Combination of de-novo assembly of massive sequencing reads with classical repeat prediction improves identification of repetitive sequences in Schistosoma mansoni

Julie Mireille Lepesant, David Roquis, Rémi Emans, Céline Cosseau, Nathalie Arancibia, Guillaume Mitta, Christoph Grunau

Combination of *de-novo* assembly of massive sequencing reads with classical repeat prediction improves identification of repetitive sequences in *Schistosoma mansoni*

Julie M.J. Lepesant [ab], David Roquis [ab], Rémi Emans [ab], Céline Cosseau [ab], Nathalie Arancibia [ab], Guillaume Mitta [ab] and Christoph Grunau [ab*]

a) Université de Perpignan Via Domitia, Perpignan, F-66860, France

b) CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan, F-66860, France

*) corresponding author :

Christoph Grunau, christoph.grunau@univ-perp.fr, Université de Perpignan Via Domitia, UMR 5244 CNRS Ecologie et Evolution des Interactions (2EI), 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

Tel : +33 468662180

Fax : +33 468662281

Abstract

The genome of the parasitic platyhelminth *Schistosoma mansoni* is composed of approximately 40% of repetitive sequences of which roughly 20% correspond to transposable elements. When the genome sequence became available, conventional repeat prediction programs were used to find these repeats, but only a fraction could be identified. To exhaustively characterize the repeats we applied a new massive sequencing based strategy: we re-sequenced the genome by next generation sequencing, aligned the sequencing reads to the genome and assembled all multiple-hit reads into contigs corresponding to the repetitive part of the genome. We present here, for the first time, this *de-novo* repeat assembly strategy and we confirm that such assembly is feasible. We identified and annotated 4,143 new repeats in the *Schistosoma mansoni* genome. At least one third of the repeats are transcribed. This strategy allowed us also to identify 14 new microsatellite markers, which can be used for pedigree studies. Annotations and the combined (previously known and new) 5,420 repeat sequences (corresponding to 47% of the genome) are available for download (http://methdb.univ-perp.fr/downloads/).

Keywords

Schistosoma mansoni; repetitive sequences; massive sequencing; *de novo* assembly

Despite their abundance, repetitive sequences of the genome are often considered as “junk”, “selfish”, or “parasitic” DNA that is tolerated by the genome but has no biological or evolutionary functions. This view is about to change. In 2005, Shapiro and von Stenberg discussed the importance of the repetitive sequences for the establishment of the frontiers between heterochromatin and euchromatin, and their influence on homologous and nonhomologous recombination (Shapiro and von Sternberg, 2005). Depending on their position, repetitive sequences can play a part in activation or repression of gene transcription (Goodier and Kazazian, 2008). Some repeats have important structural functions such as telomeric repeats or the long satellite blocks that make up the centromeres of mammals and insects (Kejnovsky et al., 2009). In some cases, transcription of repeats and subsequent processing into small RNA was described. These transcripts are involved in heterochromatization (small heterochromatin inducing RNA - shiRNA) (Reinhart and Bartel, 2002). Transposable elements, constituting a substantial share of the repetitive DNA, are known to have an impact on the genome evolution. Some were even selected to play a precise role in the cell (“domesticated repeats”) (Shapiro and von Sternberg, 2005). Taken together, repetitive elements can no longer be considered as a side-aspect of the genome and deserve a deeper investigation.

Schistosoma mansoni is a parasitic plathyhelminth responsible for intestinal schistosomiasis. This parasitic human disease ranks second only to malaria in terms of parasite-induced human morbidity and mortality, with more than 200 million infected people. The economic burden caused by the disease is tremendous as, for example, people disabled by the disease have limited job performances and are less likely to contribute to the local development. It was estimated that schistosomiasis burden represents 25 - 50 million disability-adjusted life-years (DALY) (King, 2010). The life cycle of the parasite is characterized by passage through two obligatory hosts: a fresh-water snail (*Biomphalaria* species, depending on the geographical location) as intermediate host, and humans or rodents as the final host. Miracidia infect the snail and transform into primary and secondary sporocysts, from which cercariae, capable of infecting the human host, are released

into the water. Based on RepeatScout data, the genome of *S. mansoni* was thought to contain approximately 40% repetitive sequences (Berriman et al., 2009), of which roughly 20% correspond to transposable elements (Simpson et al., 1982). Over the last 30 years, roughly a dozen repetitive sequences have been identified by classical molecular biology methods e.g. (Copeland et al., 2003; Copeland et al., 2006). When the genome sequence became available, conventional repeat prediction programs were used to identify additional repetitive sequences. These 1,225 repeat sequences are available from the J. Craig Venter Insitute Institute (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/s_mansoni/preliminary_annotation/homology_evidence/sma1.repeats.gz) and their naming convention suggests that RepeatScout was used for prediction. Fifty-five repeats were available in GenBank. At this point, when we had re-sequenced the genome by massively sequencing (next-generation sequencing, NGS) our mapping results suggested that a large number of additional repeats must exist in the *S. mansoni* genome. We reasoned that by combining alignment information to identify reads that correspond to multiple locations on the genome and short-read assembly it should be possible to indentify all repeats in a genome without *a priori* information.

A Brazilian strain (Bre) and a Guadeloupean strain (GH2), maintained respectively in their sympatric *B. glabrata* strain, were used in this study. Miracidia and eight-week adult worms were recovered as described before (Theron et al., 1997) and kept at -80°C. The French Ministère de l'Agriculture et de la Pêche and French Ministère de l'Education Nationale de la Recherche et de la Technologie provided permit A 66040 to our laboratory for experiments on animals and certificate for animal experimentation (authorization 007083, decree 87–848) for the experimenters. Housing, breeding and animal care followed the national ethical requirements. Genomic DNA was extracted from 10 adult couples using the phenol-chloroform protocol. For total RNA purification, three independent preparations of each larvae and adults were used. For the larval stages, RNA was extracted from 10,000 miracidia using 500 µl Trizol (Invitrogen™). Ten adult couples were solubilized in 500µl Trizol with a MagNA Lyser and Green beads (Roche). RNA was treated with

DNA-free (Ambion #cat: AM1907) for 45 minutes at 37°C, followed by inactivation of the enzyme using the inactivation reagent. PCR of 28s rDNA was used to test for genomic DNA contaminations. First strand cDNA was synthesized using 20 ng of the total RNA preparation, in a final volume of 20µl with 200 U of RevertAid (Fermentas, #cat: G2101). To assemble the repeat genome, we used a total of 38,004,342 36-bp single-end reads generated on a Genome Analyzer II (Illumina) according to the manufacturer's protocol, at the MGX and Oregon State University sequencing facilities. Sequences are available at the NCBI sequence read archive (study accession numbers SRA012151.6 and SRA043796.1). Reads were aligned to the reference genome v.3.1 with SOAP2/SOAPaligner (Li et al., 2009) evoking the -r 0 (not repeats) and -u (write unmapped reads into a file) options. The rationale behind this approach was that in this case, SOAP would only align reads with a single occurrence in the genome. All other reads correspond either to unknown sequences or occurring more than once, *i.e.* are repetitive. The 12,535,613 unmapped sequence reads (33% of total) were then assembled using Velvet 0.7.01 (Zerbino, 2010) with the -cov_cutoff 4 -min_contig_length 100 options resulting in 8,608 contigs. A long read assembler (Sequencher version 4.5 (Gene Codes) min match=93%, min overlap=60 bp) was used to produce finally 8,594 contigs. Each repeat was assembled individually and therefore the assemblies may be composed of two or more distinct, but very similar repeats. First pass annotation of the 8,594 presumed repeat contigs was done with Blast (Altschul et al., 1990), Censor/Repbase (Kohany et al., 2006), TEclass (Abrusan et al., 2009) and Tandem Repeats Finder (Benson, 1999). Blast2GO v2.4.8 (Conesa et al., 2005) was used to carry out various types of BLAST searches (conditions in Supplementary Table 1 and results in Table 1). CENSOR (http://www.girinst.org/censor/) (Kohany et al., 2006) and the Repbase Update (http://www.girinst.org/repbase/) (Jurka et al., 2005) were applied to find sequences sharing similarity to known repeats. Parameters were: *Sequence source*: all; *Forced translated search*: no; *Search for identity*: no; *Mask simple repeats*: yes; *Mask pseudogenes*: yes. Results were evaluated according to the 80/80/80 principle (80% of identity on 80% of the sequence spanning a minimum of 80 bp) (Wicker et al., 2007). The web application TEclass

(http://www.compgen.uni-muenster.de/tools/teclass/) (Abrusan et al., 2009) was used to predict potential transposable elements in a sequence with default parameters. Finally we used Tandem Repeats Finder (TRfinder, http://tandem.bu.edu/trf/trf.html) (Benson, 1999) to identify tandem repeats. Parameters were optimized for highest sensitivity and specificity (respectively 81% and 97%) using *in silico* generated training sequences: *Minimum alignment score*: 30; *alignment parameters*: 2-7-7. TRfinder was also used to find sequences with a period of 2, 3 or 4 bp that could serve as new microsatellite markers. Candidates were verified to have only one occurrence in the genome, to be polymorphic (by comparison with trace files used for the genome assembly) and it was checked if they were located in a gene.

For confirmation of *in-silico* results, PCRs were carried out in a final volume of 25 µL containing 0.2 µmol of each oligonucleotide primer (Supplementary Table 2), 0.2 mmol of each dNTP (Promega), 1.25 U of GoTaq polymerase (Promega, #cat: M3175) used with the recommended buffer and completed to the final volume with DNase-free water (95 °C for 10 min followed by 35 cycles at 95 °C for 30 s, 53°C for 30 s, 72 °C for 1 min and a final extension at 72 °C for 10 min). The PCR products were separated by electrophoresis through a 2% TBE agarose gel. Real-time quantitative PCR analyses were performed using the LightCycler 2.0 system (Roche Applied Science) and LightCycler Fast-start DNA Master SYBR Green I kit (Roche Applied Science) with 2.5 µl of cDNA in a final volume of 10 µl (3 mM $MgCl_2$, 0.5 µM of each primer, 5 µl of master mix). The primers were designed with the LightCycler Probe design software or the Primer3Plus web based interface (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). The following protocol was used: 95 °C for 10 minutes; 40 cycles: 95 °C for 10 seconds, 60 °C for 5 seconds, 72°C for 16 seconds; melting curve, 60-95 °C with a heating rate of 0.1 °C/s and continuous fluorescence measurement, and a cooling step to 40 °C. For each reaction, the crossing point (Ct) was determined with the “second derivative method” of the LightCycler Software 3.3. PCR reactions were done in quadruplicate and the mean value of Ct was calculated. 28s rRNA was used as an internal control and the amplification of a unique band was verified by electrophoresis

through 2% agarose gels for each qPCR product. Primer sequences and expected PCR product sizes are listed in Supplementary Tables 2 and 3. For all qPCR, efficiency was at least 1.95.

Our assembly of SOAP-sorted massive sequencing reads delivered 8,594 contigs. Contigs smr_2181, smr_2685, smr_2733, smr_3000, smr_3595, smr_3826, smr_4022, and smr_6227 – that we tested by PCR - showed a band at the predicted size indicating that assembly is correct in most cases. For two contigs (smr_3000 and smr_3826) a supplementary band was present, with a molecular weight twice as high as the major band, suggesting repetition in that these fragments correspond to an amplicon of 2 tandem repeats in the genome. Among the 8,594, we clearly identified 6,531 (76%) as repeats via *in silico* analysis using the following criteria: two or more occurrences in the reference genome and no Blast annotation (against the nr database from NCBI) related to a known gene or protein (with the exception of proteins typical of transposable elements, such as transposase, reverse transcriptase or GAG polyprotein), using an e-value cut-off of $1.0E^{-30}$. Using the information obtained from the Blast done on the nr database, we identified 306 contigs related to genes or gene families, which include 40 mitochondrial genes. A Blast search against the reference genome of *S. mansoni,* which allowed us to count the number of occurrences of each repeat, showed that 1,332 sequences (other than the ones identified as genes) were unique, and 230 were absent from the reference genome. All Blast conditions are listed in Supplementary Table 1. At this point of our analysis, a total of 10 sequences are suspected to be possible contamination, as they are neither present in the reference genome, nor in the trace files, but match perfectly sequences from other organisms. Four sequences correspond to rodents and six to small freshwater organisms such as *Hydra magnipapillata*. It is likely that the sequences are due to contaminations from host tissue and spring-water used for parasite culture. Out of the 6,531 repeated sequences, TRfinder detected 516 containing tandem repeats. Two thousand four hundred and forty five repeated sequences matched fully or partially to already known repeats, *i.e.* those present in the RepeatScout generated database of repeat predictions. In conclusion, assembly of repeats is possible by our new approach and so far, 4,143 unknown repeats have been identified. Based on the

combined results of Blast and Censor we classified the 6,531 repeats into 9 groups, and into 5 with TEClass. This categorization is based on the hierarchy of classes of repeats suggested by the Censor results (list of classes and subclasses available on the GIRI website, http://www.girinst.org/censor/help.html). The majority of repeat annotations belongs to class I and class II (retro) transposons (Table 1b). For 1,921 of the 6,531 repeats (29%) we found ESTs through Blast searches (Supplementary Table 1), indicating that a large part is transcribed. Blast searches in stage specific ESTs revealed a homogeneous distribution among the life cycle stages (data not shown). For seven repeats with 0 to 52 EST hits, we verified transcription in miracidia and adults. Six (smr_2181, smr_2685, smr_2733, smr_3595, smr_4022 and smr_6227) showed low, but significant above background transcription, while for a single repeat (smr_7590) no transcription was found (Figure 1). There was no correlation between the number of EST hits and transcription level measured by qPCR. We used TRfinder, Blast searches against the trace file database and Blast searches against the *S. mansoni* genome followed by visual inspection to identify 14 new polymorphic microsatellite markers. Five of them are located in predicted gene coding regions and the remaining 9 are probably neutral markers (Supplementary Table 4). To generate a combined database we used the abovementioned 1,225 predicted repeats. We then analyzed the 55 repeats available in GenBank and identified three duplicates: Sm_SR-AB2 (AF025674.1), Sm_SR1_pol_3 (U66331.1) and Sm_Sinbad_iS4-T (AY965073.1). For Sm_salmonid (AY834402.1) we have shown previously (Grunau and Boissier, 2010) that it is not present in *S. mansoni*. Consequently, we removed these 4 repeats from the database, and added the remaining 51 repeats sequences available in GenBank and a tandem repeat earlier identified in our laboratory (TandemRepeat_266) to the combined database (1,225 predicted + 55 GenBank – 3 duplicates – 1 wrong + 1 previously identified = 1277). In total, our new *Schistosoma mansoni* repeat database contains 5,420 sequences. The workflow and a summary of results are shown in figure 2. This repeat database was used for the annotation of assembly 3.1 of the *S.mansoni* genome. We employed RepeatMasker (Smit et al., 1996) evoking the -cutoff 250 -norna options. The repeatmasking was done with the

original Sanger/TIGR library and with our combined library. Repeatmasker identified repeats in 47.40% of the genome. Fragmentation and overlapping repeatmasking were considerably lower with our new library resulting in annotation of 623,983 repeats compared to 881,451 obtained with the previous Sanger/TIGR data. Examples that show the improved quality of annotation are shown in supplementary figure 1. During the preparation of the manuscript, version 5.2 of the *S. mansoni* genome became available in which scaffold redundancy was reduced (Protasio et al., 2012). We reasoned that some of our new repeats might be absent from the new assembly and repeated the blast search with an e-value of 1e-30. We removed all repeats that did not match to the genome 5.2 or that matched only once resulting in 3,145 different repeat sequences. Repeatmasking showed that 47.73% is repetitive, *i.e.* essentially the same value as for the 3.1 assembly. Repeat annotations were converted into GFF and loaded on an in-house genome browser for visualization. The repeats are available as fasta files and annotations as GFF files (http://methdb.univ-perp.fr/downloads/).

We describe here - to our knowledge for the first time - a *de novo* assembly strategy for repetitive sequences making use of a filter option in short read alignment programs such as SOAP. Short reads that match to more than one region of the reference genome can be handled in two different ways by these programs. Either, they are randomly positioned to the multiple matching loci, or they are not at all aligned. We made use of this latter option and employed the alignment program as a filter for repetitive sequences. A potential caveat of this method is that unique sequences that for some reasons are not present in the reference genome (or not aligned to it) will also initially be included in the sequence bin that will serve as the input for the repeat assembly process. These sequences must be eliminated after assembly by alignment to the genome. The rationale behind this approach is that true repeats will be detected - by their very nature - more than once in the genome. All other sequences must be removed. This strategy allowed us to identify 4,143 new repeats in the *S. mansoni* genome in addition to the 1,225 unique predicted and the 52 experimentally found repeat sequences available in GenBank or through our own work. The strategy can naturally also be used for other species for which a reference sequence and massive sequencing reads are available.

We then annotated the repeats with conventional annotation programs in order to classify them. It is interesting to note that a substantial proportion of the repeats (at least 30% but probably more) is transcribed. It is tempting to speculate that this transcription has a biological function as observed in other species. Further work will be necessary to explore this option in schistosomes. Finally, our approach allowed us to identify 14 new polymorphic microsatellite markers, of which 9 are probably neutral markers. Microsatellite markers are popular tools in many fields of biology such as pedigree studies or epidemiology. While being of great interest, they are notoriously difficult to find, and during the last decade only 43 were discovered (Supplementary Table 4). Our new markers could bring this number to 57, providing a larger choice that will be useful in situations where the so far used markers are not sufficiently polymorphic.

Acknowledgements

The authors are grateful to Michael Freitag (Oregon State Univesity) for library construction and sequencing of data SRA043796.1, and to the staff of the MGX-Montpellier GenomiX (MGX) sequencing center for generation of data SRA012151.6. Patricia Ruy (SchistoDB, Centro de Pesquisas Rene Rachou, Belo Horizone, Brazil) provided support for Blast against *S.mansoni* EST databases. This work received funding from the French National Agency for Research (ANR), project ANR-2010-BLAN-1720-01 (EPIGEVOL). The funding source had no involvement in the study design and publication decision. The authors declare to have no conflict of interest.

References

Abrusan, G, Grundmann, N, DeMester, L,Makalowski, W, 2009. TEclass--a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25, 1329-1330.

Altschul, SF, Gish, W, Miller, W, Myers, EW,Lipman, DJ, 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Benson, G, 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573-580.

Berriman, M, et al., 2009. The genome of the blood fluke Schistosoma mansoni. Nature 460, 352-358.

Conesa, A, Gotz, S, Garcia-Gomez, JM, Terol, J, Talon, M,Robles, M, 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676.

Copeland, CS, Brindley, PJ, Heyers, O, Michael, SF, Johnston, DA, Williams, DL, Ivens, AC,Kalinna, BH, 2003. Boudicca, a retrovirus-like long terminal repeat retrotransposon from the genome of the human blood fluke Schistosoma mansoni. Journal of Virology 77, 6153-6166.

Copeland, CS, Lewis, FA,Brindley, PJ, 2006. Identification of the Boudicca and Sinbad retrotransposons in the genome of the human blood fluke Schistosoma haematobium. Mem Inst Oswaldo Cruz 101, 565-571.

Goodier, JL,Kazazian, HH Jr, 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23-35.

Grunau, C,Boissier, J, 2010. No evidence for lateral gene transfer between salmonids and schistosomes. Nature Genetics 42, 918-919.

Jurka, J, Kapitonov, VV, Pavlicek, A, Klonowski, P, Kohany, O,Walichiewicz, J, 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462-467.

Kejnovsky, E, Hobza, R, Cermak, T, Kubat, Z,Vyskot, B, 2009. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity 102, 533-541.

King, CH, 2010. Parasites and poverty: the case of schistosomiasis. Acta Trop 113, 95-104.

Kohany, O, Gentles, AJ, Hankus, L,Jurka, J, 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7, 474.

Li, R, Yu, C, Li, Y, Lam, TW, Yiu, SM, Kristiansen, K,Wang, J, 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.

Protasio, AV, et al., 2012. A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke Schistosoma mansoni. PLoS Neglected Tropical Diseases 6, e1455.

Reinhart, BJ,Bartel, DP, 2002. Small RNAs correspond to centromere heterochromatic repeats. Science 297, 1831.

Shapiro, JA,von Sternberg, R, 2005. Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80, 227-250.

Simpson, AJ, Sher, A,McCutchan, TF, 1982. The genome of Schistosoma mansoni: isolation of DNA, its size, bases and repetitive sequences. Mol Biochem Parasitol 6, 125-137.

Smit, AFA, Hubley, R,Green, P, 1996. RepeatMasker Open-3.0.

Theron, A., Pages, J. R.,Rognon, A., 1997. Schistosoma mansoni: distribution patterns of miracidia among Biomphalaria glabrata snail as related to host susceptibility and sporocyst regulatory processes. Experimental Parasitology 85, 1-9.

Wicker, T, et al., 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews. Genetics 8, 973-982.

Zerbino, DR, 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11, Unit 11.5.

Figure legends

Figure 1: Transcription level of arbitrarily chosen repeats in miracidia and adults of *S.mansoni* measured by RT-qPCR (4 replicates). Transcription is expressed in fold of 28S rRNA that served as reference.

Figure 2: Schematic representation of the workflow that led to the identification and annotation of the repeat genome of *S. mansoni*.

Suppl. Figure 1:

Arbitrarily chosen examples illustrating the improved repeat annotation with the new repeat database (bottom: “new repeat database”) compared to the previous solely prediction-based database (middle: “Repeats”). Differences indicated by grey rectangles.

A) The newly identified repeat smr_6601 exists 170 times in the genome. Gene predictions (e.g. Smp_122190) at the same genomic locations are therefore probably an artifact.

B) Several previously as unique described regions are in fact repetitive. The example shows also lower redundancy in the new data.

C and D) Large regions of this scaffold were thought to be unique and (C) to contain genes. Our data show that the region is entirely (C) or largely (D) composed of repeats.

Table 1a : Summary of combined Blast and Censor results

Repeat class	Number	%
Class I (Retrotransposons)	565	8.6
Class I (non-LTR Retrotransposons)	720	11
Class II (DNA transposons)	239	3.7
SINEs	31	0.5
Endogenous retroviruses	18	0.3
Interspersed repeats	21	0.3
Pseudogenes, snRNA, rRNA, tRNA	38	0.6
Unknown	4899	75
total	6531	

Table 1b : Summary of TEClass results

Repeat class	Number	%
Class I (Retrotransposons)	1841	28
Class I (LTR Retrotransposons)	2067	32
Class I (non-LTR Retrotransposons)	70	1
Class II (DNA transposons)	1563	24
LINEs	398	6
Unknown	592	9
total	6531	

Supplementary Table 1: Databases used for BLAST searches and cut-off e-value

Database	BLAST	e-value
nr (NCBI) *http://www.ncbi.nlm.nih.gov/*	X	1.00E-30
Sma Repbase v2 *http://www.sanger.ac.uk/*	N	1.00E-30
nr (NCBI) *http://www.ncbi.nlm.nih.gov/*	N	1.00E-30
Reference genome *S. mansoni*, strain NMRI versions 3.1 and 5.2 *http://www.sanger.ac.uk/*	N	1.00E-30
Chromatograms used for the assembly of reference genome *http://www.sanger.ac.uk/*	N	1.00E-30
ESTs *S. mansoni* *http://www.ncbi.nlm.nih.gov/*	N	1.00E-15
Reference genome *S. japonicum*, strain Anhui *http://lifecenter.sgst.cn/*	N	1.00E-20
ESTs *S. japonicum* *http://lifecenter.sgst.cn/*	N	1.00E-15

Supplementary Table 2: Primers used in the study for standard PCR

Name	Forward	Reverse	Size
smr_2181	GCTATAATGGGCAAGGTAAATAAGTC	AAACAACGATTCACACAAATTCA	400
smr_2685	CATTTCTCTTCCTGCAGTCCT	GATCTTACGTGAAAATCGCATC	340
smr_3000	TTGTCTTATCATTTCTCAGTTGCTTC	TAGAACTTTTATCAGGGAACGTATTCA	539
smr_3826	ATATGGTGTTGTAGCGTGTGC	TTATCCCTTCAATCTCTATTAAAAACA	579
smr_3595	CCCTTTAGGGCCTCTCTTAGC	AAATTGGTAAATACAATGGGATGTG	709
smr_4022	CGTTTGCTATTAGTGTTAGGAATCTT	GTGATAACGGAAACTAGGATGGTC	713
smr_2733	ATCAACTATACGATTCCCTAAAACA	TCACTCGATAATTCGCTTAGCC	393
smr_6227	CATAATCATAGGGGATAAATGTGAA	ATTAAACCACGTTTGTAATGAATTG	243
smr_7590	GATTTCAATTATCAGGAAAACAACTTT	CTTATAAATCGCTTTGAAAACTCTG	241

Supplementary Table 3: Primers used in the study for RT-qPCR

Name	Forward	Reverse	Size
smr_2181	AGGGTCGGTTGGTTGTCCTCTGG	ACGCTTTTGGCATTGATTTTGCGA	101
smr_2685	CACATGGTCTCCGGAGTCAAGCA	TGAGCGCGTCTGCAACACTGA	125
smr_3000	TTGTCTTATCATTTCTCAGTTGCTT	AATTGAACCAGCATTGAATATACTT	152
smr_3826	ACCAAGCCGAAAGTGAAGACACCG	CGCACAAAGGCGTCTTTCCCA	138
smr_3595	ACACATGAACCAGAAAGTAACGGCCA	TGTCCCGTCCAAGTTGGTTTCCT	113
smr_4022	TCAGTTGGGGCATAGCAGCCA	CACCGGTGTCGGCTTCGTCG	147
smr_2733	ACGTCATCCGGCAGCAACGA	GGCAGAACTCTGGTTTGCACGCT	105
smr_6227	ACGTGCGAAGAAGAGAAGAACAACCA	GCCCCGTCCATGTCGGCTTT	112
smr_7590	CATGTTTTATAAGAAAGATTACATCAGC	AAACTCTGTAAACGTTGTCTCA	154

Supplementary Table 4: Characteristics of 14 *Schistosoma mansoni* microsatellite loci.

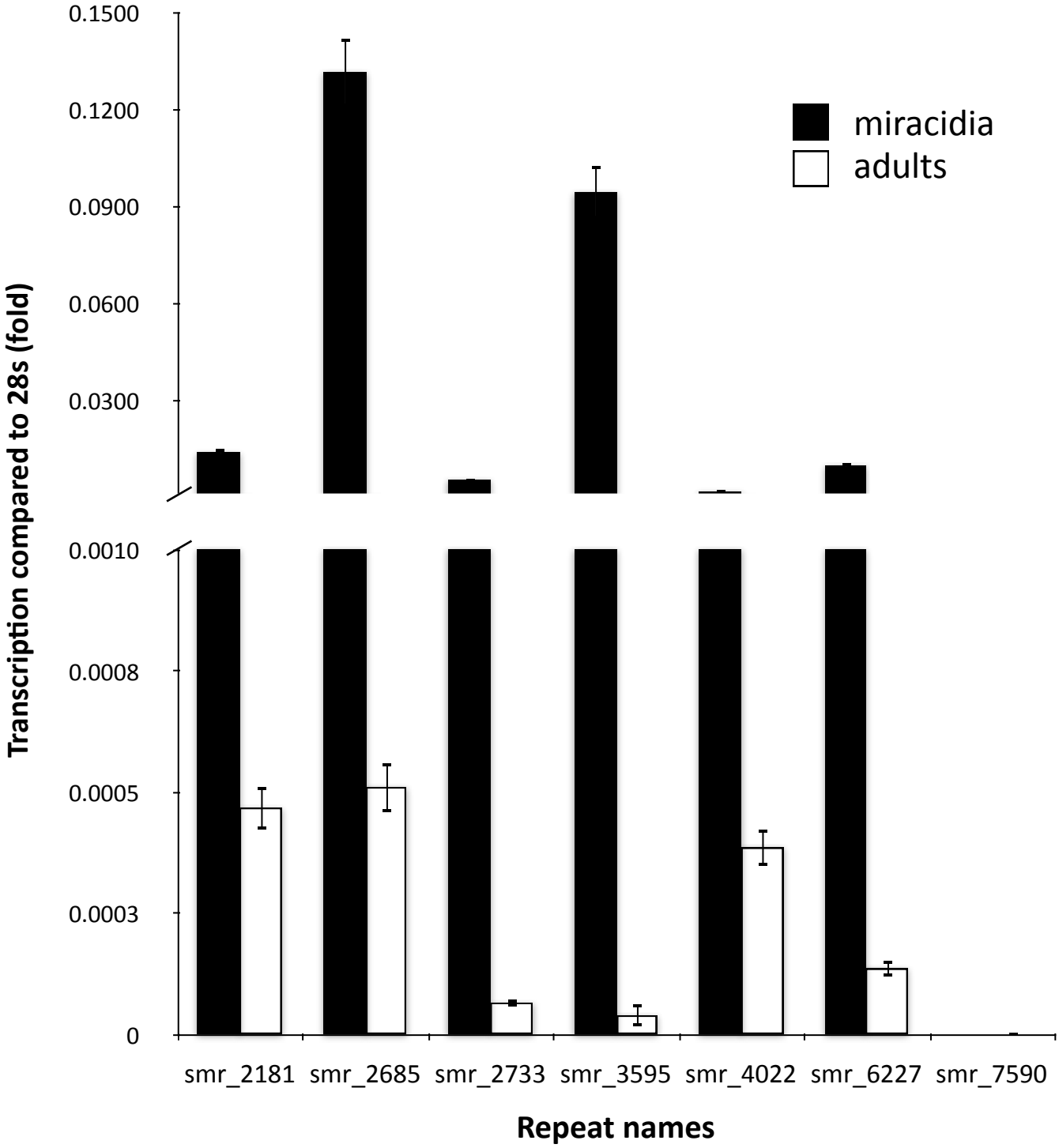
Repeat name	Genome position	Gene in vicinity
smr_93	Smp_scaff000078:1883819..1884094	
smr_534	Smp_scaff008778:843..259	
smr_1715	Smp_scaff001981:160389..160587	
smr_2501	Smp_scaff006696:1005..1124	
smr_3069	Smp_scaff011538:822..633	
smr_3778	Smp_scaff015860:304..150	
smr_3838	Smp_scaff008307:1204..419	
smr_6449	Smp_scaff008307:205..393	
smr_7342	Smp_scaff008269:1329..1531	
smr_1025	Smp_scaff018982:4667..4845	smp_194060
smr_2261	Smp_scaff002707:1010..917	smp_106750
smr_2770	Smp_scaff000617:14517..14293	smp_098420
smr_6232	Smp_scaff000018:21638..21810	smp_127980
smr_6375	Smp_scaff000018:21638..21810	smp_186020

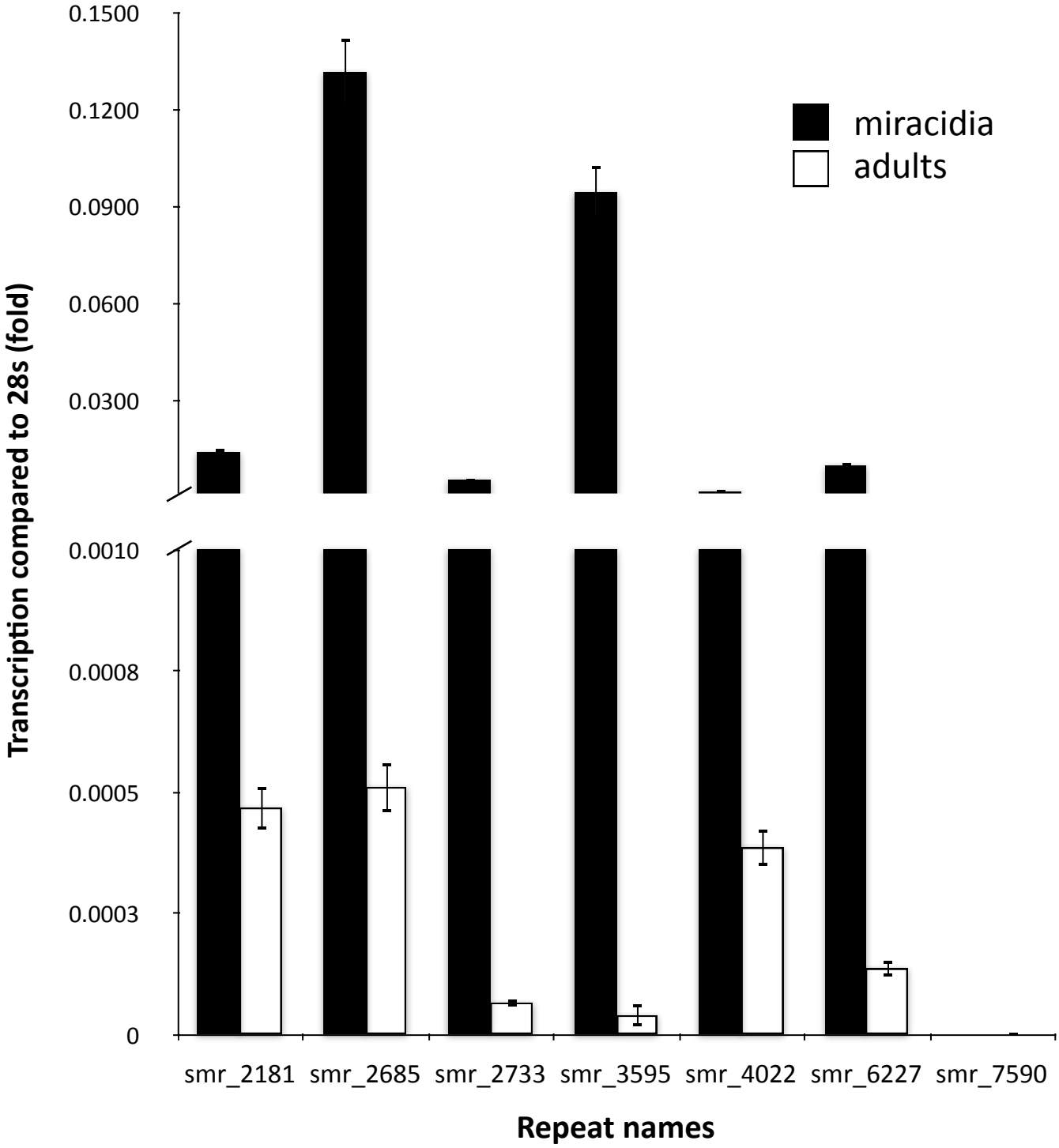
Supplementary Table 5 : Currently used microsatellite markers and corresponding PCR primers

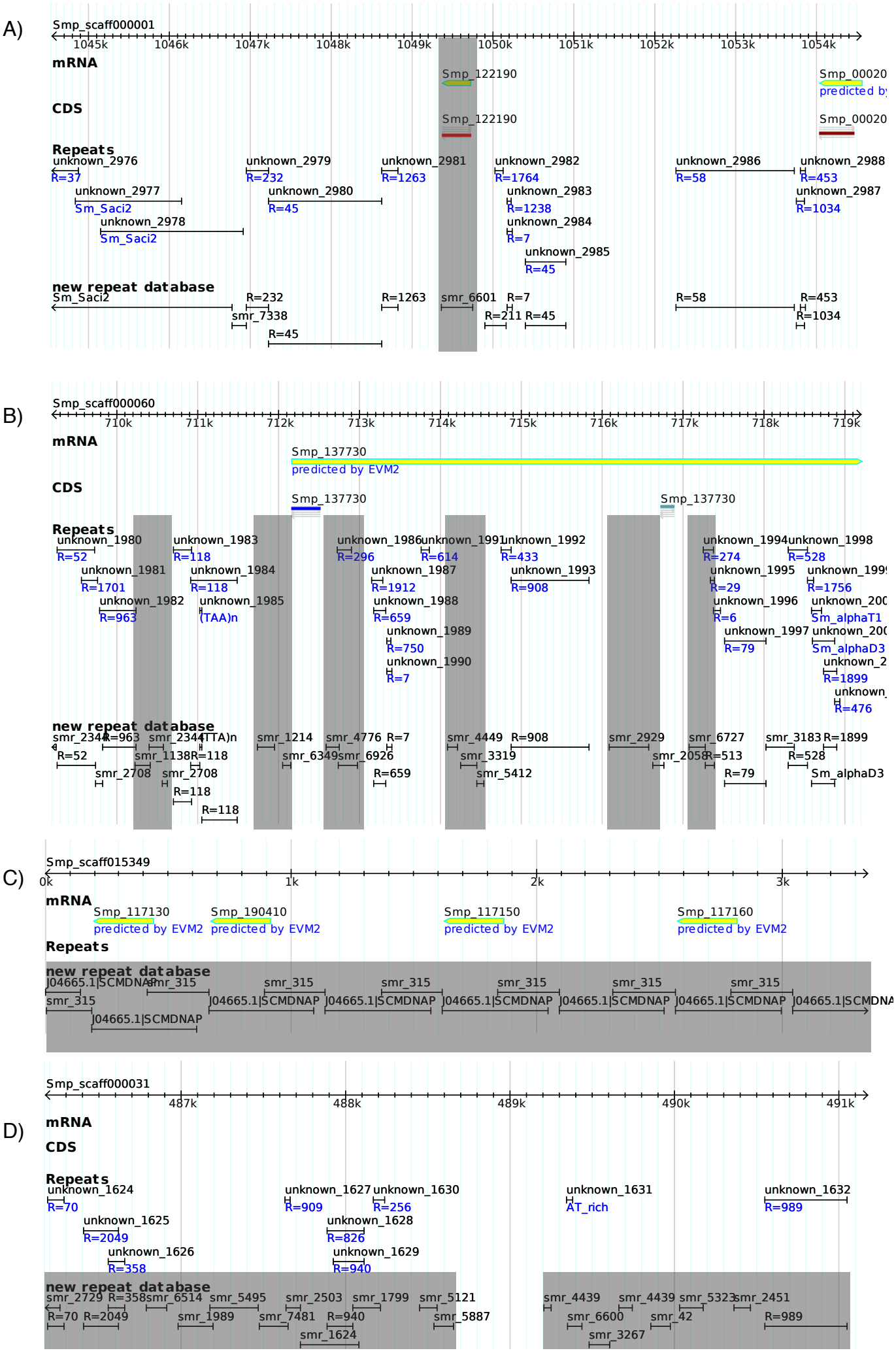
Locus	Access. number	F	R	Ref.
SmBr1	L81235	CGGAACGACAAGAAAATCAT	GAGTATACGGCTTCTTGGA	Rodrigues,N.B., Coura Filho,P., de Souza,C.P., Jannoti Passos,L.K., Dias-Neto,E. and Romanha,A.J. (2002) Populational structure of Schistosoma mansoni assessed by DNA microsatellites. Int. J. Parasitol. 32: 843-851.
SmBr2	L26968	ATCTCAAAGCCCAATACAAC	CATTTTCACTTACTGTTTATCC	
SmBr3	X77211	CATTTATTGATAATCTTTGGC	TTAACTACTTCTCACTGATG	
SmBr4	M15371	CAATTGACTATTGAAAAGGC	CGTAAAGACCACGATAAG	
SmBr5	L25065	AAACTATTCATTACTGTCGGG	GAATTACTGTCCCTTTATCTC	
SmBr6	AF009659	GAATACAGGCTATAATCTACA	CTTAACAGACATACACGC	
SmBr7	DQ137434	CGTCATCACCTTAAACATGAAC	AATCACCAATGGCAACAATCTG	Rodrigues,N.B., Silva,M.R., Pucci,M.M., Minchella,D.J., Sorensen,R., LoVerde,P.T., Romanha,A.J. and Oliveira,G. (2007) Microsatellite-enriched genomic libraries as a source of polymorphic loci for Schistosoma mansoni Molecular Ecology Notes 7: 263-265.
SmBr8	DQ448292	TAGGACAGGTTTTCCACCAA	ATGCCCACACACAAAGTAAA	
SmBr9	DQ137431	ATTCACCCATTGTCTTAAAACC	ATTGGCGTCAGTAGAAGAGATT	
SmBr10	DQ448293	CATGATCTTAGCTCAGAGAGC	GTACATTTTATGTCAGTTAGCC	
SmBr11	AC112150.4	AAGAAGTGGAGGAGGCCTTT	TTCAGTCCCTGGAACACACA	
SmBr12	DQ137724	TATATAGCAAAAGTAGTCTATATTCGTAGC	AGTAAAAACTATCCTATCCATTTCTATTG	
SmBr13	DQ137790	GTCACAGATACCTGACGAGCTG	ACTCCCCAGCAATTTGTCC	
SmBr14	DQ514536	CTGCTCATCATAGAAGTGTGGC	TCTATGTATCTACCCCACCCTATC	
SmBr15	AF325695	TATAGGACAAAACGCGGGTC	TTGGATAAACTTAGTGACTTTTC	
SmBr16	L04480	TGTGACTTTGATGCCACTGA	GGCCTGATACAATTCTCCGA	
SmBr17	AQ841039	CTGCAGGGGGAAATAGAAG	TGATCCTTTGTGCCAACA	

Locus	Access. number	F	R	Ref.
SMDA28	AF325695	CATGATCTTAG CTCAGAGAGCC	AGCCAGTATAG CGTTGATCATC	Curtis,J., Sorensen,R.E., Page,L.K. and Minchella,D.J. (2001) Microsatellite loci in the human blood fluke Schistosoma mansoni and their utility for other schistosome species Molecular Ecology Notes 1: 143-145.
SMD43	AF325697	CCCACCACAAT TTATTGATCTC	GGGTCCTCCAT TCCACTG	
SMDA23	AF325696	CCTGGTCCTAC GTTGTAGCTG	ACTTGACCTTA TTCCCCTTTCC	
SMC1	AF325694	TGACGAGGTTG ACCATAATTCT AC	AACACAGATAA GAGCGTCATGG	
SMDO11	AF325698	TGTTTAAGTCG TCGGTGCTG	ACCCTGCCAGT TTAGCGTAG	
S2-1	AI740252	TCTTTTTAAACT CTTGGCTCCTA T	GGTGAGACACA TGTTTAGTTT	Blair,L., Webster,J.P. and Barker,G.C. (2001) Isolation and characterization of polymorphic microsatellite markers in Schistosoma mansoni from Africa Molecular Ecology Notes 1: 93-95.
SCGA3	AA629514	TCTCCCTTCCCC CACCTT	TGACAGGGAAA TAGTAATGACA AC	
SATA12	AI395718	AGGTGCAACAA CCATAACATTT	CTTTGGACCGG CAGTAGC	
CA1-1	AI740374	TACTGGGGCTG GGGAGATG	AATAGTTGGAA CAGTGGCTCAT CA	
CA11-1	AI068335	TTCAAAACCAT GAGCAATAGAT AC	CAACAAACAAG AAGGCTGATTA G	
sms6 –1	AF330104	ATTACGATTGC ACAGATACTTT TG	TCCCTTTTGCCC TTTTTATTC	
sms7–1	AF330105	TCCTCCTCTCTA TTTTCTCTTTG	ATTACGATTGC ACAGATACTTT TG	
sms9 –1	AF330106	ATTACGATTGC ACAGATACTTT TG	TTTCAGAAATT TGTTTCCTCCTC	
smca7–1	AF330107	AGGAAGCTGCA TTGACTGGAG	CAAGAGCACTC TTCCAACCCT	
smca14 –1	AF330108	ACTATCTCCCT CTACCTCCCTC CC	TGATGGAGATG GTACGAAGAGA GA	
SMD25	AF202965	GATTCCCAAGA TTAATGCC	GCCATTAGATA ATGTACGTG	Durand,P., Sire,C. and Theron,A.

Locus	Access. number	F	R	Ref.
SMD28	AF202966	CATCACCATCAATCACTC	TATTCACAGTAGTAGGCG	(1995) Isolation of microsatellite markers in the digenetic trematode Schistosoma mansoni from Guadeloupe island Journal of Fish Biology 47: 29-55.
SMD57	AF202967	TCCTTGATTCCACTGTTG	GCAGTAATCCGAAAGATTAG	
SMD89	AF202968	AGACTACTTTCATAGCCC	TTAAACCGAAGCGAGAAG	
SMD94	AF202969	TAACACTCACACATACCC	AACTAATCACCCACTCTAC	
AI068335	AI068335	GTTGAGAGAGAAAAAGAAG	AGATGTTAGAAAGTGGTG	
L46951	L46951	CAAACATATACATTGAATACAG	TGAATTGATGAATGATTGAAG	
SCMSMOXII	M85305	TTCTACAATAATACCATCAAC	TTTTTTCTCACTCATATACAC	
R95529	R95529	GTGATTGGGGTGATAAAG	CATGTTTCTTCAGTGTCC	
SMU31768	U31768	TACAACTTCCATCACTTC	CCATAAGAAAGAAACCAC	
SMIMP25	X77211	CACTATACCTACTACTAATC	TCGATATACATTGGGAAG	


0.1500
0.1200
0.0900
0.0600
0.0300
0.0010
0.0008
0.0005
0.0003
0
miracidia
adults
Transcription compared to 28s (fold)
smr_2181
smr_2685
smr_2733
smr_3595
smr_4022
smr_6227
smr_7590
Repeat names


0.1500
0.1200
0.0900
0.0600
0.0300
0.0010
0.0008
0.0005
0.0003
0
miracidia
adults
Transcription compared to 28s (fold)
smr_2181
smr_2685
smr_2733
smr_3595
smr_4022
smr_6227
smr_7590
Repeat names


A)
Smp_scaff000001
1045k
1046k
1047k
1048k
1049k
1050k
1051k
1052k
1053k
1054k
mRNA
Smp_122190
Smp_00020
predicted by
CDS
Smp_122190
Smp_00020
Repeats
unknown_2976
R=37
unknown_2977
Sm_Saci2
unknown_2978
Sm_Saci2
unknown_2979
R=232
unknown_2980
R=45
unknown_2981
R=1263
unknown_2982
R=1764
unknown_2983
R=1238
unknown_2984
R=7
unknown_2985
R=45
unknown_2986
R=58
unknown_2988
R=453
unknown_2987
R=1034
new repeat database
Sm_Saci2
R=232
smr_7338
R=45
R=1263
smr_6601
R=7
R=211
R=45
R=58
R=453
R=1034
B)
Smp_scaff000060
710k
711k
712k
713k
714k
715k
716k
717k
718k
719k
mRNA
Smp_137730
predicted by EVM2
CDS
Smp_137730
Smp_137730
Repeats
unknown_1980
R=52
unknown_1981
R=1701
unknown_1982
R=963
unknown_1983
R=118
unknown_1984
R=118
unknown_1985
(TAA)n
unknown_1986
R=296
unknown_1987
R=1912
unknown_1988
R=659
unknown_1989
R=750
unknown_1990
R=7
unknown_1991
R=614
unknown_1992
R=433
unknown_1993
R=908
unknown_1994
R=274
unknown_1995
R=29
unknown_1996
R=6
unknown_1997
R=79
unknown_1998
R=528
unknown_1999
R=1756
unknown_200
Sm_alphaT1
unknown_200
Sm_alphaD3
unknown_2
R=1899
unknown_
R=476
new repeat database
smr_2344
R=963
smr_2344(TTA)n
R=52
smr_1138
R=118
smr_2708
smr_2708
R=118
R=118
smr_1214
smr_4776
R=7
smr_6349
smr_6926
R=659
smr_4449
R=908
smr_3319
smr_5412
smr_2929
smr_2058
smr_6727
R=513
R=79
smr_3183
R=1899
R=528
Sm_alphaD3
C)
Smp_scaff015349
0k
1k
2k
3k
mRNA
Smp_117130
predicted by EVM2
Smp_190410
predicted by EVM2
Smp_117150
predicted by EVM2
Smp_117160
predicted by EVM2
Repeats
new repeat database
04665.1|SCMDNAP
smr_315
smr_315
smr_315
smr_315
smr_315
smr_315
smr_315
04665.1|SCMDNAP
04665.1|SCMDNAP
04665.1|SCMDNAP
04665.1|SCMDNAP
04665.1|SCMDNAP
04665.1|SCMDNA
04665.1|SCMDNAP
Smp_scaff000031
487k
488k
489k
490k
491k
D)
mRNA
CDS
Repeats
unknown_1624
R=70
unknown_1625
R=2049
unknown_1626
R=358
unknown_1627
R=909
unknown_1630
R=256
unknown_1628
R=826
unknown_1629
R=940
unknown_1631
AT_rich
unknown_1632
R=989
new repeat database
smr_2729
R=70
R=358
smr_6514
R=2049
smr_5495
smr_1989
smr_2503
smr_7481
R=940
smr_1624
smr_1799
smr_5121
smr_5887
smr_4439
smr_4439
smr_5323
smr_2451
smr_6600
smr_42
R=989
smr_3267

Schistosoma mansoni genomic DNA

38,004,342 Illumina reads

alignment (SOAP)

uniques

non-uniques

12,535,613 Illumina reads

assembly (Velvet+Sequencher)

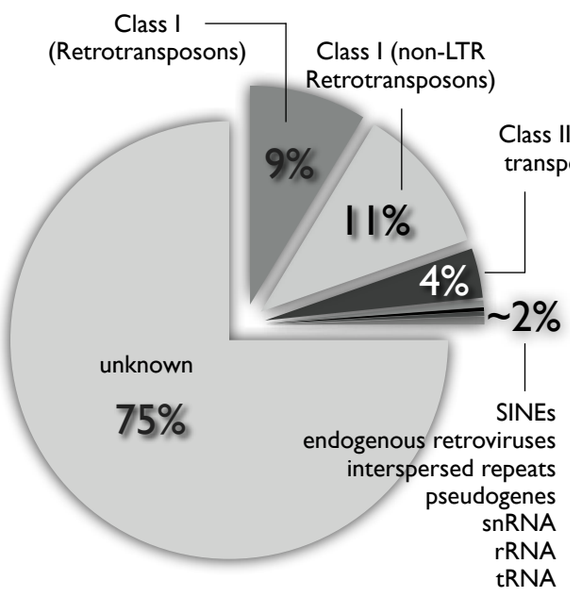
8594 contigs

Blast against *S.mansoni* genome and preliminary analysis

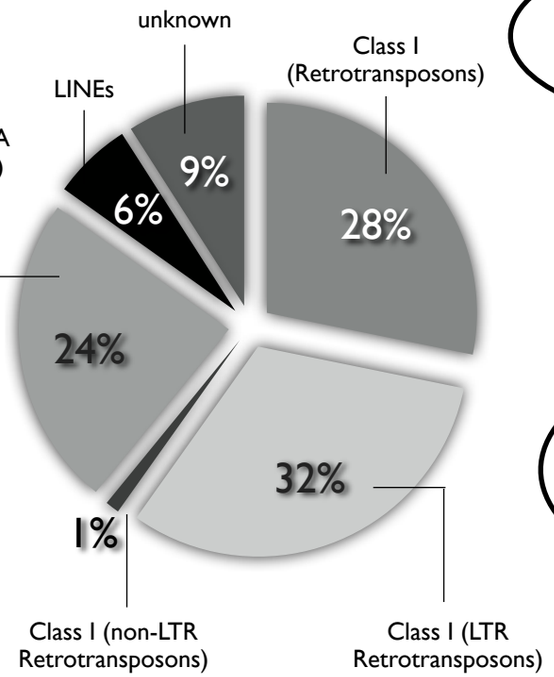
no match, single match, putative genes, mitochondrial genes ...

6531 contigs

Blast against GenBank Censor



TEClass



TRfinder

516 tandem repeats

Blast against trace files

14 polymorphic tandem repeats with single occurrence in the genome

1225 cleaned RepeatScout predictions

52 known repeats

4143 new *de-novo* assembled repeats

5420 repeats on genome assembly 3.1

3145 repeats on genome assembly 5.2