Whole-genome sequence of Schistosoma haematobium

Neil D Young, Aaron R Jex, Bo Li, Shiping Liu, Linfeng Yang, Zijun Xiao, Yingrui Li, Cinzia Cantacessi, Ross S Hail, Xun Xu, Fangyuan Ghen, 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, Huangming 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 worldwide1. No vaccines are available, and treatment relies on one drug, praziquantel2. Schistosoma haematobium has come into the spotlight as a major cause of urogenital disease, as an agent linked to bladder cancer5,7 and as a predisposing factor for HIV/AIDS6,8. The parasite is transmitted to humans from freshwater snails3, and Worms dwell in blood vessels and release eggs that become embedded in the bladder wall to elicit chronic immune-mediated disease9 and induce squamous cell carcinoma10. 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 parasites6,9. 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-based nonrandom sampling10 or of artifacts induced by multiple displacement amplification, a result that is consistent with published information11,12 (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%)6. More than half (58.5%) of the repeats were retrotransposons (at least 20 types, including LINE/RT-BoB 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 S_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 (Omegas-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 relationships16 and karyotypes16. Given the close relationship between S. haematobium and S. mansoni and the size and quality of the draft genome for S. mansoni8, we aligned S. haematobium to S. mansoni scaffolds that mapped to chromosomes (2n = 16, ZZ)7. Overall, rearrangements in S. haematobium with respect to S. mansoni were rare, with 11 inversions of syntonic blocks linked to five chromosomes (nos. 1, 3, 4, 6 and 7) (Fig. 1, Supplementary Note, Supplementary Fig. 11).

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

http://www.nature.com/ng/journal/v44/n2/full/ng.1065.html
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 markers15. 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 positions.

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 allergens–like and CD63 tetraspanin–like proteins (reported to be transcribed exclusively in the miracidium of S. mansonii) as well as two encoding CP391S–like egg proteins. Adult-enriched transcription was shown for 2,975 genes, of which 966 were significantly upregulated in females and 2,010 in males. These findings are similar to those of recent microarray studies of S. mansonii and S. japonicum, 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). Notably in the adult female, which lives within the gynecophoric canal of the male, is a major enrichment in pathways linked to hemooglobin (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 G–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. mansonii.

Figure 3: Differential transcription among adult female (F), adult male (M) and egg (E) stages of Schistosoma haematobium.
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 nematocides. 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. mansoni21 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 responses22. Chronic infections induce key changes in immune-cell populations, including a dominance of the T helper type 2 (Th2) cells and a selective loss of effector T-cell activity, against a background of regulatory T cells, alternatively activated macrophages, and Th2-inducing dendritic cells22. Much of the immunomodulatory capacity of schistosomes is attributable particularly to ES products23. 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 (Sp23 tetrapsarin), cleavage (cysteine and serine proteases) or inhibition of post-translational modification (cystatins) of host immunoglobulins; known inducers of Th2 responses (IPSE-α, ω-1, peroxiredoxin and Sm16 (also called SmSLP or SmSPO-1)); and host-defense mimcry 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 implicated24. 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 molecule 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 splicing25, differential methylation26, regulatory RNA27 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 worldwide1–28. Given the challenges in propagating these parasites, particularly S. haematobium, in the laboratory,26 the ability to sequence the genome from a single pair of worms represents an 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 transgenesis28, and will also facilitate urgently needed proteomic explorations. Published findings30 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.


Methods

Sample procurement, preparation and storage.

A laboratory strain of S. haematobium, originating from Egypt, was maintained in the Biomedical Research Institute, Rockville, Maryland31 in Bulinus truncatus (intermediate snail host) and Mesocricetus auratus (hamster; mammalian definitive host). Hamsters were each infected with 1,000 cercariae. After 50 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 previously32 and snap frozen in liquid nitrogen. S. haematobium eggs were isolated from the livers from infected hamsters33 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 protocol34, 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 quality sequences, base-calling duplicates and adaptors removed35.

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 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 algorithm35. Paired-end sequence data from the genomic DNA libraries were assembled using SOAPdenovo36. 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

http://www.nature.com/ng/journal/v44/n2/full/ng.1065.html
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 tiling 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), adapter-ligated and then sequenced on a Genome Analyzer IIx (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 SOAPdenovo30. Expressed sequence tag (EST) data for S. haematobium were used to train and validate gene models generated from the draft genome, as described previously39. Parameters for the gene-prediction model were established using a set of 985 S. haematobium genes encoding complete open reading frames (ORFs) that were predicted from an unmasked draft of the S. haematobium genome sequence. Genes were predicted30 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 predicted30. The predicted genes were merged to establish nonredundant gene sets39. 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 Gene database 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 pipeline39. 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 TMHMM2.0, and by BLASTP42 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 PSIPRED51. 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)13.

Nucleotide sequence identities in coding domains of S. haematobium and S. mansoni were established, in a pairwise manner, using the program BLASTn42, employing a permissive (E value ≤ 10−5) search strategy. In addition, using the S. mansoni genome as the reference, a genome-wide analysis of synten for one-to-one orthologs across scaffolds of >1 Mb was undertaken53 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 IBLASTx42 employing permissive (E value ≤ 10−5), moderate (≤10−15) and stringent (≤10−50) 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 (IBLASTx, 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 solo tandem repeats were identified using Tandem Repeat Finder (TRF)59 and excluded from the data. Similarly, genes with homology (BLASTn, E value ≤ 10−7) 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 log10-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. japonicum10 and S. mansoni18, 60. Homology-based
comparisons were also made at the protein level (BLASTx, 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*. 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* (BLASTx, E value ≤ 10^-5), *S. japonicum* (BLASTx, E value ≤ 10^-5), and *S. mansoni* (BLASTx, E value ≤ 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, E value ≤ 10^-5) to putative proteins of *C. sinensia*, *F. gigantica*, *F. hepatica*, *O. viverrini*, *S. japonicum*, *S. mansoni* and *Schmidtea mediterranea* 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. The aligned amino acid sequence blocks were then concatenated using FASCONCAT, 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. Second, Bayesian inference (BI) analysis was performed using Markov chain Monte Carlo analysis in MrBayes (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.

**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. 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. The essentiality of protein-coding genes of *S. haematobium* was inferred using an established approach. 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 (release FB2011.02), WormBase (release WS222) and The Mouse Genome Informatics Database (release 4.4), respectively. The molecules predicted to be essential were then compared with those predicted previously for *S. mansoni*. 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 (release 09) and DrugBank databases. Only inhibitors that bind 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.**


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 HHSN272001000005. G.O. is supported by NIH-Fogarty TW007012, FAPEMIG CB3-1181/08 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

BGU-Shenzhen, Shenzhen, China.
Bo Li, Shiqing 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 Zorolfini & 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 Robinson

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


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/secerted through the classical secretory pathway.

2. Supplementary Table 11 (766K)
   Schistosoma haematobium genes homologous to Mus musculus, subjected to probabilistic 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 (535K)
   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*.

This article is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence (http://creativecommons.org/licenses/by-nc-sa/3.0), which permits distribution, and reproduction in any medium, provided the original author and source are credited. This licence does not permit commercial exploitation, and derivative works must be licensed under the same or similar license.

*Nature Genetics*  ISSN 1061-4036  EISSN 1546-1718

© 2012 Macmillan Publishers Limited. All Rights Reserved, partner of AGORA, HINARI, OARE, INASP, ORCID, CrossRef and COUNTER