

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

► To cite this version:

Julie Mireille Lepesant, David Roquis, Rémi Emans, Céline Cosseau, Nathalie Arancibia, et al.. Combination of de-novo assembly of massive sequencing reads with classical repeat prediction improves identification of repetitive sequences in *Schistosoma mansoni*. *Experimental Parasitology*, Elsevier, 2012, 130 (4), pp.470-474. 10.1016/j.exppara.2012.02.010 . halsde-00674586

HAL Id: halsde-00674586

<https://hal.archives-ouvertes.fr/halsde-00674586>

Submitted on 27 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Combination of *de-novo* assembly of massive sequencing reads with**
2 **classical repeat prediction improves identification of repetitive**
3 **sequences in *Schistosoma mansoni***
4

5 Julie M.J. Lepesant ^{ab}, David Roquis ^{ab}, Rémi Emans ^{ab}, Céline Cosseau ^{ab}, Nathalie Arancibia ^{ab},
6 Guillaume Mitta ^{ab} and Christoph Grunau ^{ab*}

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

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

9

10 *) corresponding author :

11 Christoph Grunau, christoph.grunau@univ-perp.fr, Université de Perpignan Via Domitia, UMR
12 5244 CNRS Ecologie et Evolution des Interactions (2EI), 52 Avenue Paul Alduy, 66860 Perpignan
13 Cedex, France

14 Tel : +33 468662180

15 Fax : +33 468662281

16

17

18 **Abstract**

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

32

33

34 **Keywords**

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

36 Despite their abundance, repetitive sequences of the genome are often considered as “junk”,
37 “selfish”, or “parasitic” DNA that is tolerated by the genome but has no biological or evolutionary
38 functions. This view is about to change. In 2005, Shapiro and von Stenberg discussed the
39 importance of the repetitive sequences for the establishment of the frontiers between
40 heterochromatin and euchromatin, and their influence on homologous and nonhomologous
41 recombination (Shapiro and von Sternberg, 2005). Depending on their position, repetitive
42 sequences can play a part in activation or repression of gene transcription (Goodier and Kazazian,
43 2008). Some repeats have important structural functions such as telomeric repeats or the long
44 satellite blocks that make up the centromeres of mammals and insects (Kejnovsky et al., 2009). In
45 some cases, transcription of repeats and subsequent processing into small RNA was described.
46 These transcripts are involved in heterochromatization (small heterochromatin inducing RNA -
47 shiRNA) (Reinhart and Bartel, 2002). Transposable elements, constituting a substantial share of the
48 repetitive DNA, are known to have an impact on the genome evolution. Some were even selected to
49 play a precise role in the cell (“domesticated repeats”) (Shapiro and von Sternberg, 2005). Taken
50 together, repetitive elements can no longer be considered as a side-aspect of the genome and
51 deserve a deeper investigation.

52 *Schistosoma mansoni* is a parasitic plathyhelminth responsible for intestinal schistosomiasis. This
53 parasitic human disease ranks second only to malaria in terms of parasite-induced human morbidity
54 and mortality, with more than 200 million infected people. The economic burden caused by the
55 disease is tremendous as, for example, people disabled by the disease have limited job
56 performances and are less likely to contribute to the local development. It was estimated that
57 schistosomiasis burden represents 25 - 50 million disability-adjusted life-years (DALY) (King,
58 2010). The life cycle of the parasite is characterized by passage through two obligatory hosts: a
59 fresh-water snail (*Biomphalaria* species, depending on the geographical location) as intermediate
60 host, and humans or rodents as the final host. Miracidia infect the snail and transform into primary
61 and secondary sporocysts, from which cercariae, capable of infecting the human host, are released

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

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

88 DNA-free (Ambion #cat: AM1907) for 45 minutes at 37°C, followed by inactivation of the enzyme
89 using the inactivation reagent. PCR of 28s rDNA was used to test for genomic DNA
90 contaminations. First strand cDNA was synthesized using 20 ng of the total RNA preparation, in a
91 final volume of 20µl with 200 U of RevertAid (Fermentas, #cat: G2101). To assemble the repeat
92 genome, we used a total of 38,004,342 36-bp single-end reads generated on a Genome Analyzer II
93 (Illumina) according to the manufacturer's protocol, at the MGX and Oregon State University
94 sequencing facilities. Sequences are available at the NCBI sequence read archive (study accession
95 numbers SRA012151.6 and SRA043796.1). Reads were aligned to the reference genome v.3.1 with
96 SOAP2/SOAPaligner (Li et al., 2009) evoking the -r 0 (not repeats) and -u (write unmapped reads
97 into a file) options. The rationale behind this approach was that in this case, SOAP would only align
98 reads with a single occurrence in the genome. All other reads correspond either to unknown
99 sequences or occurring more than once, *i.e.* are repetitive. The 12,535,613 unmapped sequence
100 reads (33% of total) were then assembled using Velvet 0.7.01 (Zerbino, 2010) with the -cov_cutoff
101 4 -min_contig_length 100 options resulting in 8,608 contigs. A long read assembler (Sequencher
102 version 4.5 (Gene Codes) min match=93%, min overlap=60 bp) was used to produce finally 8,594
103 contigs. Each repeat was assembled individually and therefore the assemblies may be composed of
104 two or more distinct, but very similar repeats. First pass annotation of the 8,594 presumed repeat
105 contigs was done with Blast (Altschul et al., 1990), Censor/Repbase (Kohany et al., 2006), TEclass
106 (Abrusan et al., 2009) and Tandem Repeats Finder (Benson, 1999). Blast2GO v2.4.8 (Conesa et al.,
107 2005) was used to carry out various types of BLAST searches (conditions in Supplementary Table 1
108 and results in Table 1). CENSOR (<http://www.girinst.org/censor/>) (Kohany et al., 2006) and the
109 Repbase Update (<http://www.girinst.org/repbase/>) (Jurka et al., 2005) were applied to find
110 sequences sharing similarity to known repeats. Parameters were: *Sequence source*: all; *Forced*
111 *translated search*: no; *Search for identity*: no; *Mask simple repeats*: yes; *Mask pseudogenes*: yes.
112 Results were evaluated according to the 80/80/80 principle (80% of identity on 80% of the sequence
113 spanning a minimum of 80 bp) (Wicker et al., 2007). The web application TEclass

114 (<http://www.compgen.uni-muenster.de/tools/teclass/>) (Abrusan et al., 2009) was used to predict
115 potential transposable elements in a sequence with default parameters. Finally we used Tandem
116 Repeats Finder (TRfinder, <http://tandem.bu.edu/trf/trf.html>) (Benson, 1999) to identify tandem
117 repeats. Parameters were optimized for highest sensitivity and specificity (respectively 81% and
118 97%) using *in silico* generated training sequences: *Minimum alignment score: 30; alignment*
119 *parameters: 2-7-7*. TRfinder was also used to find sequences with a period of 2, 3 or 4 bp that could
120 serve as new microsatellite markers. Candidates were verified to have only one occurrence in the
121 genome, to be polymorphic (by comparison with trace files used for the genome assembly) and it
122 was checked if they were located in a gene.

123 For confirmation of *in-silico* results, PCRs were carried out in a final volume of 25 μ L containing
124 0.2 μ mol of each oligonucleotide primer (Supplementary Table 2), 0.2 mmol of each dNTP
125 (Promega), 1.25 U of GoTaq polymerase (Promega, #cat: M3175) used with the recommended
126 buffer and completed to the final volume with DNase-free water (95 °C for 10 min followed by 35
127 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).
128 The PCR products were separated by electrophoresis through a 2% TBE agarose gel. Real-time
129 quantitative PCR analyses were performed using the LightCycler 2.0 system (Roche Applied
130 Science) and LightCycler Fast-start DNA Master SYBR Green I kit (Roche Applied Science) with
131 2.5 μ l of cDNA in a final volume of 10 μ l (3 mM MgCl₂, 0.5 μ M of each primer, 5 μ l of master
132 mix). The primers were designed with the LightCycler Probe design software or the Primer3Plus
133 web based interface (<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>). The
134 following protocol was used: 95 °C for 10 minutes; 40 cycles: 95 °C for 10 seconds, 60 °C for 5
135 seconds, 72°C for 16 seconds; melting curve, 60-95 °C with a heating rate of 0.1 °C/s and
136 continuous fluorescence measurement, and a cooling step to 40 °C. For each reaction, the crossing
137 point (Ct) was determined with the “second derivative method” of the LightCycler Software 3.3.
138 PCR reactions were done in quadruplicate and the mean value of Ct was calculated. 28s rRNA was
139 used as an internal control and the amplification of a unique band was verified by electrophoresis

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

142 Our assembly of SOAP-sorted massive sequencing reads delivered 8,594 contigs. Contigs
143 smr_2181, smr_2685, smr_2733, smr_3000, smr_3595, smr_3826, smr_4022, and smr_6227 – that
144 we tested by PCR - showed a band at the predicted size indicating that assembly is correct in most
145 cases. For two contigs (smr_3000 and smr_3826) a supplementary band was present, with a
146 molecular weight twice as high as the major band, suggesting repetition in that these fragments
147 correspond to an amplicon of 2 tandem repeats in the genome. Among the 8,594, we clearly
148 identified 6,531 (76%) as repeats via *in silico* analysis using the following criteria: two or more
149 occurrences in the reference genome and no Blast annotation (against the nr database from NCBI)
150 related to a known gene or protein (with the exception of proteins typical of transposable elements,
151 such as transposase, reverse transcriptase or GAG polyprotein), using an e-value cut-off of $1.0E^{-30}$.

152 Using the information obtained from the Blast done on the nr database, we identified 306 contigs
153 related to genes or gene families, which include 40 mitochondrial genes. A Blast search against the
154 reference genome of *S. mansoni*, which allowed us to count the number of occurrences of each
155 repeat, showed that 1,332 sequences (other than the ones identified as genes) were unique, and 230
156 were absent from the reference genome. All Blast conditions are listed in Supplementary Table 1.

157 At this point of our analysis, a total of 10 sequences are suspected to be possible contamination, as
158 they are neither present in the reference genome, nor in the trace files, but match perfectly
159 sequences from other organisms. Four sequences correspond to rodents and six to small freshwater
160 organisms such as *Hydra magnipapillata*. It is likely that the sequences are due to contaminations
161 from host tissue and spring-water used for parasite culture. Out of the 6,531 repeated sequences,
162 TRfinder detected 516 containing tandem repeats. Two thousand four hundred and forty five
163 repeated sequences matched fully or partially to already known repeats, *i.e.* those present in the
164 RepeatScout generated database of repeat predictions. In conclusion, assembly of repeats is possible
165 by our new approach and so far, 4,143 unknown repeats have been identified. Based on the

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

192 original Sanger/TIGR library and with our combined library. Repeatmasker identified repeats in
193 47.40% of the genome. Fragmentation and overlapping repeatmasking were considerably lower
194 with our new library resulting in annotation of 623,983 repeats compared to 881,451 obtained with
195 the previous Sanger/TIGR data. Examples that show the improved quality of annotation are shown
196 in supplementary figure 1. During the preparation of the manuscript, version 5.2 of the *S. mansoni*
197 genome became available in which scaffold redundancy was reduced (Protasio et al., 2012). We
198 reasoned that some of our new repeats might be absent from the new assembly and repeated the
199 blast search with an e-value of 1e-30. We removed all repeats that did not match to the genome 5.2
200 or that matched only once resulting in 3,145 different repeat sequences. Repeatmasking showed that
201 47.73% is repetitive, *i.e.* essentially the same value as for the 3.1 assembly. Repeat annotations
202 were converted into GFF and loaded on an in-house genome browser for visualization. The repeats
203 are available as fasta files and annotations as GFF files (<http://methdb.univ-perp.fr/downloads/>).

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

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

228

229 **Acknowledgements**

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

237

238

239

References

- 240
241
242 Abrusan, G, Grundmann, N, DeMester, L, Makalowski, W, 2009. TEclass--a tool for automated
243 classification of unknown eukaryotic transposable elements. *Bioinformatics* 25, 1329-1330.
- 244 Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ, 1990. Basic local alignment search tool.
245 *Journal of Molecular Biology* 215, 403-410.
- 246 Benson, G, 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
247 *Research* 27, 573-580.
- 248 Berriman, M, et al., 2009. The genome of the blood fluke *Schistosoma mansoni*. *Nature* 460, 352-
249 358.
- 250 Conesa, A, Gotz, S, Garcia-Gomez, JM, Terol, J, Talon, M, Robles, M, 2005. Blast2GO: a universal
251 tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*
252 21, 3674-3676.
- 253 Copeland, CS, Brindley, PJ, Heyers, O, Michael, SF, Johnston, DA, Williams, DL, Ivens,
254 AC, Kalinna, BH, 2003. Boudicca, a retrovirus-like long terminal repeat retrotransposon from
255 the genome of the human blood fluke *Schistosoma mansoni*. *Journal of Virology* 77, 6153-
256 6166.
- 257 Copeland, CS, Lewis, FA, Brindley, PJ, 2006. Identification of the Boudicca and Sinbad
258 retrotransposons in the genome of the human blood fluke *Schistosoma haematobium*. *Mem*
259 *Inst Oswaldo Cruz* 101, 565-571.
- 260 Goodier, JL, Kazazian, HH Jr, 2008. Retrotransposons revisited: the restraint and rehabilitation of
261 parasites. *Cell* 135, 23-35.
- 262 Grunau, C, Boissier, J, 2010. No evidence for lateral gene transfer between salmonids and
263 schistosomes. *Nature Genetics* 42, 918-919.
- 264 Jurka, J, Kapitonov, VV, Pavlicek, A, Klonowski, P, Kohany, O, Walichiewicz, J, 2005. Repbase
265 Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110, 462-467.
- 266 Kejnovsky, E, Hobza, R, Cermak, T, Kubat, Z, Vyskot, B, 2009. The role of repetitive DNA in
267 structure and evolution of sex chromosomes in plants. *Heredity* 102, 533-541.
- 268 King, CH, 2010. Parasites and poverty: the case of schistosomiasis. *Acta Trop* 113, 95-104.
- 269 Kohany, O, Gentles, AJ, Hankus, L, Jurka, J, 2006. Annotation, submission and screening of
270 repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7, 474.
- 271 Li, R, Yu, C, Li, Y, Lam, TW, Yiu, SM, Kristiansen, K, Wang, J, 2009. SOAP2: an improved
272 ultrafast tool for short read alignment. *Bioinformatics* 25, 1966-1967.
- 273 Protasio, AV, et al., 2012. A Systematically Improved High Quality Genome and Transcriptome of
274 the Human Blood Fluke *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases* 6, e1455.

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

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

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

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

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

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

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

289

290 **Figure legends**

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

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

296 Suppl. Figure 1:

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

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

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

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

306

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 <i>S. mansoni</i> , 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 <i>S. mansoni</i> http://www.ncbi.nlm.nih.gov/ | N | 1.00E-15 |
| Reference genome <i>S. japonicum</i> , strain Anhui http://lifecenter.sgst.cn/ | N | 1.00E-20 |
| ESTs <i>S. japonicum</i> 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 | CATTTCTCTCCTGCAGTCCT | GATCTTACGTGAAAATCGCATC | 340 |
| smr_3000 | TTGTCTTATCATTTCTCAGTTGCTTC | TAGAACTTTTATCAGGGAACGTATTC A | 539 |
| smr_3826 | ATATGGTGTTGTAGCGTGTGC | TTATCCCTTCAATCTCTATTA AAAACA | 579 |
| smr_3595 | CCCTTTAGGGCCTCTCTTAGC | AAATTGGTAAATACAATGGGATGTG | 709 |
| smr_4022 | CGTTTGCTATTAGTGTTAGGAATCTT | GTGATAACGGAACTAGGATGGTC | 713 |
| smr_2733 | ATCAACTATACGATTCCCTAAAACA | TCACTCGATAATTCGCTTAGCC | 393 |
| smr_6227 | CATAATCATAGGGGATAAATGTGAA | ATTAAACCACGTTTGTAATGAATTG | 243 |
| smr_7590 | GATTTCAATTATCAGGAAAACA T | CTTATAAATCGCTTTGAAA ACTCTG | 241 |

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

| Name | Forward | Reverse | Size |
|-------------|----------------------------------|---------------------------|-------------|
| smr_2181 | AGGGTCGGTTGGTTGTCCTCTGG | ACGCTTTTGGCATTGATTTTGC GA | 101 |
| smr_2685 | CACATGGTCTCCGGAGTCAAGCA | TGAGCGCGTCTGCAAACTGA | 125 |
| smr_3000 | TTGTCTTATCATTCTCAGTTGCTT | 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 | CATGTTTTATAAGAAAGATTACATCAG C | AAACTCTGTAAACGTTGTCTCA | 154 |

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

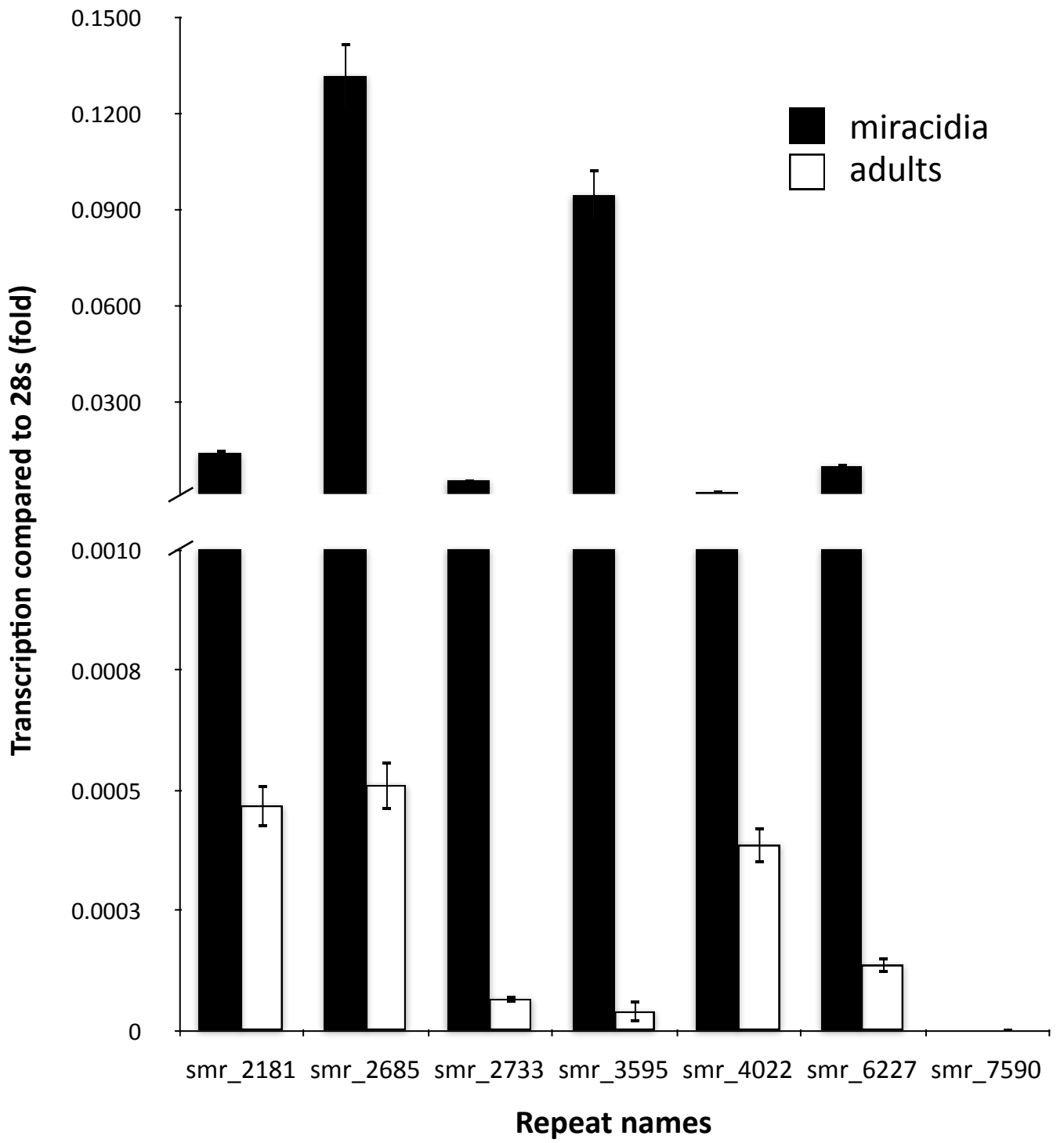
| Repeat name | Genome position | Gene in vicinity |
|--------------------|----------------------------------|-------------------------|
| smr_93 | Smp_scaff000078:1883819..1884094 | |
| smr_534 | Smp_scaff008778:843..259 | |
| smr_1715 | Smp_scaff001981:160389..160587 | |
| smr_2501 | Smp_scaff006696:1005..1124 | |
| 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 |

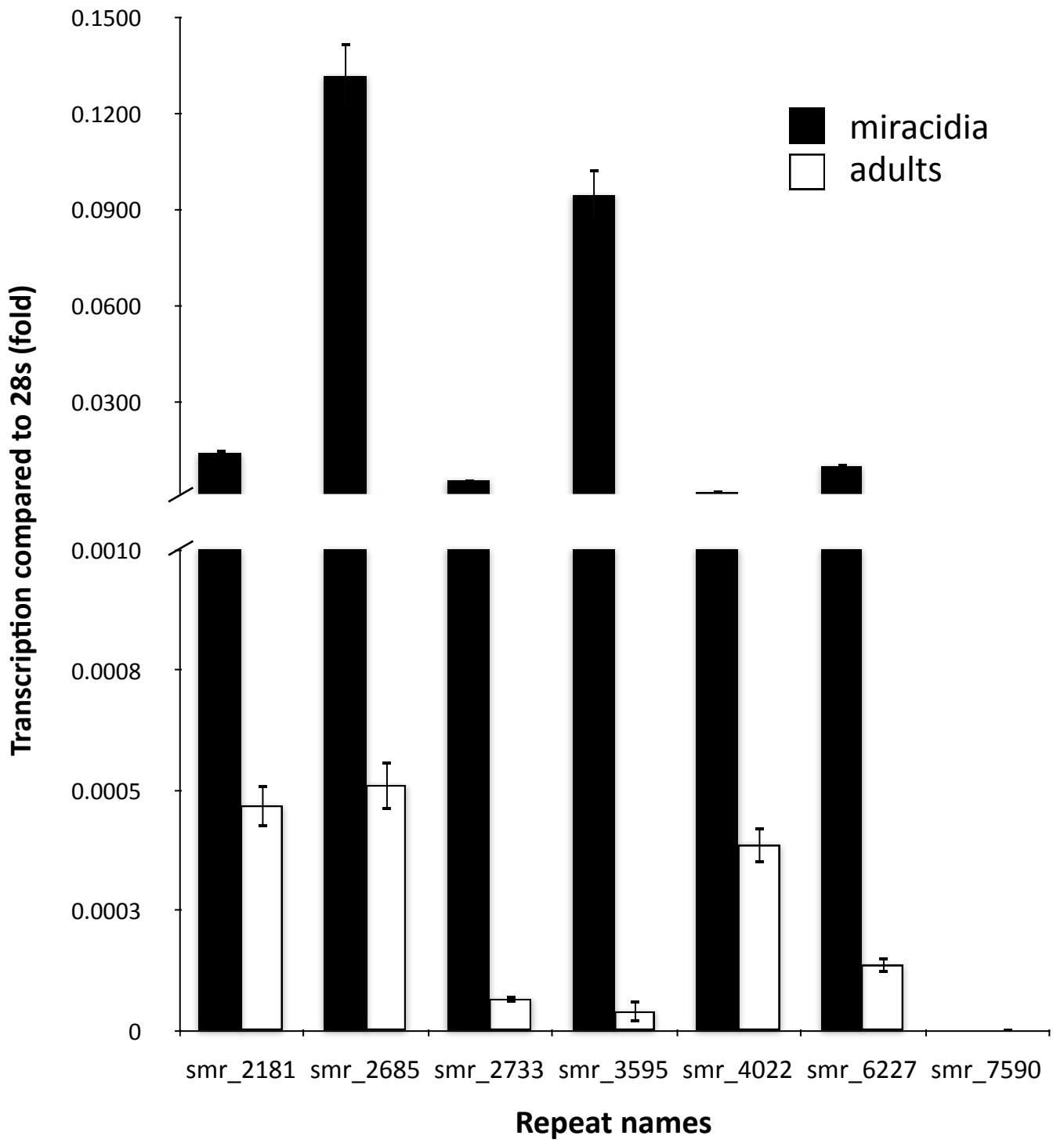
Supplementary Table 5 : Currently used microsatellite markers and corresponding PCR primers

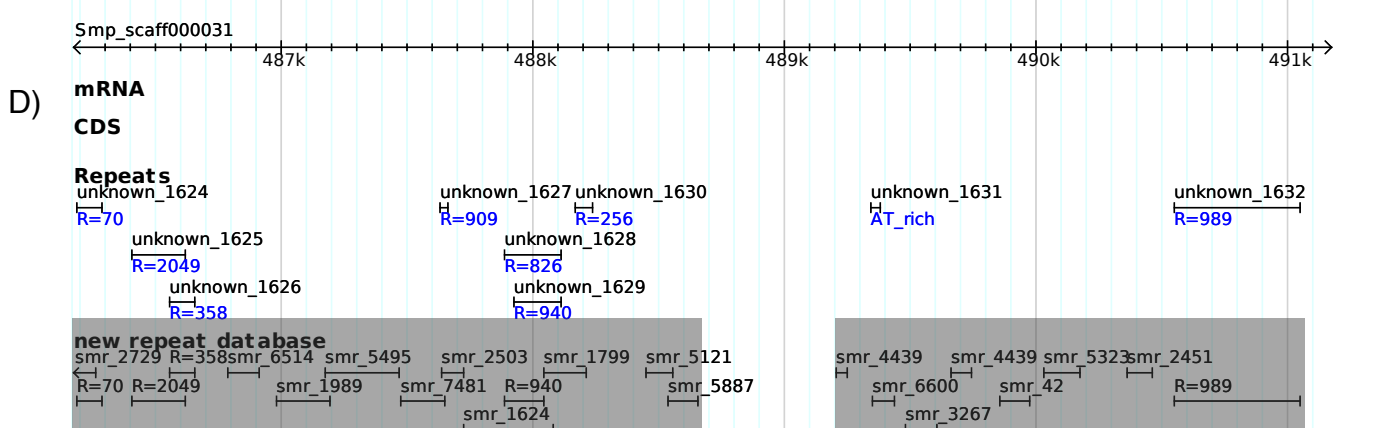
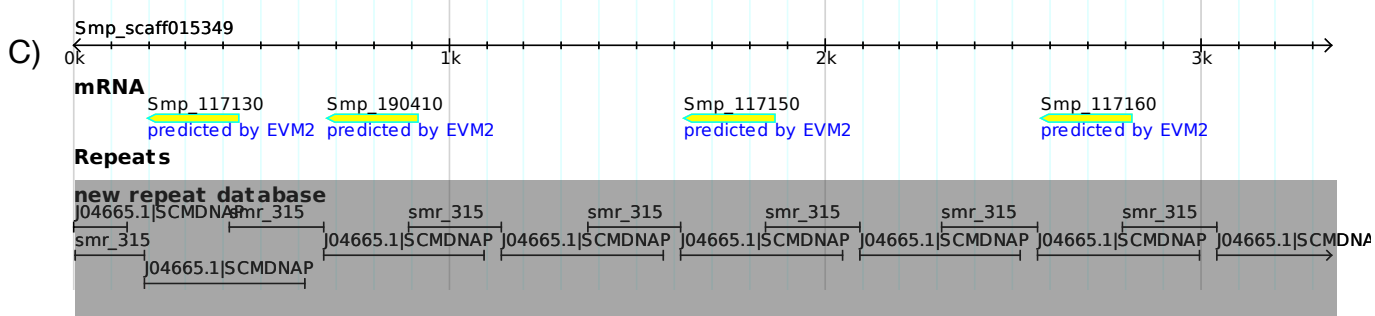
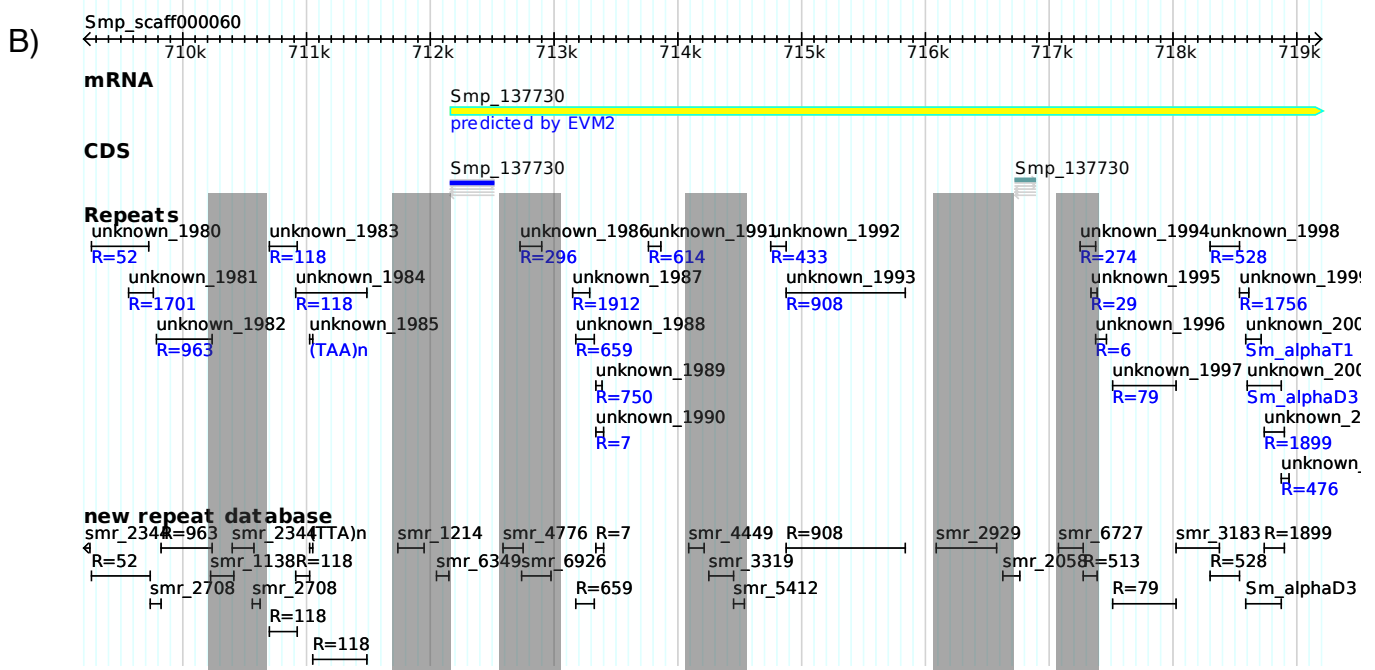
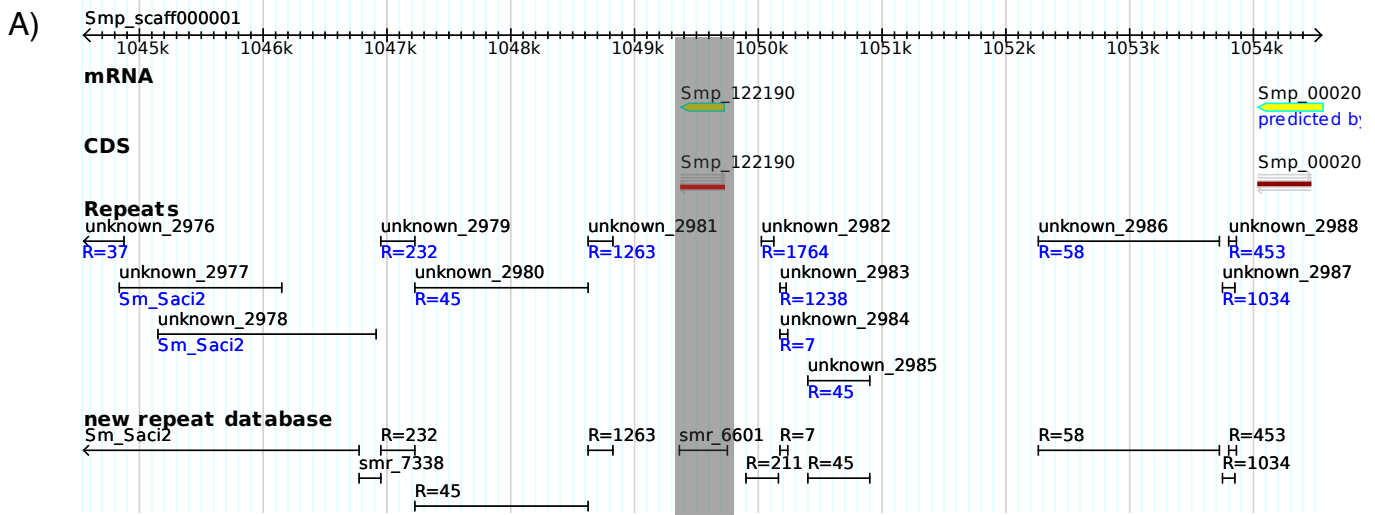
| Locus | Access. number | F | R | Ref. |
|--------|----------------|--|---------------------------------------|---|
| SmBr1 | L81235 | CGGAACGACAA GAAAATCAT | GAGTATACGGC TTCTTGGA | Rodrigues,N.B., Coura Filho,P., de Souza,C.P., Jannoti Passos,L.K., Dias- Neto,E. and Romanha,A.J. (2002) Populational structure of Schistosoma mansoni assessed by DNA microsatellites. Int. J. Parasitol. 32: 843-851. |
| SmBr2 | L26968 | ATCTCAAAGCC CAATACAAC | CATTTTCACTT ACTGTTTATCC | |
| SmBr3 | X77211 | CATTTATTGAT AATCTTTGGC | TTAACTACTTC TCACTGATG | |
| SmBr4 | M15371 | CAATTGACTAT TGAAAAGGC | CGTAAAGACCA CGATAAG | |
| SmBr5 | L25065 | AAACTATTCAT TACTGTCGGG | GAATTACTGTC CCTTTATCTC | |
| SmBr6 | AF009659 | GAATACAGGCT ATAATCTACA | CTTAACAGACA TACACGC | |
| SmBr7 | DQ137434 | CGTCATCACCT TAAACATGAAC | AATCACCAATG GCAACAATCTG | 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 | TAGGACAGGTT TTCCACCAA | ATGCCACACA CAAAGTAAA | |
| SmBr9 | DQ137431 | ATTCACCCATT GTCTTAAAACC | ATTGGCGTCAG TAGAAGAGATT | |
| SmBr10 | DQ448293 | CATGATCTTAG CTCAGAGAGC | GTACATTTTAT GTCAGTTAGCC | |
| SmBr11 | AC112150.4 | AAGAAGTGGA GGAGGCCTTT | TTCAGTCCCTG GAACACACA | |
| SmBr12 | DQ137724 | TATATAGCAAA AGTAGTCTATA TTCGTAGC | AGTAAAAACTA TCCTATCCATTT CTATTG | |
| SmBr13 | DQ137790 | GTCACAGATAC CTGACGAGCTG | ACTCCCCAGCA ATTTGTCC | |
| SmBr14 | DQ514536 | CTGCTCATCAT AGAAGTGTGGC | TCTATGTATCT ACCCACCCTA TC | |
| SmBr15 | AF325695 | TATAGGACAAA ACGCGGGTC | TTGGATAAACT TAGTGACTTTT C | |
| SmBr16 | L04480 | TGTGACTTTGA TGCCACTGA | GGCCTGATACA ATTCTCCGA | |
| SmBr17 | AQ841039 | CTGCAGGGGGA AATAGAAG | TGATCCTTTGT GCCAACA | |

| 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 <i>Schistosoma</i> <i>mansoni</i> and their utility for other schistosome species <i>Molecular</i> <i>Ecology Notes</i> 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 <i>Schistosoma</i> <i>mansoni</i> from Africa <i>Molecular</i> <i>Ecology Notes</i> 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 | TTCAAACCAT 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 | CATCACCATCA ATCACTC | TATTCACAGTA GTAGGCG | (1995) Isolation of microsatellite markers in the digenetic trematode <i>Schistosoma mansoni</i> from Guadeloupe island <i>Journal of Fish Biology</i> 47: 29-55. |
| SMD57 | AF202967 | TCCTTGATTCC ACTGTTG | GCAGTAATCCG AAAGATTAG | |
| SMD89 | AF202968 | AGACTACTTTC ATAGCCC | TTAAACCGAAG CGAGAAG | |
| SMD94 | AF202969 | TAAACTCACA CATAACC | AACTAATCACC CACTCTAC | |
| AI068335 | AI068335 | GTTGAGAGAGA AAAAGAAG | AGATGTTAGAA AGTGGTG | |
| L46951 | L46951 | CAAACATATAC ATTGAATACAG | TGAATTGATGA ATGATTGAAG | |
| SCMSMOXII | M85305 | TTCTACAATAA TACCATCAAC | TTTTTTCTCACT CATATACAC | |
| R95529 | R95529 | GTGATTGGGGT GATAAAG | CATGTTTCTTC AGTGTCC | |
| SMU31768 | U31768 | TACAACTTCCA TCACTTC | CCATAAGAAAG AAACCAC | |
| SMIMP25 | X77211 | CACTATACCTA CTACTAATC | TCGATATACAT TGGGAAG | |







Schistosoma mansoni genomic DNA

