Experimental Parasitology 130 (2012) 470–474

Contents lists available at SciVerse ScienceDirect

Experimental Parasitology

journal homepage: www.elsevier.com/locate/yexpr

Research Brief

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, David Roquis, Rémi Emans, Céline Cosseau, Nathalie Arancibia, Guillaume Mitta, Christoph Grunau *

Université de Perpignan Via Domitia, Perpignan F-66860, France
CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan F-66860, France

ARTICLE INFO

Article history:
Received 24 November 2011
Received in revised form 7 February 2012
Accepted 8 February 2012
Available online 21 February 2012

Keywords:
Schistosoma mansoni
Repetitive sequences
Massive sequencing
de novo Assembly

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 *S. 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/).

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 platyhelminth 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

* Corresponding author. Address: Université de Perpignan Via Domitia, UMR 5244 CNRS Ecologie et Evolution des Interactions (2EI), 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France. Fax: +33 468662281.
E-mail addresses: grunau@methdb.net, christoph.grunau@univ-perp.fr (C. Grunau).

0014-4894/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.exppara.2012.02.010

biology methods e.g. (Copeland et al., 2003, 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 500 μl Trizol with a MagNA Lyser and Green beads (Roche). RNA was treated with DNA-free (Ambion #cat: AM1907) for 45 min 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 20 μl 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 min; 40 cycles: 95 °C for 10 s, 60 °C for 5 s, 72 °C for 16 s; 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 8594 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 8594, 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