for major cotton genetic maps with mapped SSRs that were constructed for different crosses. Currently, CMD contains data for four genetic maps: 1 - BC_1: ((Guazuncho2 (*G. hirsutum*) x VH8-4602 (*G. barbadense*)) x Guazuncho2) (Nguyen et al., 2004; Lacape et al., 205); 2 - F_2: *G. hirsutum* race 'Palmeri" x *G. barbadense* Acc. "K101" (Rong et al., 2004); 3 - BC_1: (TM-1(*G. hirsutum*) x Hai7124 (*G. barbadense*)) x TM-1) (Han et al., 2004; Han et al., 2006; Guo et al., 2007); 4 - RIL: TM-1

(*G. hirsutum*) x 3-79 (*G. barbadense*) (Park et al., 2005). The anchored genetic markers can be viewed in several formats, including an excel spreadsheet, a database search interface, and a graphical interface for comparative visualization of SSR maps. A graphical tool CMap allows cotton genetic maps to be displayed with the number of anchored SSR markers, and the location of mapped SSRs is compared between different crosses of cotton. CMap (Generic Components for Model Organism Database (GMOD) Project [http://www.gmod.org/]) allows the user to select the map of interest and the maps for comparisons.

The CMD analysis tools page provides access to an SSR server, a CAP3 Assembly server, and a sequence similarity server that includes BLAST and FASTA search tools. SSR analysis is performed using a modified version (SSR) of a Perl script SSRIT (Temnykh et al., 2001) with parameters set to detect mono- to hexanucleotides of user specified length in sequences in FASTA format. An excel output file includes repeat(s) motif and number, SSR start/stop position, ORF start/stop position, primer pairs, SSR location relative to the ORF, and GC content of the sequence. To reduce the inherent redundancy and increase transcript length ESTs are routinely assembled into longer consensus sequences, also known as contigs. We have implemented the contig assembly program CAP3 (Huang and Madan, 1999) as an online server to allow users to assemble ESTs prior to mining the consensus sequences for microsatellites using the CMD SSR server. As more ESTs are sequenced and added to the public domain, the cotton unigene can be continually refined using the CAP3 server and mined for SSRs using the SSR server. The online BLAST and FASTA sequence similarity search servers allow users to perform homology searches between their sequences of interest and the annotated SSR sequences and primers in CMD. Our sequence similarity server, specifically designed for CMD researchers, will help users compare new sequences and

primers against existing microsatellites and help decrease redundancy of effort in developing new markers.

Future development. Future development will focus on the establishment of a standard nomenclature for cotton SSRs, adding new microsatellite data, visualization of original chromatograms for panel screened data, improving the tools and functionality of the web interface, such as an advanced search site with options for search/display categories, as well as increased collaboration with CottonDB. The annotation of the SSRs with known homology will include further classification using the gene ontology and KEGG terms. As we migrate the sequence similarity servers to a computational cluster, we plan to add the following databases: NCBI cotton ESTs, TIGR cotton gene indices, NCBI cotton genomic sequences and NCBI cotton protein sequences. When the physical map is available, users also will be able to retrieve the anchored BAC clones containing the SSRs of interest through the anchored BACs page in the map viewer. Data that are currently scheduled to be added in the near future include 700 cotton genomic SSRs from Gh project. With cotton genome sequencing in progress, the CMD also will focus on enhanced SSR data mining and analysis capabilities such as full sequence processing facilities. In addition, the CMD Advisory Board recently agreed to depositing cotton single nucleotide polymorphism (SNP) markers to the CMD with the further change of the name Cotton Microsatellite Database to a Cotton Marker Database, as well as future addition of other types of cotton markers. The first cotton SNP project was provided by Dr. Allen Van Deynze (UC Davis).

CONCLUSIONS

The CMD has been initiated to provide researchers, engaged worldwide in cotton research, with centralized access to microsatellite markers, an invaluable resource for basic and applied research in cotton breeding. As such, the CMD serves the cotton community as a major repository of the publicly available cotton microsatellite data and a unique repository

for the CMD standardized panel screened data, a key tool for systematic characterization of the SSR markers developed for cotton. CMD also provides a suite of online tools for data analysis of new and existing microsatellites. By depositing cotton single nucleotide polymorphism (SNP) markers and potentially other types of markers to the CMD, and the further change of the name Cotton Microsatellite Database to a Cotton Marker Database, CMD will enhance its positions as a major marker resource for the cotton community, and an important vehicle toward increased collaboration among cotton scientists.

REFERENCES

Blenda, A., J. Scheffler, B. Scheffler, M. Palmer, J.-M. Lacape, J. Z. Yu, C. Jesudurai, S. Jung, S. Muthukumar, P. Yellambalase, S. Ficklin, M. Staton, R. Eshelman, M. Ulloa, S. Saha, B. Burr, S. Liu, T. Zhang, D. Fang, A. Pepper, S. Kumpatla, J. Jacobs, J. Tomkins, R. Cantrell, and D. Main. 2006. CMD: a Cotton Microsatellite Database resource for *Gossypium* genomics. BMC Genomics 7:132

Connell, J.P., S. Pammi, M.J. Iqbal, T. Huizinga, and A.S. Reddy. 1998. A high throughput procedure for capturing microsatellites from complex plant genomes. Plant Mol. Biol. Rep. 16: 341-349.

Frelichowski, J.E. Jr, M.B. Palmer, D. Main, J.P. Tomkins, R.G. Cantrell, D.M. Stelly, J.Z. Yu, R.J. Kohel, and M. Ulloa. 2006. Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. Mol. Genet. Genom. 275: 479-491.

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

Han, Z.G., W.Z. Guo, X.L. Song, and T.Z. Zhang. 2004. Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol. Genet. Genom. 272: 308-327.

Han, Z.G., C. Wang, X.L. Song, W.Z. Guo, J. Gou, C. Li, X. Chen, and T.Z. Zhang. 2006. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theor. Appl. Genet. 112: 430-439.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9: 868-877.

Lacape, J.M., T.B. Nguyen, B. Courtois, J.L. Belot, M. Giband, J.P. Gourlot, G. Gawryziak, S. Roques, and B. Hau. 2005. QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* x *Gossypium barbadense* backcross generations. Crop Sci. 45: 123-140.

Liu, S., S. Saha, D. Stelly, B. Burr, and R.G. Cantrell. 2000. Chromosomal assignment of microsatellite loci in cotton. J. Hered. 91: 326-332.

Nguyen, T.B., M. Giband, P. Brottier, A.M. Risterucci, and J.M. Lacape. 2004. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. Theor. Appl. Genet. 109: 167-175.

Park, Y.H., M.S. Alabady, M. Ulloa, B. Sickler, T.A. Wilkins, J.Z. Yu, D.M. Stelly, R.J. Kohel, O.M. El-Shihy, and R.G. Cantrell. 2005. Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol. Genet. Genom. 274: 428-441.

Qureshi, S.N., S. Saha, R.V. Kantety, and J.N. Jenkins. 2004. EST-SSR: A new class of genetic markers in cotton. J. Cotton Sci. 8: 112-123.

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.

Rong, J., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. Delmonte, X. Ding, J.J. Garza, B.S. Marler, C.H. Park, G.J. Pierce, K.M. Rainey, V.K. Rastogi, S.R. Schulze, N.L. Trolinder, J.F. Wendel, T.A. Wilkins, T.D. Williams-Coplin, R.A. Wing, R.J. Wright, X. Zhao, L. Zhu, and A.H. Paterson. 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.

Taliercio, E., R.D. Allen, M. Essenberg, N. Klueva, H. Nguyen, M.A. Patil, P. Payton, A.C.M. Millena, A.L. Phillips, M.L. Pierce, B. Scheffler, R. Turley, J. Wang, D. Zhang, and J. Scheffler. 2006. Analysis of ESTs from multiple *Gossypium hirsutum* tissues and identification of SSRs. Genome 49: 306-319.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11: 1441-1452.

Yu, J.Z. 2004a. A standard panel of *Gossypium* genotypes established for systematic characterization of cotton microsatellite markers. p. 107. *In* Plant Breeding News Edition 148, an electronic newsletter of applied plant breeding sponsored by Food and Agriculture Organization of the United Nations and Cornell University.

Yu, J.Z., R. Cantrell, R. Kohel, S. Saha, J. Tomkins, A. Pepper, M. Ulloa, J. Scheffler, D. Stelly, D. Main, M. Palmer, and D. Jones. 2004b. Establishment of the standardized cotton microsatellite database (CMD) panel. *In* Proc. Beltwide Cotton Improvement Conf., San Antonio, TX. 5-8 Jan. 2004. Natl. Cotton Counc. Amer., Memphis, TN.

Yu, J.Z., R.J. Kohel, and J. Dong. 2002. Development of integrative SSR markers from TM-1 BACs. *In* Proc. Beltwide Cotton Improvement Conf., Atlanta, GA. 7-10 Jan. 2002. Natl. Cotton Counc. Amer., Memphis, TN.

Yu, J.Z., R.J. Kohel, Z. Xu, J. Dong, H.B. Zhang, D.M. Stelly, A.E. Pepper, P. Cui, and S.M. Hoffman. 2005. Integrated genetic, physical, and comparative mapping of the cotton genome. *In* Proc. Beltwide Cotton Improvement Conf., New Orleans, LA. 4-7 Jan. 2005. Natl. Cotton Counc. Amer., Memphis, TN.

Table 1. Statistics on the two main groups (EST and genomic) of the CMD cotton SSRs.

EST-SSRs			
Project	**Number**	**CMD panel screened**	***Gossypium* species**
MGHES	**84**		***G. hirsutum***
HAU	**119**		***G. barbadense***
NAU	**3250**		***G. hirsutum***
MUSS/MUCS	**1169**		***G. arboreum***
STV	**192**	**150**	***G. hirsutum***
Total	**4,814**	**150**	

Genomic SSRs			
Project	**Number**	**CMD panel screened**	***Gossypium* species**
BNL	**379**	**375**	***G. hirsutum***
CIR	**392**	**204**	***G. hirsutum***
CM	**53**		***G. hirsutum***
JESPR	**309**	**127**	***G. hirsutum***
TMB	**750**		***G. hirsutum***
MUSB	**1316**		***G. hirsutum***
DPL	**200**	**200**	***G. hirsutum***
Total	**3,399**	**906**	

Table 2. Repeat analysis of the CMD SSR-containing sequences.

	Summary		Repeat frequency (%)				
	No. sequences analysed	No. unique motifs	2 bp	3 bp	4 bp	5 bp	6 bp
EST	**4814**	**815**	**23.0**	**48.7**	**10.5**	**5.5**	**12.3**
Genomic	**3399**	**374**	**56.1**	**19.7**	**18.1**	**3.9**	**2.2**