NATURE GENETICS | LETTER OPEN

日本語要約

Whole-genome sequence of *Schistosoma haematobium*

Neil D Young, Aaron R Jex, Bo Li, Shiping Liu, Linfeng Yang, Zijun Xiong, Yingrui Li, Cinzia Cantacessi, Ross S Hall, Xun Xu, Fangyuan Chen, Xuan Wu, Adhemar Zerlotini, Guilherme Oliveira, Andreas Hofmann, Guojie Zhang, Xiaodong Fang, Yi Kang, Bronwyn E Campbell, Alex Loukas, Shoba Ranganathan, David Rollinson, Gabriel Rinaldi, Paul J Brindley, Huanming Yang *et al.*

Nature Genetics **44**, 221–225 (2012) doi:10.1038/ng.1065
Received 28 April 2011 Accepted 07 December 2011 Published online 15 January 2012

Schistosomiasis is a neglected tropical disease caused by blood flukes (genus *Schistosoma*; schistosomes) and affecting 200 million people worldwide[1]. No vaccines are available, and treatment relies on one drug, praziquantel[2]. *Schistosoma haematobium* has come into the spotlight as a major cause of urogenital disease, as an agent linked to bladder cancer[1, 3] and as a predisposing factor for HIV/AIDS[4, 5]. The parasite is transmitted to humans from freshwater snails[1]. Worms dwell in blood vessels and release eggs that become embedded in the bladder wall to elicit chronic immune-mediated disease[6] and induce squamous cell carcinoma[7]. Here we sequenced the 385-Mb genome of *S. haematobium* using Illumina-based technology at 74-fold coverage and compared it to sequences from related parasites[8, 9]. We included genome annotation based on function, gene ontology, networking and pathway mapping. This genome now provides an unprecedented resource for many fundamental research areas and shows great promise for the design of new disease interventions.

Main

We sequenced the *S. haematobium* genome from 200 ng of genomic DNA template isolated from a single, mated pair (one male and one female) of adult worms, and produced 33.5 Gb of usable sequence data (Supplementary Note, Supplementary Tables 1–3, Supplementary Figs. 1 and 2). We consistently showed low sequence heterozygosity and estimated the genome size to be 431–452 Mb. We then assembled the data and used local assemblies to close most (96.1%) of the remaining gaps, achieving a final assembly of 385 Mb (365 contigs; N50 scaffold size of 307 kb) (Table 1 and Supplementary Note). The GC content (mean: 34.3%) was similar to that of *Schistosoma mansoni* and *Schistosoma japonicum* (Table 1). After assembly, all usable reads were realigned to scaffolds to assess single-base accuracy of the assembled genome sequence. We did not find evidence of GC-biased nonrandom sampling[10] or of artifacts induced by multiple displacement amplification, a result that is consistent with published information[11, 12] (Supplementary Note, Supplementary Figs. 3 and 4).

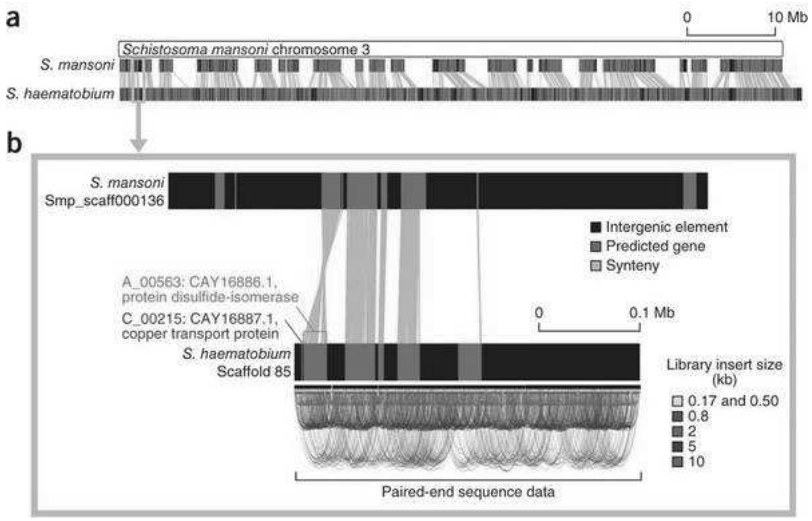
Table 1: Comparison of the *Schistosoma haematobium* genome with those of *S. mansoni* and *S. japonicum*

Comparison of the *S. haematobium* and *S. mansoni* genomes showed a similar percentage and composition of repetitive elements (Table 1, Supplementary Note, Supplementary Tables 4–6). Using both homology-based and *de novo* predictions, we estimated that 43% of the *S. haematobium* genome comprises repetitive elements, consistent with the *S. mansoni* genome (40%)[8]. More than half (58.5%) of the repeats were retrotransposons (at least 20 types, including LINE/RTE-BovB and LTR/*Gypsy*); 37% were unknown repeats, including satellites (1.9%), simple repeats (1.2%) and DNA transposons (five types; <1%). On the basis of homology, *de novo* predictions and evidence of transcription (in adult and egg stages), we inferred 13,073 protein-coding genes from the genome and included data for *S. mansoni* and *S. japonicum* for comparisons (Supplementary Note, Supplementary Table 7, Supplementary Figs. 5–7). The number of *S. haematobium* genes was consistent with those of *S. mansoni* (13,184) and *S. japonicum* (13,469), as were the gene structures. Most (9,714) *S. haematobium* genes were supported by the RNA-seq data from adult and egg stages. Comparative analyses of the complete gene set showed higher nucleotide sequence identity (mean 92%) and length match for individual coding domains between *S. haematobium* and *S. mansoni* than between *S. haematobium* and *S. japonicum* (nucleotide sequence identity 86%) or *S. mansoni* and *S. japonicum* (86%).

For the protein-coding genes (~4.4% of the *S. haematobium* genome), 96.3% had matches in nonredundant databases, 52.8% had conserved protein domains and 43% mapped to known biological pathways (Supplementary Note, Supplementary Tables 8, 9, 10, Supplementary Fig. 8). These data allowed 44% of genes to be classified by Gene Ontology (GO) terms, providing a list of terms that was consistent with *S. mansoni* and *S. japonicum*. A small percentage (2.6%) of *S. haematobium* genes was predicted to encode excretory-secretory (ES) proteins in the egg (Omega-1 and interleukin-4-inducing protein) and/or adult (including cathepsin B, heat-shock proteins, thioredoxin peroxidase, superoxide dismutase, protein disulfide isomerase and venom allergen–like proteins)[13, 14]. High-stringency genetic networking of the entire genomic data set identified major hubs of connectivity for conserved molecules associated with nucleotide and protein synthesis and degradation and with signal transduction (Supplementary Note, Supplementary Tables 11 and 12, Supplementary Fig. 9).

A genome-wide analysis revealed a significantly higher synteny between *S. haematobium* and *S. mansoni* (89.4%) than between *S. haematobium* and *S. japonicum* (51.7%) or *S. mansoni* and *S. japonicum* (67.0%). When compared to the *S. mansoni* genome, there were approximately four times more intrachromosomal rearrangements in *S. japonicum* than in *S. haematobium* (for a scaffold length of >1 Mb) (Supplementary Note, Supplementary Tables 13 and 14, Supplementary Fig. 10). These findings are consistent with present knowledge of schistosome evolutionary relationships[15] and karyotypes[16]. Given the close relationship between *S. haematobium* and *S. mansoni* and the size and quality of the draft genome for *S. mansoni*[8], we aligned *S. haematobium* to *S. mansoni* scaffolds that mapped to chromosomes (2*n* = 16; ZZ)[17]. Overall, rearrangements in *S. haematobium* with respect to *S. mansoni* were rare, with 11 inversions of syntenic blocks linked to five chromosomes (nos. 1, 3, 4, 6 and Z) (Fig. 1, Supplementary Note, Supplementary Fig. 11).

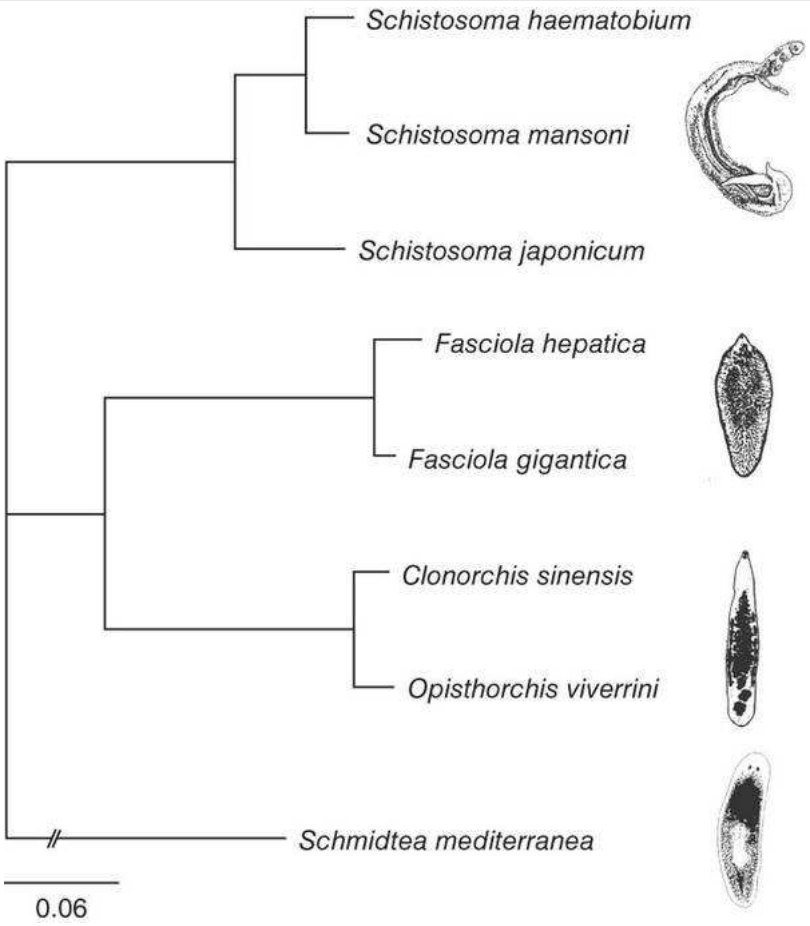
Figure 1: Synteny inferred between the *Schistosoma haematobium* and *Schistosoma mansoni* genomes.



(**a**) Large blocks of *S. haematobium* genomic scaffolds mapped physically to chromosome 3 of *S. mansoni*. (**b**) Evidence of an inversion in *S. haematobium* with reference to *S. mansoni*, supported by paired-end sequence data.

Of the proteins shared between *S. haematobium* and *S. mansoni* (*n* = 1,333) or *S. japonicum* (235), (Supplementary Note, Supplementary Table 15), only a minor portion could be assigned functional categories (using the KEGG BRITE database) linked to a wide array of different molecular groups. Of the 1,333 proteins common to *S. haematobium* and *S. mansoni*, 91 represented mainly enzymes (such as kinases, glycosyl-transferases and peptidases), cytoskeletal, DNA-repair, replication, recombination and spliceosome proteins and elements of the ubiquitin system. Of the 235 proteins common to *S. haematobium* and *S. japonicum*, 33 were linked to metabolic enzymes, cytoskeletal proteins and transcription factors or proteins in the ubiquitin complex. A subset of 73 molecules were unique to *S. haematobium* (Supplementary Note, Supplementary Table 16, Supplementary Fig. 12); although these molecules contained structural elements such as α-helices and β-sheets, none of them was similar to any presently known eukaryotic proteins or contained conserved motifs. Of the 10,880 proteins common among *S. haematobium*, *S. mansoni* and *S. japonicum*, we identified 6,142 homologs in other flatworms, including *Fasciola hepatica*, *Fasciola gigantica*, *Clonorchis sinensis* and *Opisthorchis viverrini*. Using concatenated protein sequence data inferred from a subset of 59 single-copy gene homologs (Supplementary Note, Supplementary Tables 17 and 18), we were able to provide a robust inference of the genetic relationships of socioeconomically important trematodes, in which *S. haematobium* and *S. mansoni* were most closely related, followed by *S. japonicum*, to the exclusion of other trematodes (Fig. 2). The relationship of the schistosomes was in accordance with previous studies using mitochondrial and/or nuclear DNA markers[15]. The present phylogenetic analysis extends our understanding of the evolution of key trematodes, and the approach used provides a sound basis for future, large-scale evolutionary analyses when extensive genomic and transcriptomic data sets become available for a wide range of flatworms.

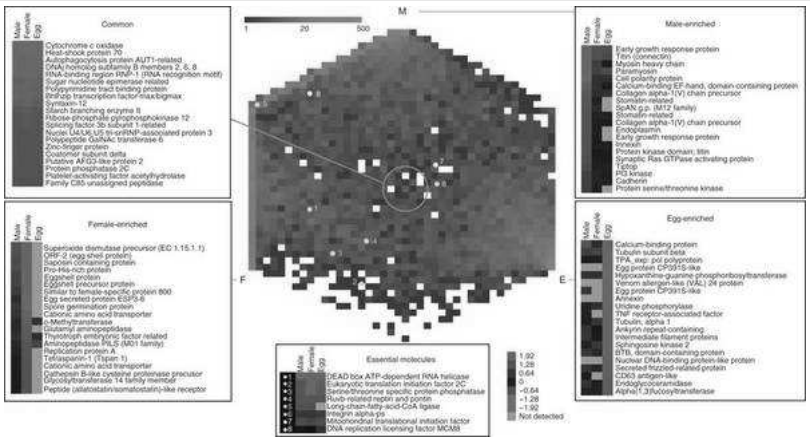
Figure 2: Genetic relationship of *Schistosoma haematobium* with other members of class Trematoda.



S. haematobium was related to trematodes (for which genomic or transcriptomic data are available) using Bayesian inference (BI) analysis of amino acid sequence data inferred from 59 single-copy genes. The topology of this tree is the same as that achieved using maximum parsimony (MP) analysis. Nodal support was absolute (100%/100%) for each branch. Branch lengths represent the numbers of amino acid substitutions per site in aligned postions.

Having assessed genetic relationships, we then proceeded to explore transcription in the adult and egg stages of *S. haematobium*, which cause disease in the human host (Fig. 3, Supplementary Note, Supplementary Tables 19,20,21,22). The number of transcribed genes was greater in adult females (10,002) and males (10,559) than in eggs (7,794). Egg-enriched transcription was observed for 1,631 genes, including those encoding venom allergen–like and CD63 tetraspanin–like proteins (reported to be transcribed exclusively in the miracidium of *S. mansoni*) [18] as well as two encoding CP391S-like egg proteins. Adult-enriched transcription was shown for 2,975 genes, of which 965 were significantly upregulated in females and 2,010 in males. These findings are similar to those of recent microarray studies of *S. mansoni* and *S. japonicum* [18, 19], although our interpretation is guarded, at this stage, as the animal hosts for parasite production and analytical methods differed among studies. Indeed, given the substantial depth of the present RNA-Seq data set (compared with microarray), we were able to accurately profile enriched biological pathways in the different stages and sexes of *S. haematobium* that clearly reflect its biology (Supplementary Note, Supplementary Fig. 13). Notable in the adult female, which lives within the gynecophoric canal of the male, is a major enrichment in pathways linked to hematophagia (including superoxide dismutase, saposin, cathepsin B and ferritin) and egg production (including lipid metabolism, protein synthesis and eggshell-specific proteins), whereas the pathways enriched in the adult male (pertaining largely to myosin, paramyosin, tropomyosin and troponin) relate to its strong need to maintain a highly developed musculature for motility, feeding, attachment and grasping the female *in copula*. In contrast, pathway enrichment in the egg stage relates predominantly to egg-specific proteins and a secreted glycoprotein that induces immunoglobulin-E–dependent production of interleukin-4 in basophils and is probably related to immune modulation or alteration in the host, as has been proposed for *S. mansoni* [20].

Figure 3: Differential transcription among adult female (F), adult male (M) and egg (E) stages of *Schistosoma haematobium*.



Transcripts (listed by protein name) were mapped to each gene in two dimensions and their relative abundance displayed. Products of constitutively transcribed genes are shown within a central 100-pixel radius, and products of the top 20 genes transcribed in a gender- or stage-enriched manner are shown within a 25-pixel radius of each node. Molecules inferred to be essential (1–8) are indicated, and those representing proposed candidate drug targets are in yellow. Transcription is expressed as $\log_{10}$-transformed reads per kilobase per million reads (RPKM).

Having explored transcription, we then constructed an interaction network for all genes inferred to be essential and transcribed constitutively or in a developmentally regulated manner in *S. haematobium*. We prioritized six molecules (Fig. 3 and Supplementary Note, Supplementary Table 23) as prime targets for the design of new trematocides. Although a small number of drug targets was predicted in *S. haematobium* using the very stringent selection criteria in our bioinformatic pipeline, all 72 candidate drug targets inferred previously for *S. mansoni*[21] were represented in the *S. haematobium* proteome.

Schistosomes have adapted to their mammalian hosts to such an extent that they can survive for decades in a host without succumbing. They achieve their longevity by suppression, diversion and alteration of immune responses[22]. Chronic infections induce key changes in immune-cell populations, including a dominance of the T helper type 2 (T_H2) cells and a selective loss of effector T-cell activity, against a background of regulatory T cells, alternatively activated macrophages, and T_H2-inducing dendritic cells[22]. Much of the immunomodulatory capacity of schistosomes is attributable particularly to ES products[23]. In the *S. haematobium* proteome, we identified 55 molecules (20 of which were predicted to be ES proteins) with known immunomodulatory roles in other helminths (Supplementary Note, Supplementary Table 24). These include molecules linked to inhibition of antigen processing and presentation via binding (Sjc23 tetraspanin), cleavage (cysteine and serine proteases) or inhibition of post-translational modification (cystatins) of host immunoglobulins; known inducers of T_H2 responses (IPSE-α-1, ω-1, peroxiredoxin and Sm16 (also called SmSLP or SmSPO-1)); and host-defense mimicry molecules (such as C-type lectins). Interestingly, we also identified a homolog of estradiol 17β dehydrogenase (Supplementary Note, Supplementary Table 24), which has a known role in the synthesis of estradiol. Intriguingly, *S. haematobium* ES products downregulate apoptosis and stimulate wound healing, mitosis and cell migration. All of these are expected to be conducive to tumorigenesis, in which one or more estradiol-like molecules have been implicated[24]. Even though homologs of 17β dehydrogenase exist in *S. mansoni* and *S. japonicum*, the specific spatial and temporal expression associated with the synthesis of estradiol-like molecules in *S. haematobium* eggs *in situ*, in the bladder, might contribute to carcinogenesis, warranting detailed exploration. Moreover, in spite of the limited proteomic differences among schistosomes of humans, substantial variation in splicing[25], differential methylation[26], regulatory RNAs[27] and other epigenetic processes is probable. These are areas that can now be tackled readily using genome information for these schistosomes.

Much remains unknown about the fundamental biology and pathogenesis of schistosomes, which cause considerable morbidity to many millions of people and animals worldwide[1, 28]. Given the challenges in propagating these parasites, particularly *S. haematobium*, in the laboratory[29], the ability to sequence the genome from a single pair of worms represents a important step in characterizing the genomes of a diverse range of other schistosomes and neglected tropical disease pathogens, including food-borne flukes, and toward addressing fundamental and controversial questions regarding their genetics, evolution, ecology, epidemiology, pathogenesis and host-parasite relationships. The genome provides a solid foundation for future large-scale and integrated studies of gene function and essentiality, using tools such as RNA interference and transgenesis[28], and will also facilitate urgently needed proteomic explorations. Published findings[30] show that developmental stages of *S. haematobium* can be manipulated genetically and that effective gene silencing can be achieved, which now provides enormous scope for future large-scale functional genomic analyses. Unlocking the molecular biology of this and related disease pathogens of global importance will offer new insights into schistosome development, host-parasite affiliations, disease and schistosomiasis-associated bladder cancer, and will underpin the design of new diagnostic tools, anti-schistosome drugs and vaccines.

URLs.

Schistosoma genome database, http://schistodb.net/; *Schistosoma mansoni* draft genome sequence v.3.1 and *S. haematobium* expressed sequence tag libraries, ftp://ftp.sanger.ac.uk/pub4/pathogens/Schistosoma/mansoni/; *Schistosoma japonicum* draft genome, http://www.chgc.sh.cn/japonicum/Resources.html; *Schistosoma mansoni* and *S. japonicum* gene sets, http://www.genedb.org/Homepage; Gephi, http://www.gephi.org/; LASTZ, http://www.bx.psu.edu/miller_lab/; Phylogenetic Analysis Using Parsimony program (PAUP), http://paup.csit.fsu.edu/.

Methods

Sample procurement, preparation and storage.

A laboratory strain of *S. haematobium*, originating from Egypt, was maintained in the Biomedical Research Institute, Rockville, Maryland[31] in *Bulinus truncatus* (intermediate snail host) and *Mesocricetus auratus* (hamster; mammalian definitive host). Hamsters were each infected with 1,000 cercariae. After 90 d, paired adults of *S. haematobium* were collected from *M. auratus*, following the perfusion of the mesenteric and intestinal vessels using physiological saline (37 °C). Worms were prepared as described previously[32] and snap frozen in liquid nitrogen. *S. haematobium* eggs were isolated from the livers from infected hamsters[33] and washed extensively in saline. All samples were frozen at −80 °C.

Genomic DNA library construction and sequencing.

Genomic DNA (~1.5 μg) was isolated from a single pair of adult worms (that is, male and female *in copula*) of *S. haematobium* using an established protocol[34], and 200 ng was subjected to whole-genome amplification (WGA) using the REPLI-g Midi Kit (Qiagen). Total DNA amounts were determined using a Qubit fluorometer dsDNA HS Kit (Invitrogen), and DNA integrity was verified by agarose gel electrophoresis. Short-insert (170 bp and 500 bp) and mate-pair (800 bp, 2 kb, 5 kb and 10 kb) genomic DNA libraries were constructed and paired-end sequenced on a Genome Analyzer II (Illumina). The sequence data from each library were verified, and low-quality sequences, base-calling duplicates and adaptors removed[35].

Assessment of heterozygosity and genome assembly.

Genome size and heterozygosity within or between the two adult worms of *S. haematobium* used for sequencing were estimated by establishing the frequency of occurrence of individual 17-bp *k*-mers within genomic sequence data for each small-insert library (170 bp, 500 bp and 800 bp) using a modification of the Lander Waterman algorithm[35]. Paired-end sequence data from the genomic DNA libraries were assembled using SOAPdenovo[36]. Short-insert, paired-end reads were used to construct a de Bruijn-graph employing a *k*-mer of 35 bp. All paired-end reads (from short-insert and mate-pair libraries) were then aligned to the contigs to construct scaffolds, with ≥3 read pairs required to form a connection. Assembly quality and completeness were assessed based on the minimum length of sequence contigs and scaffolds of >100 bp containing 50% and 90% of the sequence

data (N50 and N90, respectively).

Analyses of the assembled genome sequence, and genome sequence alignment

All usable sequences were re-aligned to contigs using SOAP2 (ref. 37), allowing for ≤5 mismatches per read. Mapped reads were used to estimate sequencing depth and GC content. Then, the frequencies of individual bases in the assembly were counted to estimate sequence coverage. We aligned, in a pairwise manner, the genome of *S. haematobium* with that of *S. mansoni* (draft genome v.3.1; see URLs) or *S. japonicum* (draft genome; see URLs) using the LASTZ program (release 1.02.00; see URLs), employing the repeat-masker setting, to identify clusters of unique alignments with a well-defined order and orientation. The *S. mansoni* scaffolds and their *S. haematobium* counterparts were then aligned against the *S. mansoni* chromosomes using the *S. mansoni* genomic linkage map [17].

Identification and annotation of genes.

In addition to the available adult and egg EST libraries (January 2011; see URLs), a full poly(A)-selected transcriptomic sequencing approach was applied to adult and egg stages of *S. haematobium* [38]. Briefly, between 5–20 μg of total RNA was extracted and used to purify polyadenylated RNA (separately) from adult males (n = 50), females (n = 50) or eggs (n > 1,000). Complementary DNA (cDNA) was synthesized, size selected (~200 bp), adaptor-ligated and then sequenced on a Genome Analyzer II (Illumina). To facilitate gene annotation, the combined RNA-seq data generated (separately) from the male, female and egg cDNA libraries for *S. haematobium* were assembled *de novo* using SOAPdenovo [36]. Expressed sequence tag (EST) data for *S. haematobium* were used to train and validate gene models generated from the draft genome, as described previously [39]. Parameters for the gene-prediction model were established using a set of 1,355 *S. haematobium* genes encoding complete open reading frames (ORFs) that were predicted from an unmasked draft of the *S. haematobium* genome sequence. Genes were predicted [39] and based on homology to the *S. mansoni* and *S. japonicum* gene sets (both v.4; see URLs). *De novo*–assembled EST data (from this study and available from NCBI) and raw, paired-end RNA-seq data were mapped to the genome and transcripts were predicted [39]. The predicted genes were merged to establish nonredundant gene sets [39]. Subsequently, we classified the predicted *S. haematobium* genes on the basis of experimental evidence of homology to genes of other eukaryotic organisms. The following codes were used to designate confidence in gene prediction: "A_", present in the Glean data set or supported by RNA-seq, *de novo* prediction and homology to a gene representing one or more other eukaryotes; "B_", supported by RNA-seq data and *de novo* prediction; "C_", homologous to a gene of one or more other eukaryotes; and "D_", supported by *de novo* prediction.

The *S. haematobium* gene set was then annotated on the basis of homology to sequences within public sequence, gene ontology and biological pathway databases using an established pipeline [39]. In addition, classical excretory-secretory (ES) proteins of *S. haematobium* were predicted (on the basis of the presence of a signal peptide at the N terminus) using SignalP v.3.0 (ref. 40; using both the neural network and hidden Markov models) and the absence of a transmembrane domain using TMHMM [41], and by BLASTp [42] homology-searching of the validated signal peptide database (SPD) [43] and an ES database containing published proteomic data for nematodes (*Brugia malayi* and *Meloidogyne incognita*) and trematodes (*S. mansoni*, *S. japonicum*, *O. viverrini* and *F. hepatica*) [13, 14, 44, 45, 46, 47, 48, 49, 50]. The secondary structure of genes specific to *S. haematobium* were predicted using PSIPRED [51]. For each orphan protein of *S. haematobium*, we attempted to search for homologs with known three-dimensional structures using pGenTHREADER (foldlib database; 17 June 2011) [52].

Nucleotide sequence identities in coding domains among *S. haematobium*, *S. japonicum* and *S. mansoni* were established, in a pairwise manner, using the program BLASTn [42], employing a permissive (E value ≤ 10^{-5}) search strategy. In addition, using the *S. mansoni* genome as the reference, a genome-wide analysis of synteny for one-to-one orthologs across scaffolds of >1 Mb was undertaken [53] to infer intrachromosomal rearrangements in *S. haematobium* and *S. japonicum*.

Sequence homology between/among proteins inferred for *S. haematobium*, *S. japonicum* and *S. mansoni* as well as other members of class Trematoda (including *Clonorchis sinensis* and *O. viverrini*, *F. gigantica* and *F. hepatica*) [32, 54, 55], for which transcriptomic data were available, was established using the program tBLASTx [42] employing permissive (E value ≤ 10^{-5}), moderate (≤10^{-15}) and stringent (≤10^{-30}) search strategies. A set of genes unique to *S. haematobium* was selected based on a lack of nucleotide sequence homology (BLASTn, E value ≤ 10^{-5}) to sequences in the *S. mansoni* and *S. japonicum* gene sets (January 2011; see URLs) or assembled genome scaffolds (*S. mansoni* and *S. japonicum* draft genomes; see URLs) and a lack of amino acid sequence homology (tBLASTx, E value ≤ 10^{-5}) to those of other selected trematodes (*S. mansoni*, *S. japonicum*, *C. sinensis*, *O. viverrini*, *F. gigantica* and *F. hepatica*) for which large genomic and/or transcriptomic data sets were available. *S. haematobium* genes that represented solely tandem repeats were identified using Tandem Repeat Finder (TRF) [56] and excluded from the data. Similarly, genes with homology (BLASTn, E value ≤ 10^{-5}) to transposable-like elements were identified and then excluded.

Identification and annotation of intergenic elements.

The frequencies of interspersed repeat sequence elements within the *S. haematobium* genome were assessed using an established bioinformatics pipeline [39].

Analysis of transcription.

To assess differential transcription between the sexes, and between the adult and egg stages of *S. haematobium*, the raw sequence reads derived from each, non-normalized cDNA library were mapped to ORFs predicted from the genome using SOAP2 (ref. 37). Briefly, raw sequence reads were aligned to the non-redundant transcriptomic data, such that only paired, raw sequence reads that mapped to a unique transcript ("unique reads") were retained. Paired reads that mapped to more than one transcript (designated "multi-reads") were randomly allocated to a unique transcript, such that they were recorded only once. To provide a relative assessment of transcript abundance, an equal number of mapped reads (n = 7,627,996) was selected at random, and the number of mapped reads was normalized for length (that is, reads per kilobase per million reads, RPKM) [57]. The analysis of statistical difference of transcription was determined using a method developed for serial analysis of gene expression (SAGE) and applied to RNA-seq data [58]. Statistical significance was set at a P value of ≤0.01 and, to control for errors associated with multiple pairwise comparisons, a false-discovery rate correction was applied to the data set [59]. In addition, for a gene to be classified as being differentially transcribed between any two sexes or stages, a minimum twofold difference in absolute RPKM values was required. The relative abundance of mapped reads to each gene was displayed in two-dimensional space (as x-y coordinates). Transcript densities were $\log_{10}$-transformed, binned into 20-pixel boxes and displayed in a heat map. Genes were selected within a 25-pixel radius from each node, representing male (M), female (F) and egg (E), and those constitutively transcribed genes within a central 100-pixel radius. In addition, data for transcripts enriched in adult male, adult female or egg stages of *S. haematobium* were compared with microarray data sets available publicly for respective developmental stages of *S. japonicum* [19] and *S. mansoni* [18, 60]. Homology-based

comparisons were also made at the protein level (tBLASTx, *E* value ≤ 10^{-5}).

Phylogenetic analysis.

Single-copy genes within the *S. haematobium* genome were inferred using stringent searches for nucleotide sequence homology against the genomes of *S. mansoni* and *S. japonicum*[8, 61]. An initial set of single-copy genes was generated based on *S. haematobium* genes with homology, at low stringency, to only one other *S. haematobium* (BLASTn, *E* value ≤ 10^{-5}), *S. japonicum* (BLASTx, ≤ 10^{-5}) and *S. mansoni* (BLASTx, ≤ 10^{-5}) gene or gene region (based on homology to genome scaffolds) (2,422 genes). Only the putative single-copy *S. haematobium* genes with homology (BLASTx, ≤ 10^{-60}) to putative proteins of *C. sinensis*, *F. gigantica*, *F. hepatica*[32, 54, 55], *O. viverrini*, *S. japonicum*, *S. mansoni* and *Schmidtea mediterranea*[62] were retained. The amino acid sequence conceptually translated from each of the 59 single-copy genes of *S. haematobium* was aligned with that of its inferred ortholog from representatives of the Trematoda and *S. mediterranea* (= outgroup) using T-Coffee[63]. The aligned amino acid sequence blocks were then concatenated using FASconCAT[64], resulting in an alignment over 10,113 positions (excluding gaps or missing data). The data were then subjected to phylogenetic analyses using two methods. First, maximum-parsimony analysis (MP; PAUP*4; v.4.0b10; see URLs) using the heuristic search with tree bisection and reconnection (TBR) branch swapping, the ACCTRAN option and random-taxon addition iterations and supported using bootstrap resampling[65]. Second, Bayesian inference (BI) analysis was performed using Markov chain Monte Carlo analysis in MrBayes[66] (v.3.1.2). The WAG model of amino acid replacement, with gamma distribution and a proportion of invariable sites, was selected for BI analysis using ProtTest[67].

Genetic interaction networking and drug target prediction.

Genetic interactions of *S. haematobium* genes (*n* = 3,997) homologous (BLASTx; *E* value ≤ 10^{-30}) to those of *Mus musculus* were predicted using the program Gene Orienteer[68]. The network of interactions among homologs was displayed using the force-directed layout algorithm in Gephi v.0.7beta (see URLs). Interactions were weighted according to number and confidence scores[69]. The essentiality of protein-coding genes of *S. haematobium* was inferred using an established approach[69, 70, 71]. The essentiality of molecules predicted for *S. haematobium* was then inferred on the basis of the presence of homologs in *Drosophila melanogaster* and *Caenorhabditis elegans* and/or *M. musculus*, for which targeted perturbation of the corresponding genes yields a deleterious, lethal phenotype, according to information available in FlyBase[72] (release FB2011.02), WormBase[73] (release WS222) and The Mouse Genome Informatics Database[74] (release 4.4), respectively. The molecules predicted to be essential were then compared with those predicted previously for *S. mansoni*[21]. To infer the potential of members of this subset as candidate drug targets, their sequences were compared by BLASTp (*E* value ≤ 10^{-15}) against peptides in the ChEMBL[75] (release 09) and DrugBank[76] databases. Only inhibitors that bound orthologous peptides and that passed the Rule of Three (ref. 77) were retained.

Accession numbers.

All sequence data have been deposited in a public genome resource database (SchistoDB; see URLs) with accession numbers Sha_120001 to Sha_301483; NCBI BioProject ID, PRJNA78265.

References

1. Rollinson, D. A wake up call for urinary schistosomiasis: reconciling research effort with public health importance. *Parasitology* **136**, 1593–1610 (2009).
2. Doenhoff, M.J., Cioli, D. & Utzinger, J. Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. *Curr. Opin. Infect. Dis.* **21**, 659–667 (2008).
3. Bouvard, V. *et al.* A review of human carcinogens—Part B: biological agents. *Lancet Oncol.* **10**, 321–322 (2009).
4. Hotez, P.J., Fenwick, A. & Kjetland, E.F. Africa's 32 cents solution for HIV/AIDS. *PLoS Negl. Trop. Dis.* **3**, e430 (2009).
5. Kjetland, E.F. *et al.* Association between genital schistosomiasis and HIV in rural Zimbabwean women. *AIDS* **20**, 593–600 (2006).
6. Gryseels, B., Polman, K., Clerinx, J. & Kestens, L. Human schistosomiasis. *Lancet* **368**, 1106–1118 (2006).
7. Palumbo, E. Association between schistosomiasis and cancer: a review. *Infect. Dis. Clin. Prac.* **15**, 145–148 (2007).
8. Berriman, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358 (2009).
9. The *Schistosoma japonicum* Genome Sequencing and Functional Analysis Consortium. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature* **460**, 345–351 (2009).
10. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
11. Paez, J.G. *et al.* Genome coverage and sequence fidelity of φ29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).
12. Valentim, C.L., LoVerde, P.T., Anderson, T.J. & Criscione, C.D. Efficient genotyping of *Schistosoma mansoni* miracidia following whole genome amplification. *Mol. Biochem. Parasitol.* **166**, 81–84 (2009).
13. Liu, F. *et al.* Excretory/secretory proteome of the adult developmental stage of human blood fluke, *Schistosoma japonicum*. *Mol. Cell. Proteomics* **8**, 1236–1251 (2009).
14. Mathieson, W. & Wilson, R.A. A comparative proteomic study of the undeveloped and developed *Schistosoma mansoni* egg and its contents: the miracidium, hatch fluid and secretions. *Int. J. Parasitol.* **40**, 617–628 (2010).
15. Webster, B.L., Southgate, V.R. & Littlewood, D.T. A revision of the interrelationships of *Schistosoma* including the recently described *Schistosoma guineensis*. *Int. J. Parasitol.* **36**, 947–955 (2006).
16. Short, R.B. & Menzel, M.Y. Chromosomes of nine species of schistosomes. *J. Parasitol.* **46**, 273–287 (1960).
17. Criscione, C.D., Valentim, C.L., Hirai, H., Loverde, P.T. & Anderson, T.J. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. *Genome Biol.* **10**, R71 (2009).
18. Fitzpatrick, J.M. *et al.* Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. *PLoS Negl. Trop. Dis.* **3**, e543 (2009).

19. Gobert, G.N., Moertel, L., Brindley, P.J. & McManus, D.P. Developmental gene expression profiles of the human pathogen *Schistosoma japonicum*. *BMC Genomics* **10**, 128 (2009).
20. Schramm, G. *et al*. Cutting edge: IPSE/alpha-1, a glycoprotein from *Schistosoma mansoni* eggs, induces IgE-dependent, antigen-independent IL-4 production by murine basophils *in vivo*. *J. Immunol.* **178**, 6023–6027 (2007).
21. Caffrey, C.R. *et al*. A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, *Schistosoma mansoni*. *PLoS ONE* **4**, e4413 (2009).
22. Allen, J.E. & Maizels, R.M. Diversity and dialogue in immunity to helminths. *Nat. Rev. Immunol.* **11**, 375–388 (2011).
23. Hewitson, J.P., Grainger, J.R. & Maizels, R.M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Mol. Biochem. Parasitol.* **167**, 1–11 (2009).
24. Botelho, M.C., Machado, J.C. & Correia da Costa, J.M. *Schistosoma haematobium* and bladder cancer: what lies beneath? *Virulence* **1**, 84–87 (2010).
25. Verjovski-Almeida, S. & DeMarco, R. Gene structure and splicing in schistosomes. *J. Proteomics* **74**, 1515–1518 (2011).
26. Geyer, K.K. *et al*. Cytosine methylation regulates oviposition in the pathogenic blood fluke *Schistosoma mansoni*. *Nat. Commun.* **2**, 424 (2011).
27. de Souza Gomes, M., Muniyappa, M.K., Carvalho, S.G., Guerra-Sa, R. & Spillane, C. Genome-wide identification of novel microRNAs and their target genes in the human parasite *Schistosoma mansoni*. *Genomics* **98**, 96–111 (2011).
28. Brindley, P.J., Mitreva, M., Ghedin, E. & Lustigman, S. Helminth genomics: The implications for human health. *PLoS Negl. Trop. Dis.* **3**, e538 (2009).
29. Mann, V.H., Morales, M.E., Rinaldi, G. & Brindley, P.J. Culture for genetic manipulation of developmental stages of *Schistosoma mansoni*. *Parasitology* **137**, 451–462 (2010).
30. Rinaldi, G. *et al*. Genetic manipulation of *Schistosoma haematobium*, the neglected schistosome. *PLoS Negl. Trop. Dis.* **5**, e1348 (2011).
31. Lewis, F.A., Liang, Y.S., Raghavan, N. & Knight, M. The NIH-NIAID schistosomiasis resource center. *PLoS Negl. Trop. Dis.* **2**, e267 (2008).
32. Young, N.D., Hall, R.S., Jex, A.J., Cantacessi, C. & Gasser, R.B. Elucidating the transcriptome of *Fasciola hepatica*—a key to fundamental and biotechnological discoveries for a neglected parasite. *Biotechnol. Adv.* **28**, 222–231 (2010).
33. Dalton, J.P., Day, S.R., Drew, A.C. & Brindley, P.J. A method for the isolation of schistosome eggs and miracidia free of contaminating host tissues. *Parasitology* **115**, 29–32 (1997).
34. Brindley, P.J., Lewis, F.A., McCutchan, T.F., Bueding, E. & Sher, A. A genomic change associated with the development of resistance to hycanthone in *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* **36**, 243–252 (1989).
35. Li, R. *et al*. The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
36. Li, R. *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
37. Li, R. *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
38. Cantacessi, C. *et al*. The transcriptome of *Trichuris suis*—first molecular insights into a parasite with curative properties for key immune diseases of humans. *PLoS ONE* **6**, e23590 (2011).
39. Jex, A.R. *et al*. *Ascaris suum* draft genome. *Nature* **479**, 529–533 (2011).
40. Dyrløv Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
41. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
42. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
43. Chen, Y. *et al*. SPD–a web-based secreted protein database. *Nucleic Acids Res.* **33**, D169–D173 (2005).
44. Bellafiore, S.P. *et al*. Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential. *PLoS Pathog.* **4**, e1000192 (2008).
45. Cass, C.L. *et al*. Proteomic analysis of *Schistosoma mansoni* egg secretions. *Mol. Biochem. Parasitol.* **155**, 84–93 (2007).
46. Wu, X.-J. Proteomic analysis of *Schistosoma mansoni* proteins released during *in vitro* miracidium-to-sporocyst transformation. *Mol. Biochem. Parasitol.* **164**, 32–44 (2009).
47. Hewitson, J.P. *et al*. The secretome of the filarial parasite, *Brugia malayi:* proteomic profile of adult excretory-secretory products. *Mol. Biochem. Parasitol.* **160**, 8–21 (2008).
48. Bennuru, S. *et al*. *Brugia malayi* excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. *PLoS Negl. Trop. Dis.* **3**, e410 (2009).
49. Mulvenna, J. *et al*. The secreted and surface proteomes of the adult stage of the carcinogenic human liver fluke *Opisthorchis viverrini*. *Proteomics* **10**, 1063–1078 (2010).
50. Robinson, M.W., Menon, R., Donnelly, S.M., Dalton, J.P. & Ranganathan, S. An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen *Fasciola hepatica*: proteins associated with invasion and infection of the mammalian host. *Mol. Cell Proteomics* **8**, 1891–1907 (2009).

51. Bryson, K. *et al.* Protein structure prediction servers at University College London. *Nucleic Acids Res.* **33**, W36–W38 (2005).
52. Lobley, A., Sadowski, M.I. & Jones, D.T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **25**, 1761–1767 (2009).
53. Kuzniar, A., van Ham, R.C.H.J., Pongor, S.N. & Leunissen, J.A.M. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24**, 539–551 (2008).
54. Young, N.D. *et al.* Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*. *PLoS Negl. Trop. Dis.* **4**, e719 (2010).
55. Young, N.D. *et al.* A portrait of the transcriptome of the neglected trematode, *Fasciola gigantica*—biological and biotechnological implications. *PLoS Negl. Trop. Dis.* **5**, e1004 (2011).
56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
57. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
58. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
59. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57**, 289–300 (1995).
60. Fitzpatrick, J.M. *et al.* An oligonucleotide microarray for transcriptome analysis of *Schistosoma mansoni* and its application/use to investigate gender-associated gene expression. *Mol. Biochem. Parasitol.* **141**, 1–13 (2005).
61. The *Schistosoma japonicum* Genome Sequencing and Functional Analysis Consortium. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature* **460**, 345–351 (2009).
62. Robb, S.M., Ross, E. & Sanchez Alvarado, A. SmedGD: the *Schmidtea mediterranea* genome database. *Nucleic Acids Res.* **36**, D599–D606 (2008).
63. Notredame, C., Higgins, D.G. & Heringa, J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
64. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
65. Felsenstein, J. Confidence-limits on phylogenies—an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
66. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
67. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
68. Zhong, W. & Sternberg, P.W. Genome-wide prediction of *Caenorhabditis elegans* genetic interactions. *Science* **311**, 1481–1484 (2006).
69. Cantacessi, C. *et al.* A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. *Nucleic Acids Res.* **38**, e171 (2010).
70. Cantacessi, C. *et al.* Massively parallel sequencing and analysis of the *Necator americanus* transcriptome. *PLoS Negl. Trop. Dis.* **4**, e684 (2010).
71. Doyle, M.A., Gasser, R.B., Woodcroft, B.J., Hall, R.S. & Ralph, S.A. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11**, 222 (2010).
72. Tweedie, S. *et al.* FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* **37**, D555–D559 (2009).
73. Harris, T.W. *et al.* WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38**, D463–D467 (2010).
74. Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. & Blake, J.A. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* **36**, D724–D728 (2008).
75. Warr, W. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aided Mol. Des.* **23**, 195–198 (2009).
76. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041 (2011).
77. Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877 (2003).

Download references

Acknowledgments

This project was funded by the Australian Research Council (R.B.G. and H.M.Y.) and BGI. We are grateful for other support from the National Health and Medical Research Council (NHMRC) of Australia (R.B.G.), the Australian Academy of Science, the Australian-American Fulbright Commission, Melbourne Water Corporation, the Victorian Life Sciences Computation Initiative (VLSCI) and the IBM Collaboratory. N.D.Y. (Early Career Research Fellow) and A.R.J. (CDA1 Fellow) were supported by NHMRC fellowships. We thank staff of BGI-Shenzhen, including Q. Nan, P. Na, B. Min and P. Ni for their contributions. *Schistosoma haematobium*–infected hamsters were provided by F.A. Lewis and Y.-S. Liang of the Biomedical Research Institute, under NIAID-NIH contract HHSN272201000005I. G.O. is supported by NIH-Fogarty TW007012, FAPEMIG CBB-1181/08 and CNPq 573839/2008-5.

Author information

These authors contributed equally to this work.

Neil D Young, Aaron R Jex & Bo Li

Affiliations

Faculty of Veterinary Science, The University of Melbourne, Victoria, Australia.
Neil D Young, Aaron R Jex, Cinzia Cantacessi, Ross S Hall, Andreas Hofmann, Bronwyn E Campbell & Robin B Gasser

BGI–Shenzhen, Shenzhen, China.
Bo Li, Shiping Liu, Linfeng Yang, Zijun Xiong, Yingrui Li, Xun Xu, Fangyuan Chen, Xuan Wu, Guojie Zhang, Xiaodong Fang, Yi Kang, Huanming Yang, Jun Wang & Jian Wang

Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais Instituto de Pesquisa René Rachou-Fiocruz, Belo Horizonte, Brasil.
Adhemar Zerlotini & Guilherme Oliveira

Eskitis Institute for Cell & Molecular Therapies, Griffith University, Brisbane, Queensland, Australia.
Andreas Hofmann

School of Public Health, Tropical Medicine and Rehabilitation Sciences, James Cook University, Cairns, Queensland, Australia.
Alex Loukas

Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales, Australia.
Shoba Ranganathan

Department of Biochemistry, National University of Singapore, Singapore.
Shoba Ranganathan

Natural History Museum, London, UK.
David Rollinson

Departamento de Genética, Universidad de la República, Montevideo, Uruguay.
Gabriel Rinaldi

Department of Microbiology, Immunology & Tropical Medicine, George Washington University, Washington, DC, USA.
Gabriel Rinaldi & Paul J Brindley

Contributions

R.B.G. conceived and led the project, with support from X.W., Y.L., Ju.W., Ji.W. and H.Y.; N.D.Y., A.R.J. and R.B.G. designed the experimental plan and executed or guided the bioinformatic analyses. P.J.B. and G.R. provided parasite material. Guided by N.D.Y., A.R.J. and R.B.G., B.L., X.F. and G.Z. coordinated genome assembly and annotation. L.Y. performed the genome assembly; Z.X. conducted comparative genomic analysis; A.H., S.L., Y.K. and F.C. undertook the genome annotation; X.X. and his team performed the whole-genome amplification step; C.C. and B.E.C. conducted the essentiality and drug target predictions. R.S.H. provided bioinformatic support. N.D.Y., A.R.J. and R.B.G. wrote the manuscript, with critical input from P.J.B., G.O., S.R., A.L. and D.R. and comments from the other coauthors. A.Z. and G.O. curated the genome resource database.

Competing financial interests

The authors declare no competing financial interests.

Corresponding authors

Correspondence to: Robin B Gasser or Jun Wang or Jian Wang

Supplementary information

PDF files

1. Supplementary Text and Figures (5M)
   Supplementary Figures 1–13, Supplementary Tables 1–9, 12, 13, 17 and 18 and Supplementary Note.

Excel files

1. Supplementary Table 10 (147K)
   Protein-coding genes of *Schistosoma haematobium*, predicted to be excreted/secreted through the classical secretory pathway.

2. Supplementary Table 11 (766K)
   Schistosoma haematobium genes homologous to *Mus musculus*, subjected to probablistic genetic interactions.

3. Supplementary Table 14 (37K)
   Intra-chromosomal rearrangement of one-to-one orthologues existing between/among the large (> 1 Mb) assembly scaffolds for *Schistosoma haematobium* and *S. japonicum* relative to the *S. mansoni* scaffolds.

4. Supplementary Table 15 (41K)
   Summary of *Schistosoma haematobium* biological pathways predicted to be shared with *S. mansoni and S. japonicum*.

5. Supplementary Table 16 (41K)

Genes predicted for *Schistosoma haematobium* with no homology to those currently known to be encoded in other trematode species.

6. Supplementary Table 19 (541K)
   Transcripts which are significantly enriched in the egg stage of *Schistosoma haematobium*.

7. Supplementary Table 20 (344K)
   Transcripts which are significantly enriched in the female stage of *Schistosoma haematobium*.

8. Supplementary Table 21 (635K)
   Transcripts which are significantly enriched in the male stage of *Schistosoma haematobium*.

9. Supplementary Table 22 (152K)
   Groups of protein-coding genes inferred to be significantly developmentally regulated in adult males, adult females or eggs of *Schistosoma haematobium*.

Nature Genetics ISSN 1061-4036 EISSN 1546-1718