Characterization of New *Schistosoma mansoni* Microsatellite Loci in Sequences Obtained from Public DNA Databases and Microsatellite Enriched Genomic Libraries

NB Rodrigues++, PT LoVerde*, AJ Romanha, G Oliveira//+**

Centro de Pesquisas René Rachou-Fiocruz, Av. Augusto de Lima 1715, 30190-002 Belo Horizonte, MG, Brasil *State University of New York at Buffalo, USA **Programa de Pós-Graduação e Pesquisa, Santa Casa de Belo Horizonte, MG, Brasil

In the last decade microsatellites have become one of the most useful genetic markers used in a large number of organisms due to their abundance and high level of polymorphism. Microsatellites have been used for individual identification, paternity tests, forensic studies and population genetics. Data on microsatellite abundance comes preferentially from microsatellite enriched libraries and DNA sequence databases. We have conducted a search in GenBank of more than 16,000 Schistosoma mansoni *ESTs and 42,000 BAC sequences. In addition, we obtained 300 sequences from CA and AT microsatellite enriched genomic libraries. The sequences were searched for simple repeats using the RepeatMasker software. Of 16,022 ESTs, we detected 481 (3%) sequences that contained 622 microsatellites (434 perfect, 164 imperfect and 24 compounds). Of the 481 ESTs, 194 were grouped in 63 clusters containing 2 to 15 ESTs per cluster. Polymorphisms were observed in 16 clusters. The 287 remaining ESTs were orphan sequences. Of the 42,017 BAC end sequences, 1,598 (3.8%) contained microsatellites (2,335 perfect, 287 imperfect and 79 compounds). The 1,598 BAC end sequences 80 were grouped into 17 clusters containing 3 to 17 BAC end sequences per cluster. Microsatellites were present in 67 out of 300 sequences from microsatellite enriched libraries (55 perfect, 38 imperfect and 15 compounds). From all of the observed loci 55 were selected for having the longest perfect repeats and flanking regions that allowed the design of primers for PCR amplification. Additionally we describe two new polymorphic microsatellite loci.*

Key words: *Schistosoma mansoni* - microsatellite repeats - polymorphism - DNA sequence analysis - nucleic acid database - computational biology

Schistosoma mansoni is the main agent of schistosomiasis, a public health problem that affects over 200 million people in the world (WHO 1993). The development of tools that facilitate our understanding of populational structure and genetic variations in *Schistosoma* will allow us to address important epidemiological questions. In this context, the development of microsatellites markers is essential to reach this objective.

Microsatellites are small tandemly repeated sequences (1-6 bp) widely dispersed in eukaryotic genomes (Tautz & Renz 1984). They are diploid, Mendelian inherited and show codominant alleles (Weber & May 1989, Ashley & Dow 1994). In the last decade microsatellites have become one of the most useful genetic markers for many organisms due to their genome wide distribution and high level of polymorphism (Weber & May 1989). Microsatellites have also been used for forensic studies, individual identification, paternity tests (Hagelberg et al. 1992, Lucas et al. 1997), and the study of population genetics of several species (Santos et al. 1993, Solano et al. 1997, Lougheed et al. 1999) including parasites (Durand et al. 2000, Curtis & Minchella 2000, Caccio et al. 2001, Rodrigues et al. 2002).

Microsatellite identification results mainly from library screening, sequence database searches, and microsatellite enriched library (Ashley & Dow 1994). The emergence of various genome projects in the past decade (Franco et al. 2000), among them the Schistosome Genome Network (SGN), has provided a rich source of sequences to be screened for microsatellites (Franco et al. 2000, Oliveira 2001, Oliveira & Johnston 2001). In this manuscript we describe the results of database searches and microsatellite enriched library sequencing for the identification of short simple repeats. In addition, we describe two new polymorphic loci.

MATERIAL AND METHODS

S. mansoni microsatellite-enriched genomic library construction and sequencing - Two partial genomic libraries enriched for repetitive sequences from *S. mansoni* genomic DNA were constructed by Genetic Identification Services. 'CA' or 'AT' enriched DNA fragments 300-700 bp in size were cloned into pUC19. After transformation into DH5α *Escherichia coli*, recombinant clones were selected on LB/agar plates with X-gal 800 µg, IPTG 800 µg and ampicilin 0.1 µg/ml (Life Technologies). White colonies were grown on 5 ml of LB with ampicilin (0.1 µg/ml) at 37°C with shaking and the plasmids prepared with the R.E.A.L. prep 96 plasmid Kit (Qiagen), according to the

This investigation received financial assistance from UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases, grant A00552, CNPq (521108/01-2), Fapemig and was supported, in part, by NIAID grant U01AI48828.

+Corresponding author. Fax: +55-31-3295.3115. E-mail: oliveira@cpqrr.fiocruz.br

++Fellowship from Fiocruz/Fundep

Received 18 June 2002
Accepted 15 August 2002

manufacturer's instructions. Plasmids were then sequenced in an automated DNA sequencer (ALF, Pharmacia) using the Thermo Sequenase Fluorecent Primer Kit (Amersham Biosciences) .

Search for microsatellites and primer design - DNA and cDNA sequences of *S. mansoni* derived from public databases and sequences from the two microsatellite enriched libraries were analyzed with the RepeatMasker software (Smit AFA & Green, P RepeatMasker at http//:ftp.genome.Washington.edu:80/cgi-bin/RepeatMasker) for determining regions containing simple DNA repeats. The search was conducted for sequences presenting more than six repetitions for mono-, five repetitions for di- and tri- and four repetitions for tetra and penta-nucleotides. Both groups, DNA or cDNA, sequences were clustered using BioEdit Sequence Alignment Editor software (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Sequences presenting the longest perfect repetitions and flanking regions that permitted primer design were selected for PCR primer design.

Microsatellite polymorphism detection - Microsatellite polymorphism detection was performed by PCR using 1 ng of template DNA, 0.75 U of *Taq* DNA polymerase (CENBIOT/Porto Alegre, Brazil), PCR buffer (1.5 mM $MgCl_2$, 50 mM KCl and 10 mM Tris/HCl, pH 8.3), 1.0 pmol of each primer, and 200 µM of each dNTP. PCR amplifications were carried out in a Perkin Elmer 9600 thermal cycler in a final reaction volume of 10 µl. The cycling conditions included an initial denaturation cycle at 95°C for 3 min, followed by 35 cycles of 95°C for 45 sec, 1 min for annealling (temperature depending on of each primer pair, Table I), 30 sec for extension at 72°C, and a final extension at 72°C for 5 min. Electrophoresis was carried out in an 8% polyacrylamide gel in a Mini-Protean II (Bio-Rad, Hercules, CA) using 3 µl of the amplified samples. The gel was then silver stained (Sanguinetti et al. 1994) and the image digitized using an Eagle Eye II system (Stratagene, La Jolla, CA). The sequences from which the amplification primers were designed were deposited in the GSS division of GenBank with accession numbers BH795455 (13TAGA) and BH795456 (27AAT).

RESULTS

Database search for microsatellite containing sequences - We searched a total of 16,022 ESTs and 42,017 BAC sequences deposited in GenBank. The complete list of microsatellite containing sequences can be found at http//:www.cpqrr.fiocruz.br/dna/ms. Out of the 16,022 ESTs a total of 481, 3% of the total, containing 622 microsatellites, were detected by RepeatMasker software (434 perfect, 164 imperfect and 24 compound). The longest repetition observed was 63 for a GA di-nucleotide (accession number BF936271) The most frequent repetition observed was AT, in 181 sequences. (Table II). Clustering analysis grouped 194 sequences in 63 clusters containing a maximum of 15 ESTs http//:www.cpqrr.fiocruz.br/dna/ms. We could observe polymorphisms in the number of repeats in 16 of the clusters. One example is shown in Fig. 1. The remaining 287 sequences were orphans. The complete clustering results can be found at http//:www.cpqrr.fiocruz.br/dna/ms.

Out of 42,017 BAC end sequences, a total of 1,598 (3.8%) presenting 2,701 microsatellite loci (2,335 perfect, 287 imperfect and 79 compound) were detected. The most frequent repetition observed was AT, in 897 sequences. The longest repetitions observed were 193 repetitions for a 'C' mono-nucleotide (accession number BH177237), followed by 99 repetitions of an 'A' mono-nucleotide (accession number BH189680) and of 58 and 59 repetitions of 'AAC' (accession number AL621065) and 'AAG' (accession number BH175667) respectively (Table III). Out of the 1,598 BAC end sequences, 80 were grouped into 17 contigs with 3 to 17 sequences per contig, no polymorphisms were observed within the groups. Among the remaining BAC end sequences, 536 were duplicates and 982 were orphan sequences.

Sequencing of microsatellite enriched S. mansoni genomic libraries - A total of 300 clones were sequenced from two microsatellites enriched genomic libraries, one 'AT-' and another 'CA-'enriched. Sixty seven sequences, 22% of the total, containing 109 microsatellites were detected by RepeatMasker, (56 perfect, 39 imperfect and 14 compound). The longest repetition observed was 34 of an 'AC' di-nucleotide (Table IV). A total of 36 sequences containing microsatellite loci were grouped into eleven clusters containing 2 to 12 sequences per cluster. The remaining sequences were orphans The complete list of microsatellite containing sequences derived from the two libraries can be found at http//:www.cpqrr.fiocruz.br/dna/ms.

Primer design and search for microsatellite polymorphism - Primers were designed for two different sequences (accession numbers BH795455 and BH795456). The BH795456 sequence contained an AAT tri-nucleotide repeat, and the BH795455 sequence a TAGA tetra-nucleotide repeat. Upon amplification of adult worm DNA these loci were observed to be polymorphic using DNA prepared from adult worms of the laboratory LE strain (Fig. 2).

TABLE I

Microsatellite loci analyzed in the *Schistosoma mansoni* samples

Microsatellite locus	Accession number	Repeat	Primer sequence (5'→3')	Tm	Amplicon [a]
13TAGA	BH795455	(GATA)10	GTACATTTTATGTCAGTTAGCC CATGATCTTAGCTCAGAGAGC	60°C 62°C	(122bp)
27AAT	BH795456	(AAT)9	AATCACCAATGGCAACAATCTG CGTCATCACCTTAAACATGAAC	62°C 62°C	(166bp)

a: in parenthesis the expected size of the amplicon.

TABLE II

Analysis of microsatellites found in 16,022 *Schistosoma mansoni* expressed sequence tag (EST) deposited in the GenBank™ database (19/01/01)

Repetition Unit (Number of repeats)	Number of ESTs	Classification		
Mono-nucleotide		Perfect	Imperfect	Compound
A (7 to 40)	007	004	003	000
Di-nucleotide				
AC (5 to37)	101	046	047	008
AT (5 to 27)	181	146	027	008
CT (5 to 63)	052	042	008	002
Tri-nucleotide				
AAT (5 to 17)	139	098	037	004
CAA (5 to 11)	026	022	003	001
CAG (5)	001	001	000	000
CAT (5 to12)	034	019	015	000
CCA (5 to 9)	004	003	001	000
CTT (5 to10)	018	011	007	000
TAC (5 to 9)	017	013	003	001
TCC (7 to 8)	002	002	000	000
Tetra-nucleotide				
CAAA (4 to 6)	012	012	000	000
CAGT (4 to 6)	004	004	000	000
CATA (4 to 6)	006	003	003	000
CTAA (5)	001	000	001	000
CTTA (4 to 5)	003	003	000	000
GTAC (5)	001	001	000	000
GTGA (4)	001	000	001	000
TAAA (4 to 6)	007	002	005	000
TATC (8)	005	002	003	000
Total	622	434	164	024

```
....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
     5         15         25         35         45         55         65         75         85         95        105
AA999422  AGGTGAACAATACTTTGTATTCATTATATATATATATATATATATATATATATA----TCACTAGTATTGTCATTGC-ATTATTGATAATATTATCATTATGAACAACGA
AA999423  AGGTGAACAATACTTTGTATTCATTATATATATATATATATATATATATATATA------TCACTAGTATTGTCATTGCCATTATTGATAATATTATCATTATGAACAACGA
BF936107  AGGTGAACAATACTTTGTATTCATTATATATATATATATATATATATATATATATCACTAGTATTGTCATTGCCATTATTGATAATATTATCATTATGAACAACGA
```

Fig. 1: polymorphism observed in a TA di-nucleotide repeat. The GenBank accession numbers are found on the left side of the sequence. The ORFs of these genes code for an unknown protein.

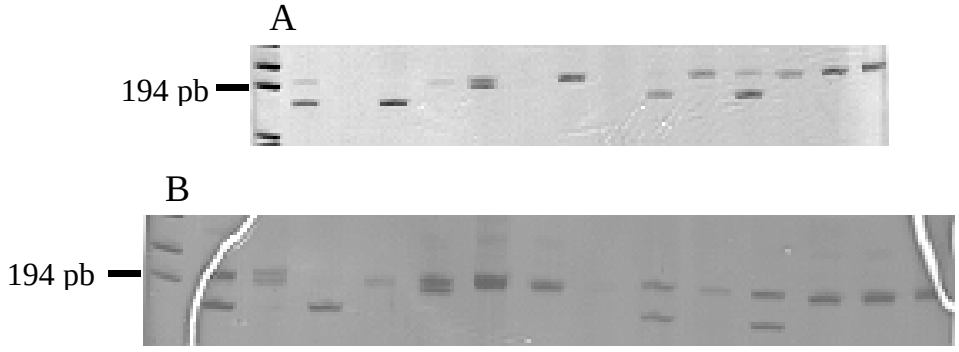

Fig. 2: polymorphism detected in two microsatellite loci, A: BH795455 containing TAGA repeats and B: BH795456 containing AAT repeats using DNA extracted from 14 individual adult worms of the LE strain, respectively. In A two amplification failed and in B one failed.

TABLE III

Analysis of microsatellites found in 42,017 *Schistosoma mansoni* BAC-end sequences deposited in the GSS division of GenBank™ database (01/11/01)

Repetition Unit (Number of repeats)	Number of BAC end sequences	Classification		
Mono-nucleotide		Perfect	Imperfect	Compound
A (10 to 99)	421	348	066	007
C (10 to 193)	274	189	083	002
Di-nucleotide				
AC (5 to 24)	189	165	018	006
AG (5 to 26)	139	128	003	008
AT (5 to 39)	897	841	037	019
CG(5 to 15)	019	015	004	000
Tri-nucleotide				
AAC (5 to 58)	028	026	001	001
AAG (5 to 59)	018	017	001	000
AAT (5 to 14)	348	304	037	007
ACC (5 to 9)	005	003	000	002
ACG (5)	002	002	000	000
ACT (5 to 11)	021	017	003	001
AGC (5 to 12)	004	004	000	000
AGG (5 to 10)	004	002	001	001
AGT (5 to 44)	105	060	025	020
ATG (5 to 36)	059	052	006	001
CGT (5 to 10)	002	002	000	000
Tetra-nucleotide				
AAAC (5 to 6)	004	004	000	000
AAAT (5 to 6)	028	028	000	000
AAGT (5 to24)	048	047	001	000
AATC (5 to 6)	008	008	000	000
AATG (5 to 15)	029	029	000	000
AATT(5)	001	001	000	000
ACAG (6 to 7)	002	002	000	000
ACGC (5 to 6)	002	002	000	000
ACTG (5 to 6)	003	003	000	000
AGAT (5 to 6)	002	002	000	000
AGCT (6)	002	002	000	000
AGGG (5)	001	001	000	000
AGTC (5 to 7)	014	013	001	000
ATAC (5)	001	001	000	000
ATCT (5 to 11)	012	010	000	002
ATGT (5 to 6)	002	000	000	002
CAGT(5)	001	001	000	000
CATA(5)	001	001	000	000
CTTT(5)	002	002	000	000
GTCA(7)	001	001	000	000
Penta-nucleotide				
TAAGT (6 to 12)	002	002	000	000
Total	2701	2335	287	079

DISCUSSION

The approach used here to identify microsatellite loci consisted in searching for repetitive *S. mansoni* DNA sequences, using RepeatMasker, in sequences deposited in GenBank and in sequences obtained from microsatellite-enriched libraries. The use of microsatellite enriched libraries generated a much greater relative number of microsatellite containing sequences (22%) as compared to ESTs (3%) or BAC end sequences searching (3.8%). However, there are costs of library construction and clone sequencing to be considered. This result contrasts with Durand et al. (2000), who used data from database and cloned libraries and selected less than 1% of sequences with microsatellite repeats. This difference may be due to the date the sequences were obtained, a large number of sequences were deposited in the last year, and/or to the tools used for searching. Our results show that the most common repeat found in *S. mansoni* was AT, which is in agreement with the high AT content (65%) reported for the schistosome genome (Hillyer 1974).

As described by Oliveira and Johnston (2001), clustered sequences can be useful for polymorphism identifi-

TABLE IV

Analysis of microsatellites found in 300 *Schistosoma mansoni* sequences from two *Schistosoma mansoni* microsatellite enriched libraries

Repetition Unit (Number of repeats)	Number of sequences	Classification		
Di-nucleotide		Perfect	Imperfect	Compound
AC (5 to 34)	073	036	027	010
CT (5 to 8)	002	000	002	000
TA (5 to 8)	002	000	001	001
Tri-nucleotide				
ATC (10)	001	001	000	000
ACC (5 to 10)	009	009	000	000
ATG (5 to 6)	002	001	001	000
ACT (5 to 21)	005	000	004	001
AAT (5 to 16)	007	004	003	000
Tetra-nucleotide				
ACTC (5)	002	002	000	000
AGAC (5 to 6)	003	002	000	001
GCGT(4 to 6)	001	000	000	001
ATCT	002	001	001	000
Total	109	056	039	014

cation and provide additional sequence information for primer design. The ESTs deposited in the GenBank come from a variety of strains possibly resulting in the observed polymorphism in ESTs clusters. This observation was not made in relation to BAC end sequences, probably because they were generated from a much less diverse source of DNA.

In conclusion, the increasing amount *S. mansoni* sequence information present in public databases can be used for the screening of possible polymorphic markers. These polymorphic genetic markers will be essential tools for research that will address questions on *S. mansoni* population structure and the genetic epidemiology of schistosomiasis.

REFERENCES

Ashley MV, Dow BD 1994. The use of microsatellite analysis in population biology: background, methods and potential applications. *Mol Ecol Evolut 69*: 185-201.

Caccio S, Spano F, Pozio E 2001. Large sequence variation at two microsatellite loci among zoonotic (genotype C) isolates of *Cryptosporidium parvum*. *Int J Parasitol 31*: 1082-1086.

Curtis J, Minchella DJ 2000. *Schistosoma* population genetic structure: when clumping worms is not just splitting hairs. *Parasitol Today 16*: 68-71.

Durand P, Sire C, Theron A 2000. Isolation of microsatellite markers in the digenetic trematode *Schistosoma mansoni* from Guadeloupe island. *Mol Ecol 9*: 997-998.

Franco GR, Valadão AF, Azevedo V, Rabelo EML 2000. The *Schistosoma* gene discovery program: state of the art. *Inter J Parasitol 30*: 453-463.

Hagelberg E, Gray IC, Jeffreys AJ 1992. Identification of the skeletal remains of a murder victm by DNA analysis. *Nature 352*: 427-429.

Hillyer GV 1974. Buoyant density and thermal denaturation profiles of schistosome DNA. *J Parasitol 60*: 725-727.

Lougheed SC, Gibbs HL, Prior KA, Weatherhead PJ 1999. Hierarchical patterns of genetic population structure in black rat snakes (*Elaphe obsoleta obsoleta*) as revealed by microsatellie DNA analysis. *Evolution 53*: 1995-2001.

Lucas M, Munoz C, Pintado E, Solano F 1997. Highly informative single-stranded conformation polymorphism (SSCP) of short tandem repeats in DNA identification. *J Forensic Sci 42*: 118-120.

Oliveira G 2001. *Schistosoma* gene discovery project update. *Trends Parasitol 17*: 108-109.

Oliveira G, Johnston DA 2001. Mining the schistosome DNA sequence database. *Trends Parasitol 17*: 501-503.

Rodrigues NB, Coura-Filho P, Souza CP, Jannoti Passos LK, Dias-Neto E., Romanha AJ 2002. Populational structure of *Schistosoma mansoni* assessed by DNA microsatellites. *Inter J Parasitol 32*: 843-851.

Sanguinetti CJ, Dias-Neto E, Simpson AJG 1994. Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. *BioTechniques 17*: 915-918.

Santos FR, Pena SDJ, Epplen JT 1993. Genetic and population study of a Y-linked tetra-nucleotide repeat DNA polymorphism with a simple non-isotopic technique. *Hum Genet 90*: 655-656.

Solano P, Duvallet G, Dumas V, Cuisance D, Cuny G 1997. Microsatellite markers for genetic population studies in *Glossina palpalis* (Diptera: Glossinidae). *Acta Trop 65*: 175-180.

Tautz D, Renz M 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res 12*: 4127-4139.

Weber JL, May PE 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet 44*: 388-396.

WHO-World Health Organization 1993. *The Control of Schistosomiais,* Second report of the WHO Expert Committee, Tech Rep Ser 830, Geneva.